PACIFIC SYMPOSIUM ON

BIOCOMPUTING 2025

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2025

Kohala Coast, Hawaii, USA,
4 – 8 January 2025

Edited by

Russ B. Altman
Stanford University, USA

Lawrence Hunter
University of Chicago, USA

Marylyn D. Ritchie
University of Pennsylvania, USA

Tiffany Murray
Stanford University, USA

Teri E. Klein
Stanford University, USA

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI • TOKYO

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Online ISSN: 2335-6936
Print ISSN: 2335-6928

Library of Congress Control Number: 2024949502

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

BIOCOMPUTING 2025
Proceedings of the Pacific Symposium

ISBN 978-981-98-0702-4 (ebook)
ISBN 978-981-98-0701-7 (print)

WORKSHOPS

HYBRID SESSION/WORKSHOP

ERRATUM

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2025

2025 marks the 30th Pacific Symposium on Biocomputing (PSB)! As always, we gather on the Big Island to share the latest progress and challenges in biocomputing. In honor of the 30th PSB, we are excited to present a session reviewing the history and scientific impact of the meeting—and the field. We hope you will agree that PSB has impact far beyond what might be expected of a relatively small annual meeting. There has been some suggestion that the rise of Artificial Intelligence (AI) in the last few years may be "hype" and that the promise and impact of AI is overstated. We agree that it is true that in some areas the discussions of AI's promise and impact may be hyperbolic. However, there seems to be little doubt that the impact of AI on science and engineering is profound and has already accelerated discovery in clear ways. One needs to look no further than the 2024 Nobel Prizes where laureates were rewarded for their pioneering work in AI. In Physics, Geoffrey Hinton and John Hopfield were recognized "*for foundational discoveries and inventions that enable machine learning with artificial neural networks.*" In Chemistry, David Baker was recognized *"for computational protein design"* while Demis Hassibis and John Jumper were recognized *"for protein structure prediction."* Of course, the awards rightly go to the scientists, but the awards also illustrate the power and impact that AI is having on science. And the transformation of capabilities is not limited to chemistry and physics but extends to all areas of science and engineering. Every field is taking advantage of tools that can find patterns in data that are not obvious to humans and can generate novel outputs based on deep statistical models of the latent relationships within large datasets. The PSB community celebrates this exciting period of accelerated capabilities. A quick review of the sessions at this year's meeting shows that important uses of AI and machine learning in precision medicine, medical communications, genomics, imaging and health equity are catalyzing progress in these critical areas. Our community has not engaged in hype, but in the responsible use of amazing power tools that allow us to continue addressing the most pressing problems facing biology and medicine. The next thirty years promises to be as amazing as the previous thirty!

In addition to being published by World Scientific and indexed in PubMed, the proceedings from all PSB meetings are available online at http://psb.stanford.edu/psb-online/. Since 1996, all PSB papers are indexed in PubMed. These papers are routinely cited in archival journal articles and routinely represent important early contributions in new subfields—many times before there is an established literature in more traditional journals; for this reason, many papers have garnered hundreds of citations.

The social media handle for PSB is @PacSymBiocomp and the hashtag for PSB 2025 is #PSB25.

The efforts of a dedicated group of session organizers have produced an outstanding program. The sessions of PSB 2025 and their hard-working organizers are as follows:

AI and Machine Learning in Clinical Medicine: Generative and Interactive Systems at the Human-Machine Interface
Organizers: Jonathan Chen, Roxana Daneshjou, Dokyoon Kim, Joseph D. Romano, Fateme Nateghi Haredasht, Geoff Tison

Precision Medicine: Multi-modal and multi-scale methods to promote mechanistic understanding of disease
Organizers: Yana Bromberg, Steven E. Brenner, Hannah Carter

Translating Big Data Imaging Genomics Findings to the Individual: Prediction of Risks and Outcomes in Neuropsychiatric Illnesses
Organizers: Peter Kochunov, Li Shen, Paul M. Thompson, Zhongming Zhao

Overcoming Health Disparities in Precision Medicine
Organizers: Kathleen Barnes, Harris Bland, Francisco De La Vega, Todd L. Edwards, Keolu Fox, Alexander Ioannidis, Eimear Kenny, Rasika Mattias, Bogdan Pasaniuc, Jada Benn Torres, Digna R Velez Edwards

We are excited to present four workshops in which investigators with a common interest come together to exchange results and new ideas in a format that is more informal than the peer-reviewed sessions. For this year, the workshops and their organizers are:

All Together Now: Data Work to Advance Privacy, Science, and Health in the Age of Synthetic Data
Organizers: Lindsay Fernandez-Rhodes, Jennifer K. Wagner

Command Line to PipeLine: Cross-Biobank Analyses with Nextflow
Organizers: Anurag Verma, Lindsay Guare, Katie Cardone, Christopher Carson, Zachary Rodriguez

Leveraging Foundational Models in Computational Biology: Validation, Understanding, and Innovation
Organizers: Steven Brenner, Brett Beaulieu-Jones

Opportunities and Pitfalls with Large Language Models for Biomedical Annotation
Organizers: Fabio Rinaldi, Jin-Dong Kim, Zhiyong Lu, Cecilia Arighi

Finally, we are pleased to have a hybrid session/workshop (with some features of both a session and a workshop):

Earth Friendly Computation: Applying Indigenous Data Lifecycles in Medical and Sovereign AI
Organizers: Keolu Fox, Krystal Tsosie, Kaja Wasik, Alec Calac, Alexander Ioannidis, Eric T. Dawson

The PSB 2025 keynote speakers are Kevin B. Johnson (Science keynote) and Sharon F. Terry (Ethical, Legal and Social Implications keynote).

Tiffany Murray has managed the peer review process and assembly of the proceedings since 2001 and plays a key role in many aspects of the meeting. We are grateful for the long-time support of the National Institutes of Health[1], ISCB, and Cleveland Institute for Computational Biology. PSB 2025 is thankful for the support of Khosla Ventures and PGxAI. This year, the co-chairs are pleased to provide financial support for the opening reception, to thank the community for thirty years of support for PSB. The Research Parasite Awards benefit from support from GigaScience, Jeff Stibel, Mr. and Mrs. Stephen Canon, and Drs. Casey and Anna Greene. The Research Symbiont Awards benefit from support from the Wellcome Trust and the DragonMaster Foundation.

In honor of the 30th Anniversary of PSB, we are pleased to support the local Waikaloa Dry Forest (https://www.waikoloadryforest.org/) dedicated to the preservation of the wiliwili tress in some of the roughest terrain in Hawai'i.

We are particularly grateful to the PSB staff Tiffany Murray, Al Conde, Paul Murray, Mark Woon, Liam Mulhall, Randy Soares, Zach Ritchie, BJ Morrison McKay, Cynthia Paulazzo, Victoria Soares, Jackson Miller, Heather Miller, and Meghann Risell for their assistance. We also acknowledge the many busy researchers who reviewed the submitted manuscripts on a very tight schedule. The partial list following this preface does not include many who wished to remain anonymous, and of course we apologize to any who may have been left out by mistake.

We look forward to a great meeting and to seeing you on the Big Island. Aloha!

Pacific Symposium on Biocomputing Co-Chairs,
October 9, 2024

Russ B. Altman
Departments of Bioengineering, Genetics, Medicine & Biomedical Data Science, Stanford University

Lawrence Hunter
Department of Pharmacology, University of Colorado Health Sciences Center

Marylyn D. Ritchie
Department of Genetics and Institute for Biomedical Informatics, University of Pennsylvania

Teri E. Klein
Departments of Biomedical Data Science, Medicine & Genetics, Stanford University

Thanks to the reviewers…

Finally, we wish to thank the scores of reviewers. PSB aims for every paper in this volume to be reviewed by three independent referees. Since there is a large volume of submitted papers, paper reviews require a great deal of work from many people. We are grateful to all of you listed below and to anyone whose name we may have accidentally omitted or who wished to remain anonymous.

Bhim Adhikari
Giuseppe Albi
Tiffany Amariutta
Fatemeh Amrollahi
Pooya Ashtari
Shamini Ayyadhury
Berardino Barile
Oliver Bear Don't Walk IV
Mike Beer
Jada Benn Torres
Erik Bergstrom
Harris Bland
Carly Bobak
Mayla Boguslav
Mary Regina Boland
Aritra Bose
Joeseph Breeyear
Yana Bromberg
Emidio Capriotti
Carlos Cardenas-Iniguez
Hannah Carter
Andrea Castro
Jui-Hsuan Chang
Kewei Chen
Li Chen
Jianlin Cheng
Tishya Chhabra
Dana Crawford
Matteo d'Antonio
olivia daigle
Roxana Daneshjou
Jishnu Das
Conor Davenport
Francisco De la Vega
Klest Dedja
David Dorr
Duy Duong-Tran
Hyrum Eddington
Rachel Edgar
Todd Edwards
Rob Gallo
Si Gao
Emily Getzen
Pedram Golnari
Bryan Gopal
Carsten Görg
Ben Greenbaum
François Grolleau
Lindsay Guare
Melissa Gymrek
Paige Haffener
Steven Hart
Lifang He
Dominik Heider
Jacklyn Hellwege
Josiah Hester
William Hou
Yidi Huang
Maui Hudson
Pedro Ilídio
Alex Ioannidis
Elizabeth Jasper
Brian Johnson
Linda Kachuri
Nicole Karcher
Monica Keith
Nikhil Khankari
Jaesik Kim
Mansu Kim
Adam Klie
Alice Kochunov
Peter Kochunov
Kuldeep Kumar
Rachit Kumar
Bill La Cava
David Laub
Jonghyun Lee
Ko-han Lee
Jake Leiby
Joshua Levy
Xi Li
Jia Liu
Jingyu Liu
Zheng Liu
Ivan Lopez
Stephen Ma
Tianzhou (Charles) Ma
Yizhou Ma
Yizhou Ma
Pranav Mani
Nick Marshall
Graham McVicker
Michelle Mello
Doug Meyer
Nigreisy Montalbo
Daniel Mas Montserrat
Keith Morse
Shizhuo Mu
Felipe Nakano
Xia Ning
Fabricio Olivetti di Franca
Joseph Orr
Kivil Ozturk
Guarav Pandey
Laxmi Parida
Bogdan Pasaniuc
Bogdan Pasaniuc
Margaret Pearce
Minoli Perera
Karl Pichotta
Malvika Pillai
Arturo Pineda
Delaram Pouyabahar
R. Prabakaran
Katrina Prantzalos
Gang Qu
Consuelo Quinto

Predrag Radivojac
Ojas Ramwala
Antonio Raussel
Brooke Rhead
Barak Richman
Nuno Rocha Nene
Rany Salem
Fateme Samea
Sriram Sankararaman
Tom Savage
TJ Sears
Shefali Setia Verma
Lauren Severance
Amir Shamaei
Supreeth Shashikumar
Ruhollah Shemirani
Megan Shuey
Pankhuri Singhal
Beckett Sterner
Quan Sun
Riley Taitingfong
Ha Tang
Kekoa Taparra
Kira Taylor
Fasil Tekola Ayele
Emel Timucin
Dipak Upadhyaya
Ryan Urbanowicz
Eli Van Allen
Digna R. Velez Edwards
Marijana Vujkovic
Ugrucan Vurgun
Cameron T. Waller
Hanyin Wang
Jing Wang
Lei Wang
Peter Washington
Zhi Wei
Qiutin Wen
Alex Wenzel
Jason Williams
Scott Williams
Fang Xiang Wu
Guangzhi Xiong
Shu Yang
Xiaohui Yao
Alaa Youssef
Dominic Yurk
Simone Zaccaria
Cyril Zakka
Aiying Zhang
Zhongming Zhao
Rui Zhu

[1]Funding for this conference was made possible (in part) by R13LM006766 from the National Library of Medicine. The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

A Comprehensive Bibliometric Analysis: Celebrating the Thirtieth Anniversary of the Pacific Symposium on Biocomputing

Rachit Kumar[1], Rasika Venkatesh[1], David Y. Zhang[1], Teri E. Klein[2], Marylyn D. Ritchie[3†]
[1]*Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA, USA*
[2]*Department of Biomedical Data Science, Medicine (BMIR) and Genetics (by courtesy), Stanford University, Stanford, CA, USA*
[3]*Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA*
Email: marylyn@pennmedicine.upenn.edu

The 2025 Pacific Symposium on Biocomputing (PSB) represents a remarkable milestone, as it is the thirtieth anniversary of PSB. We use this opportunity to analyze the bibliometric output of 30 years of PSB publications in a wide range of analyses with a focus on various eras that represent important disruptive breakpoints in the field of bioinformatics and biocomputing. These include an analysis of paper topics and keywords, flight emissions produced by travel to PSB by authors, citation and co-authorship networks and metrics, and a broad assessment of diversity and representation in PSB authors. We use the results of these analyses to identify insights that we can carry forward to the upcoming decades of PSB.

Keywords: Bibliometry; PSB Proceedings.

1. Introduction

1.1. *Overview*

The Pacific Symposium on Biocomputing (PSB) is an international conference where presentation and discussion of current research in the theory and application of computational methods in problems of biological significance take place. PSB has been held annually since 1996. PSB 2025 marks the 30th anniversary of this conference, a milestone that represents a critical opportunity to evaluate the impact the conference has had on the field of biocomputing, including the scientists in the field, and to find opportunities for growth for the future of PSB and other related conferences. This project was inspired by a similar initiative in 2015 that sought to commemorate the 20th anniversary of PSB[1].

We make use of bibliometric data available on all PSB proceedings from 1996 to 2024, which includes 1402 published papers, all of which are indexed on PubMed. Using these citations, we performed a variety of analyses, each focused on a different perspective or lens by which we reviewed the data. These analyses were inspired by the 20th anniversary review of the PSB proceedings[1] as well as some of the session topics for the PSB 2025 conference, as both a framework for the methodology as well as the topic of many of these analyses.

The session topics for PSB 2025 include:

- AI and Machine Learning in Clinical Medicine
- Earth Friendly Computation
- Precision Medicine: Multi-modal and multi-scale methods

† Corresponding author.

- Translating Big Data Imaging Genomics Findings to the Individual
- Overcoming health disparities in precision medicine

In the following sections, we discuss how we took inspiration from some of these session topics to formulate and perform a variety of analyses on the bibliometric data available from all previous PSB publications.

1.2. ***Session Topics and Mapping to Analyses***

1.2.1. *AI and Machine Learning in Clinical Medicine*

For this session topic, we used keyword analysis to explore how the topics of PSB publications have shifted from year to year along certain breakpoints or eras of PSB. This is described in more detail as the "Keywords and Topics Analysis" in Methods. It is worth noting that many of our other analyses as described below also indirectly make use of machine learning models.

1.2.2. *Earth Friendly Computation*

For this session topic, we sought to estimate the environmental impact of PSB in the form of carbon emissions, given that PSB is a conference that has been held in Hawaii every year (with the sole exception of PSB 2021, which was held virtually due to the COVID-19 pandemic). This is described as the "Emissions Analysis".

1.2.3. *Precision Medicine: Multi-modal and multi-scale methods*

For this session topic, we took inspiration from the idea of multiscale analyses in other fields such as social network analysis and applied them to citation analysis, where we look at the relationships of PSB papers and authors over time. We present this information in the form of a multimodal network that includes papers and authors as well as a co-authorship network. We further perform more traditional citation analyses. This is described as the "Citation and Authorship Analysis".

1.2.4. *Overcoming health disparities in precision medicine*

For this session topic, we wanted to perform analyses that are parallel to the growing understanding that health disparities are critical to acknowledge and address in precision medicine, specifically in the form of acknowledging the value of diverse perspectives in science and scientific discourse. Specifically, we explored the diversity of authors on PSB proceedings papers by exploring changes in representation and diversity along the axes of race and gender. This is described as the "Diversity Analysis".

2. Methods

2.1. ***Common Methods***

We acquired from PubMed the initial list of 1402 papers by using the search term " "Pac Symp Biocomput"[jour] " (outer quotes not included) and then exported a CSV of the results, which

contained information on paper titles, publication year, PMID, and authors (first initial and last names only). We spot-checked randomly-selected papers in this list against the online published conference proceedings to confirm concordance of papers and authors[2].

We then used the NCBI Entrez tools via Biopython[3] with each of the paper PMIDs to further acquire additional information from PubMed on citation PMIDs (only those to or from papers indexed in PubMed), author affiliations, paper abstract text, and full author names (where available). Notably, for papers from 1996-2004, a substantial number of authors only had first initials available.

Subsequently, we acquired dimensions.ai[4] data on PSB papers as of June 2024 to acquire overall citation counts, recent citation counts, and a machine-learning-based determination of the Australian and New Zealand Standard Research Classification 2020 Fields of Research (ANZSRC 2020 FoR)[5], which includes a hierarchical system that identifies broad categories such as "Health Sciences" (category #42) as well as more nuanced categories such as "Machine Learning" (category #4611). Papers can be assigned to multiple categories. Dimensions contained relevant information for 1367 papers out of 1402.

The analyses performed with this data are summarized in visual form in Figure 1 below and are further described in the subsequent sections.

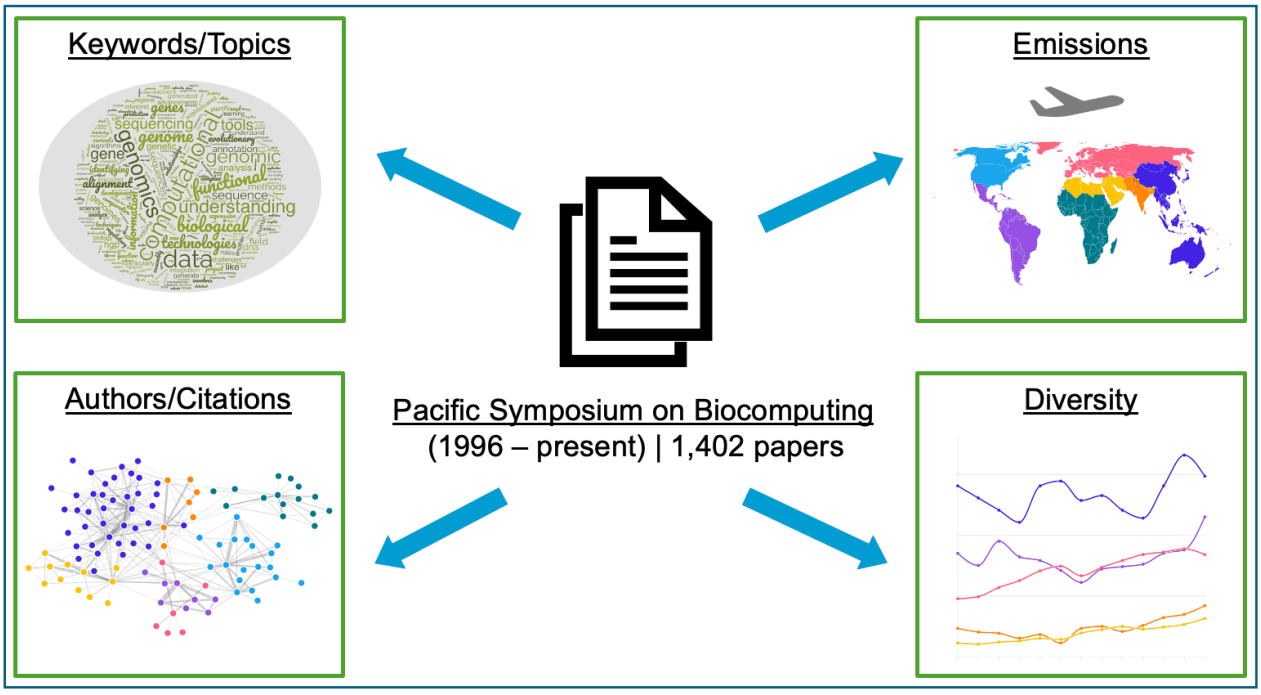


Figure 1: A graphical summary of the bibliometric analyses performed in this paper on PSB papers from 1996 to 2024, including: evaluation of keywords and topics; estimation of conference attendance emission costs; generation of author and citation co-networks; and review of author diversity.

2.2. *Keyword and Topics Analysis*

We used the assigned dimensions.ai ANZSRC 2020 FoR classifications as described in the Common Methods section as overarching paper topics. Additionally, we used KeyBERT, a tool that uses deep neural networks in the form of transformers[6]. KeyBERT generates BERT embeddings of papers and

keywords, and it then subsequently identifies the most relevant keywords for each paper. Specifically, we provide the abstract of each paper, and keywords are extracted with KeyBERT using the underlying "all-MiniLM-L6-v2" sentence transformer model. Each paper was given five keywords in this fashion, with an additional constraint added to try to make the keywords as distinct as possible using KeyBERT's MMR diversity parameter with a value of 0.7.

To interpret this information, we subdivided the papers in PSB into "eras" relating to various important milestones in the field as breakpoints:

- 1996-2003 (before the completion of the Human Genome Project[7])
- 2004-2016 (before the spike in popularity of deep learning in biomedicine, particularly transformers and LLMs)
- 2017-2024 (during the current era of an "AI" boom in biomedicine, including breakthroughs such as AlphaFold[8])

For each topic, the proportion of papers in each era as described above assigned to that topic was computed, and a bar plot was made showing the proportions for each era. Relevant topics were selected for presentation based on which topics were most nonredundant and had a critical number of papers assigned to them.

Separately, for each era, the keywords for papers in that era were lemmatized to combine singular and plural versions of the same word and then collated together to produce a word cloud using the `word_cloud` Python package[9] for each of the eras as a visualization of the most relevant keywords as aggregated across papers for each era.

2.3. *Emissions Analysis*

Using the PubMed information on author affiliations, we performed an analysis of the CO_2 equivalents that were produced as a result of flying to PSB. Specifically, we used the affiliations of the first authors of every paper and used the Google Maps Geocoding API[10] to programmatically and automatically identify the most likely latitude and longitude for each affiliation. We then used a set of data from OurAirports, an open-source and curated repository of airports around the world and their latitude and longitudes[11], to map each individual to their nearest "medium-sized" or "large-sized" airport (observing that "small-sized" airports tended to be regional or private airports) by calculating the Haversine distance[12] of each airport to each affiliation and identifying the closest such airport for each affiliation.

Once each affiliation was mapped to an airport, the Haversine distance of those airports to the Hawaii Kona airport (KOA) was computed to get a putative shortest-path flight distance. This distance was then multiplied by a constant scale factor of 0.148 kg CO_2e per passenger-kilometer (as reported by the UK Government's Department for Energy Security and Net Zero) to compute the carbon emissions of each flight (matched to each paper)[13,14].

In this analysis, we made several assumptions, some of which we recognize as unlikely (see discussion below): only the first authors fly to PSB (and they travel alone), authors fly from their closest (mapped) airport to their reported affiliation, all flights are direct to the KOA airport, all flights in the past have the same carbon efficiency as flights today, all flights take the shortest possible path according to the Haversine distance between airports, and the radius of the Earth is

generally constant at 6371 km for the purposes of computing the Haversine distance (modeling the Earth as a perfect sphere). Notably, PSB 2021 was online due to the COVID-19 pandemic, so emissions for that year were artificially zeroed out after the calculations above.

2.4. *Citation and Authorship Analysis*

Using the PubMed and dimensions.ai-acquired information on all PSB proceedings papers, we computed a variety of statistics for each paper and author in PSB as well as PSB-wide statistics. Additionally, we created an interactive network of papers and authors as well as an interactive network of coauthors. The paper-author network has edges connecting authors to the papers that they have written as well as edges connecting papers that have cited each other and has interactive nodes that allow one to see various statistics for each paper and author. This includes metrics such as the number of citations or the keywords of a paper, and the first year an author published in PSB or the total number of papers an author has published in PSB.

The coauthor network is a multigraph of nodes representing every author at PSB and edges representing their co-authorship in the three different eras of PSB as mentioned in the "Keywords and Topics Analysis" Methods section above. We identified authors uniquely by using their first initial and their last name due to limitations in the data from 1996 to 2004 (where only first initials were available). We performed some simple network analyses on the co-authorship graph: we used Louvain community detection[15,16] to identify communities of co-authors in each era; we used PageRank[17] to identify the most "central" authors for each era; we computed the "density" of "co-authorship ties" defined as a proportion of the number of co-authorship links for each era over the total number of possible links for that era (based on the authors in that era).

2.5. *Diversity Analysis*

Using the PubMed information on author first and last names, we performed an analysis of the likely genders and races of all authors to assess how the diversity of PSB has changed over time. In each of the below analyses, authors were *not* deduplicated within a given year or between years, as we sought to assess the overall diversity of published authors in PSB. For the gender analysis specifically, we took inspiration from prior work by Teich et al.[18]; however, they used a paid API that has since changed methodology to determine genders. To make our methodology more reproducible and to minimize costs, we used United States Social Security Administration data instead.

To identify gender probabilities for each name, we used available data from the United States Social Security Administration (SSA) on first names for children from 1900 to 2023[19] and the assigned gender at birth of those children, calculating a ratio of male/female for each name across all of those years. The proportions of gender probabilities for each author's first name were averaged across years and plotted, with first names not being present in the SSA data (representing names that occurred less than 5 times in every year) being dropped. Notably, PubMed and the original PSB proceedings are missing information on authors' first names for all years up to 2004 (with 100% of authors missing first names in every year up to 2004 except for 2002 and 2003, which are each missing over 50% of first names). As such, all data from the years prior to 2005 were dropped.

To identify race probabilities for each full name, we used a Python package called ethnicolr2[20], which uses deep learning models (long short-term memory models) trained on first and last names from a database of Florida voters in the United States to predict the likelihood of each name belonging to someone identifying as one of five categories: "Asian", "Hispanic", "Non-Hispanic Black", "Non-Hispanic White", and "Other"[21]. The proportions of race probabilities for each author's full name were averaged across years and plotted.

3. Results

3.1. *Keyword and Topics Analysis*

Figure 2 shows three bar plots, one for each of the broad ANZSRC 2020 FoR topics of "Biological Sciences", "Biomedical and Clinical Sciences", and "Information and Computing Sciences". We can see that the proportion of papers tagged as "Biological Sciences" decreased from ~73% in the first two eras to 60.6% in the third, the proportion of papers tagged as "Biomedical and Clinical Sciences" increased era-over-era from 3% of papers in the first era to 13.1% of papers in the second era and 24% of papers in the third, and the proportion of papers tagged as "Information and Computing Sciences" is 27.9% in the first era, 22.1% in the second era, and 33.1% in the third era.

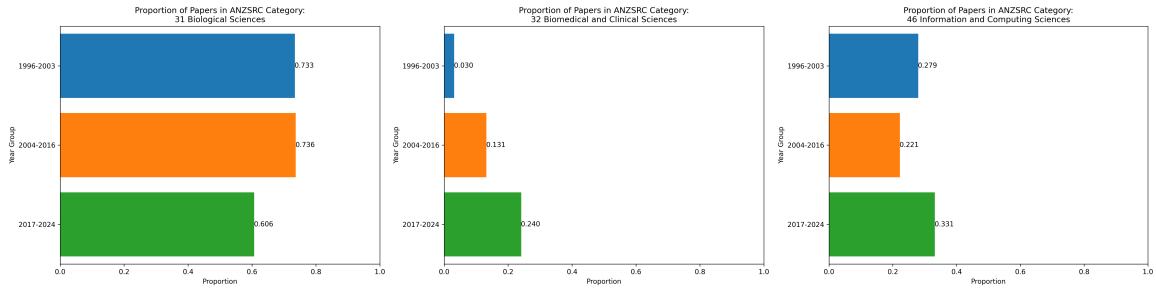


Figure 2: Proportion of papers in each of the three eras that were given the ANZSRC 2020 FoR broad categories of "Biological Sciences" (left), "Biomedical and Clinical Sciences" (middle), and "Information and Computing Sciences" (right).

Similarly to Figure 2, Figure 3 shows three bar plots for select ANZSRC subcategories - that is, categories one level lower than the broad topics as in Figure 2. The three categories shown are "Bioinformatics and Computational Biology", "Oncology and Carcinogenesis", and "Machine Learning", which are each (in order) a subcategory of the respective broad categories from above.

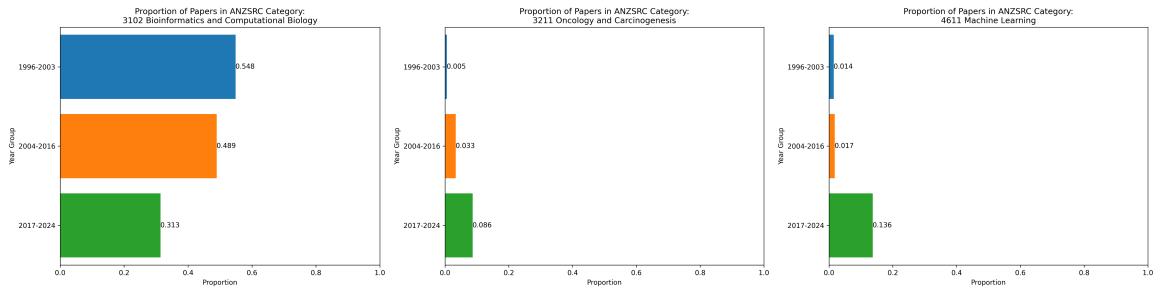


Figure 3: Proportion of papers in each of the three eras that were given the ANZSRC 2020 FoR subcategories of "Bioinformatics and Computational Biology" (left), "Oncology and Carcinogenesis" (middle), and "Machine Learning" (right).

Figure 4 shows the word clouds, one for each of the three eras. Notably, all three word clouds show many words related to genetics, genomics, and related topics with high prominence, such as "gene", "genomic", and "genome". Outside of these words, the first era word cloud shows a prominence of terms such as "alignment" and "sequence". The second era word cloud shows an increase in the prominence of "phenotype" and "annotation". The third era word cloud shows increased representation of research described by the words "neural" and "predicting". Also of note, the word "protein" was prominent in the first two word clouds, but significantly reduced in the most recent era.

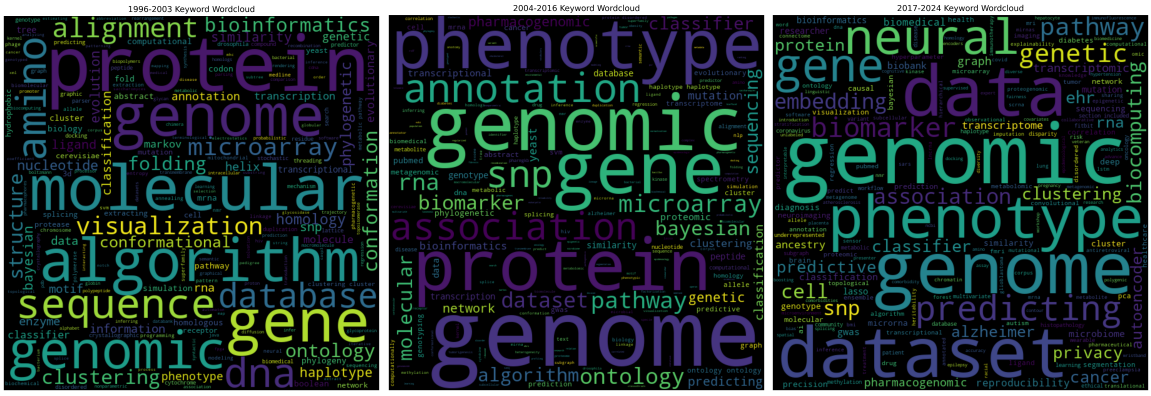


Figure 4: Word clouds for 1996-2003 (left), 2004-2016 (middle), and 2017-2024 (right).

3.2. *Emissions Analysis*

Given the assumptions and approach above, Figure 5 shows the estimated flight emissions for each year of PSB, showing the calculated average emissions per paper and the total emissions (across all papers). The total emissions for all 29 analyzed years of PSB was ~2,832,005 kg CO_2e (~2832 metric tons of CO_2e). Disregarding data from 2021 due to the online nature of PSB that year, this led to a computed average of ~2100 kg CO_2e per paper (~2.1 metric tons of CO_2e per paper) and an average of ~101 metric tons CO_2e per year of PSB.

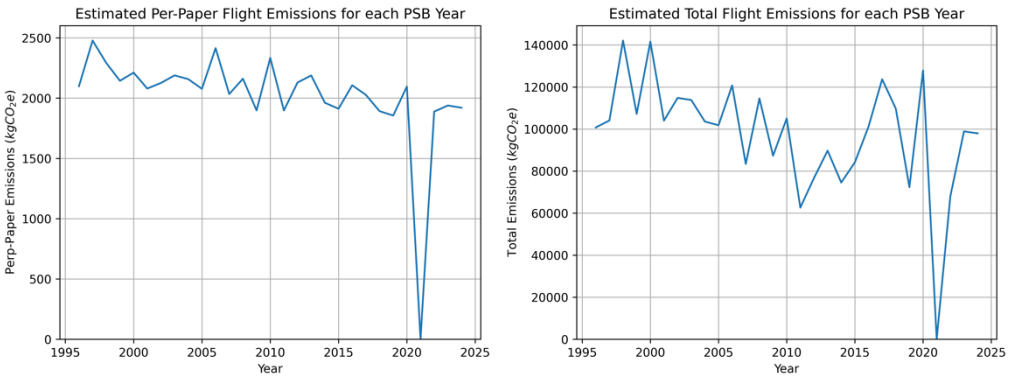


Figure 5: Emissions for each year of PSB from 1996 to 2024. (Left) The per-paper mean emissions for each year. (Right) The total emissions for each year. 2021 was held virtually due to the COVID-19 pandemic and as such had no estimated flight emissions.

3.3. *Citation and Authorship Analysis*

As of June 2024, 1367 papers out of 1402 in the PSB proceedings had citation information in dimensions.ai. The average number of citations across these papers was 20.91 (standard deviation 50.95; total 28579) with a median of 7.0 (max: 680). When normalizing by the number of years that a paper has been available up to 2025 (getting the number of times cited per year), the average is 1.42 citations/year, with a median of 0.54 citations/year. 1229 papers out of 1367 papers with citation data (~90%) published in PSB have been cited at least once.

Papers in PSB were cited, collectively, 3943 times in the past two years, with papers in the last decade receiving a larger proportion of those citations (Figure 6). PSB has, as of 2024, an h-index of 76 - that is, 76 papers have been published that received at least 76 citations. For papers in just the last 5 years (from 2020-2024), the corresponding h5-index is 13.

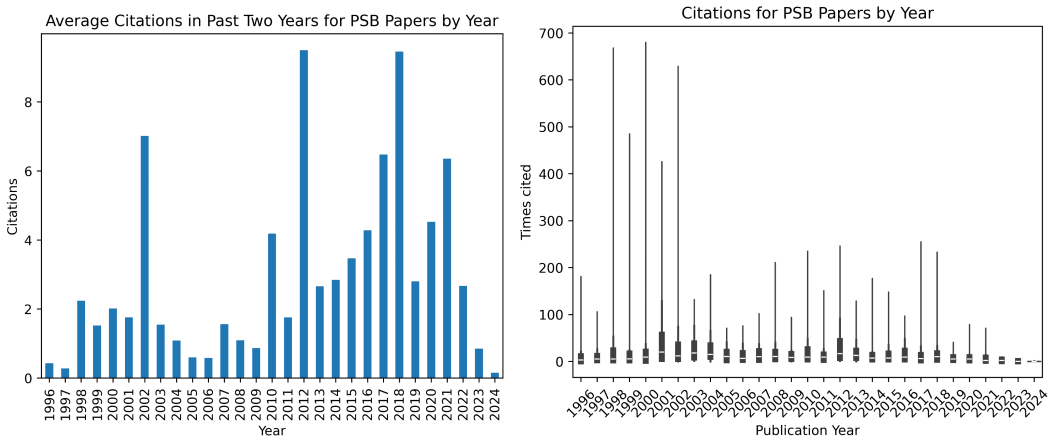


Figure 6: (Left) Average number of recent citations (in the past two years) for each paper per year of PSB from 1996 to 2024. (Right) Violin plot of the total number of citations for each year.

The interactive paper-author graph and the co-authorship graph are both available online here: https://ritchielab.org/publications/supplementary-data/psb-2025/psb-bibliometry. The best way to search for papers or authors is to select (1) “edge” (2) “from” (3) [PAPER/AUTHOR NAME] from the filter dropdowns, respectively.

From the co-authorship analysis done over eras, we find that the average size of communities of coauthors was 4.3 in 1996-2003, 7.8 in 2004-2016, and 12.2 in 2017-2024. For the same eras, the number of unique authors was 1065, 1815, and 1510, respectively. Across all years of PSB, we identified 4013 unique authors.

Based on PageRank centrality, the top 10 authors for each era are listed in Table 1:

Table 1. Top 10 authors for each era of PSB based on PageRank centrality of the co-authorship network.

1996-2003	2004-2016	2017-2024
Miyano, S	Altman, RB	Moore, JH
Altman, RB	Ritchie, MD	Crawford, DC
Takagi, T	Crawford, DC	Ritchie, MD
Hunter, L	Moore, JH	Zou, J
Dunker, AK	Cohen, KB	Tintle, N
Godzik, A	Liu, Y	Brenner, SE
Kohane, IS	Butte, AJ	Thompson, PM
Kitano, H	Chen, L	Chen, Y
Zimmer, R	Hartemink, AJ	Wall, DP
Huang, CC	Pendergrass, RA	Altman, RB

3.4. *Diversity Analysis*

Figure 7 has two line graphs, one for gender proportion and one for race and ethnicity proportions, for each year of PSB. For the gender analysis, ~32.5% of all listed authors across all years of PSB were estimated to be female. When taking the rolling mean of these proportions on a 5-year basis, we see that the earliest years of PSB of 1996-2000 had a gender proportion of ~25% while the most recent years of 2020-2024 had a gender proportion of ~35%.

For the race and ethnicity analysis, across all years of PSB, ~55.3% of all authors were estimated to be Non-Hispanic White, ~26.4% Asian, ~6.8% Non-Hispanic Black, ~5.3% Hispanic, and ~6.2% Other. When taking the rolling mean in a similar fashion to the gender analysis, we note the following changes (in the form of the mean proportion from 1996-2000 -> the mean proportion from 2020-2024): Non-Hispanic White ~62% -> ~51%; Asian ~22% -> ~30%; Non-Hispanic Black ~6.1% -> 6.5%; Hispanic ~3.4% -> 5.8%; and Other 5.7% -> 7.0%.

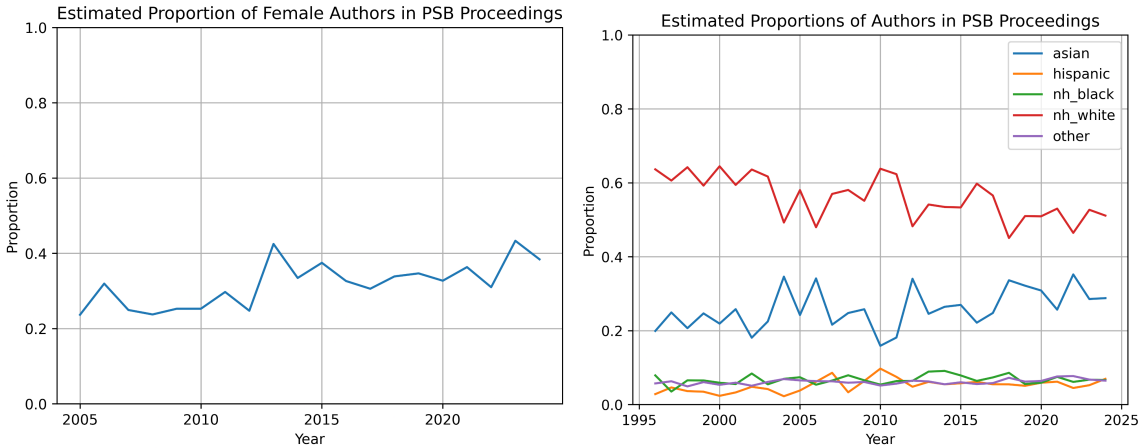


Figure 7: Line graphs of the estimated proportion of female authors (left) and the estimated proportion of authors belonging to one of the five indicated race and ethnic categories (Asian, Hispanic, Non-Hispanic Black, Non-Hispanic White, and Other).

4. Discussion

4.1. *Keyword and Topics Analysis*

As can be seen from the ANZSRC analysis in Figure 2, we see broad patterns across the eras in paper-broad topic assignments that align with the eras themselves. For example, we see a decrease in the number of papers identified as "Biological Sciences", which per the ANZSRC definitions includes more basic science and lab research with some focus on data analysis of wet-lab data[22]. This represents PSB's overall shift to more computational and big data approaches, as corroborated by the word clouds of keywords shifting slightly away from molecular data such as proteins, amino acids, and related, in favor of data and datasets. Similarly, we see an increase in research identified as "Biomedical and Clinical Sciences", which represents an increased tonal shift towards clinical data (as opposed to wet-lab or molecular data) over the years in the field of biocomputing (which encompasses computational biology, bioinformatics, biomedical informatics, and data science).

Interestingly, we see a slight decrease in the proportion of papers identified as "Information and Computing Science" from the first era to the second era, and then an increase from the second to the third era. This may reflect that, in the earliest iterations of PSB, there was a larger focus on developing methods to analyze sequencing data, as reflected in the keywords from the first word cloud showing "sequence", "amino", "alignment", and "algorithm" prominently. With the release of the Human Genome Project, such focuses became less critical and a shift occurred to more methodological *applications* such as genome-wide association studies (GWAS) and annotation analyses (which fall outside the realm of this topic, per ANZSRC), as reflected in the word cloud showing a disappearance of the aforementioned terms and the emergency of terms such as "phenotype" and "annotation". After the resurgence of machine learning and AI in biomedicine, a development of new approaches that leveraged these fields and made use of existing data became a larger focus once more - indicated in the word cloud by terms such as "data", "neural", "predicting", and "embedding" becoming more prominent.

These shifts are further reflected in the papers' subcategories assignments, as shown in Figure 3, where each plot is a subcategory of the broader categories from Figure 2, respectively. There is a consistent decrease in the number of papers described as "Bioinformatics and Computational Biology", which is curious at first for a Biocomputing conference until one recognizes that biomedical informatics is considered a distinct field that is included in the broad scope of Biocomputing. Similarly, there is an increase in clinical-adjacent research in the form of "Oncology and Carcinogenesis", which matches the broad trend of an increase in cancer research as we better understood phenotype data and with the emergence of GWAS, and these trends expectedly match the trends of their parent categories.

The final subcategory of "Machine Learning" has shown dramatic increases that align strongly with the defined eras, going from being a topic of less than 2% of papers in either of the first two eras to 13.6% of papers in the third era, reflecting the period of time in which machine learning and AI became much more strongly incorporated in biomedical research, as well as the transition of authors at PSB to more biomedical and clinical informatics research where big data allows for the training and application of more advanced and complex models.

4.2. *Emissions Analysis*

It is important to note that the assumptions that were made for the emissions analysis, as described in the methods, all lead to a likely underestimate of the true emissions produced. For example, most authors do not have access to airports that offer direct flights to Hawaii, and flights back in the 1990s produced more emissions per-passenger than flights today[23]. Additionally, PSB regularly sees approximately 200+ attendees per year, while this analysis only accounts for roughly ~40-50 of those attendees (the first authors of each accepted paper).

Despite these limitations, this analysis does highlight the fact that PSB does have a relatively high carbon footprint with an average total emission attributable to flights by just first authors of over 100 tons of CO_2e per year. Interestingly, PSB's average flight emissions per year has been decreasing, despite no notable change in the number of papers or attendees used in these calculations, which may indicate a consolidation in the number of traveling authors or an increase in the proportion of authors nearer to Hawaii. Over the years, PSB has contributed to the Hawaiian Legacy Reforestation Initiative[24] which plants koa and sandalwood trees. This is a step toward providing an offset for the carbon footprint[25].

4.3. *Citation and Authorship Analysis*

Overall, from a citation and research output perspective, PSB has been consistently impactful. With a total recorded citation count of 28579 and 90% of papers being cited at least once, PSB has contributed significantly to the body of scientific literature over the past 30 years, and continues to do so. With an h-index of 76 and an h5-index of 13, PSB remains competitive as a conference for biocomputing.

For example, the top papers by citations (Figure 6, right) are concentrated in the first decade of PSB, indicating that they have had a long and lasting impact over the years. However, the papers that have received the most citations in the last two years (Figure 6, left) are largely concentrated within the last decade of PSB, indicating PSB's consistency as a top conference in the field as time goes on, as well as its ability to best attract the cutting-edge ideas in the field of biocomputing.

Furthermore, we find that PSB has encouraged collaborations, with co-authorship networks increasing from 4.3 in its earliest years to 7.8 in the second decade and up to 12.2 in more recent years, indicating that larger groups of authors are working together in PSB. This increase occurs seemingly independently of the number of unique authors (going from 1065 to 1815 and then 1510), indicating that PSB fosters collaborations within its author network.

4.4. *Diversity Analysis*

It is important to note that this information cannot be considered definitive at any non-aggregate scale (that is, any individual level information) due to the use of computed probabilities based on machine learning models, and we recognize that the categories used do not conform to definitions outside or even inside of the USA (for race) or to nonbinary definitions (for gender). Furthermore, transgender individuals may not identify as the gender that they were assigned at birth (which is the information available from the SSA statistics used), and individuals can identify as members of multiple racial or ethnic groups. As such, we demur from drawing strong conclusions about any

individual authors and instead look primarily at population-level trends only with the caveat that this analysis is highly limited at best.

. With these considerations in mind, we do note a trend of an apparent increase in the estimated proportion of published authors that are female from ~25% in some of the first years of PSB to ~35% in recent years. This trend is relatively consistent with proportions of female authorship in other medical journals, with PSB having a slightly higher estimated representation of female authors overall[26–29], and PSB's apparent gender proportion aligns with the proportion of investigators funded by the NIH that identify as female (37% as of 2024)[30].

We note an apparent increase in the estimated proportion of authors that are one of Asian, Hispanic, Non-Hispanic Black, or Other. Correspondingly, we note an apparent decrease in the estimated proportion of authors that are Non-Hispanic White. When compared to the racial and ethnic makeup of NIH-funded investigators as of 2024, PSB has a recent estimated proportion of authors in the two subgroups that have been identified as underrepresented minorities by the NIH[31] that is similar or higher: Hispanic (NIH ~6.1%, PSB ~5.8%) and Non-Hispanic Black (NIH ~3.6%, PSB ~6.5%)[32].

5. Conclusion

Overall, PSB is a conference in the field of biocomputing that presents cutting edge research (Keyword and Topics Analysis) that is highly impactful and fosters collaboration (Citation and Authorship Analysis). Furthermore, PSB publishes papers from authors who represent a diverse range of perspectives and has improved in this regard over the years (Diversity Analysis). PSB remains committed to improving representation from a wider range of groups. We also recognize that these positive aspects of PSB do not come without an environmental cost in the form of flight emissions to travel to PSB (Emissions Analysis); however, the conference has made contributions back to the islands in the form of planting trees to offset this carbon footprint. These insights are useful as we continue to plan for PSB in coming years.

In conclusion, this paper highlights PSB's remarkable record as a leader in Biocomputing over the past thirty years, and we look forward to the future of PSB in fostering collaboration, publishing cutting edge research, and providing an avenue for continued discussions about how to best improve the landscape of biomedical research.

6. Acknowledgments

RK was partially supported by the Training Program in Computational Genomics grant from the National Human Genome Research Institute to the University of Pennsylvania (T32HG000046). DZ was supported by a fellowship from the National Heart, Lung, and Blood Institute (F30HL172382). The National Library of Medicine (R13LM006766) and the International Society for Computational Biology have provided continuous funding to PSB in support of travel awards to increase representation of broad diversity since 1996. This paper was written using data obtained on June 18, 2024, from Digital Science's Dimensions platform, available at https://app.dimensions.ai.

References

1. Hewett D, Whirl-Carrillo M, Hunter LE, Altman RB, Klein TE. A twentieth anniversary tribute to PSB. Pac Symp Biocomput Pac Symp Biocomput. 2015;1–7.

2. PSB Proceedings [Internet]. [cited 2024 Jul 31]. Available from: https://psb.stanford.edu/psb-online/

3. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinforma Oxf Engl. 2009 Jun 1;25(11):1422–3.

4. Hook DW, Porter SJ, Herzog C. Dimensions: Building Context for Search and Evaluation. Front Res Metr Anal [Internet]. 2018 Aug 23 [cited 2024 Jul 31];3. Available from: https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2018.00023/full

5. Porter SJ, Hawizy L, Hook DW. Recategorising research: Mapping from FoR 2008 to FoR 2020 in Dimensions. Quant Sci Stud. 2023 Mar 1;4(1):127–43.

6. Grootendorst M. KeyBERT: Minimal keyword extraction with BERT. [Internet]. Zenodo; 2020. Available from: https://doi.org/10.5281/zenodo.4461265

7. Austin CP. The impact of the completed human genome sequence on the development of novel therapeutics for human disease. Annu Rev Med. 2004;55:1–13.

8. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020 Jan;577(7792):706–10.

9. amueller/word_cloud: A little word cloud generator in Python [Internet]. [cited 2024 Jul 31]. Available from: https://github.com/amueller/word_cloud

10. googlemaps/google-maps-services-python [Internet]. Google Maps Platform; 2024 [cited 2024 Jul 31]. Available from: https://github.com/googlemaps/google-maps-services-python

11. Megginson D. davidmegginson/ourairports-data [Internet]. 2024 [cited 2024 Jul 31]. Available from: https://github.com/davidmegginson/ourairports-data

12. Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Maria Calò C, De Montis A, et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nat Commun. 2014 Apr 29;5:3513.

13. Greenhouse gas reporting: conversion factors 2022 [Internet]. GOV.UK. 2022 [cited 2024 Jul 31]. Available from: https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2022

14. Ritchie H, Roser M. Which form of transport has the smallest carbon footprint? Our World Data [Internet]. 2024 Mar 18 [cited 2024 Jul 31]; Available from: https://ourworldindata.org/travel-carbon-footprint

15. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019 Mar 26;9(1):5233.

16. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008 Oct;2008(10):P10008.

17. Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. [Internet]. Stanford InfoLab; 1999 Nov. Report No.: 1999–66. Available from: http://ilpubs.stanford.edu:8090/422/

18. Teich EG, Kim JZ, Lynn CW, Simon SC, Klishin AA, Szymula KP, et al. Citation inequity and gendered citation practices in contemporary physics. Nat Phys. 2022 Oct;18(10):1161–70.

19. Popular Baby Names [Internet]. [cited 2024 Jul 31]. Available from: https://www.ssa.gov/oact/babynames/limits.html

20. Chintalapati R, Laohaprapanon S, Sood G. ethnicolr2: Predict Race and Ethnicity From Name [Internet]. 2023 [cited 2024 Jul 31]. Available from: https://github.com/appeler/ethnicolr2

21. Chintalapati R, Laohaprapanon S, Sood G. Predicting Race and Ethnicity From the Sequence of Characters in a Name [Internet]. arXiv; 2023 [cited 2024 Jul 31]. Available from: http://arxiv.org/abs/1805.02109

22. Australian and New Zealand Standard Research Classification (ANZSRC), 2020 | Australian Bureau of Statistics [Internet]. 2020 [cited 2024 Jul 31]. Available from: https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/latest-release

23. Lee DS, Fahey DW, Skowron A, Allen MR, Burkhardt U, Chen Q, et al. The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. Atmospheric Environ Oxf Engl 1994. 2021 Jan 1;244:117834.

24. Hawaiian Legacy Reforestation Initiative [Internet]. [cited 2024 Jul 31]. Available from: https://legacyforest.org/

25. PSB Trees [Internet]. [cited 2024 Jul 31]. Available from: https://psb.stanford.edu/trees/

26. Brück O. A bibliometric analysis of the gender gap in the authorship of leading medical journals. Commun Med. 2023 Dec 11;3(1):1–7.

27. Krstacic JE, Carr BM, Yaligar AR, Kuruvilla AS, Helali JS, Saragossi J, et al. Academic medicine's glass ceiling: Author's gender in top three medical research journals impacts probability of future publication success. PloS One. 2022;17(4):e0261209.

28. Hart KL, Perlis RH. Trends in Proportion of Women as Authors of Medical Journal Articles, 2008-2018. JAMA Intern Med. 2019 Sep 1;179(9):1285–7.

29. Bernardi K, Lyons NB, Huang L, Holihan JL, Olavarria OA, Martin AC, et al. Gender Disparity in Authorship of Peer-Reviewed Medical Publications. Am J Med Sci. 2020 Nov;360(5):511–6.

30. NIH Data Book [Internet]. [cited 2024 Jul 31]. Available from: https://report.nih.gov/nihdatabook/report/218

31. Underrepresented Racial and Ethnic Groups | Diversity in Extramural Programs [Internet]. [cited 2024 Jul 31]. Available from: https://extramural-diversity.nih.gov/diversity-matters/underrepresented-groups

32. NIH Data Book - Data by Race [Internet]. [cited 2024 Jul 31]. Available from: https://report.nih.gov/nihdatabook/report/306

CHARTING THE EVOLUTION AND TRANSFORMATIVE IMPACT OF THE PACIFIC SYMPOSIUM ON BIOCOMPUTING THROUGH A 30-YEAR RETROSPECTIVE ANALYSIS OF COLLABORATIVE NETWORKS AND THEMES USING MODERN COMPUTATIONAL TOOLS

Leah Zhang*
Thomas Jefferson High School for Science & Technology
Alexandria, VA, USA
Email: 2025lzhang@tjhsst.edu

Sameeksha Garg*,
Department of Computer Science, Carnegie Mellon University
Pittsburgh, PA USA
Email: sameeksg@andrew.cmu.edu

Edward Zhang*, Sean McOsker, Carly Bobak, Kristine Giffin, Brock Christensen
Dartmouth College Geisel School of Medicine
Hanover, NH USA
Email: edward.b.zhang.27@dartmouth.edu, Sean.A.McOsker.GR@dartmouth.edu , carly.a.bobak@dartmouth.edu , kristine.a.giffin@dartmouth.edu , brock.c.christensen@dartmouth.edu

Joshua Levy**
Department of Computational Biomedicine, Cedars Sinai Medical Center
Los Angeles, CA USA
Email: joshua.levy@cshs.org

Founded nearly 30 years ago, the Pacific Symposium on Biocomputing (PSB) has continually promoted collaborative research in computational biology, annually highlighting emergent themes that reflect the expanding interdisciplinary nature of the field. This study aimed to explore the collaborative and thematic dynamics at PSB using topic modeling and network analysis methods. We identified 14 central topics that have characterized the discourse at PSB over the past three decades. Our findings demonstrate significant trends in topic relevance, with a growing emphasis on machine learning and integrative analyses. We observed not only an expanding nexus of collaboration but also PSB's crucial role in fostering interdisciplinary collaborations. It remains unclear, however, whether the shift towards interdisciplinarity was driven by the conference itself, external academic trends, or broader societal shifts towards integrated research approaches. Future applications of next-generation analytical methods may offer deeper insights into these dynamics. Additionally, we have developed a web application that leverages retrieval augmented generation and large language models, enabling users to efficiently explore past PSB proceedings.

Keywords: natural language processing, network analysis, Pacific Symposium on Biocomputing, topic modeling, interdisciplinary collaboration

* Denotes equal contribution as co-first authors.
** To whom correspondence should be addressed. This work is supported by DoD grant PR220927 and NIH P30CA023108 support for JL.

1. Introduction

The Pacific Symposium on Biocomputing (PSB) was co-founded in 1996 by Dr. Teri Klein, Dr. Lawrence Hunter, and Sharon Surles, originating from the Biotechnology Computing Tracks at the Hawaiian International Conference on System Sciences [1–5]. Initially, PSB aimed to provide a platform for pioneering work in databases, algorithms, interfaces, visualization, modeling, and other computational methods applied to the challenges of molecular biology. As an annual multidisciplinary scientific conference held in Hawaii, it has continuously fostered international collaboration in computational biology.

Over the past 30 years, PSB has undergone significant evolution. Each year, the conference themes are curated to address emerging and critical issues in biocomputing, driven by proposals from leading researchers in new areas. This dynamic approach, unique among scientific gatherings, makes PSB an ideal subject for examining the progression of research themes over time, thereby reflecting the evolving landscape of computational biology.

Attending PSB offers significant opportunities for career advancement, professional development, and networking. These conferences are essential for discussing cutting-edge scientific themes and acquiring new knowledge [6,7]. Beyond immediate academic and professional benefits, attendees gain exposure to new technologies and methodologies that can be implemented in their own practices and institutions. Therefore, understanding the academic impact of such conferences is crucial for appreciating their role in advancing science and practice.

Over the past thirty years, PSB has witnessed transformative changes in biocomputing. This period has seen the rise of artificial intelligence in medicine [8], the sequencing of the human genome, and advancements in precision health. Innovations such as multimodal, single-cell [9], and spatial analyses, along with vast bioimaging datasets, have revolutionized our approach to biological data. Concurrent advancements in computing speed, storage capacity, GPUs, and internet connectivity have further enabled these scientific breakthroughs.

In 1996, PSB manuscripts and presentations focused on the foundational aspects of computational biology [4]. In contrast, by 2024, the focus has shifted towards leveraging complex integrations of multimodal data and advanced computational techniques [2]. The upcoming 30th anniversary of PSB presents a prime opportunity to reflect on the evolution of research themes, highlighting the growth in collaboration and scientific impact within the community.

To comprehensively understand these developments, we have conducted a quantitative retrospective analysis of the entire history of PSB proceedings. This study spans numerous articles and abstracts presented at PSB, providing insights into the dynamic nature of biocomputing as a field. By employing advanced computational tools for this meta-analysis, we aim to elucidate the intricate patterns of research evolution, collaboration networks, and thematic shifts over the conference's history. This analysis not only underscores the importance of PSB in shaping the field but also demonstrates the power of computational methods in understanding scientific progress.

2. Methods

2.1. *Overview*

Inspired by a similar work analyzing conference themes and impact over 30 years [10], our analysis utilizes topic modeling, large language models (LLM) and network analysis to map out:

1. **Topic Modeling:** The main themes of PSB, their prevalence and evolution over time.
2. **Evolving Co-Authorship Networks:** The personal impact of participation in the symposium, examining how PSB has spurred the formation of new, transdisciplinary collaborations.
3. **Citation Networks:** The scientific impact of PSB themes, as evidenced by citation metrics, broken down by topic and reported independently.
4. **Interactive Dashboard for Perusing Prior Proceedings:** The development of a Retrieval Augmented Generation (RAG) tool as an interactive research tool for rapid access of past proceedings.

Readers can find the code used for data preprocessing and analysis as well as instructions for deploying our interactive PSB dashboard at the following GitHub repositories: https://github.com/Leahie/PSBmodel

2.2. *Examining Evolving PSB Themes through Count-Based and Neural Topic Modeling*

2.2.1. *Extraction of Proceedings Text*

We used the Beautiful Soup package to web scrape PDFs of all PSB conference proceedings, available at https://psb.stanford.edu/psb-online/ [11,12]. Each year's proceedings included documents ranging from session introductions, short abstracts, workshops, and full peer reviewed papers. Only peer reviewed papers, from the years 1996-2024, with viable linked PDF files were downloaded and parsed. Due to inconsistencies in web formatting, separate web scrapers were developed for years 1996, 1997, 1998-2001, and 2002-2024. Document parsing for all proceedings led to the extraction of information such as the link of the pdf, title of authors, for each manuscript.

2.2.2. *Text Preprocessing*

After the PDFs were downloaded, pdfplumber was used to extract the text from each manuscript [13]. A custom text processor was developed to remove accents, special figures, numbers, stopwords, extra whitespace, and words less than 2 letters. After this step, contractions were expanded, and text was converted to lowercase.

Further text processing enhanced the readability of the documents. The appearance of section numbers and words such as "abstract", "introduction", "references" — words typically found in conference proceedings and part of the PSB manuscript template — helped filter PDFs that were poorly converted to text– these same subsections were used to divide the document into three components which were subsequently saved: 1) abstract, 2) main body, and 3) references. The main body of the document included all text between the "Introduction" and "Reference" headers.

2.2.3. *Topic Modeling with LDA, BERTopic, and DTM*

After preprocessing the text, we employed three primary methods to identify and model emerging themes: Latent Dirichlet Allocation (LDA), Dynamic Topic Modeling (DTM) and BERTopic [14–16]. These techniques focused exclusively on the main body of texts spanning from 1996 to 2024, allowing for a precise analysis of words and phrases that characterize the themes and topics of the documents. By applying these methods, we were able to ascertain the prevalence of each topic across various manuscripts and authors at specific time points. This approach facilitated a detailed

exploration of the evolving landscape of themes throughout the study period, offering insights into the dynamics of topic popularity and relevance over time.

LDA is a generative probabilistic model designed to identify latent topics within a corpus of text documents. LDA assumes that each document is a mixture of topics, and each topic is a distribution of words. By using Dirichlet distributions to guide the selection of topics for each document and words from topics, LDA can effectively capture the prevalence of topics across documents and the frequency of words within topics. The model achieves this by estimating three key components: (1) the topic distribution within each document, (2) the word distribution within each topic, and (3) the topic assignment for each word in the documents. We use the python package *Tomotopy* for our LDA implementation which uses Collapsed Gibbs Sampling, a Markov Chain Monte Carlo (MCMC) method which iteratively samples the conditional distributions of latent variables allowing the model to estimate the posterior distribution of topics within the corpus[17].

For LDA, words within each topic were initially ranked by the estimated Dirichlet parameters, $\{p_i \text{ for } i = 1,2, \dots, n | \sum p_i = 1\}$. The Dirichlet parameters in our topic modeling methods do not account for the ubiquity of words, which often leads to an oversaturation by commonly used terms. To address this issue, we implemented a re-ranking strategy for these words based on their saliency and relevance, both of which reweight the importance of words by considering their document frequency. Saliency quantifies a word's relative importance by measuring how significantly it contributes to the uniqueness of a topic. Relevance, on the other hand, evaluates a word based on both its probability within a topic and its frequency across documents. This dual metric ensures a balanced assessment that enhances topic interpretability.

Term frequencies were normalized, which served as a foundation for calculating saliency and relevance for each topic. The formulas for calculating Salience, Relevance, and Frequency Normalization are outlined below [18–20]:

$$S_{k,w} = P_{k,w} \log\left(\frac{P_{k,w}}{F'_w}\right)$$

$$R_{k,w} = \lambda \log\left(P_{k,w}\right) + (1-\lambda) \log\left(\frac{P_{k,w}}{F'_w}\right)$$

$$F'_w = \frac{F_w}{\sum_w F_w}$$

Dynamic topic models were utilized alongside standard LDA in our dataset. Unlike LDA, which assumes static topics, dynamic topic models incorporate changes over time by using topic priors from previous time steps to inform the topic distributions at subsequent time steps. This approach allows for the detection of emerging or evolving topics that might otherwise be overlooked by LDA's time-averaged methodology.

BERTopic, proposed by Maarten Grootendorst, is another topic modeling technique that integrates state-of-the-art transformer models such as Bidirectional Encoder Representations from Transformers (BERT). Our BERTopic implementation generates dense sentence-level embeddings which were aggregated across each manuscript to form document-level embeddings which were subsequently clustered to derive coherent topics across documents [16]. By using transformer models like BERT, these contrived embeddings encapsulate contextual relationships between words offering a rich semantic representation of the documents, addressing the limitations of traditional topic modeling which often approaches these texts as a bag of words.

Generated high dimensionality embedding produced by these transformer models are reduced in dimensionality with techniques such as Uniform Manifold Approximation and Projection (UMAP) [21] and subsequently clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [22] which identifies dense regions in the embeddings space and groups documents together without a need for a preset amount of clusters. After clustering the embeddings, BERTopic extracts the most representative words for each cluster by ranking them using the Class Based Term Frequency-Inverse Document Frequency (c-TF-IDF). C-TF-IDF is calculated by taking the logarithm of one plus the average number of words per class divided by the frequency of word across all classes. The term frequency emphasizes words that are more frequent and the inverse document frequency captures rarely used but still important words.

$$w_{x,c} = \|tf_{x,c}\| + \log\left(1 + \frac{A}{f_x}\right)$$

The optimal number of topics for each topic modeling method was determined using the coherence metric [23], which measures the semantic similarity between high scoring words within each topic. This metric helps ensure that the topics generated are meaningful and interpretable. We utilized the coherence scores to select the number of topics that provided the highest level of interpretability while maintaining a balance with model complexity.

2.2.4. *Characterizing Topic Prevalence over Time*

To streamline the interpretation process, we opted to restrict our analysis to LDA models which did not initially account for the temporal dynamics of each topic's evolution. This approach simplifies the initial modeling by focusing solely on prevalence of thematic content without the additional complexity of temporal variation in topic content. After training, we extracted document-topic distributions for each paper, which represent the proportion of each topic within each document. These distributions were then aligned with the corresponding dates of publication or timestamps. To capture temporal trends, we computed the average topic distribution for each defined time period.

To identify overarching patterns in the evolution of topic prevalence over time, we employed K-Means clustering via the *tslearn* python package [24]. This method utilized a dynamic time warping (DTW) distance matrix of the time series data [25]. DTW is particularly adept at capturing similarities in temporal sequences, even when there are shifts or timing differences among the sequences. By applying K-Means clustering to this DTW distance matrix [26], we were able to discern and illustrate the predominant trends and shifts in topic prevalence throughout the corpus.

2.3. *Evaluating the Influence of Collaborative Networks at PSB on Research Themes*

2.3.1. *Extraction and Fuzzy Matching of Author Names*

Author names for each manuscript were extracted from the proceedings website for each year, and a database of these titles and names was established. To ensure unique identification, we employed a combination of citation analysis and relied on the Scopus database of authors. Each paper was mapped to its unique DOI and PubMed ID using CrossRef's REST API (https://api.crossref.org/swagger-ui/index.html) and MetaPub (https://pypi.org/project/metapub/) [27,28]. Then, each identifier was looked up using Pybliometrics, a python-based wrapper for the Scopus API [29]. Using Pybliometrics, each paper was mapped to its authors and each author was mapped to their Scopus ID, a unique identifier assigned to them by Scopus. This approach allowed us to account for variations in spelling and other inconsistencies that commonly occur in author name listings. By using citation data, we were able to link each paper to a unique identifier and link variations of a name to a single author.

2.3.2. *Development of Collaborative Networks over Time*

Collaborative networks were constructed annually based on co-authorships (edges) within articles published that year [30]. The attributes of each node (representing an author) were defined by the average topic distribution from Latent Dirichlet Allocation (LDA), specifically averaged across the manuscripts the authors contributed to within PSB that year. Each network represented a cross-sectional snapshot at a specific point in time, typically characterized by sparse connections due to its annual limitation.

To gain a deeper understanding of the evolving collaborative landscape, we extended our analysis to include cumulative networks. In this approach, nodes and edges from previous years were incorporated into the current year's network. This method allowed us to observe not only isolated annual interactions but also the development and persistence of collaborative ties over time.

2.3.3. *Overall Measures of Interdisciplinarity and Collaboration*

In our study, we focused on characterizing authors' topical areas of interest by analyzing their cumulative topic distributions. These distributions were derived from the topic-document matrices of all their prior publications at PSB up to but not including the current evaluation point. We hypothesized that alignment in these topical areas might influence the likelihood of future collaborations, and that this influence could vary over time.

To empirically test this hypothesis, we calculated the cosine similarity between the topical distributions of two authors, each aggregated from prior years. Cosine similarity measures the cosine of the angle between two vectors in a multidimensional space, serving as an indicator of how aligned two authors are in their prior topics of interest. To assess the potential for these authors to form a collaborative connection (or 'edge'), we employed a logistic regression model that includes an interaction term with time, using R v4.3:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 \times \text{similarity}(d_i, d_j) + \beta_2 \times t + \beta_3 \times (\text{similarity}(d_i, d_j) \times t)$$

where p_{ij} is the probability of forming an edge between authors *i* and *j*, $\text{similarity}(d_i, d_j)$ is the cosine similarity score between their prior topic distributions, and t represents the year of the

collaboration relative to the study period. This model not only quantifies the relationship between topical alignment and formation of collaborative links but also how this relationship evolves over time, permitting a dynamic analysis of factors influencing collaboration within the PSB community. Results were also stratified by the number of prior joint publications within a co-author dyad.

Furthermore, each author's ability to bridge across diverse topics was quantified using an entropy score, calculated at each timepoint, reflecting the variety and distribution of topics in their publications to that point. This score served as an indicator of an author's interdisciplinarity, suggesting their potential to contribute to and collaborate across various thematic areas.

Finally, an author's influence at each timepoint was quantified using various network centrality measures, including degree centrality, eigenvector centrality, and betweenness centrality [31]. Degree centrality measures the number of direct connections an author has, indicating their immediate influence within the network. Eigenvector centrality accounts for the influence of an author's connections, reflecting how connected they are to other highly connected authors. Betweenness centrality highlights authors who serve as bridges between different clusters or groups within the network, showcasing their role in facilitating information flow. Centrality measures were normalized based on the size of the connected component (subgraph) to which each node belongs.

As a descriptor of overall network dynamics, the final cumulative network for 2024 was analyzed using the Leiden algorithm [32]. This approach partitions the network into clusters based on the strength of the connections, ensuring that clusters are more connected internally than with other parts of the network. Each cluster was then labeled based on averaged topic distribution to that point, providing a thematic summary that reflects the predominant scholarly interests of each subgroup.

2.4. *Measuring Scholastic Impact through Citations*

Finally, the impact of PSB papers was characterized by analyzing the number of citations each paper received. For each topic identified by LDA analysis, now assigned to individual papers, we calculated the average number of citations both overall and across different time periods. This approach enabled us to determine which topics garnered the most attention and influence within the scholarly community, while accounting for the publication dates of the articles. Measures of interdisciplinarity and collaboration (*2.3.3*) were correlated with citation counts (independent variable) using linear regression modeling, adjusting for time as a covariate. The analysis was restricted to the 2005-2019 period to allow sufficient time for collaborations/topics to develop and to mitigate potential biases from lower citation counts associated with more recent publications.

2.5. *Developing Interactive Dashboard to Facilitate Review of Papers*

2.5.1. *Developing Retrieval Augmented Generation Approach*

Retrieval Augmented Generation (RAG) enhances the capabilities of large language models (LLMs) by incorporating a preliminary reference to a knowledge base before generating responses. This method is particularly beneficial when applying LLMs to specialized or highly specific domains that are not well-represented in the model's initial training data [33]. For efficient querying of PSB manuscripts, this involves augmenting the user query with a relevancy search within a vector database that contains embeddings of the knowledge base, addressing common issues such as inaccuracies or the generation of irrelevant content by the LLM.

Our RAG setup for analyzing PSB documents was implemented using LangChain [34]. Initially, papers were downloaded in PDF format and segmented into chunks of approximately 1000 words each. These segments were then transformed into vector embeddings using OpenAI's "text-embedding-3-small" model and stored within a vector database managed by Chroma. For each user query, the LangChain Merger Retriever searches this database to find and retrieve the most relevant embeddings, which are then provided as context to the LLM through the RunnablePassThrough function. This process ensures that the generated responses are both accurate and contextually relevant to the specific queries related to PSB documents.

2.5.2. *Web Application and Availability*

To facilitate user interaction with our RAG setup, we developed a web application using Streamlit [35]. This application provides a user-friendly interface for querying the PSB document database and viewing the augmented responses. The web application is accessible at https://psb-rag.streamlit.app, and the complete codebase for the RAG workflow and further reference to the application is available for public review and use on our GitHub repository. To utilize the site, users will need to provide an OpenAI API key.

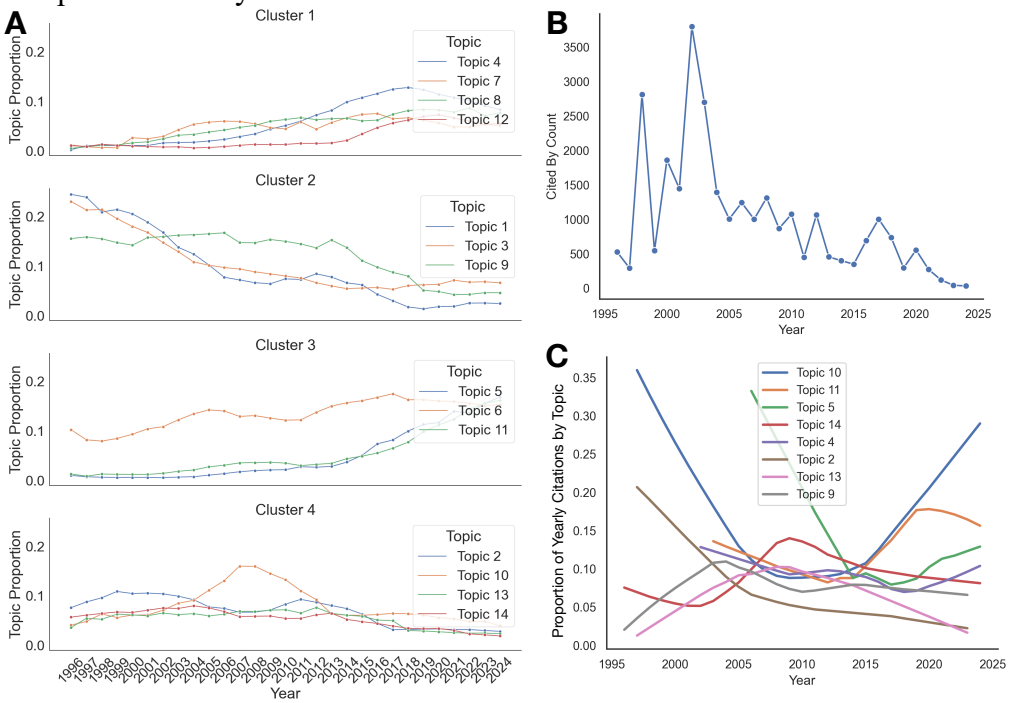


Figure 1: Topic and Citation Dynamics at PSB, 1996 to 2024: A) Prevalence of topics over time, highlighting evolving interests in specific research areas, grouped into four clusters to enhance readability. **B)** tracks the total cumulative citations of PSB publications broken down by uear, with a notable peak in the early 2000s. **C)** Proportion of yearly citations by select topics, indicating shifts in the impact of various research themes over three decades.

3. Results

3.1. *Topic Modeling Results*

Topic modeling was optimized using coherence metrics to ascertain the most interpretable number of topics for Latent Dirichlet Allocation (LDA), BERTopic, and dynamic topic models (DTM). This

approach identified 14 distinct topics using LDA, 13 with DTM, and 19 with BERTopic. Detailed topic-word distributions for all models are available in the supplementary materials hosted on our GitHub repository. These results (including LDA relevance metrics) are summarized in **Tables 1 and 2**, providing a direct comparison of the outputs from the topic modeling techniques, with complete parameters found in the supplementary. LDA topics were clustered based on their prevalence over time (**Figure 1**). While both LDA and BERTopic underwent thorough analysis, the LDA results demonstrated higher coherence, with less overlap between topics compared to BERTopic, where topics tended to show more redundancy. As a result, discussions in our main text have primarily focused on the LDA topics.

Table 1: Comparative Overview of Topic Keywords in LDA and BERTopic Models

LDA		BERTopic	
Topic	**Words**	**Topic**	**Words**
1	cancer cell tumor pathway samples cells survival pathways sub breast	**1**	protein proteins structure residues sequence structures binding set function two
2	drug drugs harm disease knowledge diseases relationships sources target meta	**2**	snps snp disease genetic population plo association gene genotype allele
3	reads peak rate sites posterior peaks read site likelihood mass	**3**	terms gene information ontology text system term used one database
4	sequences rna dna regions mutation mutations genome disordered base disorder	**4**	gene genes expression regulatory transcription network set binding motif time
5	interactions interaction proteins functional cluster clusters similarity clustering networks ppi	**5**	patient patients health clinical medical features set models using time
6	features performance learning training feature prediction classification fier trained models	**6**	cancer gene genes mutations sub tumor expression cell drug samples
7	snp snps plo genotype population allele variants populations locus genetic	**7**	network networks time graph that system state nodes are pathway
8	algorithm tree problem size optimal matrix probability proceedings trim let	**8**	tree trees taxa species number distance gene genomes two algorithm
9	text terms ontology query database relations system name language concepts	**9**	drug drugs target similarity compounds targets network based set chemical
10	this from which can each all have not our die	**10**	imaging brain features age subjects cognitive disease mci poe feature
11	residues binding structure structures energy residue motif amino motifs surface	**11**	sequence dna sequences coding length domain reads genome system gene
12	state reactions reaction activity compounds metabolic enzyme transcription molecules metabolites	**12**	cell cells immune spatial expression seq gene single crna tumor
13	clinical patients risk age health patient phenotypes causal cohort was	**13**	cancer features set mirna layer genes feature gene samples cluster
14	user software flow work visualization tools field file environment science	**14**	virus cov viral sequences protein human hiv sars proteins host
		15	drug harm gene relationships drugs xml genes heritable text ontology
		16	rac species rrna biome communities micro microbial diversity subgroup coa
		17	rna rnas sequence structure secondary sequences reads alignments base sci
		18	peptide mass spectra peak peaks peptides spectrum ion teo intensity
		19	alignment alignments sequences scoring score gap optimal length sequence path

Table 2: Evolution of Topics in Dynamic Topic Model: Progression of keywords for selected topics across three years—1996, 2010, and 2024—using DTM, illustrating the shift in research focus, such as Topic 1 evolving from basic molecular structures to complex cancer drug models. All years can be found in the supplementary material.

Topic	1996: Words	2010: Words	2024: Words
1	proton system proteins structure length molecular you	time algorithm sub interactions system class well	different performance samples sub models cancer drug
2	structure sequences molecular given site proteins solvent	problem algorithm shown networks different interactions function	table clustering cell patients disease samples values
3	time structure system molecular class structures points	graph size patterns algorithm different state are	values samples features patients models data disease
4	function sequences points course table system time	binding clustering sub different class algorithm rna	samples different use cell models values transcript
5	tree surface however sequences structure different proton	interactions clustering time structure table state algorithm	learning clinical data features predicted performance brain
6	would system surface proteins point sequences pair	possible however structure different nodes given time	data models drug age patients brain learning
7	structure system given information distance students tree	well clustering state base time algorithm harm	across patients use disease studies models graph
8	surface system sequences points structure given time	different class state size algorithm base function	data patients table features learning cell values
9	given system molecular grape die residues proteins	nodes different algorithm structure table state states	learning time drug transcript use data values
10	algorithm point database structure molecular system given	first drug use sub time class are	age effect features studies performance models brain
11	points structure time students system point sequences	interactions shown time drug size algorithm different	feature individuals performance ancestry samples spatial models
12	system would site second point die pair	nodes size class sub algorithm time samples	clinical table ancestry patients age across training
13	site system you tree value time could	drug hee different rna residues off rees	samples clinical table disease patients across clustering

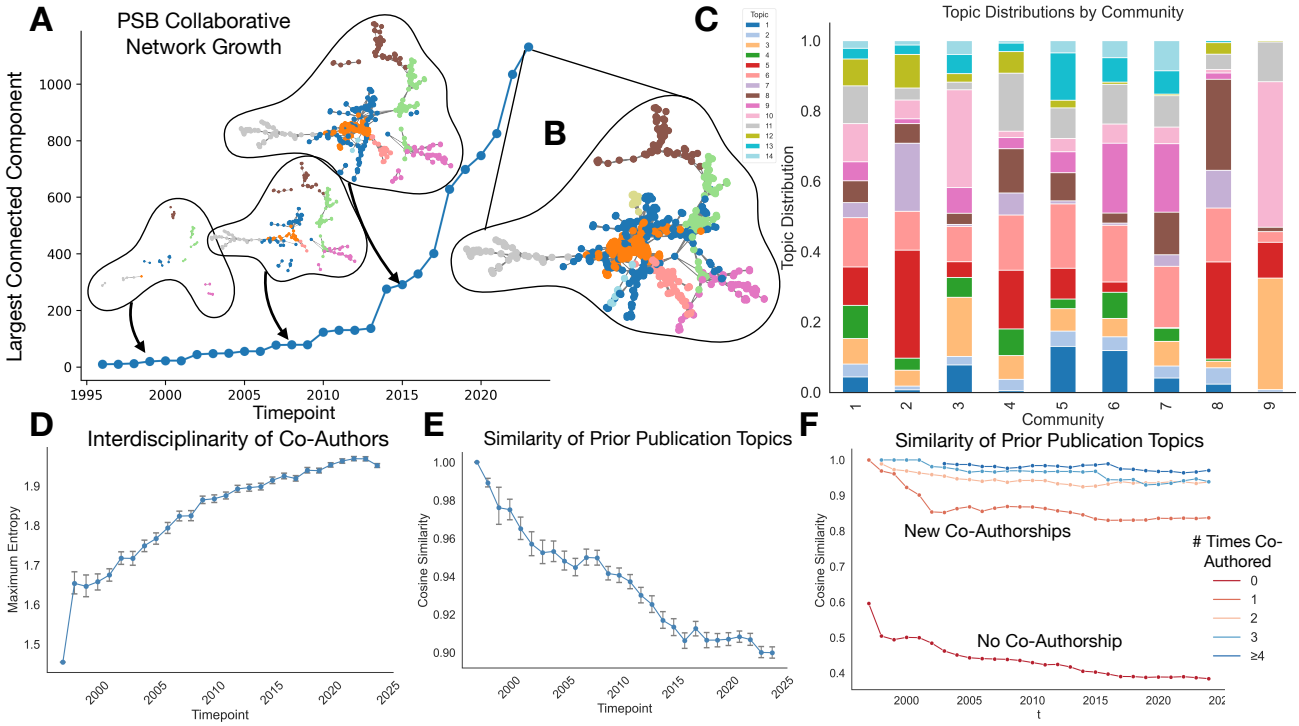


Figure 2: Analysis of Collaboration Dynamics: A) Growth of the largest connected component within the PSB collaboration network from 1996 to 2024, **B)** Visualization of the final 2024 collaborative network, with authors labeled by assigned community via the Leiden algorithm, **C)** Cumulative distribution of topics within each community, reflecting the thematic focus areas that have emerged among collaborators, **D)** Increasing interdisciplinary nature of co-authorships over time as denoted through maximal entropy of prior years' topic distribution within subsequent co-author dyads; **E)** Declining trend in cosine similarity of prior publication topics among subsequent co-author dyads; **F)** Analyzes the relationship between the frequency of co-authorship and topic similarity, showing that more frequent collaborators tend to share more similar research interests from prior years' topic similarity, while first-time collaborators often engage from more diverse thematic backgrounds with collaborators

3.2. *Collaboration Network Results*

Our analysis mapped the growth of the largest connected component in the collaboration network over time as an indicator of collaboration intensity (**Figure 2A**). Initially, in 1996, the largest group comprised of 9 co-authors. By 2003, this number had grown to 45. Significant growth occurred in 2011 and 2015, with the largest connected components increasing from 79 to 123 in 2011, and from 136 to 278 in 2015, respectively. By 2019, the component had expanded to 632 members, reaching 1147 by 2024—nearly one-third of the entire network size of 3932 PSB authors.

The resulting network was divided into 9 distinct communities, reflecting unique topical focuses as determined by average topic-document distributions among community members (**Figure 2B,C**).

Our analysis highlighted substantial shifts in the nature of collaborative ties within the PSB network. To quantify the diversity of topics present within collaborations, we calculated entropy

measurements for each co-author based on their topic distributions prior to the year of publication. These entropy values provided a numerical representation of the thematic diversity within each collaboration, illustrating the broadening scope of interdisciplinary interaction over time (**Figure 2D**). There was a gradual increase in the diversity of topics involved in collaborations, with entropy values rising steadily from the year 2000 onwards (β=0.01, $p<0.001$). This suggests that researchers are increasingly engaging in collaborations that cross traditional disciplinary boundaries.

Cosine similarity was used to assess the thematic alignment between collaborating authors over time based on prior years' aggregate topic distributions. Initially high similarity scores in the early years of the symposium have gradually decreased, suggesting that over time, collaborators are less likely to share a common research focus before co-authoring together (β=-2.8e-3, $p<0.001$). This trend is pronounced among new collaborations, where cosine similarity scores dropped by nearly 20% from 2000 to 2024, reflecting a broadening of interdisciplinary interaction (**Figure 2E,F**). Despite the decrease over time in topic alignment, prior years' topic alignment was positively associated with the likelihood of co-authorship (OR=1.8e6, $p<0.001$) and co-authors who continued to publish together maintained higher levels of topic alignment (β=0.03, $p<0.001$).

Centrality measures were computed yearly to identify key individuals within the final 2024 cumulative PSB collaboration network. These measures pinpointed those who were central in connecting larger subnetworks, reflecting their pivotal roles in fostering collaboration (**Table 3**).

Table 3: Key Influencers in the PSB Network Across Different Years, influence determined using weighted eigenvector, betweenness and degree centrality

Timepoint	Eigenvector	Betweenness	Degree
1999	Toshihisa Takagi	Subramanian Subbiah	Satoru Kuhara
	Satoru Kuhara	A. Keith Dunker	Toshihisa Takagi
	Emiko Furuichi	Satoru Kuhara	Adam Godzik
2004	Satoru Miyano	Satoru Miyano	Satoru Miyano
	David C. Kulp	Philip E. Bourne	Satoru Kuhara
	Conrad C. Huang	Adam Godzik	William Stafford Noble
2009	Bart L.R. de Moor	Satoru Miyano	Russ B. Altman
	Conrad C. Huang	Russ B. Altman	Philip E. Bourne
	Thomas E. Ferrin	Philip E. Bourne	William Stafford Noble
2014	Russ B. Altman	Marylyn D. Ritchie	Adam Godzik
	Philip E. Bourne	Russ B. Altman	Russ B. Altman
	Zoubin Ghahramani	Satoru Miyano	Philip E. Bourne
2019	Marylyn D. Ritchie	Marylyn D. Ritchie	Russ B. Altman
	Sarah A. Pendergrass	Sarah A. Pendergrass	Atul Janardhan Butte
	Shefali Setia Verma	Russ B. Altman	Jason H. Moore
2024	Marylyn D. Ritchie	Marylyn D. Ritchie	Russ B. Altman
	Shefali Setia Verma	Russ B. Altman	Lawrence E. Hunter
	Sarah A. Pendergrass	Shefali Setia Verma	Joel T. Dudley

3.3. *Citation Results*

The manuscripts published in the yearly PSB proceedings have significantly varied in their impact over time, with a notable peak in citations during the early 2000s. As illustrated in **Figure 1B**, the today's citation count for these papers shows a substantial rise around this period, followed by a gradual decline. This figure traces the number of current citations received by papers based on their publication year and does not normalize by passing time– manuscripts published earlier are more likely to have more citations. After adjusting for time, we found that articles with a higher entropy score (indicating interdisciplinarity; t=3.33, p=0.001) and lower cosine similarity (indicating formation of interdisciplinary relationship; t=-3.06, p=0.002) were associated with higher citation count. **Figure 1C** delineates the proportion of yearly citations attributable to specific topics,

assigning each manuscript the topic with the highest document-topic score. This analysis reveals that certain topics have gained or lost prominence in terms of citation impact over the years.

4. Discussion

4.1. *Topic Modeling Interpretation and Discussion*

The topics derived from BERTopic shared some commonalities with those from LDA, including areas such as pathway analysis, drug-drug polypharmacy interactions, CHIP-seq peak calling, SNPs, sequence alignment, protein-protein interactions, and biomedical ontologies. However, BERTopic covered a broader array of topics, including network analysis, COVID-19, microbiome analysis, brain imaging, spatial transcriptomics, and temporal features, showcasing its expansive thematic reach (**Table 1**). Conversely, LDA uniquely captured topics related to machine learning and residue binding, which were not present in the BERTopic set. Notably, the exclusion of rapidly emerging fields such as multimodal analysis in BERTopic was also observed, highlighting some limitations in its topic coverage. Dynamic topic models provided an evolutionary view of these topics, which were initially based on themes from 1996. Over time, these topics have notably shifted from focusing primarily on biomolecular structures and sequences to more complex areas such as clinical prediction models that integrate spatial data and RNA sequencing prediction models.

Cluster 3 highlights a marked increase in topics such as residue binding and machine learning (specifically topics 5, 6, and 7) (**Figure 1A**). The surge in these topics aligns with the rise of deep learning and sophisticated protein folding algorithms, which gained prominence nearly a decade ago [36]. This trend underscores the impact of technological advancements on driving research focus areas within bioinformatics, particularly those that leverage computational innovations.

In contrast, Cluster 2, which includes topics 1, 3, and 9, pertains to pathway analysis and biomedical ontologies. Notably, pathway analysis (topic 1) was a central theme in sessions as far back as 1996, with titles like “Genome, Pathway and Interaction Bioinformatics” and “Computation in Biological Pathways” in 1997 [37,38]. Despite their current popularity, these topics are long-established in the field rather than emerging areas. Over time, the prevalence of these foundational themes has seen a relative decrease, suggesting a shift in research focus toward newer computational techniques and applications.

4.2. *Collaborative Network Discussion and Interpretation*

The identified communities in the largest connected component from the 2024 network and their differing topic distributions highlight the symposium’s role in facilitating diverse interdisciplinary collaborations (**Figure 2**). Our results show a marked shift towards interdisciplinary collaboration at the PSB, as evidenced by increasing entropy in topic distributions and decreasing cosine similarity over time among collaborators. This evolving trend suggests that PSB participants are not only expanding their collaborative networks but are also engaging with a wider array of scientific disciplines than in previous years. The decrease in cosine similarity particularly highlights how the nature of these collaborations has evolved from close-knit, topic-specific interactions to more diverse, interdisciplinary exchanges. This shift may reflect broader changes in the field of bioinformatics, where cross-disciplinary approaches are becoming essential to tackle increasingly complex research questions [39–41].

The trend of decreasing topic similarity, especially notable among first-time collaborators, indicates that PSB is successfully fostering an environment where researchers feel encouraged to explore new collaborations outside their immediate expertise. This is crucial for driving innovation and adapting to the rapidly changing landscape of bioinformatics research. The data also suggest that while established collaborators continue to work within familiar thematic areas, there is a strong movement towards branching out into new topics.

Over time, the composition of influential members within the PSB network has evolved (**Table 3**), with recent years marking the rise of key figures, including three current editors/organizers. Their prominence might stem from consistent presence, increasing opportunities for co-authorship. While this could indicate a strategic integration of leadership roles, it might also reflect incidental outcomes of sustained participation. This observation underscores the complexities of interpreting the dynamics between leadership presence and collaborative patterns in academic networks.

4.3. *Citation Discussion and Interpretation*

It was not surprising that earlier PSB publications, especially those from around the year 2000, received more attention, as reflected by the number of cumulative citations. Our citation analysis also revealed a declining trend in the citation relevance of certain topics. For instance, LDA topic 2, which focuses on drug-drug interactions, and topic 9, covering ontologies, were highly cited in the early 2000s but have experienced a gradual decrease in citation percentage over the years. In contrast, topic 11 on protein folding has seen a noticeable increase in popularity.

The future trajectory of less frequently cited topics remains uncertain as the field evolves with new technologies. The process of these topics becoming mainstream could significantly alter their impact. Additionally, shifts in community focus—from established scholars to emerging researchers—may also influence citation patterns. The growing interdisciplinarity of the field presents another challenge, as works that span multiple disciplines sometimes struggle to connect with a well-defined audience, potentially diluting their impact [42]. Nevertheless, our citation analysis suggests that forming interdisciplinary ties, as fostered through this venue, was associated with greater scientific impact, even after adjusting for time.

5. Conclusion

The Pacific Symposium on Biocomputing stands as a premier venue in bioinformatics, embodying the forefront of convergent thinking by bringing together individuals from diverse backgrounds to address complex problems that span multiple disciplines. Through our application of quantitative NLP and network analysis methods, we have effectively mapped the scope and nature of the various themes and collaborative ties that have formed at this venue over the past 30 years. These analyses reveal not only the evolving patterns of collaboration but also highlight the increasing diversity and interdisciplinarity of the research presented at PSB. Looking ahead, we anticipate that PSB will continue to foster groundbreaking interdisciplinary research, adapting to new scientific challenges and technologies. As the field grows, the symposium will likely play a crucial role in shaping future trends in bioinformatics and computational biology. We expect that continued innovations in analytical methods will further illuminate the dynamics of collaboration and influence within this community, enhancing our understanding of how interdisciplinary interactions drive scientific progress.

References

1. Pacific Symposium on Biocomputing [Internet]. Wikipedia. 2022 [cited 2024 Jul 31]. Available from: https://en.wikipedia.org/w/index.php?title=Pacific_Symposium_on_Biocomputing&oldid=1123233788
2. Altman RB, Hunter L, Ritchie MD, Murray T, Klein TE. Pacific Symposium on Biocomputing 2024. Biocomputing 2024. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2023.
3. Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray T, Klein TE. Biocomputing 2020: Proceedings of the Pacific Symposium [Internet]. WORLD SCIENTIFIC; 2020 [cited 2019 Nov 28]. Available from: https://www.worldscientific.com/worldscibooks/10.1142/11698
4. Hunter L, Klein TE. Pacific Symposium on Biocomputing'96: Hawaii, USA, 3-6 January, 1996 [Internet]. World Scientific; 1995 [cited 2024 Jul 31]. Available from: https://books.google.com/books?hl=en&lr=&id=20soDwAAQBAJ&oi=fnd&pg=PR7&dq=pacific+symposium+on+biocomputing&ots=EjZk6sRoiC&sig=JkZyQ7UgrWd3vZ6j0xJkegwG2xM
5. Altman RB, Hunter L, Klein TE, Murray T, Dunker AK, Ritchie MD. Biocomputing 2021:Proceedings of the Pacific Symposium [Internet]. 2020 [cited 2024 Jul 31]. Available from: https://directory.doabooks.org/handle/20.500.12854/42151
6. Hauss K. What are the social and scientific benefits of participating at academic conferences? Insights from a survey among doctoral students and postdocs in Germany. Res Eval. 2020 Aug 27;rvaa018. PMCID: PMC7499794
7. Augustine EF, Steele SJ, McIntosh S, Sugarwala L, White RJ, Yousefi-Nooraie R, Zand MS, Ossip DJ. Utilizing the Un-Meeting model to advance innovative translational and team science. J Clin Transl Sci. 7(1):e176. PMCID: PMC10514683
8. Daneshjou R, Brenner SE, Chen JH, Crawford DC, Finlayson SG, Kidziński Ł, Bulyk ML. Precision Medicine: Using Artificial Intelligence to Improve Diagnostics and Healthcare. Biocomputing 2022 [Internet]. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2021 [cited 2024 Jul 31]. p. 223–230. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789811250477_0021
9. Garmire LX, Yuan GC, Fan R, Yeo GW, Quackenbush J. SINGLE CELL ANALYSIS, WHAT IS IN THE FUTURE? Biocomputing 2019 [Internet]. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2018 [cited 2024 Jul 31]. p. 332–337. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789813279827_0030
10. Peek N, Combi C, Marin R, Bellazzi R. Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. Artificial Intelligence in Medicine. 2015 Sep 1;65(1):61–73.
11. Hajba GL. Website Scraping with Python: Using BeautifulSoup and Scrapy [Internet]. Berkeley, CA: Apress; 2018 [cited 2024 Jul 31]. Available from: http://link.springer.com/10.1007/978-1-4842-3925-4
12. Patel JM. Web Scraping in Python Using Beautiful Soup Library. Getting Structured Data from the Internet [Internet]. Berkeley, CA: Apress; 2020 [cited 2024 Jul 31]. p. 31–84. Available from: http://link.springer.com/10.1007/978-1-4842-6576-5_2

13. S.V J. pdfplumber (Version 0.8.0) [Internet]. 2020. Available from: https://github.com/jsvine/pdfplumber
14. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3(Jan):993–1022. PMID: 36346659
15. Blei DM, Lafferty JD. Dynamic topic models. Proceedings of the 23rd international conference on Machine learning - ICML '06 [Internet]. Pittsburgh, Pennsylvania: ACM Press; 2006 [cited 2024 Jul 31]. p. 113–120. Available from: http://portal.acm.org/citation.cfm?doid=1143844.1143859
16. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [Internet]. arXiv; 2022 [cited 2024 Jul 31]. Available from: http://arxiv.org/abs/2203.05794
17. Chauhan U, Shah A. Topic Modeling Using Latent Dirichlet allocation: A Survey. ACM Comput Surv. 2021 Sep 17;54(7):145:1-145:35.
18. Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces [Internet]. 2014 [cited 2024 Jul 31]. p. 63–70. Available from: https://aclanthology.org/W14-3110.pdf
19. Chuang J, Manning CD, Heer J. Termite: visualization techniques for assessing textual topic models. Proceedings of the International Working Conference on Advanced Visual Interfaces [Internet]. Capri Island Italy: ACM; 2012 [cited 2024 Jul 31]. p. 74–77. Available from: https://dl.acm.org/doi/10.1145/2254556.2254572
20. Fan A, Doshi-Velez F, Miratrix L. Assessing topic model relevance: Evaluation and informative priors. Statistical Analysis. 2019 Jun;12(3):210–222.
21. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018 Sep 2;3(29):861. PMID: 33588368
22. Campello RJGB, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G, editors. Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg; 2013. p. 160–172.
23. Röder M, Both A, Hinneburg A. Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining [Internet]. Shanghai China: ACM; 2015 [cited 2024 Jul 31]. p. 399–408. Available from: https://dl.acm.org/doi/10.1145/2684822.2685324
24. Tavenard R, Faouzi J, Vandewiele G, Divo F, Androz G, Holtz C, Payne M, Yurchak R, Rußwurm M, Kolar K. Tslearn, a machine learning toolkit for time series data. Journal of machine learning research. 2020;21(118):1–6.
25. Dynamic Time Warping. Information Retrieval for Music and Motion [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007 [cited 2024 Jul 31]. p. 69–84. Available from: http://link.springer.com/10.1007/978-3-540-74048-3_4
26. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability [Internet]. Oakland, CA, USA; 1967 [cited 2024 Jul 31]. p. 281–297. Available from: https://www.google.com/books?hl=en&lr=&id=IC4Ku_7dBFUC&oi=fnd&pg=PA281&dq=+Some+Methods+for+classification+and+Analysis+of+Multivariate+Observations&ots=nQUdG-L8oP&sig=Z6DngAT2EpGdYnq-cXK49tVTDPI
27. Hendricks G, Tkaczyk D, Lin J, Feeney P. Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …; 2020;1(1):414–427.

28. Metapub [Internet]. [cited 2024 Jul 31]. Available from: https://metapub.org/
29. Rose ME, Kitchin JR. pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. SoftwareX. Elsevier; 2019;10:100263.
30. Liu X, Bollen J, Nelson ML, Van de Sompel H. Co-authorship networks in the digital library research community. Information processing & management. Elsevier; 2005;41(6):1462–1480.
31. Borgatti SP. Centrality and network flow. Social networks. Elsevier; 2005;27(1):55–71.
32. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports. Nature Publishing Group; 2019;9(1):1–12.
33. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W tau, Rocktäschel T. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459–9474.
34. Topsakal O, Akinci TC. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. International Conference on Applied Engineering and Natural Sciences [Internet]. 2023 [cited 2024 Jul 31]. p. 1050–1056. Available from: https://www.researchgate.net/profile/Oguzhan-Topsakal/publication/372669736_Creating_Large_Language_Model_Applications_Utilizing_LangChain_A_Primer_on_Developing_LLM_Apps_Fast/links/64d114a840a524707ba4a419/Creating-Large-Language-Model-Applications-Utilizing-LangChain-A-Primer-on-Developing-LLM-Apps-Fast.pdf
35. Khorasani M, Abdou M, Hernández Fernández J. Web Application Development with Streamlit: Develop and Deploy Secure and Scalable Web Applications to the Cloud Using a Pure Python Framework [Internet]. Berkeley, CA: Apress; 2022 [cited 2024 Jul 31]. Available from: https://link.springer.com/10.1007/978-1-4842-8111-6
36. Fidelis K, Grudinin S. Session introduction: AI-driven Advances in Modeling of Protein Structure. Biocomputing 2022 [Internet]. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2021 [cited 2024 Jul 31]. p. 1–9. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789811250477_0001
37. Karp P, Romero PR, Neumann E. GENOME, PATHWAY AND INTERACTIONS BIOINFORMATICS. Pacific Symposium on Biocomputing [Internet]. World Scientific; 2002 [cited 2024 Jul 31]. p. 398–399. Available from: http://psb.stanford.edu/psb-online/proceedings/psb03/intro-path.doc
38. Karp PD, Riley M. Session on Computation in Biological Pathways. Biocomputing'97-Proceedings Of The Pacific Symposium [Internet]. World Scientific; 1996 [cited 2024 Jul 31]. p. 18. Available from: https://books.google.com/books?hl=en&lr=&id=bEBPDwAAQBAJ&oi=fnd&pg=PA18&dq=%22Computation+in+Biological+Pathways%22&ots=Z30I-4GwvN&sig=8fE0E70A2sb7FN9qTQjxjCW5oSY
39. Powell WW, White DR, Koput KW, Owen-Smith J. Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences. American Journal of Sociology. The University of Chicago Press; 2005 Jan;110(4):1132–1205.
40. Romano P, Giugno R, Pulvirenti A. Tools and collaborative environments for bioinformatics research. Briefings in Bioinformatics. 2011 Nov 1;12(6):549–561.

41. Exploratory Analysis of Topic Interests and Their Evolution in Bioinformatics Research Using Semantic Text Mining and Probabilistic Topic Modeling | IEEE Journals & Magazine | IEEE Xplore [Internet]. [cited 2024 Jul 31]. Available from: https://ieeexplore.ieee.org/document/9738599
42. Yegros-Yegros A, Rafols I, D'Este P. Does Interdisciplinary Research Lead to Higher Citation Impact? The Different Effect of Proximal and Distal Interdisciplinarity. [cited 2024 Jul 31]; Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0135095

AI and Machine Learning in Clinical Medicine: Generative and Interactive Systems at the Human-Machine Interface

Fateme Nateghi Haredasht
Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
Email: fnateghi@stanford.edu

Dokyoon Kim
Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA
Email: dokyoon.kim@pennmedicine.upenn.edu

Joseph D. Romano
Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA
Email: joseph.romano@pennmedicine.upenn.edu

Geoff Tison
Center for Biosignal Research, University of California, San Francisco San Francisco, CA, USA
Email: geoff.tison@ucsf.edu

Roxana Daneshjou
Department of Biomedical Data Science, Stanford University Stanford, CA, USA
Email: roxanad@stanford.edu

Jonathan H. Chen
Department of Medicine and Center for Biomedical Informatics Research, Stanford University Stanford, CA, USA
Email: jonc101@stanford.edu

Artificial Intelligence (AI) technologies are increasingly capable of processing complex and multi-layered datasets. Innovations in generative AI and deep learning have notably enhanced the extraction of insights from both unstructured texts, images, and structured data alike. These breakthroughs in AI technology have spurred a wave of research in the medical field, leading to the

creation of a variety of tools aimed at improving clinical decision-making, patient monitoring, image analysis, and emergency response systems. However, thorough research is essential to fully understand the broader impact and potential consequences of deploying AI within the healthcare sector.

Keywords: Artificial Intelligence, clinical medicine, decision support systems, large language models.

1. Introduction

The integration of Artificial Intelligence (AI) into clinical medicine continues to expand at a rapid pace, promising transformative changes across diagnostics, treatment planning, and patient monitoring [1], [2], [3], [4], [5]. While AI technologies offer remarkable capabilities in analyzing extensive and complex medical data sets, their real-world application necessitates robust frameworks that support explainability and generalizability. These aspects are crucial for building trust among clinicians and patients alike, ensuring that AI-driven interventions are both understandable and applicable across diverse clinical environments. This session showcases pioneering research that addresses these needs, highlighting innovative solutions that aim to set new standards in the deployment of AI tools in medicine. The emergence of large language models (LLMs) and other sophisticated AI systems has propelled forward our ability to interpret and utilize medical data, promising significant improvements in patient outcomes [6], [7]. However, the deployment of such technologies must be accompanied by stringent evaluations to confirm their effectiveness and safety in real-world clinical settings [8], [9]. This includes understanding their potential biases, operational limitations, and their overall impact on clinical decision-making processes.

This year's session at the 2025 Pacific Symposium on Biocomputing (PSB), titled AI in Clinical Medicine: Towards Explainable and Generalizable AI Systems, concentrates on the latest advancements in AI that not only enhance clinical effectiveness but also prioritize transparency, adaptability, and ethical implementation in healthcare settings. Here, we highlight the accepted submissions for this session and set the stage for a discussion of AI’s role in revolutionizing medical practice, emphasizing the need for solutions that are not only technically proficient but also ethically sound and universally beneficial. As AI continues to permeate the healthcare landscape, this session provides a critical examination of both its achievements and the challenges that lie ahead in its journey from experimental algorithms to essential clinical tools.

2. Artificial Intelligence in Clinical Medicine

2.1. *AI for Clinical Decision Support and Medical Workflows*

AI has increasingly been integrated into clinical decision-making processes, providing support for tasks such as diagnostics, treatment planning, and risk prediction [10]. Decision support systems that incorporate AI can process vast amounts of clinical data in real time, offering clinicians enhanced insights into patient care [11]. These tools are particularly valuable in settings where time and precision are critical, such as emergency departments, oncology, and intensive care units.

Prince et al. 2025 present a visual analytics framework aimed at evaluating interactive AI systems in pediatric brain tumor diagnosis [12]. Their work underscores the importance of understanding how AI can support clinicians in decision-making and improve clinical workflows.
Bedi et al. 2025 introduce QUEST-AI, an innovative LLM-based system designed to generate and refine USMLE-style exam questions [13]. By automating this process, their system promises to reduce the time and cost involved in medical education.
Rao et al. 2025 tackle the challenge of error generation in radiology reports with CX-REGen, a system that uses LLMs to create synthetic errors in chest X-ray reports to improve AI model training [14].
Healey et al. 2025 introduce LLM-CGM, a benchmark for summarizing continuous glucose monitor (CGM) data using AI, with applications in enhancing diabetes management [15].
Lastly, Godeme et al. 2025 investigate the use of synthetic text for developing NLP models to support peer supporters [16]. This study demonstrates that AI-generated synthetic text can effectively augment training datasets, which enhances the fidelity of peer support tools. Their findings emphasize the utility of AI in improving both training and support mechanisms for peer-assisted health interventions.

2.2. *Improving AI Models for Critical Healthcare Tasks*

To ensure that AI models can handle the complexity and variability of healthcare data, improving their generalizability and performance is crucial. Healthcare environments are diverse, and AI models must perform well across different populations, institutions, and clinical settings. Additionally, AI tools need to handle both structured data, such as lab results and vital signs, and unstructured data, like clinical notes, to offer comprehensive support [17], [18]. Several papers in this session focus on optimizing the performance and utility of AI models in healthcare.
Shashikumar and Nemati 2025 present a comparative study of LLMs in sepsis prediction, demonstrating that smaller models can achieve performance levels comparable to larger ones, thus offering more resource-efficient solutions [19].
Wang et al. 2025 explore the use of LLMs in cancer registry coding, where AI models are applied to streamline and enhance the accuracy of reporting in real-world hospital settings [20].
Weissenbacher et al. 2025 developed an NLP-based system to evaluate the appropriateness of pediatric antibiotic prescriptions, contributing to improved antibiotic stewardship practices in healthcare [21].

2.3. *Ethical and Regulatory Considerations in AI Deployment*

The rapid development of AI in healthcare presents significant ethical and regulatory challenges. AI systems must be designed and deployed in ways that prioritize patient safety, privacy, and equity. As AI becomes more embedded in clinical workflows, it is critical to ensure that these systems comply with existing regulations and adapt to evolving legal frameworks [22]. Regulatory agencies, such as the FDA, are tasked with ensuring that AI technologies are safe, effective, and accessible to all patients [23].

Rincon et al. 2025 explore the evolving regulatory landscape in healthcare AI, focusing on how recent Supreme Court decisions could impact the authority of regulatory agencies like the FDA [24]. Their analysis highlights the potential implications of regulatory uncertainty for the healthcare industry.
In a related study, Levy et al. 2025 investigate the use of AI to predict suicide risk in veterans, integrating both structured and unstructured EHR data [25]. Their work underscores the ethical importance of using AI responsibly in sensitive areas like mental health.

2.4. *Generalizability and Validation of AI Models*

Ensuring that AI models are generalizable and can be validated across different clinical settings is essential for their widespread adoption [26]. AI systems trained on data from a single institution or population often struggle to perform well in other settings due to variations in patient demographics, clinical practices, and data collection methods. Generalizability is crucial for developing AI tools that can be deployed in diverse healthcare environments without compromising accuracy or fairness [27].
In this session, Banerjee et al. 2025 introduce a multi-site validation framework to test the robustness of radiology AI models across different populations and institutions, addressing the critical need for AI systems that can generalize beyond their training data [28].
Xiong et al. 2025 propose i-MedRAG, an iterative Retrieval-Augmented Generation (RAG) system designed to improve medical question-answering by incorporating follow-up queries, further enhancing AI's ability to handle complex clinical cases [29].
Ramwala et al. 2025 present ClinValAI, a cloud-based framework for the external validation of AI models in medical imaging, ensuring that these models meet high standards of performance and fairness [30].
Lastly, Keat et al. 2025 introduce PGxQA, a resource for evaluating the performance of LLMs on pharmacogenomic question-answering tasks [31]. This benchmark is designed to assess the ability of LLMs to provide clinically accurate information related to pharmacogenomics, which is crucial for ensuring these AI tools' safety and effectiveness when used in personalized medicine applications.

3. Conclusion

The papers presented in this session demonstrate the expanding role of AI in clinical medicine. They showcase a range of applications designed to improve diagnostic accuracy, enhance decision support, and address ethical and regulatory challenges. As AI continues to integrate into healthcare, the need for rigorous validation, regulatory oversight, and ethical deployment becomes increasingly important. These contributions highlight the promise of AI while addressing the ongoing challenges of ensuring that these systems are safe, explainable, and generalizable across diverse clinical environments.

References

[1] M. Khalifa and M. Albadawy, "Artificial Intelligence for Clinical Prediction: Exploring Key Domains and Essential Functions," *Comput. Methods Programs Biomed. Update*, vol. 5, p. 100148, Jan. 2024, doi: 10.1016/j.cmpbup.2024.100148.

[2] S. Bindra and R. Jain, "Artificial intelligence in medical science: a review," *Ir. J. Med. Sci. 1971* -, vol. 193, no. 3, pp. 1419–1429, Jun. 2024, doi: 10.1007/s11845-023-03570-9.

[3] L. Weidener and M. Fischer, "Artificial Intelligence in Medicine: Cross-Sectional Study Among Medical Students on Application, Education, and Ethical Aspects," *JMIR Med. Educ.*, vol. 10, no. 1, p. e51247, Jan. 2024, doi: 10.2196/51247.

[4] H. Akhlaghi *et al.*, "Machine learning in clinical practice: Evaluation of an artificial intelligence tool after implementation," *Emerg. Med. Australas.*, vol. 36, no. 1, pp. 118–124, 2024, doi: 10.1111/1742-6723.14325.

[5] V. D. Karalis, "The Integration of Artificial Intelligence into Clinical Practice," *Appl. Biosci.*, vol. 3, no. 1, Art. no. 1, Mar. 2024, doi: 10.3390/applbiosci3010002.

[6] S. Reddy, "Evaluating large language models for use in healthcare: A framework for translational value assessment," *Inform. Med. Unlocked*, vol. 41, p. 101304, Jan. 2023, doi: 10.1016/j.imu.2023.101304.

[7] E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review," *Diagn. Pathol.*, vol. 19, no. 1, p. 43, Feb. 2024, doi: 10.1186/s13000-024-01464-7.

[8] C. T.-T. Chang *et al.*, "Red Teaming Large Language Models in Medicine: Real-World Insights on Model Behavior," *medRxiv*, pp. 2024–04, 2024.

[9] S. Gottlieb and L. Silvis, "How to Safely Integrate Large Language Models Into Health Care," *JAMA Health Forum*, vol. 4, no. 9, p. e233909, Sep. 2023, doi: 10.1001/jamahealthforum.2023.3909.

[10] A. F. Hernandez and C. J. Lindsell, "The Future of Clinical Trials: Artificial to Augmented to Applied Intelligence," *JAMA*, vol. 330, no. 21, pp. 2061–2063, Dec. 2023, doi: 10.1001/jama.2023.23822.

[11] R. Khera *et al.*, "AI in Medicine—JAMA's Focus on Clinical Outcomes, Patient-Centered Care, Quality, and Equity," *JAMA*, vol. 330, no. 9, pp. 818–820, Sep. 2023, doi: 10.1001/jama.2023.15481.

[12] E. W. Prince, T. C. Hankinson, and C. Gorg, "A Visual Analytics Framework for Assessing Interactive AI for Clinical Decision Support," presented at the Pacific Symposium on Biocomputing (PSB)., 2025.

[13] S. Bedi *et al.*, "QUEST-AI: A System for Question Generation, Verification, and Refinement using AI for USMLE-Style Exams," 2025. doi: 10.1101/2023.04.25.23288588.

[14] V. M. Rao, S. Zhang, J. Acosta, and P. Rajpurkar, "CX-REGen: Generating Complete, Consistent, and Clinically Relevant Errors Within Chest X-Ray Reports," presented at the Pacific Symposium on Biocomputing (PSB)., 2025.

[15] E. Healey and I. Kohane, "LLM-CGM: A Benchmark for Large Language Model-Enabled Summarization of Time Series Data from Wearable Continuous Glucose Monitors for Conversational Diabetes Management," *Pacific Symposium on Biocomputing (PSB).*, 2025.

[16] J. Godeme *et al.*, “Artificial Allies: Validation of Synthetic Text for Peer Support Tools through Data Augmentation in NLP Model Development,” *Pacific Symposium on Biocomputing (PSB).*, 2025.
[17] W. S. Shah, Z. Elkhwesky, K. M. Jasim, E. F. Y. Elkhwesky, and F. F. Y. Elkhwesky, “Artificial intelligence in healthcare services: past, present and future research directions,” *Rev. Manag. Sci.*, vol. 18, no. 3, pp. 941–963, Mar. 2024, doi: 10.1007/s11846-023-00699-w.
[18] I. Ghebrehiwet, N. Zaki, R. Damseh, and M. S. Mohamad, “Revolutionizing personalized medicine with generative AI: a systematic review,” *Artif. Intell. Rev.*, vol. 57, no. 5, p. 128, Apr. 2024, doi: 10.1007/s10462-024-10768-5.
[19] S. P. Shashikumar and S. Nemati, “A Prospective Comparison of Large Language Models for Early Prediction of Sepsis,” . *Pacific Symposium on Biocomputing (PSB).*, 2025.
[20] C.-K. Wang, M.-S. Huang, Y.-H. Yang, V. S. Tseng, and H.-J. Dai, “Using Large Language Models for Efficient Cancer Registry Coding in the Real Hospital Setting: A Feasibility Study,” *Pacific Symposium on Biocomputing (PSB).*, 2025.
[21] D. Weissenbacher *et al.*, “Automated Evaluation of Antibiotic Prescribing Guideline Concordance in Pediatric Sinusitis Clinical Notes,” 2025. doi: 10.1101/2024.08.09.24311714.
[22] G. Karimian, E. Petelos, and S. M. A. A. Evers, “The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review,” *AI Ethics*, vol. 2, no. 4, pp. 539–551, Nov. 2022, doi: 10.1007/s43681-021-00131-7.
[23] T. Gundersen and K. Bærøe, “The Future Ethics of Artificial Intelligence in Medicine: Making Sense of Collaborative Models,” *Sci. Eng. Ethics*, vol. 28, no. 2, p. 17, Apr. 2022, doi: 10.1007/s11948-022-00369-2.
[24] N. Rincon, S. Gerke, and J. K. Wagner, “Implications of An Evolving Regulatory Landscape on the Development of AI and ML in Medicine,” presented at the Pacific Symposium on Biocomputing (PSB)., 2025.
[25] J. Levy, M. Dimambro, A. Diallo, J. Gui, B. Shiner, and M. Levis, “Investigating the Differential Impact of Psychosocial Factors by Patient Characteristics and Demographics on Veteran Suicide Risk Through Machine Learning Extraction of Cross- Modal Interactions,” presented at the Pacific Symposium on Biocomputing (PSB)., 2025.
[26] J. Yang, A. A. S. Soltan, and D. A. Clifton, “Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening,” *Npj Digit. Med.*, vol. 5, no. 1, pp. 1–8, Jun. 2022, doi: 10.1038/s41746-022-00614-9.
[27] M. D. Abràmoff *et al.*, “Considerations for addressing bias in artificial intelligence for health equity,” *Npj Digit. Med.*, vol. 6, no. 1, pp. 1–7, Sep. 2023, doi: 10.1038/s41746-023-00913-9.
[28] O. Banerjee *et al.*, “A Multi-Site Framework for Validating Automatic Metrics in Radiology Generative AI,” *Pacific Symposium on Biocomputing (PSB).*, 2025.
[29] G. Xiong, Q. Jin, X. Wang, M. Zhang, Z. Lu, and A. Zhang, “Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions,” *Pacific Symposium on Biocomputing (PSB).*, 2025.
[30] O. A. Ramwala *et al.*, “ClinValAI: A framework for developing Cloud-based infrastructures for the External Clinical Validation of AI in Medical Imaging,” presented at the Pacific Symposium on Biocomputing (PSB)., 2025.

[31] K. Keat *et al.*, “PGxQA: A Resource for Evaluating LLM Performance for Pharmacogenomic QA Tasks,” presented at the Pacific Symposium on Biocomputing (PSB)., 2025.

A Visual Analytics Framework for Assessing Interactive AI for Clinical Decision Support

Eric W. Prince† and Todd C. Hankinson

Department of Neurosurgery, University of Colorado Anschutz Medical Campus
Aurora, Colorado 80045, USA
†*E-mail: Eric.Prince@CUAnschutz.edu*

Carsten Görg

Department of Biostatistics and Informatics, Colorado School of Public Health
Aurora, Colorado 80045, USA

Human involvement remains critical in most instances of clinical decision-making. Recent advances in AI and machine learning opened the door for designing, implementing, and translating interactive AI systems to support clinicians in decision-making. Assessing the impact and implications of such systems on patient care and clinical workflows requires in-depth studies. Conducting evaluation studies of AI-supported interactive systems to support decision-making in clinical settings is challenging and time-consuming. These studies involve carefully collecting, analyzing, and interpreting quantitative and qualitative data to assess the performance of the underlying AI-supported system, its impact on the human decision-making process, and the implications for patient care. We have previously developed a toolkit for designing and implementing clinical AI software so that it can be subjected to an application-based evaluation. Here, we present a visual analytics framework for analyzing and interpreting the data collected during such an evaluation process. Our framework supports identifying subgroups of users and patients based on their characteristics, detecting outliers among them, and providing evidence to ensure adherence to regulatory guidelines. We used early-stage clinical AI regulatory guidelines to drive the system design, implemented multiple-factor analysis and hierarchical clustering as exemplary analysis tools, and provided interactive visualizations to explore and interpret results. We demonstrate the effectiveness of our framework through a case study to evaluate a prototype AI-based clinical decision-support system for diagnosing pediatric brain tumors.

Keywords: Clinical Decision Making; AI-Supported Interactive Decision Making; Evaluation Studies; Visual Analytics Framework.

1. Introduction

Artificial Intelligence (AI) can transform healthcare decision-making by quickly analyzing large amounts of data and improving diagnostic accuracy and patient outcomes.[1] However, ethical and legal implications, transparency of AI algorithms, and integration into existing workflows present challenges that require careful management.[1,2] Although AI has been increasingly used to support decision-making across various fields, more studies are needed to safely and

efficiently enhance human judgment and interpretation. Achievieng this goal requires the evaluation of AI systems on their algorithmic performance and their impact on humanistic aspects.

A comprehensive and systematic approach is needed to assess the impact of AI on decision making, particularly in high-risk settings such as healthcare.[3,4] For example, a recent study demonstrated that GPT-4V frequently presents flawed medical rationales in cases where it makes the correct final choices regarding the interpretation of radiologic imaging.[5] Examples of clinical experts' interactions with AI systems[6,7] reveal a gap in understanding AI's impact on humanistic aspects of clinical decision-making.

This gap extends to developing objective and precise techniques to evaluate AI technologies' safety and predictive precision.[5] Such evaluation techniques are still a bottleneck in the translational pipeline from a prototype tool to clinical deployment.[8] Our research highlights the substantial challenges related to implementing AI in high-risk decision-making scenarios within healthcare. We present robust and scalable exploratory analysis methods for evaluating AI systems and facilitating their broader acceptance and implementation in healthcare decision-making.

Monitoring clinical AI software effectively ensures performance, compliance, innovation, and better patient outcomes through data analysis and personalized medicine. Our proposed framework uses regulatory guidelines and statistical methods to assess system factors. Developing clinical AI software requires a structured framework: defining the problem, collecting and preparing data, developing and evaluating the model, and implementing and monitoring it. This comprehensive approach ensures that the software addresses specific clinical tasks, uses relevant data, integrates into the clinical workflow, and stays up-to-date.

We introduce a scalable framework implemented as an interactive software solution to analyze AI's impact in high-risk clinical decision-making scenarios. Its goals include identifying subgroups, detecting outliers, and supporting compliance with regulations. We integrate methodologies from multiple fields, including factor analysis, hierarchical clustering, adherence to regulatory guidelines, and interactive visualizations, to thoroughly analyze and enhance AI effectiveness in clinical decision-making. An end-to-end evaluation framework can enhance healthcare decision-making by improving AI's effectiveness, facilitating its implementation, and promoting adherence to regulatory guidelines, potentially leading to better patient outcomes. We demonstrate the utility and effectiveness of our framework through a case study assessing a prototype AI-based clinical decision-support system for the diagnosis of pediatric brain tumors.

2. Background

AI-assistance for Clinical Interpretation on Radiographic Images of CNS Tumors

As a use case in high-risk clinical decision-making, we look to AI support for diagnosing and managing central nervous system (CNS) tumors. In this context, experts use demographics, clinical presentation, imaging, and molecular information[9] for tumor diagnosis. AI systems can support efficient detection, diagnosis, staging, prognosis, and treatment planning of brain

tumors, among other applications.[9,10] These clincial decisions are only sometimes clear-cut and can require significant resource allocation. It is generally agreed upon that AI has ample room to support clinical decision-making in this context.[6,11,12] However, when considering the humanistic aspects of clinical AI support, it is becoming increasingly apparent that AI has a heterogeneous impact on human decision-makers.[6,11,12] Human experts may exhibit automation bias or neglect, where they overweight and underweight the AI prediction relative to their own, respectively.[6] Therefore, assessing the effect AI assistance has on decision-makers at the system level is essential. It is important to note that although AI has the potential to enhance clinical decision-making significantly, it also brings challenges that need to be addressed. These challenges include data-related issues, digital inequity gaps, bias, and the need for robust governance frameworks that balance safety and innovation.[10]

Consensus Statements and Guidelines for Clinical AI

In the healthcare sector, specific guidelines have been established to rigorously evaluate the clinical impact of AI, ensuring standards for transparency and ethical adherence. These guidelines contrast with those of other sectors, such as finance. Frameworks such as TRIPOD-AI[13] and CONSORT-AI[14] provide structured recommendations for preclinical and clinical AI trials; they emphasize standardized reporting and detailed intervention analysis. The DECIDE-AI[15] guidelines serve a critical role in bridging the preclinical and clinical AI trial phases.

DECIDE-AI targets early-stage clinical evaluations of AI-driven decision-support systems, emphasizing the importance of assessing clinical utility, safety, and ergonomic factors to prepare for broader clinical trials. Developed through international consensus involving experts from diverse areas, these guidelines are pivotal in ensuring that AI technologies are safely and effectively integrated into clinical practices. We used DECIDE-AI to drive the design of our framework, aligning our evaluation methods with best practices for early-stage AI assessment in healthcare.

Visual Analysis of Qualitative Data

The qualitative data analysis software landscape mainly features commercial products, with a notable deficit in advanced open-source options tailored for specialized fields such as clinical AI. Although feature-rich, commercial software like NVivo and ATLAS.ti are expensive and designed for broader use, making them less suitable for niche research areas with limited budgets and cases.

We introduce a new visual data analysis tool designed specifically for early-stage clinical AI evaluations to address this gap. It offers a cost-effective, scalable solution for clinical AI studies, enhancing user-centered evaluations and supporting the development of tailored clinical AI applications.

3. Analytical Objective, Experimental Data, Regulatory Guidelines, and Interface Design

We previously presented a framework for designing, implementing, and evaluating clinical AI tools from an implementation science perspective.[16] Here, we introduce a new framework for

the analysis phase of application-based studies. Specifically, we consider (a) how users can interact with AI systems to make sense of patient data so that they can make effective care decisions and (b) how we monitor these AI systems for safety and efficacy (Figure 1).

We emphasize an exploratory and holistic mindset when interpreting the results of AI evaluation studies. Our framework provides an overview of the data collected in the experiment, complemented by secondary views that can display various facets that detail aspects of the experimental data. We strive for simplicity and efficiency, integrating a minimalistic user interface and implementing linked-view mechanisms for seamless visual filtering. Below, we present the experimental data used to inform our design choices and describe the design of the primary and secondary views.

Following regulatory guidelines, such as DECIDE-AI, to lead analysis is essential when developing AI for clinical decision support. Figure 1 depicts some of these guidelines as black-and-white text boxes. This structured method improves system development, ensures healthcare compliance, and thoroughly evaluates AI integration. It is important for creating efficient, secure systems. Our framework supports identifying personas and patterns in evaluation data and aligns with the DECIDE-AI guidelines.

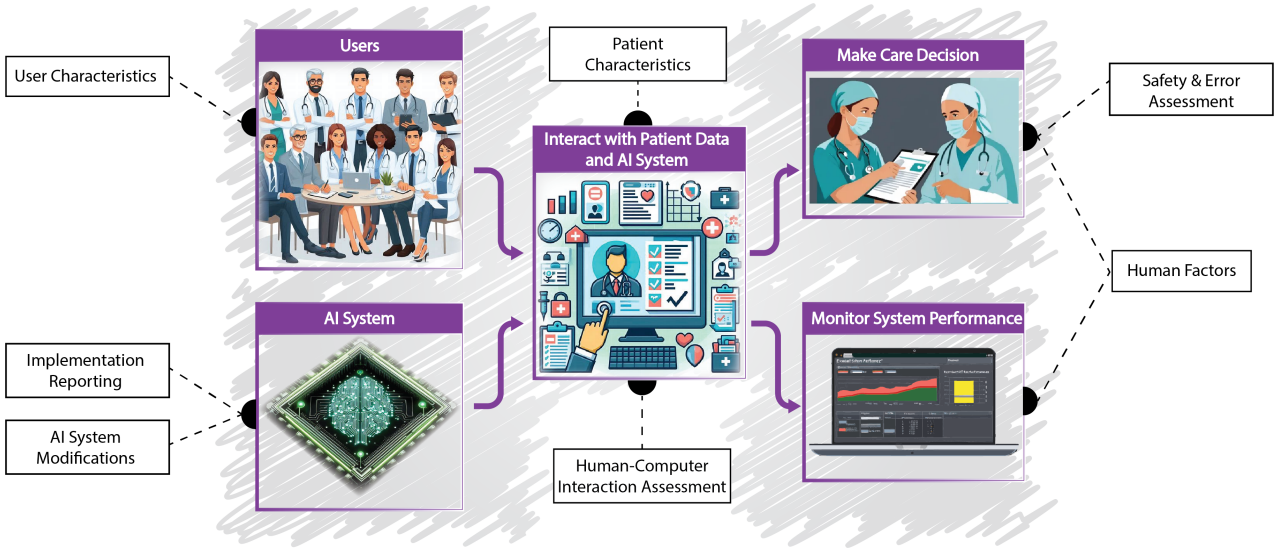


Fig. 1. Conceptual depiction of how users and interactive AI systems come together to make care decisions while monitoring system performance. DECIDE-AI themes guiding clinical AI evaluation are shown in black-and-white text boxes.

Analytical Objective: Identifying Personas and Patterns

Our analytical initiative is focused on coarse but comprehensive data exploration. This task plays a significant role in the initial phases of clinical AI system development to support identifying user personas and patient subgroups, as well as detecting patterns of human-AI agreement and disagreement. This exploration aims to guide the development of AI systems tailored to their users and environments, resulting in more personalized and relevant applica-

tions. During this stage, it is essential to acknowledge the subtle interactions between users and AI systems that can yield valuable insights into refining AI algorithms for optimal performance in real-world clinical settings.

Experimental Data

When evaluating AI's role in clinical decision-making, it is important to adopt a comprehensive approach that considers user-centered design and patient-related data. This approach is centered around the collection of a broad range of categorical and continuous variables describing both the performance of the AI system and the interactions of the user with the system and patient data.

In user-centered design, categorical data (nominal and ordinal) is essential for categorizing and understanding user interactions and experiences. Nominal data can reveal usage patterns and tool preferences, such as user roles (e.g., doctors, nurses, administrators) and AI tool types (e.g., diagnostic aid, treatment planner). Ordinal data, such as user satisfaction ratings or task difficulty levels, can provide insight into the usability and effectiveness of AI tools. Meanwhile, continuous data, including interval and ratio data, provide quantitative user engagement and tool performance measures. Interval data, such as response times or system up-time, and ratio data, such as usage counts, session lengths, or error rates, provide precise metrics to track changes over time or after modifications.

In addition to user-centered data, it is equally important to gather patient-related data. Categorical patient data, such as diagnosis (e.g., Central Nervous System (CNS) tumor), treatment type (e.g., surgery, radiation therapy, chemotherapy), and genetic markers, offer essential insights into the patient's health status and the complexity of their case. Similarly, continuous data points such as tumor size, biomarker levels, and treatment response (e.g., tumor size changes or patient symptoms over time) play a pivotal role in providing precise and quantifiable measures of the patient's condition and treatment progress.

Taking into account both user and patient data, the AI tool can be designed to provide a more holistic and personalized user experience. It can cater to the user's specific tasks, such as diagnosing a CNS tumor or monitoring a patient's response to treatment, thereby enhancing the tool's effectiveness and usability in the clinical setting. This comprehensive data collection and consideration approach is fundamental in the user-centered design and evaluation of AI tools in clinical settings. It ensures that the tool meets the user's needs and improves patient outcomes, which is the ultimate goal of healthcare delivery. Thus, collecting and considering diverse data types is fundamental in evaluating and optimizing AI in a clinical setting. It also aids in mitigating biases and improving the fairness and equity of AI-driven clinical decisions.

Themes of Regulatory Guidelines Driving the Design

To achieve our analytical objective, we lean on the themes and guidelines in the DECIDE-AI framework.[15] Each theme is tailored to glean critical insights during the early phases of AI system development and deployment in healthcare settings. The remainder of this subsection provides a summary of each of these themes.

User Characteristics Analysis. This theme involves collecting and assessing demographic and clinical data from healthcare providers to develop practical AI solutions that meet diverse user needs. This strategy enhances the system's versatility and facilitates its acceptance and integration in clinical settings. The theme aligns with DECIDE-AI guidelines 9A and 9B.

Implementation Reporting. This theme analyzes user interaction with the AI system and its impact on clinical workflows, focusing on user engagement and system acceptance. The goal is to ensure that the AI improves existing workflows and is easily integrated into clinical settings. This theme aligns with DECIDE-AI guidelines 10A and 10B.

AI System Modifications. To maintain the AI system's effectiveness and meet its users' needs, it is imperative to document all modifications made during the study and analyze their impact on the system's outcomes. This theme is essential for the system's continued evolution and clinical efficacy. It corresponds with DECIDE-AI guideline 11.

Human-Computer Interaction Assessment. This theme assesses user agreement and compliance with AI recommendations, focusing on improving trust and system reliability. By analyzing deviations, developers can refine the AI to better meet user expectations and ensure its recommendations are practical for integration into daily operations. This theme aligns with DECIDE-AI guideline 12.

Safety and Error Analysis. This theme focuses on identifying and addressing errors, malfunctions, potential risks, and observed harm in the AI system to safeguard patient safety. Vigilant monitoring and mitigation ensure compliance with healthcare regulations and ethical technology deployment in clinical settings. This theme aligns with DECIDE-AI guidelines 13A and 13B.

Human Factors Analysis. This theme combines usability testing and learning curve evaluations to ensure the AI system is practical and accessible from initial use to complete competence. Meeting practical needs and improving user experiences provides high user adoption and satisfaction and aligns with DECIDE-AI guidelines 14A and 14B.

Due to space constraints, we focus on the themes of Implementation Reporting, Human-Computer Interaction Assessment, and Human Factor Analysis in the remainder of this paper.

Interface Design

This section outlines our interface design, which follows a top-down conceptual approach. The UI is organized into primary and secondary views, creating a light, focused layout that enhances interaction. This hierarchical structure improves user efficiency by providing coarse overviews and allowing granular analysis of selected topics. Interactive views enable dynamic data filtering, enriching the user experience with structured navigation and focused content.

Design of Primary View. Figure 2 shows an example of the primary view, detailed in the case study in Section 4. Our visual analytics framework supports a two-stage, top-down analytical process for handling complex clinical datasets. The initial analysis uses a full-screen plotting window for broad pattern recognition and preliminary insights.

The primary view is configured with interactive functionalities to transition to the finer data exploration phase. These include dynamic linking capabilities between primary and secondary data views and providing contextual information via tooltips. Such features are in-

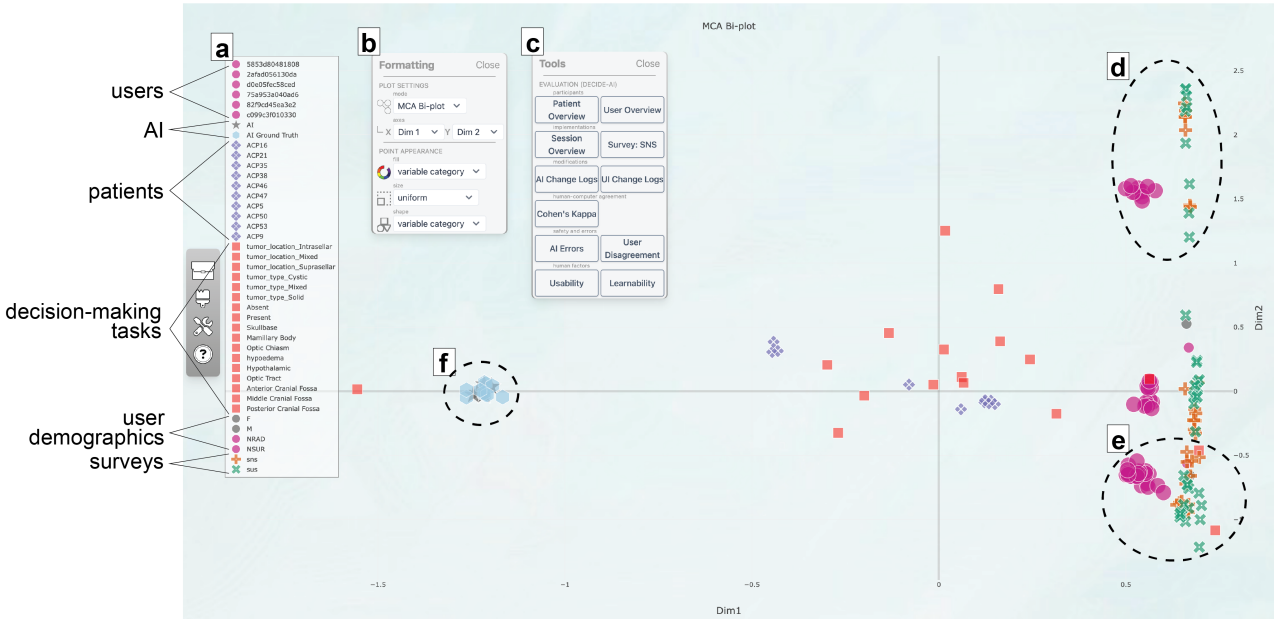


Fig. 2. Overview of the primary view displaying a factor analysis (i.e., MCA) bi-plot of clinical AI user study experimental data. (a) Legend for the plot depicting entities and variables within the study dataset. Entities and variables are annotated based on whether they are a user, AI agent, patient, decision-making task, demographic, or survey value. (b) The Formatting panel controls the display of the primary view and the appearance of marks. Marks are currently double-encoded for color and shape, showing variable categories. (c) The Tools panel contains buttons to toggle secondary views. The buttons are organized according to DECIDE-AI guidelines. (d, e) Examples of participants that represent 2 Personas. (f) Grouping of AI-predicted and AI-ground truth values.

dispensable for users focusing on detailed data inspections, where precision in isolating and scrutinizing data segments is necessary. Interactive tools like zoom, adjustable filters, and data point selection (rectangular or lasso) enhance query specificity, streamline workflows, and deepen analysis.

The minimalist toolbar of the interface, shown adjacent to Figure 2a on the left side, maintains simplicity by housing navigational buttons like project, formatting, tools, and help. This design choice preserves an intuitive navigation structure while supporting extensive functionality, minimizing cognitive load for the analyst.

For this example, our primary view layout is determined using multiple correspondence analysis (MCA). This type of factor analysis optimally suits the assessment of nominal categorical data like surveys. We expand on other factor analyses in our Discussion below. Using a unified graphical interface with dual-coding (glyphs and colors) helps understand the relationships between clinicians' behaviors, patient data, and AI insights.

The configuration of the primary view (Figure 2b), therefore, elevates the analytical capabilities required in clinical settings and aligns with rigorous academic data processing and visualization standards. Designing to meet users' operational and cognitive needs supports nuanced data exploration, which is essential for advancing AI in healthcare and evaluating its impact on clinical decision-making.

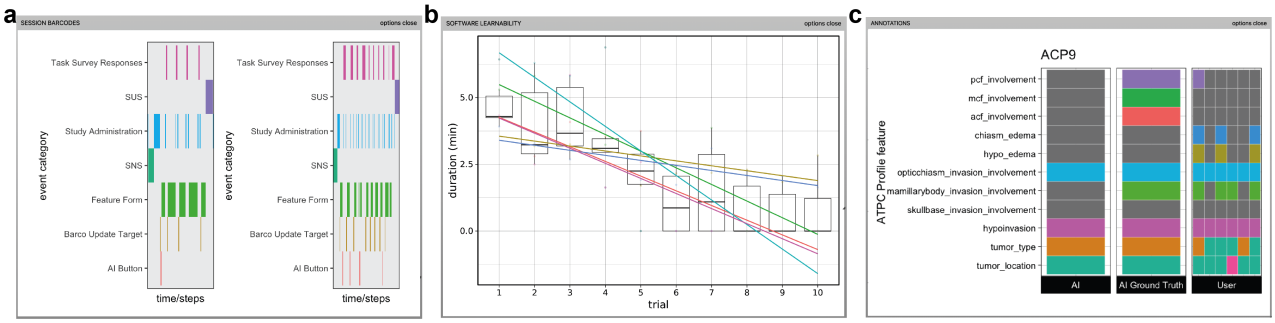


Fig. 3. Comparative Analysis of User Sessions and AI Interactions. (a) This panel illustrates interactions across two distinct sessions, capturing surveys, administrative actions (e.g., 'Next' button clicks), and specialized tasks (e.g., survey responses). The utility of AI features is examined through variations in image viewing (Barco Update Target) and AI button use, indicating differing reliance on AI tools between sessions. (b) This diagram shows a trend of decreasing task completion times, indicating improved user proficiency with system utilization over the session. (c) A heatmap highlights alignment and discrepancies in decision-making annotations between AI predictions, AI ground truth, and user selections. This visualization is instrumental in evaluating the AI's alignment with user decisions and overall influence on decision-making.

Design of Secondary Views. The secondary views in our visual analytics framework are intricately designed to complement the primary view by providing enhanced functionality for detailed and task-specific analysis, as illustrated in Figure 3. These views, which can be triggered from the toolbar menu (Figure 2c) have been developed with particular considerations to support the themes within DECIDE-AI required to effectively evaluate AI interactions within clinical contexts.

The decision to implement popup windows for secondary views is purposeful. It is designed to preserve the primary interface's clarity while enabling access to advanced data inspection when required. This approach allows users to engage with complex data sets without cluttering the primary view, facilitating user-controlled complexity in the visualization environment. Such a design is critical for tasks requiring focused analytical attention on specific data subsets while maintaining sight of the broader analytical context.

The secondary views utilize responsive SVG display widgets, which are pivotal for the dynamic visualization of intricate, multi-dimensional data typical in clinical analytics. These widgets are essential for detailed data relationship analyses, especially for interactions between patient data and AI outputs, as they allow users to interactively manipulate visual elements.

The ability to resize and reposition popup windows empowers users to tailor the analytical workspace to their specific needs or preferences, enhancing the ergonomics of data analysis. This flexibility is essential during analyses such as cross-referencing multiple data sources or adjusting visual layouts to better interpret data correlations and trends. Combining secondary views with the primary view allows for both broad and detailed questions to be addressed simultaneously.

4. Case Study

To demonstrate our framework, we used data from a previously conducted evaluation study of a prototype AI-based clinical decision-support system for the diagnosing pediatric brain tumors. For context, we summarize the system and study here; details are provided in our previous work.[17]

Development and Evaluation Study of Interactive AI Clinical Decision Support Software

User Interface and Radiology Workstation Simulator. We collaborated with clinical partners in a step-by-step design study, collecting visualization samples, conducting user interviews, and improving designs based on feedback. In the initial phase, we used visualizations to display the performance of the AI model. In the second phase, we visualized predictions for clinicians using existing tools and conducted a user study. After immersing ourselves in the clinical environment, we refined our task specifications and created initial prototypes in the third phase. Finally, we built a radiology reading terminal and implemented basic AI interfaces as web applications in the fourth phase. Additional details are provided in our previous publication.[18]

AI Model Backend. The study used the ATPC50 dataset from the Advancing Treatment for Pediatric Craniopharyngioma (ATPC) international multi-institutional consortium in North America, which included information from 50 ACP patients.[17] The study focused on patients' initial presentations, utilizing imaging data from preoperative CT and MRI scans, with radiographic features annotated by a certified neuroradiologist. The AI model thoroughly preprocessed DICOM inputs by resizing images, adjusting contrast, and simulating different patient positions. The data was rescaled to the JPEG range and then processed using ResNet V2 techniques. The study also included using a variational autoencoder for data reconstruction and deep learning classifiers for diagnostic analysis.[17,19]

Experimental Study Design. The study recruited six post-residency faculty attending clinicians (three females and three males) from Children's Hospital Colorado, focusing on those specializing in neurosurgery and neuroradiology. Participants were recruited via email and scheduled for individual 30-minute sessions over a two-week period to accommodate their busy schedules.

At the start of each session, participants shared demographic information, were introduced to the study's goals, and completed the Subjective Numeracy Scale (SNS) survey. They then received a step-by-step guide to the AI decision support tool through ten instructional slides. Participants used radiologic images of CNS tumors to annotate an 11-point feature profile of a pediatric CNS tumor known as Adamantinomatous Craniopharyngioma, both with and without AI support. These feature profiles were completed within the software as a form with checkboxes.

Participants engaged with AI in two forms. The first was a passive AI assistant that flagged a checkbox if the user selected a value that was different from the AI prediction. The second was a direct AI assistant that provided users with the AI-predicted feature profile and a list of other patients that the AI model suggested were similar, based on L1 distance between

prediction vectors. At the end of the session, participants provided feedback by completing the System Usability Scale (SUS) survey. We collected survey data, feature predictions, and system interaction logs.

Exploration and Interpretation of Data from the Evaluation Study

We now describe how our new framework for assessing interactive AI for clinical decision support was used to analyze the data collected in our evaluation study.

Our analysis, which encompassed survey responses and system interaction logs, distinguished two primary user personas: the 'Tech Novice Numerate' (Figure 2d) and the 'Confident Numerate' (Figure 2e). The 'Tech Novice Numerate' users displayed moderate numerical skills but struggled to navigate the AI system, indicating a pressing need for improvements in interface design and enhanced user training. In contrast, the 'Confident Numerate' users, who demonstrated high numerical proficiency, expressed concerns about the system's consistency, suggesting potential reliability and user acceptance issues.

An in-depth examination of the utilization of AI tools revealed significant variances in the degree of dependency on AI support, as observed through differential usage of the "Barco Update" feature for additional image views and "AI button" interactions. Additionally, a chronological analysis of task completions, encompassing SUS, SNS, and Feature Form responses, shed light on the users' learning trajectories and the system's adaptability throughout the session.

In Figure 2f, the overlay of points for AI predicted values and the AI ground truth suggests a high degree of agreement between the AI model's predictions and the annotations made by a board-certified clinical expert, considered the 'ground truth' in this context. This expert is highly skilled and certified in the task at hand within this specific use case.

The fact that the AI model aligns closely with the ground truth annotator indicates that the model has learned to mimic the decision-making process of this particular expert quite accurately. However, it is essential to note that this expert may have interpretations that differ from other experts in the field. This is a common occurrence in many professional fields, including clinical practice, where different experts may have slightly different interpretations or approaches based on their training, experience, and personal biases.

Collaboration among human experts in clinical practice is crucial. Discussing interpretations with colleagues can help reach a consensus or understand different viewpoints, which can help mitigate discrepancies between different human experts. This collaborative approach is particularly important in the context of the figures, as it can help reconcile differences between the annotations that fall into Figure 2f (where the AI and the ground truth annotator agree) versus those in Figure 2d or 2e (where there may be disagreement).

Understanding potential biases in the AI model is essential for evaluating clinical AI devices in real-world settings. If the AI model consistently aligns with one expert (the ground truth annotator, in this case), it may indicate that the model is biased towards that individual's interpretations. These models need to generalize well across different experts and not just mimic the decisions of one individual.

This granular analysis of user-system engagement deepens our understanding of behavioral dynamics and provides actionable insights for targeted enhancements in AI system design and interface. These empirical findings emphasize the critical role of user-centered design in developing intuitive and reliable clinical decision-support tools, enhancing system functionality, and fostering greater user trust and satisfaction in healthcare AI applications.

5. Discussion and Lessons Learned

Integrating AI in decision-making across high-stake sectors underscores a transformative shift towards data-driven practices. However, deploying these AI systems, particularly in sensitive areas such as healthcare care, requires an approach that couples algorithmic insights with indispensable human judgment. The predominant reliance on commercial products often leaves gaps in affordability and customization, especially in specialized fields such as clinical AI. By introducing a visual analytics framework purposely built for clinical AI applications, we propose a solution tailored to meet these unique requirements. Our framework can advance analysis capabilities by interpreting data from clinical user studies and increasing accessibility and practical relevance, reducing dependency on costly and often overly complex tools.

Consideration of the humanistic aspects of clinical AI evaluation is essential for several reasons. Real-world scenarios vary significantly from controlled experiments, making evaluating AI tools with diverse patient populations and varying data quality across discrete clinical tasks to ensure their generalizability. Evaluations help to identify and mitigate biases inherent in healthcare systems, ensuring fairness and equity. Real-world testing is vital in revealing potential safety issues and unintended consequences, guaranteeing that AI tools perform accurately and reliably in clinical settings. Furthermore, realistic evaluations consider how AI integrates into existing workflows, including integration challenges, user experience, and impact on efficiency. Involving clinicians and patients in the evaluation process provides valuable insights into user acceptance, trust, and willingness to adopt AI tools, informing necessary improvements. Finally, adhering to regulatory guidelines, such as DECIDE-AI, significantly enhances the robustness and generalizability of clinical AI tools by emphasizing fundamental principles. These include risk assessment and benefit analysis in real-world contexts, encouraging external validation and independent testing, assessing clinical utility, and promoting transparency through clear documentation.

Basing a clinical AI evaluation method on factor analysis can enhance scalability and accommodate diverse data types. Evaluative efforts for clinical AI systems can generate a large volume of multiple data types. Empirical tools used in this field often involve survey methods that can gather character descriptions of users (e.g., demographics), information about system usability and a way to measure how well users can complete the specific task supported by the system. In addition, continuous numeric data is also relevant in this space with aspects like predictive probabilities from the AI model, system response time, user interaction metrics, and human error rates. Understanding and considering all aspects of the evaluation, including patient data, human expert judgment, and AI software interactions is important. This comprehensive understanding is what produces robust tools that fundamentally improve patient care.

Factor analysis is a method for identifying latent factors, or underlying variables, in observed data. Factor analysis uses the correlation structure amongst observed variables to model fewer unobserved, latent variables known as factors. Researchers use this statistical method when subject-area knowledge suggests that latent factors cause observable variables to covary. For instance, we can evaluate an expert's diagnostic prediction using patient data accessed in software and compare it to the validated diagnosis. The prediction may need to be corrected due to unobservable software interaction patterns, which are observable in factor analysis. By capturing shared variance, it simplifies complex relationships among variables, aiding in the simplification of data analysis. This method is also scalable, enabling the efficient handling of large datasets by reducing dimensionality and making computations more manageable.

Factor analysis can accommodate diverse data types, including continuous and categorical data, allowing for incorporating survey responses (categorical) and continuous data into factor models. For example, Multiple Factor Analysis (MFA) is a multivariate method used to study tables where a group of individuals is described by a set of variables, which can be quantitative and qualitative and are structured in groups. It is an extension of Principal Component Analysis (PCA) for quantitative variables, Multiple Correspondence Analysis for qualitative variables, and Factor Analysis of Mixed Data for variables that belong to both types.

We implemented factor analysis using MCA for our framework because the data from our evaluation study were mainly qualitative. However, many other factor analysis methods are available, such as nonlinear PCA, which handles mixed data types more effectively.[20] The selection of the factor analysis method is flexible, and we will explore this area further in future work to identify more sophisticated representations of this complex experimental context.

Effectively evaluating AI requires a delicate balance between realism and controlled experiments to ensure robustness and practical applicability. Multiple facets are involved in ensuring robust clinical AI software. One approach starts with simulated environments to understand fundamental behavior in controlled settings, allowing for controlled variation while maintaining reproducibility and gradually transitioning to real-world data. Standardized benchmark datasets can provide a baseline for performance comparison in controlled experiments, although it is essential to recognize their limitations in representing real-world complexity. Another valuable strategy is transfer learning, which entails training models on controlled data and fine-tuning them on real-world data to bridge the gap between controlled and realistic contexts. Field studies conducted in clinical settings with actual users are essential for observing how AI tools impact workflows, patient outcomes, and user satisfaction. Adversarial testing is also important, introducing realistic challenges such as noisy data and adversarial attacks during controlled experiments to reveal vulnerabilities and test robustness. When used collectively, these strategies contribute to a comprehensive and balanced approach to AI evaluation. This approach ensures that all aspects of AI performance are thoroughly tested and evaluated, providing a fair and thorough assessment of the system's capabilities.

An example of the need to consider the reality of deployment in contrast with controlled experiments and statistical analysis can be seen in our study. Factor analysis is useful for evaluating user studies, especially with structured questionnaires and surveys. It identifies relationships between variables, simplifies data, and highlights key factors influencing responses.

This helps researchers understand patterns in feedback and make informed decisions about tool design and functionality. However, it mainly focuses on statistical relationships and might miss nuances in user interactions. For example, in our study, all participants consistently experienced passive AI, but active AI was less used, likely due to the flawed concept requiring comparisons without prior knowledge. This added complexity and confusion, which would not be evident through factor analysis alone. To address such issues, deeper qualitative investigations are necessary. These can include user interviews, observational studies, and detailed feedback sessions to understand the context and reasons behind user behaviors. This approach provides richer insights beyond statistical analysis, ensuring that AI tools are usable and practical in clinical settings. Combining quantitative and qualitative methods can lead to a more comprehensive evaluation and refinement of AI support systems.

In conclusion, while these strategies have advantages and potential challenges, they all play an important role in ensuring the practical evaluation of clinical AI tools. By proactively considering these points and addressing potential critiques, we can work towards more robust, ethical, and effective AI in healthcare.

References

1. S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, A. Aldairem, M. Alrashed, K. Bin Saleh, H. A. Badreldin *et al.*, Revolutionizing healthcare: the role of artificial intelligence in clinical practice, *BMC medical education* **23**, p. 689 (2023).
2. R. Khera, A. J. Butte, M. Berkwits, Y. Hswen, A. Flanagin, H. Park, G. Curfman and K. Bibbins-Domingo, Ai in medicine—jama's focus on clinical outcomes, patient-centered care, quality, and equity, *Jama* (2023).
3. C. A. Longhurst, K. Singh, A. Chopra, A. Atreja and J. S. Brownstein, A call for artificial intelligence implementation science centers to evaluate clinical effectiveness (2024).
4. Y. Park, G. P. Jackson, M. A. Foreman, D. Gruen, J. Hu and A. K. Das, Evaluating artificial intelligence in medicine: phases of clinical research, *JAMIA open* **3**, 326 (2020).
5. Q. Jin, F. Chen, Y. Zhou, Z. Xu, J. M. Cheung, R. Chen, R. M. Summers, J. F. Rousseau, P. Ni, M. J. Landsman, S. L. Baxter, S. J. Al'Aref, Y. Li, A. Chen, J. A. Brejt, M. F. Chiang, Y. Peng and Z. Lu, Hidden flaws behind expert-level accuracy of multimodal gpt-4 vision in medicine, *npj Digital Medicine* **7**, p. 190 (Jul 2024).
6. N. Agarwal, A. Moehring, P. Rajpurkar and T. Salz, *Combining human expertise with artificial intelligence: Experimental evidence from radiology*, tech. rep., National Bureau of Economic Research (2023).
7. F. M. Calisto, J. Fernandes, M. Morais, C. Santiago, J. M. Abrantes, N. Nunes and J. C. Nascimento, Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis, in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023.
8. S. Reddy, W. Rogers, V.-P. Makinen, E. Coiera, P. Brown, M. Wenzel, E. Weicken, S. Ansari, P. Mathur, A. Casey *et al.*, Evaluation framework to guide implementation of ai systems into healthcare settings, *BMJ health & care informatics* **28** (2021).
9. B. H. Kann, A. Hosny and H. J. Aerts, Artificial intelligence for clinical oncology, *Cancer Cell* **39**, 916 (2021).
10. S. Khalighi, K. Reddy, A. Midya, K. B. Pandav, A. Madabhushi and M. Abedalthagafi, Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment, *NPJ Precision Oncology* **8**, p. 80 (2024).

11. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in *Proceedings of the AAAI conference on artificial intelligence*, (01)2019.
12. D. C. Chan, M. Gentzkow and C. Yu, Selection with variation in diagnostic skill: Evidence from radiologists, *The Quarterly Journal of Economics* **137**, 729 (2022).
13. G. S. Collins, P. Dhiman, C. L. A. Navarro, J. Ma, L. Hooft, J. B. Reitsma, P. Logullo, A. L. Beam, L. Peng, B. Van Calster *et al.*, Protocol for development of a reporting guideline (tripod-ai) and risk of bias tool (probast-ai) for diagnostic and prognostic prediction model studies based on artificial intelligence, *BMJ open* **11**, p. e048008 (2021).
14. X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, A. K. Denniston, H. Ashrafian, A. L. Beam, A.-W. Chan, G. S. Collins, A. D. J. Deeks *et al.*, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension, *The Lancet Digital Health* **2**, e537 (2020).
15. B. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim *et al.*, Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai, *bmj* **377** (2022).
16. E. Prince, T. C. Hankinson and C. Görg, Easl: A framework for designing, implementing, and evaluating ml solutions in clinical healthcare settings, in *Machine Learning for Healthcare Conference*, 2023.
17. E. W. Prince, D. M. Mirsky, T. C. Hankinson and C. Görg, Impact of ai decision support on clinical experts' radiographic interpretation of adamantinomatous craniopharyngioma, in *AMIA Annual Symposium Proceedings*, 2024 (to be released November 2024).
18. E. W. Prince, T. C. Hankinson and C. Görg, The iterative design process of an explainable ai application for non-invasive diagnosis of cns tumors: A user-centered approach, in *2023 Workshop on Visual Analytics in Healthcare (VAHC)*, 2023.
19. E. W. Prince, R. Whelan, D. M. Mirsky, N. Stence, S. Staulcup, P. Klimo, R. C. Anderson, T. N. Niazi, G. Grant, M. Souweidane *et al.*, Robust deep learning classification of adamantinomatous craniopharyngioma from limited preoperative radiographic images, *Scientific reports* **10**, p. 16885 (2020).
20. M. Linting and A. Van der Kooij, Nonlinear principal components analysis with catpca: a tutorial, *Journal of personality assessment* **94**, 12 (2012).

QUEST-AI: A System for Question Generation, Verification, and Refinement using AI for USMLE-Style Exams

Suhana Bedi[1,*,†], Scott L. Fleming[1,2*], Chia-Chun Chiang[3], Keith Morse[4], Aswathi Kumar[5], Birju Patel[5], Jenelle A. Jindal[6], Conor Davenport[7], Craig Yamaguchi[7], Nigam H. Shah[1,8,9,10]

[1]*Department of Biomedical Data Science, Stanford University, Stanford, CA, USA*
[2]*SmarterDx, Inc. Brooklyn, NY, USA*
[3]*Department of Neurology, Mayo Clinic, Rochester, MN*
[4]*Department of Pediatrics, Stanford School of Medicine, Stanford, CA, USA*
[5]*Department of Medicine, Stanford School of Medicine, Stanford, CA, USA*
[6]*Center for Biomedical Informatics Research, Stanford, CA, USA*
[7]*Western University of Health Sciences, Pomona, California, USA*
[8]*Human-Centered Artificial Intelligence Institute, Stanford University, Stanford, CA, USA*
[9]*Clinical Excellence Research Center, Stanford School of Medicine, Stanford, CA, USA*
[10]*Technology and Digital Solutions, Stanford Health Care, Palo Alto, California, USA*

[*]*Equal Contribution*
[†]*Email: suhana@stanford.edu*

The United States Medical Licensing Examination (USMLE) is a critical step in assessing the competence of future physicians, yet the process of creating exam questions and study materials is both time-consuming and costly. While Large Language Models (LLMs), such as OpenAI's GPT-4, have demonstrated proficiency in answering medical exam questions, their potential in generating such questions remains underexplored. This study presents QUEST-AI, a novel system that utilizes LLMs to (1) generate USMLE-style questions, (2) identify and flag incorrect questions, and (3) correct errors in the flagged questions. We evaluated this system's output by constructing a test set of 50 LLM-generated questions mixed with 50 human-generated questions and conducting a two-part assessment with three physicians and two medical students. The assessors attempted to distinguish between LLM and human-generated questions and evaluated the validity of the LLM-generated content. A majority of exam questions generated by QUEST-AI were deemed valid by a panel of three clinicians, with strong correlations between performance on LLM-generated and human-generated questions. This pioneering application of LLMs in medical education could significantly increase the ease and efficiency of developing USMLE-style medical exam content, offering a cost-effective and accessible alternative for exam preparation.

Keywords: USMLE; Medical Education; Large Language Models; Artificial Intelligence; GPT-4; Exam Question Generation; Automated Assessment; Medical Exam Preparation; Question Validity; Medical Licensing Examination

1. Introduction

Every year, over 100,000 medical students take the United States Medical Licensing Examination (USMLE), administered by the National Board of Medical Examiners (NBME).[1] This rigorous examination is crucial for ensuring the competence of future physicians. However, generating the exam questions and related preparation materials is a manual process, which is both time-consuming and costly. On average, each student spends over $4,000 on buying USMLE-related study materials.[2] The high costs and substantial effort associated with producing these materials are the primary drivers of the cost, and offer a great opportunity for technological intervention.

The quality of these exam questions plays a critical role in medical education and the training of future healthcare professionals. These exams assess key clinical knowledge and decision-making skills, which directly influence how prepared medical students are to handle real-world patient care. Ensuring the accuracy and biological relevance of the questions is vital for maintaining high standards in healthcare, as the competence of future physicians ultimately impacts patient outcomes.

The adoption of Artificial Intelligence (AI) in healthcare is rapidly increasing, driven by advancements in Generative AI and especially, Large Language Models (LLMs) such as OpenAI's GPT-4.[3,4,5] LLMs have been explored for various use cases in medicine, including generating clinical notes, summarizing patient records, and providing decision support.[6,7,8] Numerous studies have demonstrated the proficiency of these models in answering USMLE questions, achieving over 80% accuracy on the USMLE Step 2 Clinical Knowledge (CK) exam.[9] Despite their success in answering exam questions, there is limited research on the use of LLMs for *generating* medical exam questions, particularly for the USMLE. To address this gap, we introduce QUEST-AI, an autonomous system powered by LLMs that (1) generates USMLE-style questions based on in-context examples, (2) verifies the system-generated questions using an ensemble of LLMs, and (3) refines any questions identified as incorrect. The system is evaluated with the assistance of physicians and medical students.

We began by prompting GPT-4 to generate 50 questions inspired by sample questions from the USMLE Step 2 Clinical Knowledge (CK) exam. Then, we used aggregated predictions from an ensemble of diverse LLMs to flag incorrect questions. Finally, we prompted GPT-4 again to correct the flawed questions. In order to evaluate the quality of questions generated using our approach, we constructed a test set containing our 50 system-generated questions randomly interspersed with 50 human-generated sample questions. Three physicians and two medical

students engaged in a twofold assessment: (1) they attempted to distinguish between the system-generated and human-generated USMLE-style questions, and (2) they assessed the validity of the system-generated questions and answers.

To our knowledge, ours is the first study to generate, verify, and refine USMLE-style questions using LLMs (Figure 1). This shift from answering questions to generating questions represents a novel application of AI in medical education, with the potential to revolutionize exam content development.

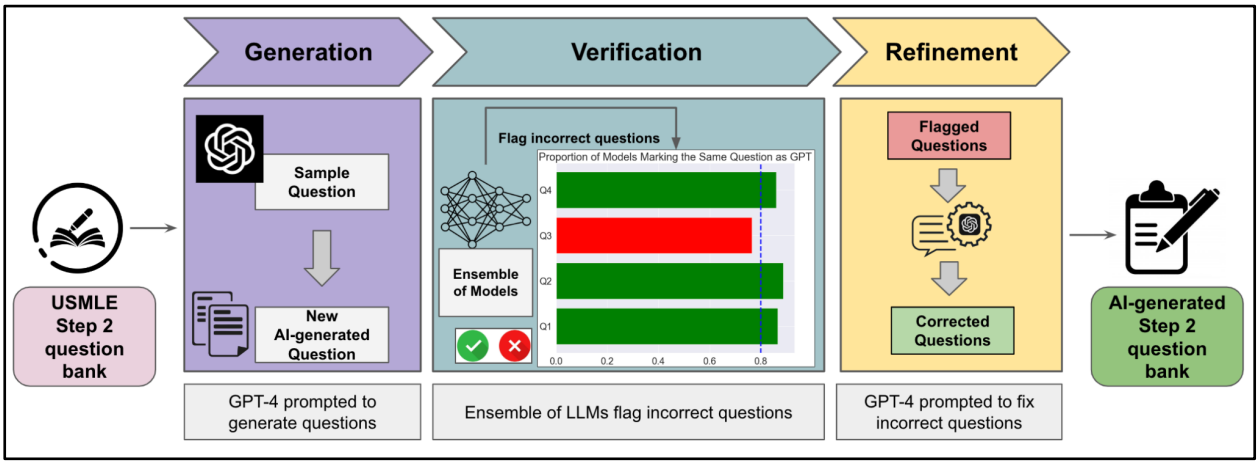


Figure 1: **The QUEST-AI System for Generation, Verification, and Refinement of USMLE-Style Questions:** This figure illustrates the process used by QUEST-AI to generate, verify, and refine USMLE-style questions. The process begins with GPT-4 generating questions using sample questions from the USMLE Step 2 question bank as in-context examples. An ensemble of LLMs then processes these questions, flagging any incorrect ones based on their ensembled predictions. Finally, GPT-4 refines the flagged questions, resulting in a high-quality, system-generated Step 2 question bank.

2. Related Work

The use of Large Language Models (LLMs) in healthcare and education has seen considerable growth and innovation.

2.1. *LLMs in Healthcare*

LLMs have become prominent in healthcare due to their advanced natural language processing capabilities, allowing them to handle large datasets and generate accurate, contextually relevant text.[10] Bedi et al[11] provide a systematic review of LLM applications across various healthcare tasks, including diagnosis[12], report generation[13], treatment recommendations[14], and clinical referrals.[14,15] While these studies demonstrate the potential of LLMs in clinical settings, few have

explored their application in the educational domain, specifically for training future healthcare professionals. This study builds on these advancements, applying LLMs to a novel task: automated medical exam question generation.

2.2. *LLMs in Education*

LLMs have shown great promise in education, particularly in providing real-time support and feedback across a range of subjects, such as math [16], law [17] and medicine [4].

Recent research has applied LLMs for automatic question generation, improving educational content and assessment quality. Laverghetta Jr. and Licato demonstrated the use of GPT-3 for cognitive assessments, and Tran et al. applied GPT-4 to generate high-quality multiple-choice questions (MCQs) for computing courses. [18] [19] However, these efforts focus on general education, with little attention given to specialized medical education, particularly for high-stakes exams like the USMLE.

2.3. *LLMs in Medical Education*

There has been growing interest in using LLMs to generate medical exam questions due to their potential to reduce the burden on educators and streamline content creation. A systematic review by Artsi et al. discovered a total of eight studies that explored LLMs like GPT-3.5 and GPT-4 for producing valid multiple-choice questions (MCQs) across various medical disciplines, including neurosurgery, internal medicine, and dermatology.[20] While these studies demonstrate the feasibility of LLMs in medical education, they also highlight limitations such as inaccuracies, lower complexity in generated questions, and a lack of rigorous evaluation of content quality and validity, particularly for high-stakes exams like the USMLE. [21] [22] [23] [24] [25]

To address these gaps, our study evaluates GPT-4's ability to generate USMLE Step 2 CK-style exam questions. We provide insights into the practical applications of AI in medical education and its potential to enhance the accessibility and quality of exam preparation materials. By presenting a fully autonomous system for generating, verifying, and refining USMLE-style questions, we aim to demonstrate the capacity of LLMs to generate high-quality exam content, thereby improving the development and accessibility of medical education resources.

3. Methods

3.1. *Data collection and generation*

We randomly selected a set of 50 human-generated questions from a bank of 120 publicly available USMLE Step 2 CK test sample questions, ensuring that these questions did not include associated images or abstracts[26]. This was done to maintain a controlled and uniform format for comparison purposes.

For system-generated questions, we employed a prompt chaining approach with GPT-4 as shown in Figure 2. We started with a human-generated USMLE CK test question-answer pair, which was included in the initial prompt to GPT-4. The model then generated an explanation of why the given answer was correct and the others were incorrect. This original question, along with the system-generated explanation, were used in a follow-up prompt instructing GPT-4 to generate another USMLE Step 2 CK-style question in a similar format. This method ensured the generated questions closely matched the format, style, and complexity of the human-generated ones, promoting consistency and reducing deviations from the desired standards.

After generating a set of system-generated questions, we compiled these alongside the human-generated ones and randomly shuffled them to create a comprehensive 100-question set. This randomization was crucial to ensure an unbiased evaluation.

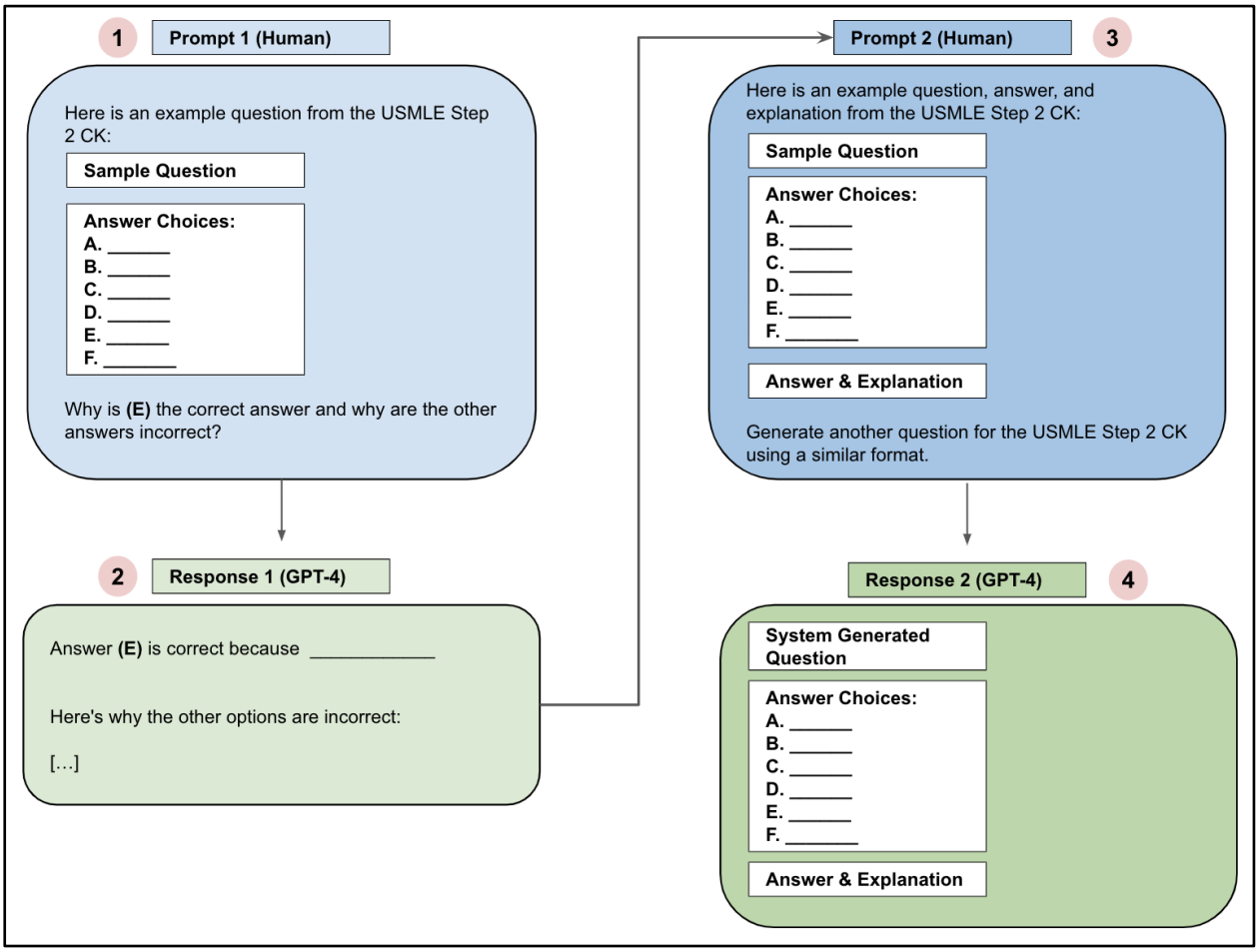


Figure 2: Prompt chaining strategy for question generation: First, we provide GPT-4 with an example question from the USMLE CK exam and ask why a specific option is correct and why others are incorrect. Once GPT-4 generates a response, we create a new prompt incorporating this response and the original question, then ask GPT-4 to generate another question in a similar format.

3.2. *Evaluation by Physicians*

A group of three licensed, practicing physicians and two medical students were tasked with evaluating the 100-question set. They were instructed to:

1. Choose the single best answer to each question without consulting any external reference.
2. Guess whether each question was generated by humans or GPT-4.

In a separate task, three physicians reviewed the 50 system-generated exam questions to evaluate their correctness, using any available external references. They recorded the type of errors found in the system-generated questions and the time taken to make their determinations. The two

phases of the study, marked by the different tasks performed by the medical specialists, are illustrated in Figure 3.

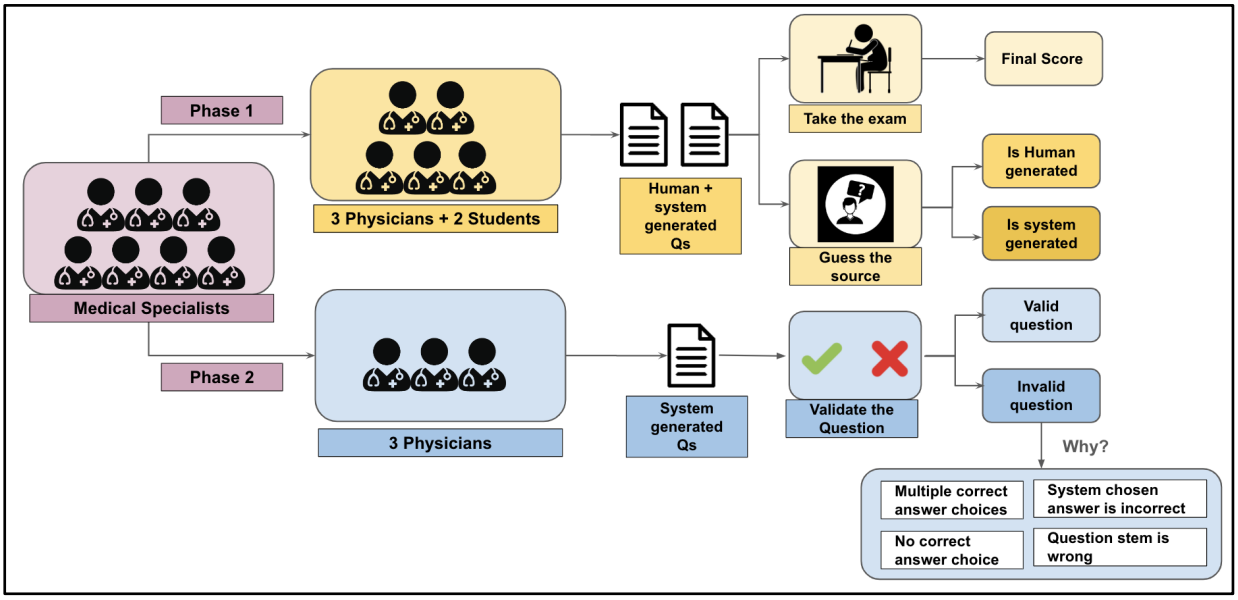


Figure 3: Evaluation Process by Medical Specialists: In Phase 1 of the study, three physicians and two medical students attempted a USMLE exam that included both real and system-generated questions, tasked with choosing the best answer for each question and identifying which questions were system-generated. In Phase 2, three physicians evaluated the system-generated question-answer pairs to determine their validity. For invalid questions, they categorized the issues into four types: multiple correct answer choices, no correct answer choice, the system-chosen answer choice is incorrect, or the question stem is incorrect.

3.3. *Evaluation by LLMs*

An ensemble of five LLMs from the Hugging Face hub[27] (a public repository of models) was selected for evaluation based on the models' performance in public open LLM leaderboards[28] and community support: Meta-Llama-3-70B-Instruct from Meta, Mixtral-8x22B-Instruct-v0.1 from Mistral AI, Qwen2-72B-Instruct from Alibaba, Phi-3-medium-4k-Instruct from Microsoft, and llama-2-70b-chat from Meta. Each of these models brings unique strengths due to variations in their architectures and training datasets, providing diverse perspectives on the task of identifying the best answer to system-generated USMLE-style questions.

To evaluate the validity of the system-generated question-answer pairs, we tasked the ensemble with selecting the best answer. A simple majority-based classifier was constructed to flag potentially flawed questions: if any of the models disagreed with GPT-4's selected best answer, the question was flagged for further review. This design is grounded in ensemble learning theory [29], which posits that combining predictions from multiple models can improve accuracy and

reliability by reducing individual model bias and variance . The assumption here is that even a single disagreement may indicate a potential error, ambiguity, or inconsistency in the question or answer choices. This model heterogeneity strengthens the system's ability to detect flaws by increasing the likelihood of identifying subtle inconsistencies.

Conversely, when all models in the ensemble agreed with GPT-4's answer, the question was considered less likely to be flawed. This consensus-based approach is consistent with voting schemes in ensemble learning, where agreement across multiple models typically indicates a high confidence prediction. [30]

3.4. *Categorization and Post-Hoc Editing by GPT-4*

GPT-4 was prompted to categorize each question-answer pair in the 100-item set into one of 18 categories outlined in the USMLE content outline[31]. This categorization aimed to evaluate whether the system-generated questions fell into the same categories as the original human-generated questions used in the prompts. A physician then reviewed these category assignments for each question to verify their validity. We caveat the evaluation of category assignment validity by noting that the ground truth categories per question are not made publicly available by the NBME and the team of physicians and medical students that imputed item categories have no affiliation with NBME/USMLE.

For the system-generated questions deemed incorrect by the ensemble of LLMs and physicians, we conducted a post-hoc editing stage. During this stage, we asked GPT-4 to first identify why a specific question was wrong and then modify or correct the mistakes present in the flagged questions. This was done to assess GPT-4's capability to improve its responses based on iterative feedback and refinement. A physician manually evaluated each corrected exam item to assess whether the corrections led to a valid result.

The data collection and analysis were conducted in multiple phases from April 2023 to December 2023, and an updated evaluation analysis was completed in July 2024. This evaluation process aimed to provide insights into the practical applications of AI in generating high-quality exam questions and to assess the potential of GPT-4 in contributing to medical education. Data and code for reproducing our results can be found at https://github.com/som-shahlab/gpt4usmle.

4. Results

4.1. *Evaluation by Physicians*

The average accuracy of the three physicians and two medical students in distinguishing between human-generated and GPT-4-generated questions was 51.8% (95% CI: 46.1% to 57.5%), nearly equivalent to random guessing. This suggests that there were no identifiable systematic differences between the system-generated and human-generated questions. Furthermore, the Pearson correlation between participants' performance on system-generated and human-generated questions was statistically significant (ρ = 0.952, P-value = 0.012). This strong correlation indicates that system-generated questions are comparable to human-generated questions in their ability to assess the clinical knowledge and skills examined in USMLE-style questions, providing a reliable and consistent measure of performance across both question types.

On a separate task where three physician reviewers were asked to validate the 50 AI generated questions, 32 (64%) questions were deemed "correct" by all reviewers, while 18 (36%) were deemed "incorrect" by at least one reviewer. The reasons for labeling exam items as "incorrect" included "Multiple correct answer choices" (n=9), "AI-chosen answer is incorrect" (n=6), and "No correct answer choice" (n=3). These findings highlight specific areas where the system-generated questions fell short and suggest areas for further refinement in the AI's question generation capabilities.

Reviewers spent, on average, 3.21 minutes (95% CI 2.73 to 3.69) reviewing each system-generated exam item for correctness. This quick evaluation time highlights a significant potential efficiency advantage, as it is substantially faster than drafting a question from scratch, which typically involves extensive research, drafting, and revision.

4.2. *Evaluation by LLMs*

All LLMs within our LLM ensemble achieved adequate performance on the human-generated USMLE-style exam questions (see Table 1). Our proposed LLM ensemble classifier was able to discriminate between invalid system-generated questions with an Area under the Receiver-Operator Characteristic curve (AUROC) of 0.79. We considered an item to be classified by the model as "flawed" if any one of the 5 LLMs in the ensemble disagreed with GPT-4 on the best answer choice. Of the 18 system-generated question-answer pairs deemed flawed by clinician reviewers, our approach correctly flagged 15 (Recall = 15/18 = 0.83). Overall, our approach flagged 25 system-generated question-answer pairs as flawed (Precision = 15/25 = 0.60). Of the 25 system-generated questions not flagged by our approach, 22 were deemed valid by clinicians. See Table 2.

Table 1: Performance of LLMs in model ensemble on human- and system-generated USMLE-style questions. All models performed reasonably well (examinees typically must answer approximately 60% of items correctly to achieve a passing score on the USMLE)[32]

Model	% Correct on human-generated questions	% Correct on system-generated questions
Meta-Llama-3-70b-Instruct	0.80	0.80
Mixtral-8x22B-Instruct-v0.1	0.80	0.78
Phi-3-medium-4k-instruct	0.76	0.80
Qwen2-72B-Instruct	0.80	0.82
llama-2-70b-chat_huggingface	0.60	0.62

Table 2: Confusion matrix for the LLM ensemble used to determine whether system-generated questions are potentially invalid by analyzing whether all LLMs agree with GPT-4 on the best answer (not flagged) or at least one LLM disagrees with GPT-4 on the best answer (flagged).

	Flagged as invalid by LLM ensemble	Not flagged by LLM ensemble
Deemed invalid by clinician reviewers	15	3
Deemed valid by clinician reviewers	10	22

4.3. *Categorization and Post-Hoc Editing by GPT-4*

For the categories assigned to each question by GPT-4, 8 questions were assigned invalid content category labels, while the remaining 92 questions were assigned appropriate labels. This outcome shows that GPT-4 generally performed well in classifying question categories, although it occasionally struggled to differentiate between Behavioral Health and Social Sciences. This challenge might be addressed by clarifying that Behavioral Health pertains to psychiatry and mental health topics, whereas Social Sciences covers medical ethics, interpersonal health, and health system quality improvement.

Additionally, 16 out of the 50 questions matched the category of their corresponding sample question. This suggests that GPT-4 introduces a degree of variability and diversity in its generated questions. Rather than merely replicating existing content, GPT-4 demonstrates the ability to create new and varied material. A breakdown of categories can be seen in the Supplementary section at - https://www.medrxiv.org/content/10.1101/2023.04.25.23288588v2

For post-hoc editing, the questions deemed incorrect by at least one reviewer were passed through GPT-4. The model was asked to classify why a question-answer pair was incorrect and then to provide a corrected version. Impressively, for 9 out of 18 questions (50%), GPT-4 identified the same reason for incorrectness as the physician reviewers. For 11 of these 18 questions (61%), GPT-4 was able to correct its original mistake, resulting in a valid exam item. This demonstrates GPT-4's capability not only to generate questions but also to accurately diagnose issues with them and offer corrections.

5. Conclusion

With ever-increasing costs of medical education, medical student debt, and a looming physician shortage[33], there is an urgent need for cost-effective and easily accessible medical exam preparation resources. We designed QUEST-AI, a first-of-its-kind system that can improve access to high-quality USMLE-style questions by using LLMs to generate candidate exam questions, flag invalid candidate items, and correct flawed exam items. While performance of the system is not perfect, clinician evaluation suggests that (1) a significant majority of exam items generated using our approach are valid; (2) candidate performance on items generated using our approach correlates strongly with performance on human-generated USMLE-style questions; and (3) our system can be used to generate exam across a variety of content categories. This offers a promising solution for decreasing the cost and time required to generate USMLE-style questions. This in turn could reduce both the costs for exam preparation materials that debt-burdened medical students face and the costs for generating new exam items that non-profit organizations like the National Board of Medical Examiners face.

6. Limitations

There are several important limitations to our system to consider when assessing whether it can be used in medical education.

First, with respect to our evaluation, the medical specialists who attempted to select the best answer on the evaluation set of 50 system-generated and 50 human-generated questions were not MD

students (the primary audience that would benefit from such a system); they were practicing MDs who had already passed the USMLE Step 2 CK exam and DO students who would take a different but similar exam as part of their training. This was by design: we wanted to ensure that no assessor would recognize the exam items in the publicly available NBME-provided USMLE-style practice exam. Otherwise, their ability to distinguish between human- and system-generated questions would be overly optimistic. Additional study is needed to understand whether our results translate to the primary population of interest, namely MD students preparing to take the USMLE Step 2 CK exam.

Second, the clinicians who determined whether or not the system-generated exam items were valid were not expert exam writers nor were they affiliated with the NBME. It is quite possible that system-generated exam items deemed valid by our panel of clinicians would be considered invalid by NBME-employed expert exam writers, and vice versa.

Third, there was no threshold for which our LLM ensemble-based flagging system was able to correctly recall *all* the system-generated exam items deemed invalid (except for if we trivially flagged all the items as invalid). There were 3 of 18 items deemed invalid for which all 5 LLMs in the ensemble agreed with GPT-4's best answer selection (thus the question was not flagged) but where at least one clinician deemed the overall exam item to be invalid. This suggests that, were this system to be used entirely autonomously, it could generate flawed exam items. This has important ethical implications that should be considered and potentially addressed with improved methods before releasing the tool to the broader public.

Finally, the number of system-generated questions was relatively small, with only 50 questions included in the study. While this sample size provided useful insights for an initial evaluation, it limits the generalizability of the findings. A larger set of questions is needed to provide a more comprehensive assessment of the system's performance across different content areas and question formats. Increasing the sample size in future studies will also allow for a more detailed evaluation of additional metrics and improve the statistical power of the results.

7. Funding and Conflicts of Interest

This work is supported by the Mark and Debra Leslie endowment for AI in Healthcare; the Stanford University Department of Medicine; Stanford Healthcare; and the Stanford Medicine Program for AI in Healthcare. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the

funding bodies. N.H.S. is a cofounder of Prealize Health and Atropos Health; reports funding from the Gordon and Betty Moore Foundation; and served on the board of the Coalition for Healthcare AI (CHAI). J.A.J. is the founder of Jindal Neurology, Inc. and paid per diem as a physician with Kaiser Permanente, South San Francisco, CA. The other authors declare no competing financial interests. No proprietary NBME data or information were used in the study.

References

1. Performance Data. Accessed July 20, 2024. https://www.usmle.org/performance-data

2. Bhatnagar V, Diaz SR, Bucur PA. The Cost of Board Examination and Preparation: An Overlooked Factor in Medical Student Debt. *Cureus*. 2019;11(3):e4168.

3. Stafie CS, Sufaru IG, Ghiciuc CM, et al. Exploring the Intersection of Artificial Intelligence and Clinical Healthcare: A Multidisciplinary Review. *Diagnostics (Basel)*. 2023;13(12). doi:10.3390/diagnostics13121995

4. Lee P, Goldberg C, Kohane I. *The AI Revolution in Medicine: GPT-4 and Beyond*. Pearson; 2023.

5. Goldberg CB, Adams L, Blumenthal D, et al. To do no harm - and the most good - with AI in health care. *Nat Med*. 2024;30(3):623-627.

6. Du X, Novoa-Laurentiev J, Plasaek JM, et al. Enhancing Early Detection of Cognitive Decline in the Elderly: A Comparative Study Utilizing Large Language Models in Clinical Notes. *medRxiv*. Published online May 6, 2024. doi:10.1101/2024.04.03.24305298

7. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024;30(4):1134-1142.

8. Skryd A, Lawrence K. ChatGPT as a Tool for Medical Education and Clinical Decision-Making on the Wards: Case Study. *JMIR Form Res*. 2024;8:e51346.

9. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. Published online March 20, 2023. Accessed July 20, 2024. http://arxiv.org/abs/2303.13375

10. Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus*. 2023;15(5):e39305.

11. Bedi S, Liu Y, Orr Ewing L. A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs). *MedRxiv*. Published online May 7, 2024. doi:10.1101/2024.04.15.24305869

12. Pagano S, Holzapfel S, Kappenschneider T, et al. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. *J Orthop Traumatol*. 2023;24(1):61.

13. Zhou Z. Evaluation of ChatGPT's Capabilities in Medical Report Generation. *Cureus*. 2023;15(4):e37589.

14. Wang Z, Zhang Z, Traverso A, Dekker A, Qian L, Sun P. Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach. *Quant Imaging Med Surg*. 2024;14(2):1602-1615.

15. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 Assistance in Optimizing Emergency Department Radiology Referrals and Imaging Selection. *J Am Coll Radiol*. 2023;20(10):998-1003.

16. Zheng Y, Hongyi Y, Chuanqi T, Wei W, Songfang H. How well do Large Language Models perform in Arithmetic tasks? *ArXiv*. Published online March 16, 2023. https://arxiv.org/abs/2304.02015

17. Cui J, Li Z, Yan Y, Chen B, Yuan L. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *ArXiv*. Published online June 28, 2023. https://arxiv.org/abs/2306.16092

18. Laverghetta A Jr, Licato J. Generating Better Items for Cognitive Assessments Using Large Language Models. In: Kochmar E, Burstein J, Horbach A, et al., eds. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics; 2023:414-428.

19. Tran A, Angelikas K, Rama E, Okechukwu C, Smith DH, MacNeil S. Generating Multiple Choice Questions for Computing Courses Using Large Language Models. In: *2023 IEEE Frontiers in Education Conference (FIE)*. IEEE; 2023:1-8.

20. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. *BMC Med Educ*. 2024;24(1):354.

21. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLoS One*. 2023;18(8):e0290691.

22. Ayub I, Hamann D, Hamann CR, Davis MJ. Exploring the Potential and Limitations of Chat Generative Pre-trained Transformer (ChatGPT) in Generating Board-Style Dermatology Questions: A Qualitative Analysis. *Cureus*. 2023;15(8):e43717.

23. Sevgi UT, Erol G, Doğruel Y, Sönmez OF, Tubbs RS, Güngor A. The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. *Neurosurg Rev*. 2023;46(1):86.

24. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. *Med Teach*. 2024;46(5):657-664.

25. Biswas S. Passing is Great: Can ChatGPT Conduct USMLE Exams? *Ann Biomed Eng*. 2023;51(9):1885-1886.

26. Step 2 CK sample test questions. Accessed July 29, 2024. https://www.usmle.org/prepare-your-exam/step-2-ck-materials/step-2-ck-sample-test-questions

27. Jain SM. Hugging Face. In: Jain SM, ed. *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*. Apress; 2022:51-67.

28. Open LLM Leaderboard 2 - a Hugging Face Space by open-llm-leaderboard. Accessed July 25, 2024. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

29. Dietterich TG. Ensemble Methods in Machine Learning. In: *Multiple Classifier Systems*. Lecture notes in computer science. Springer Berlin Heidelberg; 2000:1-15.

30. Kuncheva LI. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons; 2004.

31. Step 2 CK content outline & specifications. Accessed July 29, 2024. https://www.usmle.org/prepare-your-exam/step-2-ck-materials/step-2-ck-content-outline-specifications

32. Bulletin of information. Accessed July 27, 2024. https://www.usmle.org/bulletin-information/scoring-and-score-reporting

33. How much does it cost to attend medical school? Here's a breakdown. Students & Residents. Accessed July 25, 2024. https://students-residents.aamc.org/premed-navigator/how-much-does-it-cost-attend-medical-school-here-s-breakdown

ReXErr: Synthesizing Clinically Meaningful Errors in Diagnostic Radiology Reports

Vishwanatha M. Rao[†1], Serena Zhang[†1], Julian N. Acosta[1], Subathra Adithan[2], Pranav Rajpurkar[1]

[1]*Department of Biomedical Informatics, Harvard Medical School Boston, MA 02115, USA*
[2]*Department of Radiodiagnosis, Jawaharlal Institute of Postgraduate Medical Education and Research, India*
E-mail: : vishwanatha.rao@pennmedicine.upenn.edu, serena2z@stanford.edu, julian_acosta@hms.harvard.edu, subathra.a@jipmer.edu.in

Accurately interpreting medical images and writing radiology reports is a critical but challenging task in healthcare. Both human-written and AI-generated reports can contain errors, ranging from clinical inaccuracies to linguistic mistakes. To address this, we introduce ReXErr, a methodology that leverages Large Language Models to generate representative errors within chest X-ray reports. Working with board-certified radiologists, we developed error categories that capture common mistakes in both human and AI-generated reports. Our approach uses a novel sampling scheme to inject diverse errors while maintaining clinical plausibility. ReXErr demonstrates consistency across error categories and produces errors that closely mimic those found in real-world scenarios. This method has the potential to aid in the development and evaluation of report correction algorithms, potentially enhancing the quality and reliability of radiology reporting.

Keywords: Radiology Report Generation; Chest X-Rays; LLMs; Chat-GPT; Error Injection; Synthetic Data.

1. Introduction

Radiology reports provide crucial information for clinical decision-making and patient outcomes.[1] However, creating radiology reports is an intensive process, and requires a trained specialist to analyze medical images and write in-depth medical reports.[2,3] In human-written reports, errors can arise due to various factors such as fatigue, high case volumes or complexity. These errors may include misinterpretation of imaging findings, incomplete documentation of relevant clinical information, and inconsistencies in terminology and language usage. In addition to such inaccuracies, the subjective nature of radiological interpretation leaves room for errors, which may go unnoticed until they impact patient care.[4,5]

Recently, there has been a significant push towards automating the creation of these reports using deep learning. While current approaches to generating radiology reports have, in some cases, succeeded in creating complete and clinically relevant reports,[6–9] automated report generation presents its own set of challenges stemming from inherent biases within algorithms, model constraints, and limitations in the data used. Errors can range from references to non-

[†]Authors contributed equally to this work.

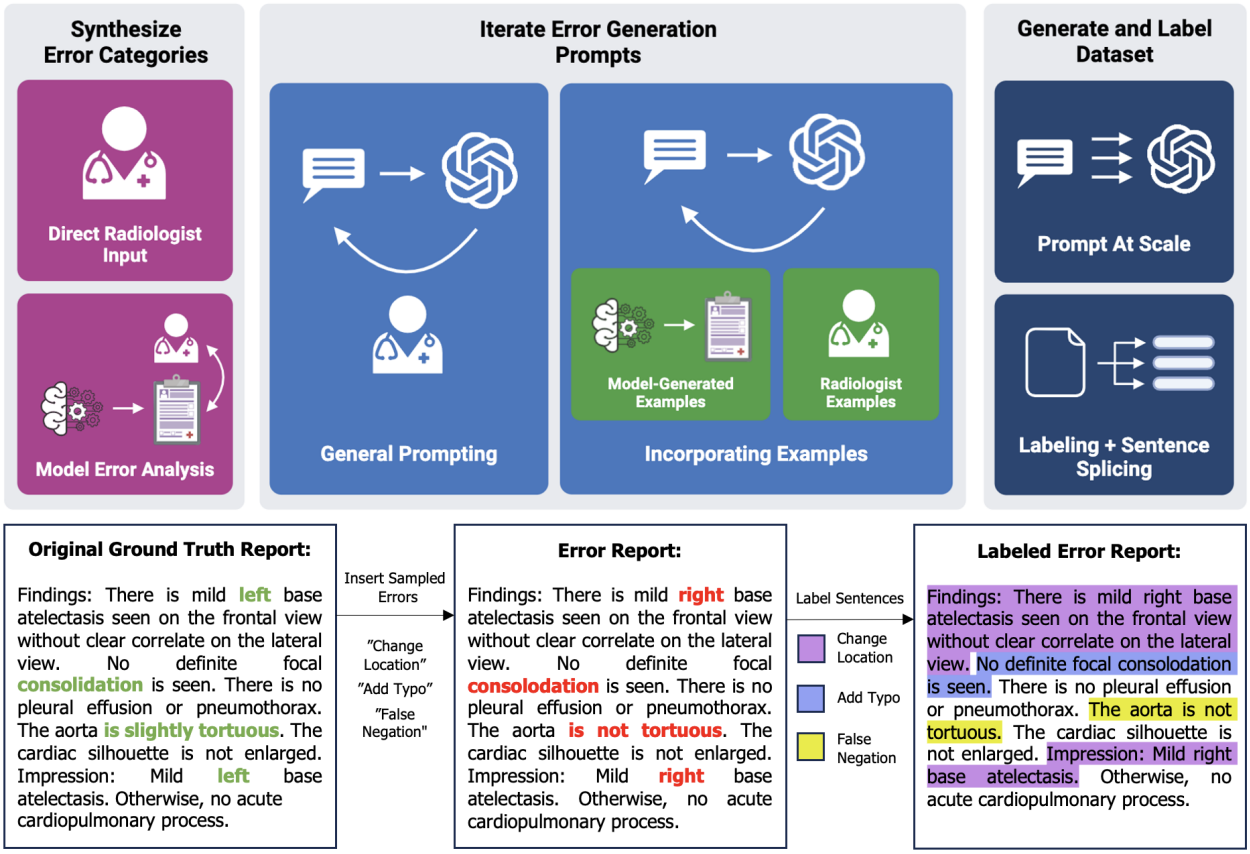


Fig. 1. Summary of ReXErr error generation pipeline. The bottom panel provides an example of applying ReXErr to a sample radiology report.

existing priors, which are easier to detect, to false predictions or omissions, which are much more problematic clinically and often go unnoticed.[10,11] The prevalence of errors, both in radiologist-written as well as AI-based reports, leaves a great need for more comprehensive tools that can screen for and correct them. Throughout this paper, we present the Chest X-Ray Report Error (ReXErr) method that can generate errors at a report and sentence level. ReXErr offers a novel pipeline to synthesize plausible errors that capture the breadth and diversity of errors made by humans and models and can thus be used to generate data to train and adapt error correction algorithms. Figure 1 outlines an overview of the error generation process.

2. Related Work

2.1. *Error Classification in Radiology Reports*

The 12-category framework developed by Kim and Mansfield, based on an evaluation of 1,269 errors, offers a foundation for understanding and classifying errors in human-generated reports and is the most frequently used for human-error analysis.[4,12] Most of the errors in this classification system fall under two types: missed findings (under-reading, satisfaction of search, etc.) and interpretation errors (finding attributed to wrong cause/clinical entity due to faulty

Table 1. Summary of the errors incorporated within the ReXErr pipeline.

Error Type	Error Category	Specific Errors
AI Generated Report Errors	Content Addition	Add Medical Device False Prediction False Negation
	Linguistic Quality	Add Repetitions Add Contradictions
	Context-Dependent	Change Name of Device Change Position of Device Change Severity Change Location Change Measurement
Human Errors	Content Addition and Context-Dependent	Human error - similar to above
	Linguistic Quality	Change to Homophone Add Typo

reasoning, lack of knowledge, etc.), with each of the 12 classifications focusing on the cause for such an error to occur..[5]

Errors from report generation models differ, with more specific issues including hallucinated references to prior studies, and have their own categorization framework. One example is the framework developed by Yu et al. to analyze common errors in model-generated radiology reports, aiming to create metrics that account for these errors and improve alignment with clinician feedback.[13] Their framework includes six categories: "False prediction of finding", "Omission of finding", "Incorrect location/position of finding", "Incorrect severity of finding", "Mention of comparison that is not present in the reference impression", and "Omission of comparison describing a change from a previous study." They develop a dataset, ReXVal, which contains annotations on clinically significant and insignificant errors under their six category framework for AI generated radiology reports with respect to ground-truth reports. Another dataset, Refisco, was created to categorize the errors commonly made in retrieval-based report generation models by their severity level and then correct each error using either deletion, substitution, or insertion of a line.[14] Both datasets provide different error categorization frameworks specific for AI-generated reports, offering error-report ground truth pairs with clinician annotations. However, their limited size (200 and 60 reports, respectively) underscores the need for more extensive datasets that contain error reports and ground truth pairs.

2.2. *Synthetic Data Generation for Radiology Reports*

Synthetic data generation is emerging as a valuable tool in radiology reporting research, addressing challenges of data scarcity. Recent studies have demonstrated its potential in various applications. Zhao et al.[15] generated modified reports with revision instructions, aiding in the training of instruction-based report revision systems. Hyland et al.[16] used GPT to paraphrase MIMIC dataset reports, expanding their training set for report generation models. Others have leveraged large language models to selectively modify radiology reports, addressing various clinical and research needs such as removing prior medical history references and standardizing report structures.[17,18] Most similarly, Asiimwe et al.[19] created a synthetic error

report set for developing an error detection and correction model. They intentionally introduced errors into radiology reports, focusing on four out of the six error categories defined by Yu et al.'s framework. This process resulted in the creation of 120,000 pairs of error-containing and error-free reports, which serve as training data for their model. These advancements highlight the growing importance of synthetic data in improving radiology reporting systems by enabling large-scale, precise data generation. However, it also reveals the need for a more comprehensive dataset that captures a wider range of diverse and complex errors.

Building upon these advancements, our study utilizes synthetic data generation to create a large-scale dataset that incorporates a broader range of errors. We expand on the framework established by Yu et al., addressing all six major categories of AI-generated errors, while also introducing additional subtypes such as device-related errors. Furthermore, our dataset addresses linguistic quality issues in both human- and AI-generated reports. This comprehensive approach allows us to create a more diverse and robust error dataset, providing a valuable resource for developing and evaluating advanced radiology reporting systems.

2.3. *Applications in Error Detection and Report Correction*

Our comprehensive error dataset has significant potential applications in advancing both error detection and report correction in radiology. In error detection, research has progressed from simple matching techniques for specific issues like laterality errors to more sophisticated methods using LSTM and BERT-based models.[20–22] Recent studies have even shown GPT-4's capability to identify common error categories (omission, insertion, spelling, and side confusion).[23] Our dataset, encompassing a wider range of error types, could further enhance these detection models.

In report correction, efforts have focused on addressing specific types of hallucinations in AI-generated reports, such as false references to non-existent prior scans.[24,25] The emerging task of report revision aims to refine existing reports through instruction prompts, as demonstrated in recent multi-functional foundation models.[6,15] Such an error-rich dataset could serve as a valuable resource for training and evaluating these correction and revision systems, potentially improving their ability to handle a diverse array of error types.

Furthermore, our dataset could be utilized as negative examples in reinforcement learning algorithms to enhance AI model performance, or to validate automatic evaluation metrics like RadCliQ and FineRadScore.[26] This broad applicability underscores the potential impact of our error injection method and resulting dataset in advancing the accuracy and reliability of radiological reporting systems.

3. Methods

We created a streamlined pipeline to inject errors into radiology reports, which can be used downstream to generate large datasets and train models for the identification and revision of incorrect radiology reports. We demonstrate error generation with the ReXErr pipeline using reports from the MIMIC-CXR train, dev, and test sets.[27] This pipeline supports two main tasks: report correction and sentence-level entailment. For both tasks, sentences are classified into three categories: correct (0), error (1), and neutral (2). Neutral sentences reference past

Table 2. Baseline prompting description for each error category.

Error	Baseline Instruction / Description
Add Medical Device	Add sentences that could be part of a radiology report regarding the presence of one or more devices such as these: pacemaker, central venous line, NG tube, ET tube, ICD.
Change Name of Device	If there is a medical device present in the report, change the name of the medical instrument to a different name that is clinically plausible.
Change Position of Device	If there is a medical device location present in the report, change the position of the medical instrument to a different position that is clinically plausible.
Change Severity	Change the severity of a finding in the report in a manner that makes clinical sense (e.g., change 'mild' to 'moderate').
Change Location	Change the location or anatomy of a finding in the report in a manner that is still clinically accurate (e.g., change 'right' to 'left' or 'lateral' to 'medial'; always modifying a sentence).
False Prediction	Add a finding that is not present in the report (either adding a sentence or modifying a sentence to insert).
False Negation	Change a particular finding from the report from present to absent by changing a sentence to indicate absence of the positive finding.
Change Measurement	If there is a measurement for a device/finding present, change the units of measurement (e.g., change 'cm' to 'mm') or change the value of the measurement to a different but still reasonable value (e.g. change '4.9 cm' to '5.8 cm').
Add Opposite Sentence	Add/alter a statement that is the opposite of another statement earlier in the same report.
Add Repetitions	Add repetitions of sentences present within the report.
Change to Homophone	Change a word in the report to a homophone of that word.
Add Typo	Add a typographical error in the report.

reports, findings, or scans and are categorized separately, as algorithms would not be able to determine their accuracy without additional context.

Report correction: Our pipeline generates paired ground truth and error reports, with each error report containing three errors sampled from 12 possible error categories. Sampling three errors per report provides a balanced representation of diverse error types while maintaining a degree of similarity to the original report and has been used prior in the literature.[23] We also separately specify the three error categories used in generating each report.

Sentence-level entailment: We provide a separate pipeline to create a sentence-level error categorization by splicing pairs of sentences from ground truth and error reports. Each pair includes the original sentence and its error version, their label (see categories below), type of error injected (error class) and sequence in the original report (index). Maintaining the sequential detail can help sentence-level entailment models developed upon data generated through ReXErr use contextual information from previous sentences to identify errors such as repetitions and contradictions.

3.1. *Error Categories*

Three board-certified radiologists were consulted in synthesizing the final list of errors included within this generation protocol. The errors fall under two broad categories: AI generated report errors and human errors. We further identify three sub-categories of errors: content addition, context-dependent, and linguistic quality errors. Each of the 12 final error categories fall under one of these subcategories and one of the two broader categories. The particular errors were determined in careful collaboration with radiologists; specifically, we used a set of radiologist-annotated reports generated from a current state of the art model to determine the most salient automated generation errors, and consulted radiologists directly to gain a sense for

human errors.[6] We incorporate all six major categories of AI-generated errors established by Yu et al.[13] The content addition and context-dependent errors observed in human-generated reports closely parallel those found in AI-generated reports. Additionally, we introduce a set of errors that address linguistic quality issues present in both human- and AI-generated reports, thereby creating a comprehensive error classification system. Table 1 contains a summary of the errors implemented.

3.2. *Data Synthesis*

After extensive iteration and feedback from clinical experts, we developed a comprehensive pipeline for introducing plausible errors into radiology reports using GPT-4o.[28] GPT-4o was chosen given its high performance relative to price. We define "plausible" errors as those that either a human or an AI model could realistically make. The pipeline employs a sophisticated sampling strategy to inject errors across all three categories within each report. Context-dependent errors are only introduced when the associated context is present, as determined by regex-based labeling that searches for specific keywords in each report. For instance, errors related to changing the location and type of medical devices are only injected if a device is mentioned in the report. The regex keywords for each category are constructed through a combination of clinician input and analysis of radiology report terms used in the dataset. Our approach balances the need for diverse and plausible errors while maintaining the overall structure and believability of the reports. The problem formulation for the injection of errors across all three categories is represented in Equation 1, where E_C, E_A, and E_L represent context-dependent, content addition, and linguistic quality errors respectively. T refers to the tags present, where $T \in \{\text{“device”}, \text{“measurement”}, \text{“location”}, \text{“severity”}\}$.

$$P(E_c, E_A, E_L \mid T) = P(E_c \mid T) \times P(E_A) \times P(E_L) \tag{1}$$

The probability of selecting both the content addition (E_A) and linguistic quality (E_L) errors are shown below in Equation 2. A and L both represent the number of individual errors present within the content addition and linguistic quality error categories respectively across both the AI and human groups. In our case, A would be 3 and L would be 4, where L includes the linguistic quality errors in both the AI and human error categories.

$$P(E_A) = \frac{1}{|A|} \qquad\qquad P(E_L) = \frac{1}{|L|} \tag{2}$$

The probability of selecting a context-dependent error given a particular tag is given by Equation 3 below, where the error for the context-dependent error category is sampled across the other categories if no relevant context is present. In the case where multiple tags are provided, the probability of selecting a particular context-dependent error given a tag depends on the normalized weight assigned to the tag ($w'(t_i)$) as well as the total number of context-dependent errors associated with each tag ($E(t)$).

$$P(E_c \mid T = t_i) = \begin{cases} \frac{w'(t_i)}{\sum_{t \in T} w'(t) \times E(t)} & \text{if } T \neq \emptyset \\ \frac{1}{|A|+|L|} & \text{if } T = \emptyset \end{cases} \tag{3}$$

Table 3. Examples of ground truth and error report generated through the ReXErr pipeline.

Ground Truth	Error Report	Errors Injected
Findings: Findings: The patient is status post median sternotomy and CABG. The heart size is top normal. The mediastinal and hilar contours are unremarkable. Bilateral calcified pleural plaques are seen diffusely which limits assessment of the underlying pulmonary parenchyma. No focal consolidation, pleural effusion or pneumothorax is clearly demonstrated. There are no acute osseous abnormalities. **Impression:** Bilateral calcified pleural plaques indicative of prior asbestos exposure. No definite acute cardiopulmonary abnormality otherwise noted.	**Findings:** The patient is status post median sternotomy and CABG. The heart size is enlarged. The mediastinal and hilar contours are unremarkable. Right calcified pleural plaques are seen diffusely which limits assessment of the underlying pulmonary parenchyma. No focal consolidation, pleural effusion or pneumothorax is clearly demonstrated. There is a suspected left clavicle fracture. **Impression:** Right calcified pleural plaques indicative of prior asbestos exposure. There is a moderate left pleural effusion. No definite acute cardiopulmonary abnormality otherwise noted.	'change location', 'false prediction', 'add contradiction'
Findings: Single frontal view of the chest provided. There is no focal consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette is normal. Again seen are multiple clips projecting over the left breast and remote left-sided rib fractures. No free air below the right hemidiaphragm is seen. **Impression:** No acute intrathoracic process.	**Findings:** Single frontal view of the chest provided. There is know focal consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette is normal. Again seen are multiple clips projecting over the left breast and remote left-sided rib fractures. There is an ET tube present in the trachea. No free air below the right hemidiaphragm is seen. No free air below the right hemidiaphragm is seen. **Impression:** No acute intrathoracic process.	'add repetitions', 'add medical devices', 'change to homophone'
Findings: There is mild-to-moderate cardiomegaly, not significantly changed compared with prior study. There is no pneumothorax. A newly placed endotracheal tube ends 4.3 cm above the carina. An NG tube is seen ending in the stomach with its tip and side ports beyond the margin of imaging. **Impression:** 1. Severe acute pulmonary edema. 2. Endotracheal tube ending 4.3 cm above the carina.	**Findings:** There is mild-to-moderate cardiomegaly, not significantly changed compared with prior study. There is no pneumothorax. A newly placed endotracheal tube ends 4.3 mm above the carina. An NG tube is seen ending in the stomach with its tip and side ports beyond the margin of imaging. **Impression:** 1. No pulmonary edema. 2. Endotrakheal tube ending 4.3 cm above the carina.	'change measurement', 'false negation', 'add typo'

The weights assigned to each tag $w(t_i)$ was calculated based on the frequency of each tag present within the reports through the equations shown below. Each weight is equal to the inverse of the prevalence of its respective tag. The weights are then normalized to $w'(t)$.

$$w(t) = \frac{1}{f(t)} \qquad W = \sum_{t \in T} w(t) \qquad w'(t) = \frac{w(t)}{W} \tag{4}$$

Based on this sampling strategy, GPT-4o was then used to inject the errors. Table 2 summarizes the baseline instructions given for each error type. Appendix A contains the complete long-form prompt used to prompt GPT, whereas Appendix B contains the particular prompts for each error category, including the examples for the relevant errors that use them.

3.3. *Sentence Level Error Generation Process*

Once the error reports were generated, each report was split into individual sentences and mapped based on sentence similarity to their corresponding ground truth sentence. We used Llama 3.1 to identify the error type in each sentence and screen for prior reports.[29] Llama 3.1 was chosen instead of GPT-4o for error relabeling due to its sufficient accuracy and greater

Table 4. Examples of ground truth and error sentences generated through the ReXErr sentence splicing and labeling pipeline.

Original Sentence	Error Sentence	Label	Error Class	Index
Findings: Comparison is made to previous study from ___.	Findings: Comparison is made to previous study from ___.	2	Not Applicable	0
There is a right-sided PICC line with distal lead tip at the cavoatrial junction.	There is a right-sided PICC line with distal lead tip at the mid SVC.	1	Change Position of Device	1
There has been removal of the right-sided chest tube.	There has been removal of the right-sided chest tube.	0	Not Applicable	2
There remains a curvilinear tubular device projecting over the mediastinum.	There remains a curvilinear tubular device projecting over the mediastinum.	0	Not Applicable	3
This has been seen on multiple images.	This has been seen on muitiple images.	1	Add Typo	4
There is persistent opacity at the left mid lung field and left-sided pleural effusion which is stable.	There is persistent opacity at the left mid lung field and left-sided pleural effusion which stable.	1	Add Typo	5
There is no pulmonary edema.	There is no pulmonary edema.	0	Not Applicable	6
The right lung is relatively clear.	The right lung is relatively clear.	0	Not Applicable	7
	The patient has had placement of an endotracheal tube.	1	Add Medical Device	8

cost-efficiency. The model was prompted to produce a Python dictionary with two keys: "label" and "error class." The "label" key indicated whether the sentence was correct (0), erroneous (1), or neutral (2), while the "error class" key specified the error type, if applicable. The "Add Repetition" error category was excluded, as repetition is only relevant at the report level, and "Add Opposite Sentence" was reclassified as "False Prediction." In cases where a new sentence was added, the original sentence field was left blank, and for omitted sentences, the error report sentence was left blank. Through this methodology, we are able to provide side-by-side comparisons between individual sentences and their associated error sentences. The order of sentences within the original report is maintained, including the position of particular added or omitted error sentences. The sentences were manually reviewed to ensure the accuracy of the sentence splicing.

3.4. *Validating Error Injection Pipeline*

In order to validate the quality and efficacy of our error injection pipeline, we analyze the projected frequency of every single error category injected across the MIMIC train, dev, and test subsets. Furthermore, a clinician reviewed 100 paired original and error-injected reports to determine the fraction of error reports which are plausible AI-generated or human-written reports. This was done to determine whether the synthesized error reports contain language atypical to radiology reports or very obvious modifications and statements that are not medically plausible which might limit the utility of the synthetic error reports.

Table 5. Distribution of errors inserted across the MIMIC train, dev, and test sets using the ReXErr methodology.

Error Category	Train (%)	Dev (%)	Test (%)
Add Medical Device	33.33	33.32	33.33
Change Name of Device	13.64	13.47	18.91
Change Position of Device	13.64	13.47	18.91
Change Severity	28.71	29.88	30.18
Change Location	38.07	37.07	23.26
False Prediction	33.33	33.32	33.33
False Negation	33.33	33.32	33.33
Change Measurement	5.93	6.10	7.01
Add Opposite Sentence	25.00	24.97	24.99
Add Repetitions	25.00	24.97	24.99
Change to Homophone	25.00	24.97	24.99
Add Typo	25.00	24.97	24.99

4. Results

4.1. *Strengths and Limitations of ReXErr*

The ReXErr pipeline was found to proficiently generate errors across all of the error categories listed for the majority of radiology report inputs. It is able to create multiple types of errors in the same report, with variation within each error subtype as well. These errors closely mimic those found in real-world report generation scenarios. Table 3 includes three examples of error reports generated using our report-level error injection pipeline, while Table 4 presents several examples of the sentence-level error generation process, along with the error labeling scheme. Despite ReXErr's ability to generate errors within the findings and impressions sections, there are still limitations in its ability to maintain consistency in the error injections across both sections. For example, while the first example in Table 3 is handled well, others such as the measurement change in the third example show discrepancies.

4.2. *Consistency Across Error Types*

ReXErr also demonstrates reasonable consistency in distribution of errors inserted across the MIMIC train, dev, and test sets. Certain errors, including "change measurement", "change name of device", and "change position of device" are injected less frequently in the dataset due to their reduced prevalence in the original reports. While the weighting mechanism used during sampling helped augment this discrepancy, this quantitative analysis highlights key areas for targeted improvements in developing more robust error injection and correction methods. Table 5 outlines the frequencies of each error type across the train, dev, and test sets, with each value representing the percentage of reports within the given set containing that specific error. Notably, these percentages are relatively consistent across the three different splits.

4.3. *Plausibility of Errors*

Lastly, ReXErr was found to predominantly inject plausible errors within reports. Plausible errors are mistakes that could reasonably occur in real-world radiology practice, while implausible errors involve anatomical impossibilities or fundamental misunderstandings of medical

principles that would otherwise never be made. Examples of implausible errors include substituting one medical device for another inappropriately (replacing "pacemaker" with "ET tube" in "A pacemaker is present with leads in the right ventricle"), or attributing findings to anatomical structures beyond the chest x-ray image. In the sample of 100 ground truth and error-injected reports reviewed by a board-certified clinician, 83 of the modified reports were found to be plausible, while only 17 contained errors that were implausible in AI-generated or human reports.

5. Discussion

Throughout this paper, we present ReXErr, a new pipeline designed to generate clinically relevant and plausible errors. Despite ReXErr's demonstrated capability to inject diverse errors, certain limitations that may hinder its use. Firstly, the applicability downstream models trained on data generated using ReXErr depends heavily on the quality and clinical relevance of the errors generated. While the majority of ReXErr-generated errors were plausible, we found 17 out of 100 augmented reports to contain implausible errors, meaning that the prompting methodology could be further improved before implementation on a larger scale.

Another potential limitation is the error sampling approach. ReXErr's sampling strategy does not account for nested compound errors, where errors can belong to more than one category, or cases where a sentence can contain multiple errors. Depending on how prevalent such errors are in actual human or AI generated reports, the absence of these errors could negatively impact ReXErr's downstream utility. Furthermore, downstream models may struggle to discriminate errors made in AI-generated text as ReXErr only adds errors to human-generated reports. Even though the errors themselves are sampled amongst errors commonly made by AI models, their addition to human generated text may not make them as representative as errors that were added to AI-generated text.

Lastly, future pipelines could benefit from more extensive downstream model testing using preliminary data generated. For example, while GPT-4o was chosen for its high performance and affordability, other open-source LLMs may yield more robust errors, and downstream testing would help elucidate which models can generate the most effective synthetic errors. Moreover, downstream testing would help determine whether the changes made to reports are significant enough for models to discern, as in some cases, the errors added are very minor.

6. Conclusion

Synthesizing accurate radiology reports is both difficult and time consuming, even for medical professionals. While automated AI generation approaches are promising in alleviating this workload and more efficiently generating comprehensive reports, they are liable to frequent errors across report content, linguistics, and consistency. Throughout this paper, we present the novel ReXErr method for generating annotated errors on both a report and sentence level. Developed with radiologists, ReXErr captures common AI and human errors in a representative and plausible manner, therefore offering a promising avenue for the development of report screening and correction algorithms as well as improving the accuracy of existing report generation approaches.

Acknowledgments

We would like to thank Dr. John Farner and Dr. Rohit Reddy for their valuable clinical input into the error categories and prompts chosen.

References

1. M. J. Côté and M. A. Smith, Forecasting the demand for radiology services, *Health Systems* **7**, 79 (2018), ISBN: 2047-6965 Publisher: Taylor & Francis.
2. B. I. Reiner, N. Knight and E. L. Siegel, Radiology reporting, past, present, and future: the radiologist's perspective, *Journal of the American College of Radiology* **4**, 313 (2007), ISBN: 1546-1440 Publisher: Elsevier.
3. A. Al Yassin, M. S. Sadaghiani, S. Mohan, R. N. Bryan and I. Nasrallah, It is About" Time": Academic Neuroradiologist Time Distribution for Interpreting Brain MRIs, *Academic Radiology* **25**, 1521 (2018), ISBN: 1076-6332 Publisher: Elsevier.
4. M. A. Bruno, E. A. Walker and H. H. Abujudeh, Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction, *Radiographics* **35**, 1668 (2015), ISBN: 0271-5333 Publisher: Radiological Society of North America.
5. A. P. Brady, Error and discrepancy in radiology: inevitable or avoidable?, *Insights into Imaging* **8**, 171 (February 2017).
6. H.-Y. Zhou, S. Adithan, J. N. Acosta, E. J. Topol and P. Rajpurkar, A Generalist Learner for Multifaceted Medical Image Interpretation (May 2024), arXiv:2405.07988 [cs].
7. T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno and I. Ktena, Towards generalist biomedical ai, *NEJM AI* **1**, p. AIoa2300138 (2024), ISBN: 2836-9386 Publisher: Massachusetts Medical Society.
8. C. Wu, X. Zhang, Y. Zhang, Y. Wang and W. Xie, Towards generalist foundation model for radiology, *arXiv preprint arXiv:2308.02463* (2023).
9. T. Tanida, P. Müller, G. Kaissis and D. Rueckert, Interactive and Explainable Region-guided Radiology Report Generation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , 7433 (2023).
10. P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andia, C. Tejos, C. Prieto and D. Capurro, A survey on deep learning and explainability for automatic report generation from medical images, *ACM Computing Surveys (CSUR)* **54**, 1 (2022), ISBN: 0360-0300 Publisher: ACM New York, NY.
11. P. Sloan, P. Clatworthy, E. Simpson and M. Mirmehdi, Automated Radiology Report Generation: A Review of Recent Advances, *IEEE Reviews in Biomedical Engineering* (2024), Publisher: IEEE.
12. Y. W. Kim and L. T. Mansfield, Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors, *American journal of roentgenology* **202**, 465 (2014).
13. F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad and A. Y. Ng, Evaluating progress in automatic chest x-ray radiology report generation, *Patterns* **4** (2023), ISBN: 2666-3899 Publisher: Elsevier.
14. K. Tian, S. J. Hartung, A. A. Li, J. Jeong, F. Behzadi, J. Calle-Toro, S. Adithan, M. Pohlen, D. Osayande and P. Rajpurkar, ReFiSco: Report Fix and Score Dataset for Radiology Report Generation, *PhysioNet* (2023).
15. B. N. Zhao, X. JIANG, X. Luo, Y. Yang, B. Li, Z. Wang, J. Alvarez-Valle, M. P. Lungren, D. Li and L. Qiu, Large Multimodal Model for Real-World Radiology Report Generation (September 2023).
16. S. L. Hyland, S. Bannur, K. Bouzid, D. C. Castro, M. Ranjit, A. Schwaighofer, F. Pérez-García, V. Salvatelli, S. Srivastav and A. Thieme, MAIRA-1: A specialised large multimodal model for radiology report generation, *arXiv preprint arXiv:2311.13668* (2023).

17. O. Banerjee, H.-Y. Zhou, S. Adithan, S. Kwak, K. Wu and P. Rajpurkar, Direct preference optimization for suppressing hallucinated prior exams in radiology report generation, *arXiv preprint arXiv:2406.06496* (2024).
18. O. Banerjee, A. Saenz, K. Wu, W. Clements, A. Zia, D. Buensalido, H. Kavnoudias, A. S. Abi-Ghanem, N. E. Ghawi, C. Luna *et al.*, Rexamine-global: A framework for uncovering inconsistencies in radiology report generation metrics, *arXiv preprint arXiv:2408.16208* (2024).
19. A. C. Asiimwe, D. Surís, P. Rajpurkar and C. Vondrick, Image-conditioned autocorrection in medical reporting (2024).
20. Y. H. Lee, J. Yang and J.-S. Suh, Detection and Correction of Laterality Errors in Radiology Reports, *Journal of Digital Imaging* **28**, 412 (August 2015).
21. J. Zech, J. Forde, J. J. Titano, D. Kaji, A. Costa and E. K. Oermann, Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models, *Annals of translational medicine* **7** (2019), Publisher: AME Publications.
22. D. Min, K. Kim, J. H. Lee, Y. Kim and C. M. Park, RRED: a radiology report error detector based on deep learning framework, *Proceedings of the 4th Clinical Natural Language Processing Workshop* , 41 (2022).
23. R. J. Gertz, T. Dratsch, A. C. Bunck, S. Lennartz, A.-I. Iuga, M. G. Hellmich, T. Persigehl, L. Pennig, C. H. Gietzen, P. Fervers, D. Maintz, R. Hahnfeldt and J. Kottlors, Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy, *Radiology* **311**, p. e232714 (April 2024).
24. V. Ramesh, N. A. Chi and P. Rajpurkar, Improving radiology report generation systems by removing hallucinated references to non-existent priors, in *Machine Learning for Health*, (PMLR, 2022).
25. O. Banerjee, H.-Y. Zhou, S. Adithan, S. Kwak, K. Wu and P. Rajpurkar, Direct Preference Optimization for Suppressing Hallucinated Prior Exams in Radiology Report Generation (June 2024), arXiv:2406.06496 [cs].
26. A. Huang, O. Banerjee, K. Wu, E. P. Reis and P. Rajpurkar, FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores (May 2024), arXiv:2405.20613 [cs].
27. A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark and S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, *Scientific data* **6**, p. 317 (2019), ISBN: 2052-4463 Publisher: Nature Publishing Group UK London.
28. O. Team, Hello GPT-4o, *OpenAI* (May 2024).
29. M. A. Team, Introducing Llama 3.1: Our most capable models to date, *Meta AI* .

7. Appendix

Please find the appendix here: `https://drive.google.com/file/d/15dCVF8yh8i6UIOaS_m5-biA28fCSiQjh/view?usp=sharing`

LLM-CGM: A Benchmark for Large Language Model-Enabled Querying of Continuous Glucose Monitoring Data for Conversational Diabetes Management

Elizabeth Healey[†]
Program in Health, Sciences, and Technology, Massachusetts Institute of Technology, Cambridge, MA 02138, USA
[†]*E-mail: ehealey@mit.edu*

Isaac Kohane
Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

Over the past decade, wearable technology has dramatically changed how patients manage chronic diseases. The widespread availability of on-body sensors, such as heart rate monitors and continuous glucose monitoring (CGM) sensors, has allowed patients to have real-time data about their health. Most of these data are readily available on patients' smartphone applications, where patients can view their current and retrospective data. For patients with diabetes, CGM has transformed how their disease is managed. Many sensor devices interface with smartphones to display charts, metrics, and alerts. However, these metrics and plots may be challenging for some patients to interpret. In this work, we explore how large language models (LLMs) can be used to answer questions about CGM data. We produce an open-source benchmark of time-series question-answering tasks for CGM data in diabetes management. We evaluate different LLM frameworks to provide a performance benchmark. Lastly, we highlight the need for more research on how to optimize LLM frameworks to best handle questions about wearable data. Our benchmark is publicly available for future use and development. While this benchmark is specifically designed for diabetes care, our model implementation and several of the statistical tasks can be extended to other wearable device domains.

Keywords: Large Language Models, Human-AI Interaction, Diabetes, Time Series

1. Introduction

Large language models (LLMs) have demonstrated tremendous promise in transforming how information is automatically distilled and extracted. In clinical medicine, there has been much excitement about how LLMs can transform the way doctors and patients interact with healthcare systems.[1–4] Recent literature has demonstrated the ability of LLMs to extract medical information and provide clinical summaries,[5–8] even using medical images as input.[9,10] These advances have the potential to dramatically change the way that patients and clinicians interact with medical data. Despite these advances, there has been less focus on how LLMs can be used to extract information from time-series data from patient-owned medical devices.

In diabetes management, patient interpretation and understanding of their data is key to making behavioral modifications. In recent years, the use of wearable continuous glucose monitors (CGMs) for diabetes management has increased.[11] These devices are worn on the body and measure interstitial blood glucose approximately every 5 minutes. These devices allow patients to view both their real-time and retrospective data on their smart devices. The insights gained from CGM data are important for helping patients make behavioral and treatment modifications to manage their diabetes.[12] While several applications exist where patients can view their retrospective data, some patients may find the interpretation of CGM data to be challenging.[13]

In this work, we develop a benchmark of CGM question-answering (QA) tasks: LLM-CGM. Figure 1 shows a schematic of the ideal system for LLM-enabled QA for CGM data. In this setup, the user could ask a question about their CGM data, and receive a written answer in return, thus transforming the way patients interact with their data.

Our contributions can be summarized as follows:

(1) We outline four categories of tasks for CGM QA. We articulate subtasks describing potential natural language queries about the data for each task. For each subtask, we include sample question queries. The final benchmark contains a total of 30 questions.
(2) We provide a module to get the empirical answer questions in the benchmark from any raw CGM data for evaluation.
(3) We implement three distinct baseline approaches to LLM QA of time-series data and show the performance on the benchmark tasks.
(4) We evaluate our benchmark using synthetic and real CGM data of up to 14 days in length.

LLM-CGM can be accessed at https://github.com/lizhealey/LLM-CGM and can be leveraged to evaluate future iterations of LLMs for diabetes.

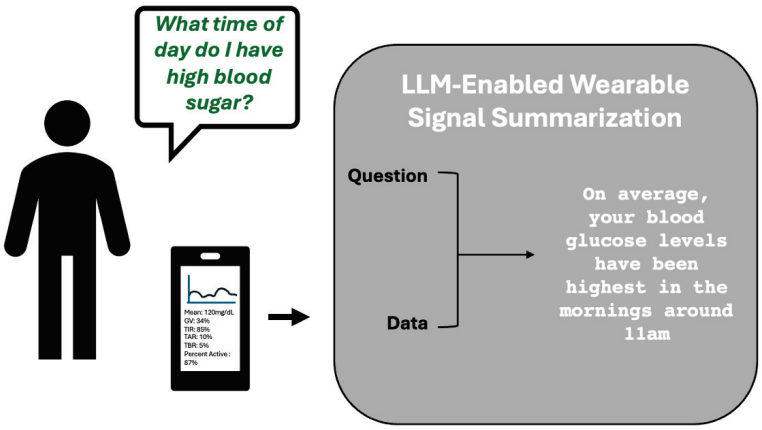


Fig. 1. Illustrative overview.

2. Related Work

2.1. *LLMs of Time Series Interpretation*

Recent work has investigated using LLMs for time-series data analysis,[14] with a subset of this space focusing on how LLMs can be used to interpret and understand time series data.[15] In the medical domain, there has been interest in building benchmarks for question-answering (QA) tasks for wearable data. ECG-QA provides a QA dataset with a benchmark for 70 questions related to electrocardiogram interpretation.[16] The Personal Health Large Language Model (PH-LLM) was developed to provide insights on sleep and fitness goals from wearable data.[17] Similar work was recently published by Merrill et al., where they proposed a Personal Health Insights Agent (PHIA),[18] which leverages code generation and information retrieval to respond to questions about data from wearable devices, such as step count. Our work builds upon this previous work by providing a benchmark for wearable health data interpretation with tasks specific for CGM data.

2.2. *Diabetes Technology*

Interest has also increased in using LLMs to enhance diabetes management through education and personal coaching.[19,20] A previous randomized control trial investigated using voice-based AI to help patients with T2D manage their insulin,[21] and they found that the AI application benefited patients' glycemic control. Other work has investigated a conversational health agent for patients with diabetes, incorporating carbohydrate information and guidelines.[22] Recently, a few works have investigated using LLMs, such as GPT-4,[23] to summarize CGM data.[24,25] These works have explored how LLMs are capable of interpreting CGM data to produce easily understandable summaries. Given the recent interest in the development of diabetes chatbots, there is a need for further investigation of how to optimize LLMs for the analysis of CGM data. Our work fills this gap by presenting a benchmark for conversational queries about CGM data and a preliminary evaluation of different LLM frameworks.

3. Methods

3.1. *Benchmarking Tasks*

Queries of CGM data can have either objective or subjective answers. Many QA tasks for CGM are subjective and depend on the specific patient circumstances. For example, a query of "Is my blood glucose control good?" is subjective and requires consideration of the patient's medical context. In this work, we focus on CGM tasks that can primarily be answered objectively.

Figure 2 gives an overview of the four task categories and subtasks, with example questions. The tasks are broken down into categories that are delineated by both the computational processes required to get an answer and the domain knowledge necessary to understand the task. Many of the questions are inspired by guidelines from the American Diabetes Association (ADA) on glycemic control[26] and current frameworks for analyzing CGM data.[27] Table 1 shows the 30 questions included in this benchmark that are distributed across the task categories. While there are many more types of questions that patients may want to ask about their data, the purpose of these 30 questions was to provide a foundational baseline for a range of query types.

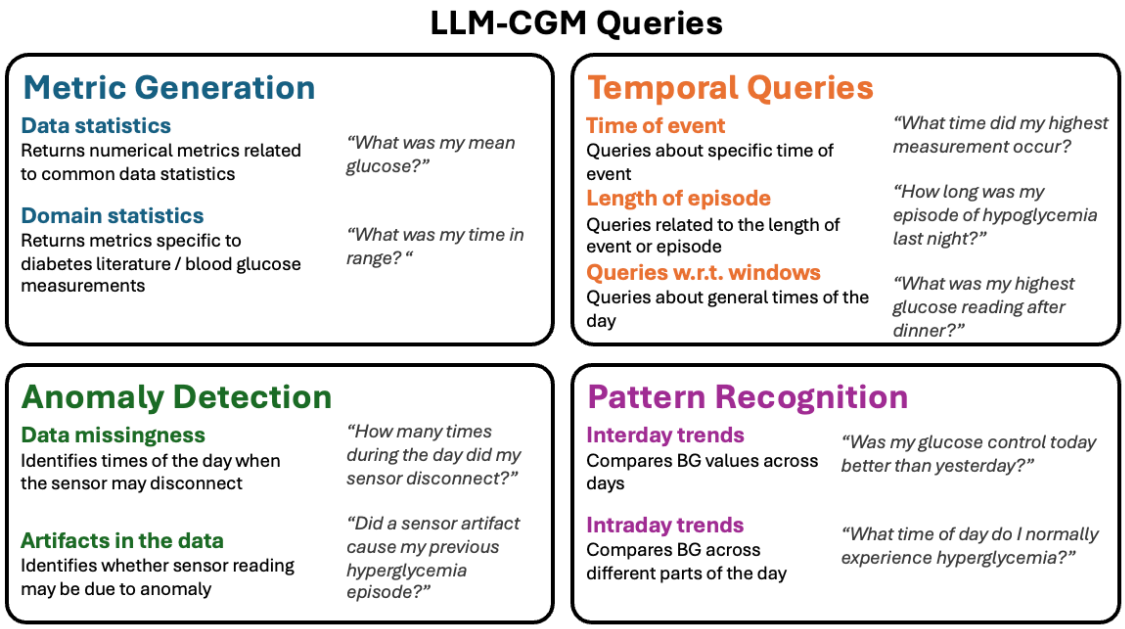


Fig. 2. Benchmarking tasks by category and subcategory

3.2. *Task Evaluation*

Our curated list of 30 tasks has Python-generated solutions. We include the full list of tasks and how they are evaluated in Table 1. Given any comma-separated value (CSV) file as input with a column for the CGM values and timestamp, we automatically compute the answers to the queries using the definitions in the table. For some tasks, the quantitative answer can be subjective. For example, some of the questions depend on the period in which breakfast and dinner are defined. These queries are noted in the table.

3.3. *Model Framework*

In our analysis, we use GPT-4[23] to generate text responses. We test three different frameworks designed to analyze CGM data using GPT-4 that serve as baselines. The details for each model and prompt framework can be found in Figure 3 and we also describe each below.

(1) **LLM-Text:** LLM-Text is a naive implementation where the CGM data and time stamps are inputted to the language model as text as part of the prompt.

(2) **LLM-Code:** LLM-Code is a framework implemented in Python with three main steps. This framework was inspired by recent work examining the ability of GPT-4 to analyze data.[30] In their work, they create a framework where the language model writes code that is automatically executed. We adapt that approach to our setting. In the first step, the

Table 1. LLM-CGM Benchmark Queries and Solutions. The colors correspond to benchmark task categories.

	User Question	Ground Truth Answer
Q1	What was my mean glucose?	Mean of glucose readings
Q2	What was my maximum glucose?	Maximum of glucose readings
Q3	What was the standard deviation of my glucose?	Standard deviation of glucose readings
Q4	What was my minimum glucose?	Minimum of glucose readings
Q5	What was my percent time in range?	Percent time between 70 mg/dL and 180mg/dL
Q6	What was my percent time in hyperglycemia?	Percent time above 180 mg/dL
Q7	What was my percent time in hypoglycemia?	Percent time below 70mg/dL
Q8	What was my glycemic variability?	Standard deviation divided by mean of glucose readings
Q9	What was my percent time in severe hyperglycemia?	Percent of time spent above 250 mg/dL
Q10	What is my estimated A1C?	Using estimated average glucose formula[28]
Q11	What was my percent time in severe hypoglycemia?	Percent time spent below 54 mg/dL
Q12	What time was my blood glucose highest?	Date and time when blood glucose was max
Q13	What day was my glucose control the most out of range?	Day with greatest absolute time outside of range 70-180mg/dL
Q14	What time of the day was my blood glucose lowest?	Date where minimum glucose reached
Q15	When did my most recent episode of hypoglycemia occur?	Time of most recent hypoglycemia episode
Q16	How long was my last episode of hypoglycemia?	Length of most recent period where glucose was consistently below 70mg/dL
Q17	What was my longest time spent in hyperglycemia?	Longest period where glucose was over 180mg/dL
Q18	How many times did I experience hypoglycemia?	Number of episodes where glucose was less than 70mg/dL
Q19	What was my mean overnight blood glucose?	Mean glucose from 12am to 6am**
Q20	What meal of the day did I have the highest blood glucose?	Time window with max glucose where breakfast is 6am-11am, lunch is 11am-4pm, dinner is 5pm-9pm**
Q21	Did I have noctural hypoglycemia?	Yes if blood glucose was less than 70mg/dL between 12am and 6am**
Q22	What was my highest glucose reading during dinner?	Maximum glucose any day between 5pm and 10pm**
Q23	Is there any missingness in the data?	Yes if there are gaps between data longer than 5 minutes
Q24	How many times did my sensor disconnect ?	Number of gaps greater than 5 minutes
Q25	Was my low blood glucose likely due to sensor error?	Yes if reading less than 70 mg/dL due to sensor anomaly*
Q26	Are there any artifacts in the CGM data?	Yes if there was a sensor anomaly in data causing observed glucose reading*
Q27	Was my glucose control today better than yesterday?	Yes if mean glucose on current day was better than previous day**
Q28	Was my time in range improved this week compared to last week?	Yes if time in range for the most recent week was better than the previous week*
Q29	Was my max glucose lower today than yesterday?	Yes if the maximum glucose on most recent day was lower than the previous day
Q30	Did I spend less time in hypoglycemia this week than last week?	Yes if total minutes in hypoglycemia for the most recent week was less than the previous week*
	*Not included in this evaluation	** May be subjective

LLM writes a Python script that begins by loading a CSV file with the CGM data. We then program LLM-Code to automatically execute the Python script and produce text in a new file. The final answer is obtained from the text file.

(3) **LLM-CodeChain:** Our workflow leverages the *create_csv_agent*() from Langchain[29] that allows the use of a Python tool. This allows the agent to write and run code to analyze the CSV file. We use Langchain to connect to OpenAI's GPT-4 model.[23] The agent takes the preprocessed CSV file as input, along with a prompt. The output is a generated narrative and the log of computations. This framework is most similar to recent work PHIA,[18] where the LLM can iterate through a thought-action chain.

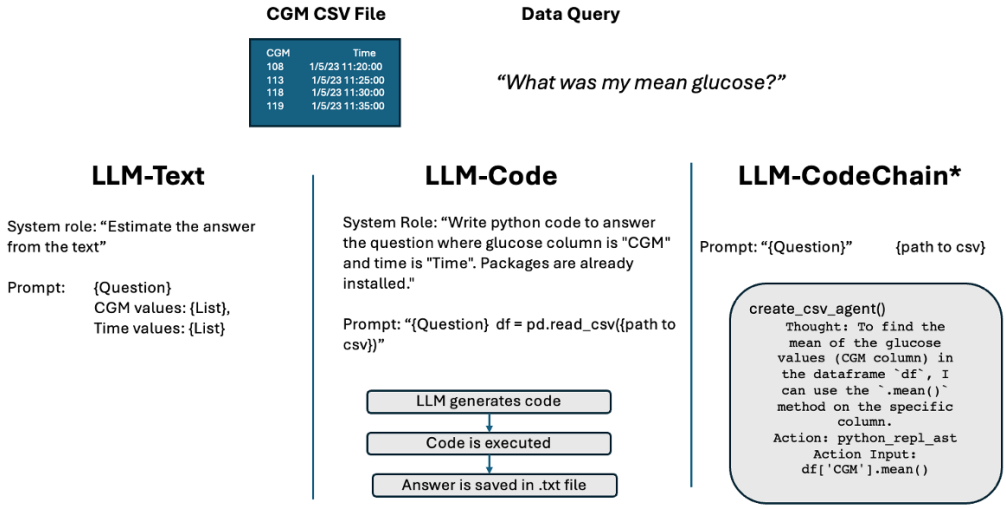


Fig. 3. Model and prompt frameworks included in benchmark for testing and evaluation. LLM-CodeChain leverages builtin functions in Langchain[29]

Prompts: Figure 3 shows the prompts used as input for each of the model frameworks. The prompts always include the query and, depending on the model framework, some information about the context of the data. Future evaluations should include retrieval-augmented generation, where the prompt includes information about diabetes, including definitions of terms and instructions on how to analyze the data.

Technical Specifications: For all model implementations, we generate text using OpenAI's GPT-4.[23] Our repository enables the testing of multiple models; however, for this paper, all experiments were done using the model "gpt-4-0125-preview", with the temperature set to .1.

3.4. *Simulated Data*

While there are many available datasets with CGM data from T1D, many require a data-use agreement to be signed. Since uploading data to open-source LLMs conflicts with the terms of these agreements, we curated our own CGM dataset using an FDA-accepted T1D patient simulator.[31] We generated five different cases of roughly 14 days of CGM data sampled every five minutes. The simulator used was generated from an open-source Python patient simulator.[32] The characteristics of this dataset are visualized in Table 2 and Figure 4. By using the patient simulator, we were able to curate a dataset with variable glycemic control. Simulated cases had significantly varying glycemic signatures and characteristics, with some patients spending a majority of their time in healthy glucose range, and with some individuals spending less than 50% of the time in healthy glucose range.

3.5. *Real Data*

We also used publicly available real CGM data,[33] that was collected from individuals with diabetes, pre-diabetes, and no diabetes. For this work, we only use five individuals in our analysis to demonstrate the performance of LLMs on various CGM QA tasks. This subset included three individuals with pre-diabetes and two without diabetes. The characteristics of this dataset can be visualized in Table 2 and Figure 4.

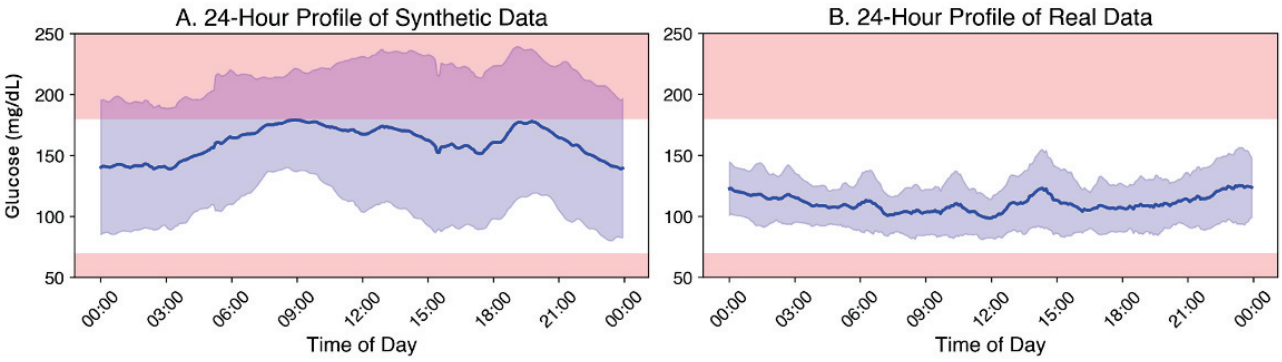


Fig. 4. Data included in benchmark: (A) 24-hour mean and standard deviation of 5 cases from synthetic data simulating patients with T1D. (B) 24-hour mean and standard deviation from 5 cases from the real dataset[33]

Table 2. Characteristics of data. We show the mean value for each of the statistics, as well as the minimum value in the dataset and maximum value in the dataset.

	Synthetic T1D Data (n=5)			Real Data (n=5)		
	Mean	Min	Max	Mean	Min	Max
Number of data points	4033.0 (0.0)	4033	4033	1875.2 (171.814)	1779	2180
Average glucose (mg/dL)	168.085 (25.887)	130.298	196.627	108.052 (7.021)	97.013	116.556
Glucose management indicator	7.331 (0.619)	6.427	8.013	5.895 (0.168)	5.631	6.098
Coefficient of variation	0.3 (0.04)	0.242	0.354	0.172 (0.037)	0.135	0.225
Minimum glucose (mg/dL)	53.15 (10.846)	43.888	71.121	65.0 (4.528)	58	69
Maximum glucose (mg/dL)	352.354 (63.41)	267.212	400	192.8 (33.937)	144	234
Percent time sensor active	1.0 (0.0)	1	1	0.465 (0.043)	0.441	0.541
Percent time in range (70mg/dL-180mg/dL)	0.644 (0.175)	0.472	0.901	0.987 (0.011)	0.975	0.997
Percent time above range 1 (>180mg/dL)	0.349 (0.177)	0.099	0.526	0.006 (0.008)	0	0.02
Percent time above range 2 (>250mg/dL)	0.094 (0.087)	0.001	0.224	0.0 (0.0)	0	0
Percent time below range 1 (<70mg/dL)	0.007 (0.011)	0	0.026	0.004 (0.005)	0.001	0.012
Percent time below range 2 (<54mg/dL)	0.001 (0.002)	0	0.005	0.0 (0.0)	0	0

4. Results

In Table 3, we show the results categorized by the model type and the task categories. The questions are shown individually across all cases, with total scores also listed for each task category. We found that for simpler tasks, such as metric generation, performance was high. Errors were often caused by a misinterpretation of the task. For example, when computing

glycemic variability, the LLM would return the standard deviation, not the coefficient of variation (Q8). The more complicated tasks had higher error rates. This was seen through anomaly detection tasks and pattern recognition tasks. We also note that the performance of LLM-Code compared to LLM-CodeChain varied depending on the tasks.

Table 4 gives examples of incorrect answers by framework. During our evaluation, there were many times when the model did not produce an answer. This was often due to an error in the original code. For these instances, instead of rerunning the example, we counted the instances as inaccurate. These instances often occurred for tasks that were complicated, and the model output suggested the limitation was due to inadequate information. For most tasks, LLM-Code outperformed LLM-CodeChain. A notable limitation with LLM-Code is that code is only written once, so the agent has no ability to rewrite code based on the output. This is seen as an example in Table 4 where the length of the most recent episode of hypoglycemia was not able to be computed. However, for some of the more complicated temporal queries, LLM-CodeChain outperforms LLM-Code for the real cases.

Performance for the anomaly detection tasks and pattern recognition tasks were particularly low. This was due to the fact that the computations necessary to answer these was more complicated than to those of the other tasks. Without any information in the prompt about what to execute, the LLM fails to answer correctly most of the time. Additionally, the prompts did not include any information on what day "today" was, impairing the performance.

We do not show the results for the LLM-Text framework due to the fact that there was very poor performance for most of the tasks. The data used in our evaluation had CGM traces of up to 14 days in length. This caused the token size of the model input to be extremely large and the LLM struggled to return even basic estimates. We expect that the performance could likely increase with smaller amounts of CGM data. An example output of LLM-Text to Q1 is seen in Table 4.

There was some subjectivity when grading whether or not the LLM outcome was accurate. For example, some numerical results were rounded, or within a very small margin of error. For questions that returned percentages and values, answers were marked as correct if they were equivalent when rounded to the nearest whole number. For questions related to meal times, such as Q19 and Q22, answers were marked correct if they were within 10mg/dL of the solution. We omitted four questions in the analysis presented in the paper. We omitted Q28 and Q30 since they are dependent on how a week is defined. We also omitted Q25 and Q26 because the data we used had no documented artifacts.

5. Conclusions

In this work, we developed a benchmark for LLM-enabled CGM QA tasks. We hope that this work promotes further investigation of conversational agents for diabetes management. Our work highlighted the potential for innovation of LLM frameworks for wearable data analysis. LLM-Code and LLM-CodeChain both involved leveraging Python to analyze the data based on the LLM output. LLM-Code was limited by the fact that it was designed only to be able to write one Python script. We suspect for more complex tasks, LLM-CodeChain has benefits that should be further investigated.

Table 3. Table shows the fraction of CGM cases with correct answer for each question. Results are broken down by the model framework used (LLM-Code vs LLM-CodeChain) and the data type

Metric Generation	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**
LLM-Code Synth (n=5)	1	1	1	1	.8	.8	.8	0	1	1	1
LLM-Code Real (n=5)	1	.8	1	1	1	1	1	0	1	1	1
LLM-Code Total (n=10)	1	.9	1	1	.9	.9	.9	0	1	1	1
LLM-Codechain Synth (n=5)	1	1	1	1	.2	1	1	0	1	1	1
LLM-Codechain Real (n=5)	1	1	1	1	0	1	1	0	1	1	1
LLM-Codechain Total (n=10)	1	1	1	1	.1	1	1	0	1	1	1
Temporal Queries	**Q12**	**Q13**	**Q14**	**Q15**	**Q16**	**Q17**	**Q18**	**Q19**	**Q20**	**Q21**	**Q22**
LLM-Code Synth (n=5)	1	1	.8	.8	.6	0	.2	.8	.4	1	.6
LLM-Code Real (n=5)	1	1	.8	.6	.4	.8	.4	.8	.2	1	.6
LLM-Code Total (n=10)	1	1	.8	.7	.5	.4	.3	.8	.3	1	.6
LLM-Codechain Synth (n=5)	1	0	1	.8	.2	.2	.2	.4	0	.4	0
LLM-Codechain Real (n=5)	1	.2	1	1	.2	.2	.4	.6	.8	1	0
LLM-Codechain Total(n=10)	1	.1	1	.9	.2	.2	.3	.5	.4	.7	0
Anomaly Detection	**Q23**	**Q24**		**Pattern Recognition**					**Q27**	**Q29**	
LLM-Code Synth (n=5)	.8	1		LLM-Code Synth (n=5)					0	.4	
LLM-Code Real (n=5)	0	0		LLM-Code Real (n=5)					0	.6	
LLM-Code Total (n=10)	.4	.5		LLM-Code Total (n=10)					0	.5	
LLM-Codechain Synth (n=5)	.8	.2		LLM-Codechain Synth (n=5)					0	0	
LLM-Codechain Real (n=5)	0	0		LLM-Codechain Real (n=5)					.4	.4	
LLM-Codechain Total(n=10)	.4	.1		LLM-Codechain Total(n=10)					.2	.2	

Table 4. Error Examples

Framework	Error Examples
LLM-Text	Q1: "Given the extensive list of values, a precise calculation here isn't feasible without computational tools. However, we can provide a rough estimate by observing the range and distribution of values. The glucose values range from lows around 137 to highs of 322. Most values seem to cluster in the mid-100s to mid-200s, with occasional peaks above 300. Given this distribution, it's reasonable to estimate that the mean glucose level might fall somewhere in the mid to upper 100s, possibly around the 180 to 220 range."
LLM- Code	Q17: "Based on the data provided, your longest time spent in hyperglycemia was 23 hours and 35 minutes." Q16: "If your episode of hypoglycemia is ongoing, it's not possible to determine its duration until it has concluded."
LLM- CodeChain	Q20: "The period of the day when you had the highest blood glucose was at 07:45:00 on January 9, 2024" Q24: "Without further information on how sensor disconnections are indicated in the data, it's not possible to determine the number of times the sensor disconnected based on the provided information."

This study had several limitations. We used a general purpose model that had not been fine-tuned on any diabetes guidelines as our baseline. In our work, we showed baselines for the performance of GPT-4 in answering these questions. Future work should investigate different models, as well as different prompting techniques. Future work should also investigate performance on different data. In this work, we used a mix of synthetic data and real data. The performance of these frameworks may vary with real CGM data that is different than what was tested. In particular, the performance may vary based on the length of the data being analyzed.

There are particular safety concerns when developing and implementing LLMs for diabetes management. Even in the absence of LLM-generated medical advice, incorrect assessment of glucose data could cause patients to incorrectly dose insulin and put them at risk for life-threatening hypoglycemia. While the framework we proposed in this work is a promising research direction, incorrect answers pose a safety risk. These safety risks should inform future model development and evaluation. Lastly, future work should explore how clinicians and patients evaluate the output of these LLMs. While this work focused on benchmarking the accuracy of QA tasks, there is much to be investigated to determine the clinical utility of LLM-

enabled CGM analysis. The 30 questions in this benchmark were included to demonstrate the breadth of questions that could be asked about CGM data. In the future, the benchmark will expand with questions derived from patients themselves.

6. Acknowledgments

This work was supported by the National Science Foundation Graduate Research Fellowship Program under grant number 2141064.

References

1. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan and D. S. W. Ting, Large language models in medicine, *Nat. Med.* **29**, 1930 (August 2023).
2. N. H. Shah, D. Entwistle and M. A. Pfeffer, Creation and adoption of large language models in medicine, *JAMA* **330**, 866 (September 2023).
3. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera Y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam and V. Natarajan, Large language models encode clinical knowledge, *Nature* **620**, 172 (August 2023).
4. T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev, S. Azizi, K. Singhal, Y. Cheng, L. Hou, A. Webson, K. Kulkarni, S. Sara Mahdavi, C. Semturs, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, A. Karthikesalingam and V. Natarajan, Towards conversational diagnostic AI (January 2024).
5. D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly and A. S. Chaudhari, Adapted large language models can outperform medical experts in clinical text summarization, *Nat. Med.* **30**, 1134 (April 2024).
6. M. Agrawal, S. Hegselmann, H. Lang, Y. Kim and D. Sontag, Large language models are Few-Shot clinical information extractors (May 2022).
7. L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau, C. Weng and Y. Peng, Evaluating large language models on medical evidence summarization, *NPJ Digit Med* **6**, p. 158 (August 2023).
8. K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. O. Sabel, J. Ricke and M. Ingrisch, ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports, *Eur. Radiol.* **34**, 2817 (May 2024).
9. S. Lee, W. J. Kim, J. Chang and J. C. Ye, LLM-CXR: Instruction-Finetuned LLM for CXR image understanding and generation (May 2023).
10. M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, A. V. Parwani, A. Zhang and F. Mahmood, A visual-language foundation model for computational pathology, *Nat. Med.* **30**, 863 (March 2024).
11. G. Cappon, M. Vettoretti, G. Sparacino and A. Facchinetti, Continuous glucose monitoring sensors for diabetes management: A review of technologies and applications, *Diabetes Metab. J.* **43**, 383 (August 2019).
12. N. Ehrhardt and E. Al Zaghal, Continuous glucose monitoring as a behavior modification tool, *Clin. Diabetes* **38**, 126 (April 2020).
13. K. Mackett, H. Gerstein and N. Santesso, Patient perspectives on the ambulatory glucose profile

report for type 1 diabetes management in adults: A national online survey, *Can J Diabetes* **47**, 243 (April 2023).
14. X. Zhang, R. R. Chowdhury, R. K. Gupta and J. Shang, Large language models for time series: A survey, *arXiv [cs.LG]* (February 2024).
15. E. Fons, R. Kaur, S. Palande, Z. Zeng, S. Vyetrenko and T. Balch, Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark, *arXiv [cs.CL]* (April 2024).
16. J. Oh, G. Lee, S. Bae, J.-M. Kwon and E. Choi, ECG-QA: A comprehensive question answering dataset combined with electrocardiogram, *arXiv [q-bio.QM]* (June 2023).
17. J. Cosentino, A. Belyaeva, X. Liu, N. A. Furlotte, Z. Yang, C. Lee, E. Schenck, Y. Patel, J. Cui, L. D. Schneider, R. Bryant, R. G. Gomes, A. Jiang, R. Lee, Y. Liu, J. Perez, J. K. Rogers, C. Speed, S. Tailor, M. Walker, J. Yu, T. Althoff, C. Heneghan, J. Hernandez, M. Malhotra, L. Stern, Y. Matias, G. S. Corrado, S. Patel, S. Shetty, J. Zhan, S. Prabhakara, D. McDuff and C. Y. McLean, Towards a personal health large language model (June 2024).
18. M. A. Merrill, A. Paruchuri, N. Rezaei, G. Kovacs, J. Perez, Y. Liu, E. Schenck, N. Hammerquist, J. Sunshine, S. Tailor, K. Ayush, H.-W. Su, Q. He, C. Y. McLean, M. Malhotra, S. Patel, J. Zhan, T. Althoff, D. McDuff and X. Liu, Transforming wearable data into health insights using large language model agents (June 2024).
19. G. G. R. Sng, J. Y. M. Tung, D. Y. Z. Lim and Y. M. Bee, Potential and pitfalls of ChatGPT and Natural-Language artificial intelligence models for diabetes education, *Diabetes Care* **46**, e103 (March 2023).
20. B. Sheng, Z. Guan, L.-L. Lim, Z. Jiang, N. Mathioudakis, J. Li, R. Liu, Y. Bao, Y. M. Bee, Y.-X. Wang, Y. Zheng, G. S. W. Tan, H. Ji, J. Car, H. Wang, D. C. Klonoff, H. Li, Y.-C. Tham, T. Y. Wong and W. Jia, Large language models for diabetes care: Potentials and prospects, *Sci Bull (Beijing)* **69**, 583 (March 2024).
21. A. Nayak, S. Vakili, K. Nayak, M. Nikolov, M. Chiu, P. Sosseinheimer, S. Talamantes, S. Testa, S. Palanisamy, V. Giri and Others, Use of Voice-Based conversational artificial intelligence for basal insulin prescription management among patients with type 2 diabetes: A randomized clinical trial, *JAMA Network Open* **6**, e2340232 (2023).
22. M. Abbasian, Z. Yang, E. Khatibi, P. Zhang, N. Nagesh, I. Azimi, R. Jain and A. M. Rahmani, Knowledge-Infused LLM-Powered conversational health agent: A case study for diabetes patients (February 2024).
23. OpenAI, GPT-4 technical report (March 2023).
24. C. Martinez-Cruz, J. F. G. Guerrero, J. L. L. Ruiz, A. J. Rueda and M. Espinilla, A first approach to the generation of linguistic summaries from glucose sensors using GPT-4, in *Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023)*, (Springer Nature Switzerland, 2023).
25. E. Healey, A. Tan, K. Flint, J. Ruiz and I. Kohane, Leveraging large language models to analyze continuous glucose monitoring data: A case study, *medRxiv* (April 2024).
26. G. Assessment, 6. glycemic targets: standards of medical care in diabetes—2022, *Diabetes Care* **45**, p. S83 (2022).
27. L. Czupryniak, G. Dzida, P. Fichna, P. Jarosz-Chobot, J. Gumprecht, T. Klupa, M. Mysliwiec, A. Szadkowska, D. Bomba-Opon, K. Czajkowski, M. T. Malecki and D. A. Zozulinska-Ziolkiewicz, Ambulatory glucose profile (AGP) report in daily care of patients with diabetes: Practical tips and recommendations, *Diabetes Ther.* **13**, 811 (April 2022).
28. D. M. Nathan, J. Kuenen, R. Borg, H. Zheng, D. Schoenfeld, R. J. Heine and A1c-Derived Average Glucose Study Group, Translating the A1C assay into estimated average glucose values, *Diabetes Care* **31**, 1473 (August 2008).
29. H. Chase, "langchain-experimental 0.0.40" (2023), Version 0.0.40, Software available from

https://pypi.org/project/langchain-experimental/0.0.40/.
30. L. Cheng, X. Li and L. Bing, Is GPT-4 a good data analyst?, *arXiv [cs.CL]* (May 2023).
31. C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev and C. Cobelli, The UVA/PADOVA type 1 diabetes simulator: New features, *J. Diabetes Sci. Technol.* **8**, 26 (January 2014).
32. J. Xie, Simglucose v0.2.1 . (2018) [Online]. Available: https://github.com/jxx123/simglucose.
33. H. Hall, D. Perelman, A. Breschi, P. Limcaoco, R. Kellogg, T. McLaughlin and M. Snyder, Glucotypes reveal new patterns of glucose dysregulation, *PLoS Biol.* **16**, p. e2005143 (July 2018).

Artificial Allies: Validation of Synthetic Text for Peer Support Tools through Data Augmentation in NLP Model Development

Josué Godeme

Research Computing and Data Services, Information, Technology & Consulting, Dartmouth College
Hanover, NH 03784, USA
E-mail: josue.f.godeme.26@dartmouth.edu
www.dartmouth.edu

Julia Hill

Department of Psychiatry, Geisel School of Medicine, Dartmouth College
Hanover, NH 03784, USA
E-mail:Julia.clark.hill@gmail.com

Stephen P. Gaughan, Wade J. Hirschbuhl, Amanda J. Emerson, Christian Darabos, Carly A. Bobak

Research Computing and Data Services, Information, Technology & Consulting, Dartmouth College
Hanover, NH 03784, USA
Email: Stephen.P.Gaughan, Wade.J.Hirschbuhl, Amanda.J.Emerson, Christian.Darabos, Carly.A. Bobak@dartmouth.edu

Karen L. Fortuna

Department of Psychiatry, Geisel School of Medicine, Dartmouth College
Hanover, NH 03784, USA
E-mail: Karen.L.Fortuna@dartmouth.edu

This study investigates the potential of using synthetic text to augment training data for Natural Language Processing (NLP) models, specifically within the context of peer support tools. We surveyed 22 participants—13 professional peer supporters and 9 AI-proficient individuals—tasked with distinguishing between AI-generated and human-written sentences. Using signal detection theory and confidence-based metrics, we evaluated the accuracy and confidence levels of both groups. The results show no significant differences in rater agreement between the two groups ($p = 0.116$), with overall classification accuracy falling below chance levels (mean accuracy = 43.10%, $p < 0.001$). Both groups exhibited a tendency to misclassify low-fidelity sentences as AI-generated, with peer supporters showing a significant bias ($p = 0.007$). Further analysis revealed a significant negative correlation between errors and confidence among AI-proficient raters ($r = -0.429$, $p < 0.001$), suggesting that as their confidence increased, their error rates decreased. Our findings support the feasibility of using synthetic text to mimic human communication, with important implications for improving the fidelity of peer support interventions through NLP model development.

Keywords: Synthetic text generation; Natural language processing; Peer support; Signal detection theory; AI-generated content; Rater agreement; Fidelity classification.

1. Introduction

Peer support specialists play a crucial role in the mental health care system.[1–4] These individuals, who have lived experiences of mental health conditions, provide emotional, social, and practical assistance to others facing similar challenges. The peer support movement has grown significantly, with peer support specialists becoming an integral part of mental health services due to their ability to engage and support individuals in ways that complement traditional clinical interventions.[1–4] This form of support is particularly important for adults with serious mental illnesses, who often face high rates of morbidity and reduced life expectancy due to poorly managed health conditions.[5]

Despite the proven benefits of peer support,[3,4,6] there is a gap in tools that can assist peer supporters in delivering consistent and high-quality care. An ideal tool would not only aid in real-time fidelity monitoring but also enhance the delivery of evidence-based practices.[7,8] Kadakia et al's[9,10] prior work has shown promise in this area, utilizing a deep learning model trained on data from both recorded peer support conversations and Reddit to classify high-fidelity peer support techniques. This approach demonstrated that natural language processing (NLP) could be used to scale and ensure the fidelity of digital peer support interventions.

It has previously been established that improving data quality and quantity is a critical step in improving deep learning model accuracy, particularly for NLP models.[11,12] However, in our application, accessing mental health data is often difficult, and transcription of interactions can be labor-intensive and prone to errors.[13–15] Furthermore, deep learning NLP algorithms typically require large amounts of high-quality data to perform optimally.[16,17] Previously, researchers have demonstrated that LLM generated text can be used to improve the performance of NLP-related tasks, including text classification.[18–20] Hence, we hypothesize that large language models (LLMs) can be used to generate synthetic data that closely mimic real peer support mental health sessions, thereby enhancing the fidelity classification of peer support interventions.

In this study, we seek to demonstrate the feasibility of using synthetic data to mimic human-written content effectively in the peer-supporter context. We also aim to validate that peer supporters, as well as individuals who are professionally engaged in working with LLMs (dubbed AI-proficient non-peer supporters) struggle to reliably distinguish between LLM-generated sentences and real human sentences. This research will contribute to the understanding of synthetic data validation and its potential to support the development of robust tools for peer supporters, ultimately enhancing the quality of mental health care.

2. Methods

2.1. *Original Data Collection*

Collection and transcription of the original human generated conversations are described in Kadakia et al.[9] In short, anonymized records of peer support conversations from the PeerTECH platform where manually recorded verbatim.[9] High-fidelity and low-fidelity sentences are defined in the context of adherence to best practices for peer support in mental

health.[9,10,21] High-fidelity sentences refers to interactions that strictly follow established protocols and best practices, ensuring comprehensive and consistent delivery of peer support. These interactions typically include elements such as active listening, empathy, validation of experiences, and appropriate use of self-disclosure.[9,10,21] Low-fidelity sentences, on the other hand, denote interactions that deviate from these best practices, potentially lacking in one or more critical aspects of effective peer support.[9,10,21] Such deviations might include inadequate listening, insufficient emotional engagement, or inappropriate self-disclosure, which can undermine the effectiveness of the support provided.

2.2. *Synthetic Text Generation*

To generate the synthetic text, we utilized OpenAI's Application programming interface (API), specifically the GPT-4 Turbo model,[22] which was the most advanced model available at the time of the study. The process aimed to produce 10,000 sentences, which should provide a robust training set for downstream NLP modeling.[23] The GPT-4 Turbo model was configured with a temperature setting of 0.9.

The generation process involved three key components: a system prompt, a specific prompt, and user profiles. Two distinct system prompts were employed to generate transcripts demonstrating both high- and low-fidelity practices in peer support conversations. The specific prompt was constructed using characteristics of both the peer supporter and the patient. For the peer supporter, the prompt included their age, gender, personality traits, mental health history, and the topic of the support session. For the patient, the prompt specified their age, gender, personality traits, and their diagnosed mental health condition. This structured approach ensured that each generated conversation was contextualized with specific demographic and psychological information for both participants.

Example Prompt 1
Peer Supporter - Age: 35, Gender: female, Traits: compassionate, insightful, Mental Health History: post-traumatic stress disorder, Session Topic: coping with trauma
Patient - Age: 29, Gender: male, Traits: distrustful, struggling, Mental Health Issue: trauma recovery

Example Prompt 2
Peer Supporter - Age: 41, Gender: female, Traits: calm, reassuring, Mental Health History: post-traumatic stress disorder, Session Topic: managing triggers
Patient - Age: 30, Gender: female, Traits: jumpy, anxious, Mental Health Issue: post-traumatic stress disorder

All data manipulation and analysis were conducted in R version 4.3.2,[24] with extensive use of the `tidyverse` suite of packages[25] for data manipulation, cleaning, and visualization. A total of 154 API calls were executed, resulting in the generation of 10,736 sentences, exceeding the initial target of 10,000 sentences.

2.3. *Synthetic Text Validation*

To evaluate the accuracy and confidence of human raters in distinguishing between human-generated and synthetic text, we randomly selected 100 sentences. These sentences were categorized based on their origin and fidelity: 17 high-fidelity and 14 low-fidelity sentences were human-generated, while 43 high-fidelity and 26 low-fidelity sentences were synthetic. High-fidelity refers to adherence to best practices for peer support in mental health providing, while low-fidelity indicates lesser adherence.

We recruited two types of raters: AI-proficient non-peer supporters from Information, Technology, and Consulting at Dartmouth College, and peer professionals recruited through social media calls and email lists. Raters rated their confidence in how each sentence was generated, using the options: Definitely Human, Maybe Human, I Don't Know, Maybe AI, and Definitely AI. Responses were collected using the *Qualtrics* survey platform (Qualtrics, Provo, UT).

The survey data, which included ratings from AI-proficient non-peer supporters and peer supporters was rated on a scale from 1 to 5, where 1 represented "Definitely AI" and 5 represented "Definitely Human." Confidence ratings were assigned numerical values: 100 for "Definitely," 60 for "Maybe," and 0 for "I don't know."

To evaluate rater performance, several metrics were calculated:

1. **Percentage Agreement**: For each sentence, the percentage agreement among all raters was calculated by determining the proportion of ratings that matched the most common rating.

$$\text{Agreement} = \left(\frac{\sum_{i=1}^{n} I(r_i = \text{mode}(r))}{n} \right) \times 100$$

where r_i represents the rating of the i-th rater, mode(r) is the most common rating among all raters, and n is the total number of raters.

2. **Group-Specific Agreement**: The percentage agreement was calculated separately for AI-proficient non-peer supporters and peer supporters to understand agreement within each group.

3. **Weighted Accuracy**: Weighted accuracy was determined by comparing each rating to the true origin of the sentence and adjusting for confidence levels.

$$\text{Weighted Accuracy} = \left(\frac{\sum_{i=1}^{n} w_i \cdot I(r_i = \text{true origin})}{n} \right) \times 100$$

where w_i is the weight assigned based on the confidence level of the i-th rating, r_i is the rating of the i-th rater, and true origin indicates whether the sentence is human or AI-generated.

4. **Percentage of Errors**: The percentage of incorrect ratings was calculated by determining the proportion of ratings that deviated from the true origin of the sentence.

5. **Average Confidence**: The average confidence level for each sentence was calculated by averaging the confidence scores provided by the raters.

Summary statistics were generated to provide an overview of the overall agreement, accuracy, error rates, and confidence levels among the different rater groups.

We also calculated accuracy, errors, and confidence at the rater level. For each rater, the

percentage of sentences judged as human that were actually human was calculated:

$$\text{Percentage Judged Human (Actual Human)} = \left(\frac{\sum_{i=1}^{n} I(r_i \in \{4,5\} \wedge t_i = \text{Human})}{\sum_{i=1}^{n} I(r_i \in \{4,5\})} \right) \times 100$$

and the percentage of sentences judged as human that were actually AI was calculated:

$$\text{Percentage Judged Human (Actual AI)} = \left(\frac{\sum_{i=1}^{n} I(r_i \in \{4,5\} \wedge t_i = \text{Synthetic})}{\sum_{i=1}^{n} I(r_i \in \{4,5\})} \right) \times 100$$

2.4. *Statistical Analysis*

We performed various statistical tests to evaluate differences in rater performance and confidence. Paired t-tests[26,27] compared the accuracy, confidence, and agreement between peer supporters and AI-proficient non-peer supporters. A one-sample t-test assessed whether the overall accuracy differed significantly from 50%. An independent t-test evaluated the overall agreement among all raters.

Correlation tests[28] examined the relationship between errors and confidence levels for both rater groups and for sentences labeled with low-fidelity. Specifically, correlations between errors and confidence for AI-proficient non-peer supporters and peer supporters were assessed, as well as for low-fidelity sentences.

To compare the proportion of AI judgments between low and high-fidelity sentences, paired t-tests were performed separately for AI-proficient non-peer supporters and peer supporters. Paired t-tests[26,27] were also conducted to compare the percentage of judgments that were actually human versus AI for each rater type.

2.5. *Signal Detection Analysis*

To evaluate the ability of raters to distinguish between human-generated and synthetic text, we calculated signal detection measures. Weights for definite and maybe confidence levels were defined, assigning a weight of 1 for definite judgments and 0.6 for maybe judgments.

For each rater, we calculated the signal detection theory (SDT) measures,[29] including the signal detection score (d'), beta, and criterion (c).

For each sentence, the counts of hits (true positives), false alarms (false positives), misses (false negatives), and correct rejections (true negatives) were determined based on the ratings and true origin. Specifically, for sentences with a true origin of human, hits were defined as ratings of "Definitely Human" or "Maybe Human," and false alarms were defined as ratings of "Definitely AI" or "Maybe AI." For sentences with a true origin of synthetic, hits were defined as ratings of "Definitely AI" or "Maybe AI," and false alarms were defined as ratings of "Definitely Human" or "Maybe Human."

The SDT measures were calculated using the `psycho` package in R.[30] The d' score was computed as:

$$d' = \Phi^{-1}(\text{hit rate}) - \Phi^{-1}(\text{false alarm rate})$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution.

For each rater, the weighted SDT measures (d', beta, and c) were calculated separately for human and synthetic origins. The combined measures for each rater were used to perform

t-tests to compare against a hypothesized mean of 0. T-tests[26,27] for the combined d' scores, beta, and c values were conducted to determine if there was a significant ability to distinguish between human and synthetic text.

2.6. *Insight Calculation (Meta d')*

To evaluate rater insight, we calculated the meta d' score, which measures a rater's metacognitive ability to discriminate between their correct and incorrect judgments.[31,32]

The meta d' score was calculated using the negative log-likelihood optimization approach. Specifically, we minimized the negative log-likelihood to find the meta d' value that best describes the observed data. The optimization was performed using the L-BFGS-B method,[33] ensuring the parameter estimates stayed within reasonable bounds.

The steps to calculate meta d' included:

(1) Aggregating the ratings data for each rater to count the occurrences of each confidence level (0, 60, 100) for human and synthetic sentences.
(2) Defining the negative log-likelihood function based on the signal detection theory model parameters.
(3) Using the `optimx` package[34] to optimize the parameters and calculate the meta d' score.
(4) Extracting and summarizing the meta d' scores for each rater.

2.7. *Sensitivity and Specificity Analysis*

We calculated the sensitivity and specificity for each rater to evaluate their ability to correctly identify human-generated text (with human as the positive case). The analysis was performed using R with the `dplyr`,[35] `tidyr`,[25] `purrr`,[36] `ggplot2`,[37] and `pROC`[38] packages.

Area under the receiver operating characteristic (AUROC) curves were calculated for each rater. AUROC curves were plotted for the best, worst, and median raters.[39,40]

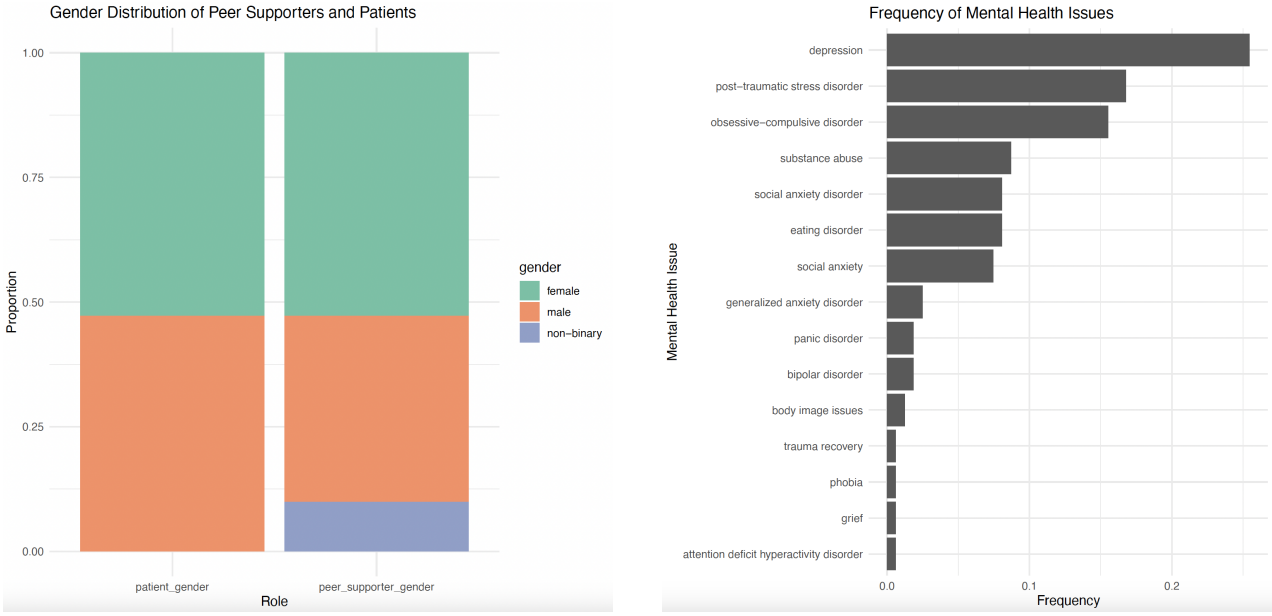
2.8. *Code Availability Statement*

The code used in this study is publicly available on GitHub at [https://github.com/FrejusGdm/Synthetic-Text-Validation-Karen-Fortuna].

3. Results

The age of peer supporters ranged from 26 to 45 years (M = 32.5, SD = 4.2), while patients' ages ranged from 19 to 45 years (M = 29.3, SD = 5.7). The most common mental health issues addressed were depression (22.4%), social anxiety (14.9%), and obsessive-compulsive disorder (13.0%). These are shown in Figure 1.

We recruited 9 AI-proficient non-peer supporters professionals and 13 professional peer supporters to complete the survey (n=22). The mean agreement across all raters was 27.97 (95% CI: 25.55, 30.39). There was no significant difference in the levels of agreement between AI-proficient non-peer supporters and peer supporters ($p = 0.12$). The overall accuracy of raters was lower than what would be expected by random chance, with a mean accuracy of 43.10



A **B**

Fig. 1: Demographic Distributions of Peer Supporters and Patients. (A) Gender distribution among peer supporters and patients. (B) Frequency of various mental health issues reported by patients.

(95% CI: 41.11, 45.09; $p < 0.001$ for a two-sided t-test). Within this, AI-proficient non-peer supporters demonstrated higher accuracy (48.62%) compared to peer supporters (36.41%; mean difference -12.21$p < 0.001$) and reported higher confidence levels (mean difference -13.30$p < 0.001$), although the overall confidence was generally low, with a mean confidence score of 47.75 (95% CI: 45.61, 49.89). These relationships are illustrated in Figure 2 (A)-(C).

Overall, we found that errors and confidence were not significantly correlated ($p = 0.08$). However, this overall trend masks important differences between groups and conditions. Among AI-proficient non-peer supporters, there was a significant negative correlation between errors and confidence ($r = -0.43$, 95% CI: -0.55, -0.25; $p < 0.001$), indicating that as confidence increased, errors decreased. In contrast, for peer supporters, the correlation between errors and confidence was not significant ($r = -0.19$, 95% CI: -0.37, 0.01; $p = 0.06$). These results are shown in Figure 2 (D).

When examining sentences labeled with low-fidelity, the correlation between errors and confidence for peer supporters was not significant ($r = -0.03$, 95% CI: -0.34, 0.29; $p = 0.87$). However, for AI-proficient non-peer supporters, there was a significant negative correlation ($r = -0.33$, 95% CI: -0.58, -0.02; $p = 0.04$) in low-fidelity sentences. These results are shown in Figures 2 (E) and (F).

In high-fidelity sentences, peer supporters exhibited a significant negative correlation between errors and confidence ($r = -0.38$, 95% CI: -0.58, -0.14; $p = 0.003$). Similarly, AI-proficient non-peer supporters showed a significant negative correlation ($r = -0.51$, 95% CI: -0.68, -0.30; $p < 0.001$).

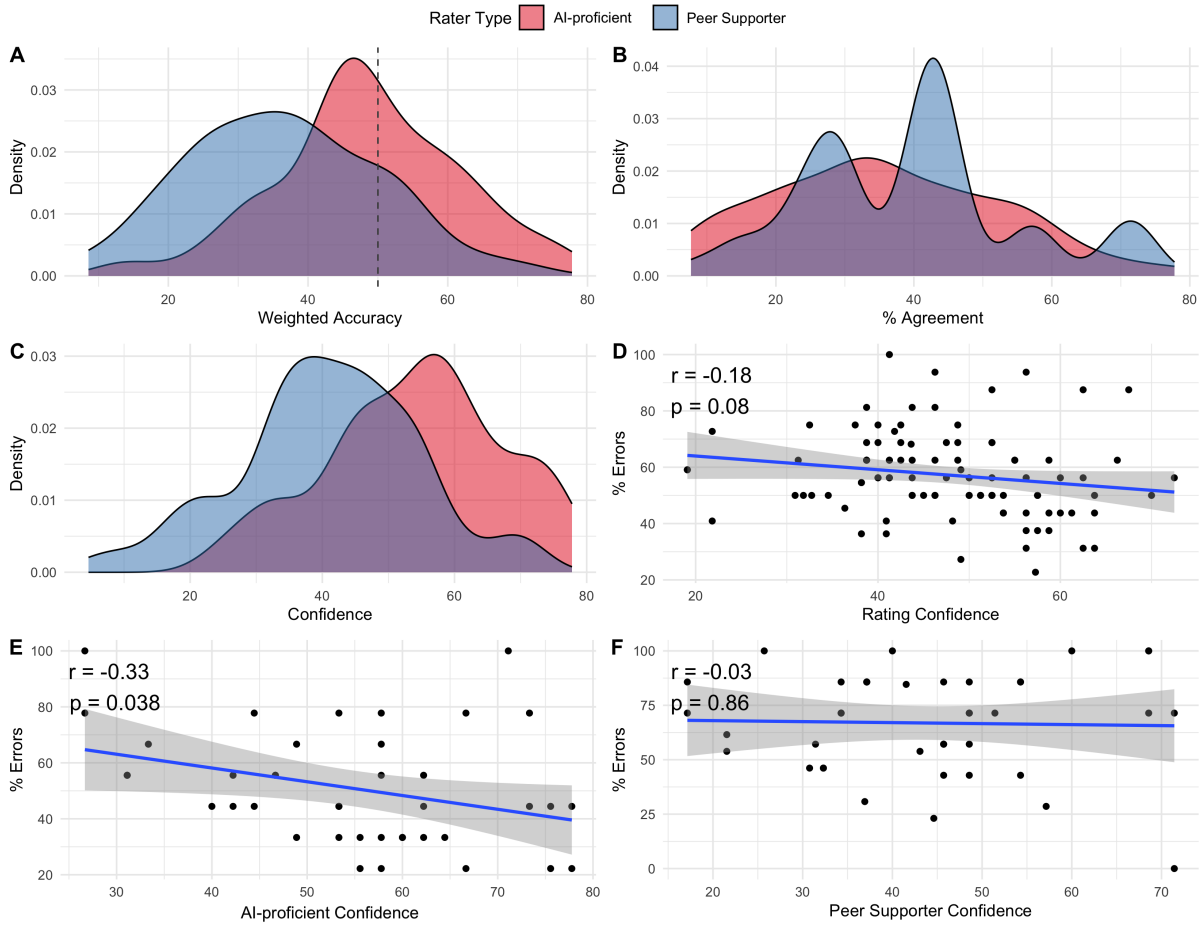


Fig. 2: Analysis of rater performance and confidence. (A) Weighted accuracy for AI-proficient non-peer supporters and peer supporters, with a 50% accuracy line indicated. (B) Percentage agreement among AI-proficient non-peer supporters and peer supporters. (C) Rating confidence. (D) Scatter plot with fitted line and 95% confidence intervals showing rating confidence by percentage errors. (E) Same as (D) for AI-proficient non-peer supporters with low-fidelity ratings. (F) Same as (D) for peer supporters with low-fidelity ratings.

Peer support raters were more likely to assume that low-fidelity sentences were AI-generated compared to high-fidelity sentences. This difference in proportions was statistically significant ($p = 0.007$), with a difference in proportions ranging from 0.02 to 0.10. For AI-proficient non-peer supporters, the tendency to assume low-fidelity sentences were AI-generated was also observed, although the difference was only borderline significant ($p = 0.05$), with a difference in proportions ranging from 0.00 to 0.09. These findings indicate that both peer supporters and AI-proficient non-peer supporters are more inclined to classify low-fidelity sentences as AI-generated, though this tendency is more pronounced and statistically significant among peer supporters.

The bar plot in Figure 3 (A) visualizes the percentage of sentences judged as "Human" by two different groups of raters: AI-proficient non-peer supporters and peer supporters, for

sentences that were actually AI-generated (AI). Of all sentences rated as "Human" by AI-proficient non-peer supporters, 66.7% were AI-generated. This percentage was higher in the peer supporter group, with 83.3% of sentences rated as "Human" being AI-generated. Statistical tests revealed that for AI-proficient non-peer supporters, the tendency to judge AI sentences as "Human" was borderline statistically significant ($p = 0.05$). In contrast, Peer Supporters showed a statistically significant tendency to judge AI-generated sentences as "Human" ($p < 0.001$). These findings indicate a tendency for both AI-proficient non-peer supporters and Peer Supporters to be deceived by AI-generated content, with Peer Supporters being particularly susceptible. This could perhaps indicate AI hyperrealism[31]— where even trained individuals are frequently unable to distinguish AI from human-generated text.

Signal detection theory was applied to evaluate the ability of raters to distinguish between human-generated and AI-generated text. The d' (d-prime) score is a measure of a rater's ability to discriminate between signal (human-generated text) and noise (AI-generated text), where a higher d' indicates better discrimination ability. Our analysis revealed that the combined mean d' score was significantly greater than zero (mean = 0.39, 95% CI: 0.22, 0.55; $p < 0.001$), indicating that detection is occurring among raters.

In addition to d', we also evaluated beta (β) and criterion (c), which provide insights into the decision-making biases of the raters. A positive beta (β) indicates a conservative response bias, meaning raters are less likely to label sentences as human. The combined mean beta (β) was significantly greater than zero (mean = 1.37, 95% CI: 1.17, 1.57; $p < 0.001$), suggesting a strong bias towards not labeling sentences as human. Similarly, the combined mean criterion (c) was significantly greater than zero (mean = 0.62, 95% CI: 0.44, 0.80; $p < 0.001$), reinforcing the notion of a reluctance to label sentences as human.

We also calculated the meta d' score, referred to as insight, based on the raters' confidence levels. The meta d' score measures a rater's metacognitive ability to discriminate between their correct and incorrect judgments. Our results showed that the combined mean meta d' score was not significantly different from zero (mean = −0.39, 95% CI: -1.37, 0.59; $p = 0.42$). This near-zero insight score indicates that raters do not have a reliable metacognitive awareness of their accuracy in distinguishing between human and AI-generated sentences, suggesting that their confidence levels do not effectively reflect their true performance.

Plotting the signal detection score against the insight score allows us to identify how detection and insight are interrelated. As shown in Figure 3 (B), we observe only two raters (9%) with both good insight and good detection in the top-left quadrant of the plot.

Sensitivity and specificity were calculated with regards to the raters' ability to discern human-written sentences and are displayed in Figure 2(C). No rater achieved both a sensitivity and specificity greater than 0.7, indicating that none of the raters were highly proficient at correctly identifying human-written sentences while also correctly rejecting AI-generated ones.

We calculated the area under the receiver operating characteristic curve (AUROC) for each rater, and a boxplot of AUROC scores across AI-proficient non-peer supporters and peer supporters is shown in Figure 2(D). The mean AUROC for peer supporters was 0.59 (95% CI: 0.52, 0.67), while the mean AUROC for AI-proficient non-peer supporters was 0.61 (95% CI: 0.56, 0.66).

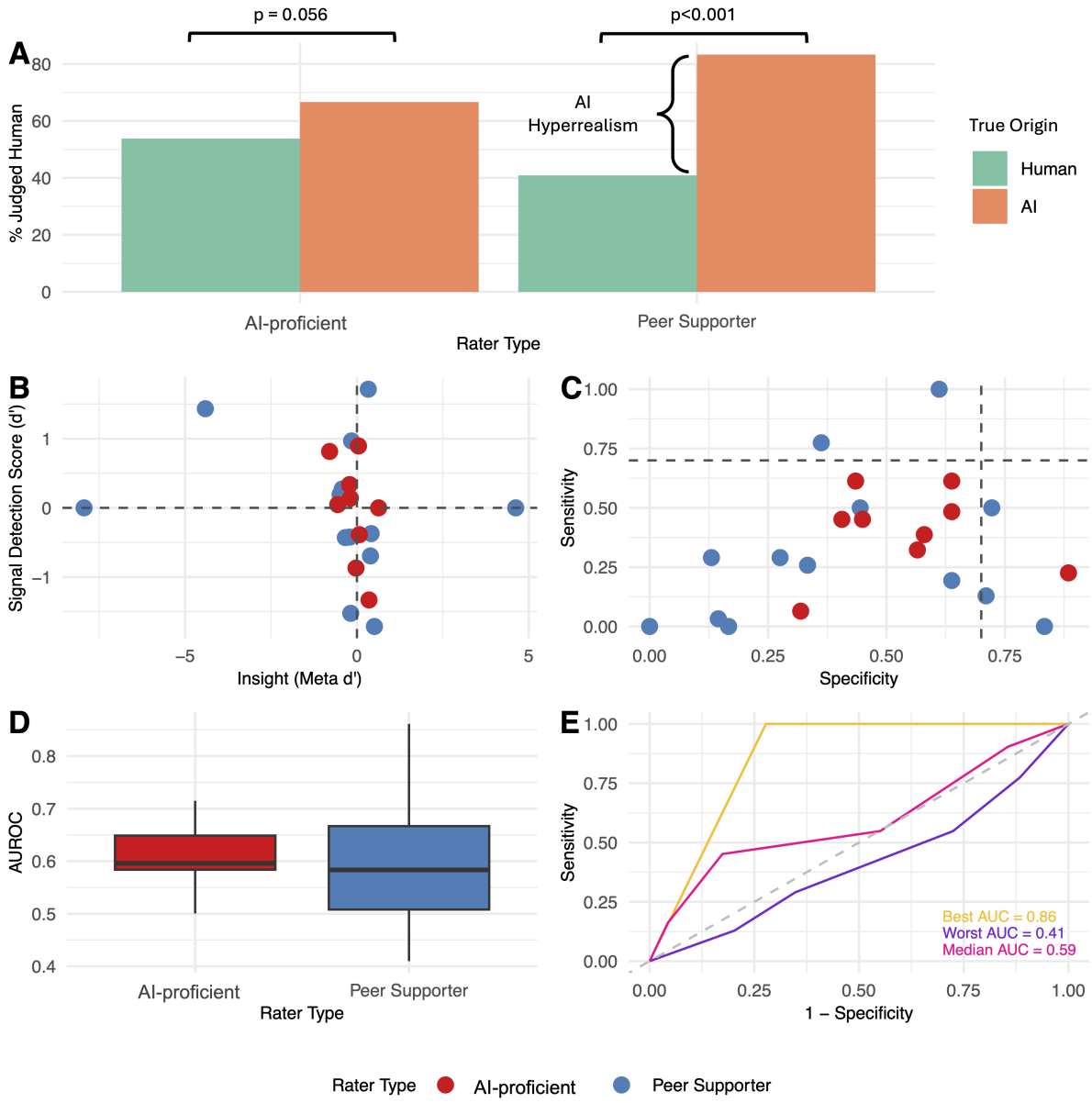


Fig. 3: Comparison of rater performance and insights. (A) Bar chart showing the percentage of ratings within each rater type judged as human, grouped by the true origin (AI vs Human generation). Low and high fidelity sentences are pooled. (B) Plot of rater insight, calculated as Meta d' (as described by[31]), and signal detection score (d') using the `psycho`[30] package. (C) Scatter plot of sensitivity versus specificity of a human rating for a true human label by rater. (D) Boxplot of AUROC scores by rater type. (E) AUROC for the best, worst, and median rater.

The top AUROC calculated (Peer Supporter), the bottom AUROC calculated (also Peer Supporter), and the rater with an AUROC closest to the median (AI-proficient non-peer supporters) are shown in Figure 2(E). These results highlight the variability in rater performance

and suggest that, overall, raters struggled to consistently distinguish between human and AI-generated text.

4. Discussion

There is a clear need for targeted training programs to enhance peer supporters' ability to critically evaluate their performance during peer-support calls, allowing them to improve the quality of care they are producing.[9,10] This aligns with recommendations by Naslund et al.[41] philosophies on the importance of digital literacy in mental health support contexts. To support peer supporters in self-evaluating and improving their job performance, digital tools can play a crucial role, but such tools require access to large amounts of high-quality data.

Given the difficulty in obtaining sufficient real-world data, large language models (LLMs) offer a promising solution by generating synthetic data, as evidenced by the low detection accuracy (43.10%) in our study, where the AI-generated text closely mimicked human-created content.

The significant difference in accuracy between AI-proficient non-peer supporters and peer supporters, with the AI-proficient group demonstrating higher accuracy, is expected. However, the overall low accuracy for both groups underscores the challenges in reliably detecting AI-generated content, even for those with technical expertise.

This raises the possibility that exposure to AI in professional settings may confer some advantage in detecting synthetic content. However, the performance gap was small, which suggests that even AI-exposure may not be sufficient to reliably distinguish between human and AI-generated text in all cases. This brings into question how evaluators are selected for similar studies, as familiarity with AI might not always correlate with better performance in validation tasks. Future research should consider how varying levels of AI-exposure might impact evaluators' ability to assess synthetic text, and whether additional training or background knowledge is required for optimal evaluation.

The tendency of both peer supporters and AI-proficient non-peer supporters to classify low-fidelity sentences as AI-generated more often than high-fidelity sentences is particularly interesting. This suggests that the quality or adherence to best practices in peer support conversations might be a key factor in how text is perceived.

The promising results of this study, reflected in the low detection accuracy, suggest that synthetic text could be effectively integrated into training data for automatic feedback algorithms designed for peer supporters. However, it is essential to carefully consider the ethical implications and ensure that the human element, which is crucial to peer support, is maintained.[42]

4.1. *Limitations*

There are several limitations to this analysis. Firstly, the small sample size ($n = 22$) limits further generalizing our findings. However, the effects observed achieve statistical significance, which suggests that the findings are robust despite the sample size. Secondly, our analysis was based on the classification of individual sentences without additional context. While this serves our goal of creating technologies to highlight sentences of high- and low-fidelity, it is

reasonable to expect that providing more context around each sentence might yield different results, as raters could potentially make more accurate judgments with more information.

Another limitation is the potential bias introduced by the specific demographic and professional backgrounds of our raters, which may not be representative of the broader population of peer supporters and AI-proficient individuals. Additionally, the inherent variability in individual raters' experiences and familiarity with AI-generated content could influence their performance and confidence levels.

Testing was conducted on a single model, which restricts our ability to generalize the findings across LLMs that may perform differently. Furthermore, we did not conduct a qualitative analysis of the synthetic data, which could have provided deeper insights into its linguistic quality, semantic accuracy, stylistic consistency, and realism. A more detailed assessment, such as annotation by professional peer supporters, could offer valuable perspectives on the text's quality and its alignment with human communication in similar contexts.

We did not evaluate the synthetic data for downstream tasks, leaving its practical application in real-world settings unexplored. This remains an important area for future work, and in our next follow-up study, we plan to investigate how synthetic data can be integrated into various downstream tasks, including its potential to enhance peer-support tools and other applications in similar domains.

Despite these limitations, our findings support the hypothesis that synthetic data generation for augmentation is feasible. The validation efforts indicate that both AI-proficient non-peer supporters and peer supporters struggle to reliably distinguish between human and AI-generated text, suggesting that AI-generated synthetic data can effectively mimic human-written content. This finding has promising implications for the use of synthetic data to augment training datasets and improve the performance of fidelity classification algorithms.

5. Conclusion

This study demonstrates the potential for using LLMs in synthetic text generation to create diverse datasets of peer support conversations, encompassing both high- and low-fidelity examples. Our findings reveal that both our test groups had below 50% in distinguishing synthetic text from human-created content, underscoring the sophisticated nature of current AI language models. Importantly, this work does not aim to replace human peer supporters with AI chatbots, but instead lays the groundwork for developing an automated feedback system to enhance peer support training and quality assurance. The synthetic sentences generated provide a rich dataset for training AI models to classify the quality of peer support interactions, potentially offering real-time feedback to supporters.

Future work might focus on developing and validating an AI-based feedback algorithm using our synthetic dataset, exploring its ethical implications, and investigating the long-term impacts of AI-assisted training on peer support outcomes. Ultimately, this study represents a significant step towards leveraging AI to enhance, rather than replace, human-delivered peer support, contributing to improved mental health support services.

Acknowledgements

We acknowledge Information, Technology, and Consulting (ITC) at Dartmouth College and Research Computing and Data Services (RC) team for providing the high-performance computational expertise and resources crucial to our research project's success. Their support and dedication significantly contributed to our progress and outcomes.

References

1. R. Cooper, K. Saunders, A. Greenburgh *et al.*, The effectiveness, implementation, and experiences of peer support approaches for mental health: a systematic umbrella review, *BMC Medicine* **22**, p. 72 (2024).
2. F. Schneider, M. Erhart, W. Hewer, L. AK Loeffler and F. Jacobi, Mortality and Medical Comorbidity in the Severely Mentally Ill, *Deutsches Ärzteblatt International* **116**, 405 (2019).
3. S. White, R. Foster, J. Marks *et al.*, The effectiveness of one-to-one peer support in mental health services: a systematic review and meta-analysis, *BMC Psychiatry* **20**, p. 534 (2020).
4. C. S. T. Yim, J. H. L. Chieng, X. R. Tang, J. X. Tan, V. K. F. Kwok and S. M. Tan, Umbrella review on peer support in mental disorders, *International Journal of Mental Health* **52**, 379 (2023).
5. K. Barnhouse, S. Clark and J. Waters Davis, Special population: Adults with severe and persistent mental health disorders, in *Chronic Illness Care*, eds. T. Daaleman and M. Helton (Springer, Cham, 2023)
6. D. Smit, C. Miguel, J. Vrijsen, B. Groeneweg, J. Spijker and P. Cuijpers, The effectiveness of peer support for individuals with mental illness: systematic review and meta-analysis, *Psychological Medicine* **53**, 5332 (2023).
7. C. Collins-Pisano, M. Johnson, G. Mois, J. Brooks, A. Myers, A. Muralidharan, M. Storm, M. Wright, N. Berger, A. Kasper *et al.*, Core competencies to promote consistency and standardization of best practices for digital peer support: focus group study, *JMIR Mental Health* **8**, p. e30221 (2021).
8. K. L. Fortuna, M. Venegas, E. Umucu, G. Mois, R. Walker and J. M. Brooks, The future of peer support in digital psychiatry: promise, progress, and opportunities, *Current treatment options in psychiatry* **6**, 221 (2019).
9. A. Kadakia, S. Preum, A. Bohm and K. Fortuna, Investigating the fidelity of digital peer support: A preliminary approach using natural language processing to scale high-fidelity digital peer support, in *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 5: HEALTHINF*, (SCITEPRESS – Science and Technology Publications, Lda., 2023).
10. K. Fortuna, A. Wright, G. Mois *et al.*, Feasibility, acceptability, and potential utility of peer-supported ecological momentary assessment among people with serious mental illness: a pilot study, *Psychiatric Quarterly* **93**, 717 (2022).
11. Y. Roh, G. Heo and S. E. Whang, A survey on data collection for machine learning: a big data-ai integration perspective, *IEEE Transactions on Knowledge and Data Engineering* **33**, 1328 (2019).
12. Y. Cui, W. Che, T. Liu, B. Qin, S. Wang and G. Hu, Revisiting pre-trained models for chinese natural language processing, *arXiv preprint arXiv:2004.13922* (2020).
13. E. Watson, S. Fletcher-Watson and E. J. Kirkham, Views on sharing mental health data for research purposes: qualitative analysis of interviews with people with mental illness, *BMC Medical Ethics* **24**, p. 99 (2023).
14. F. Bernardi, D. Alves, N. Crepaldi, D. Yamada, V. Lima and R. Rijo, Data quality in health

research: Integrative literature review, *Journal of Medical Internet Research* **25**, p. e41446 (2023).
15. R. Syed, R. Eden, T. Makasi, I. Chukwudi, A. Mamudu, M. Kamalpour, D. Kapugama Geeganage, S. Sadeghianasl, S. Leemans, K. Goel, R. Andrews, M. Wynn, A. ter Hofstede and T. Myers, Digital health data quality issues: Systematic review, *Journal of Medical Internet Research* **25**, p. e42615 (2023).
16. A. R. Luca, T. F. Ursuleanu, L. Gheorghe, R. Grigorovici, S. Iancu, M. Hlusneac and A. Grigorovici, Impact of quality, type and volume of data used by deep learning models in the analysis of medical images, *Informatics in Medicine Unlocked* **31**, p. 100911 (2022).
17. Y. K. A. Baqraf, P. Keikhosrokiani and M. Al-Rawashdeh, Evaluating online health information quality using machine learning and deep learning: A systematic literature review, *Digital Health* **9** (2023).
18. A. Rosenbaum, S. Soltan, W. Hamza, A. Saffari, M. Damonte and I. Groves, Clasp: Few-shot cross-lingual data augmentation for semantic parsing, in *AACL-IJCNLP 2022*, 2022.
19. A. Rosenbaum, S. Soltan, W. Hamza, Y. Versley and M. Boese, Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging, in *COLING 2022*, 2022.
20. H. Zhao, H. Chen, T. A. Ruggles, Y. Feng, D. Singh and H.-J. Yoon, Improving text classification with large language model-based data augmentation, *Electronics* **13**, p. 2535 (2024).
21. M. Chinman, S. McCarthy, C. Mitchell-Miland, K. Daniels, A. Youk and M. Edelen, Early stages of development of a peer specialist fidelity measure, *Psychiatric Rehabilitation Journal* **39**, 256 (2016).
22. OpenAI, ChatGPT API `https://openai.com/blog/chatgpt`, (2022).
23. S. Gholizadeh and N. Zhou, Model explainability in deep learning based natural language processing (2021), arXiv preprint.
24. R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2021).
25. H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo and H. Yutani, *Welcome to the tidyverse*, (2019). R package version 1.3.1.
26. D. W. Zimmerman, A note on interpretation of the paired-samples t test, *Journal of Educational and Behavioral Statistics* **22**, 349 (1997).
27. Student, The probable error of a mean, *Biometrika* **6**, 1 (1908).
28. K. Pearson, Note on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London* **58**, 240 (1895).
29. H. Stanislaw and N. Todorov, Calculation of signal detection theory measures, *Behavior Research Methods, Instruments, & Computers* **31**, 137 (1999).
30. D. Makowski, *psycho: Procedures for Psychological, Psychometric, and Personality Research*, (2018). R package version 0.6.0.
31. E. J. Miller, B. A. Steward, Z. Witkower, C. A. M. Sutherland, E. G. Krumhuber and A. Dawel, AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones, *Psychological Science* **34**, 1390 (2023).
32. B. Maniscalco and H. Lau, A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings, *Consciousness and Cognition* (2012), journal homepage: www.elsevier.com/locate/concog.
33. D. Liu and J. Nocedal, On the limited memory bfgs method for large scale optimization, *Mathematical Programming* **45**, 503 (1989).
34. J. C. Nash and R. Varadhan, Unifying optimization algorithms to aid software system users: optimx for r, *Journal of Statistical Software* **43**, 1 (2011).

35. H. Wickham, R. François, L. Henry and K. Müller, *dplyr: A Grammar of Data Manipulation*, (2021). R package version 1.0.7.
36. L. Henry and H. Wickham, *purrr: Functional Programming Tools*, (2022). R package version 0.3.5.
37. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, (2016). R package version 3.3.5.
38. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez and M. Müller, *pROC: Display and Analyze ROC Curves*, (2022). R package version 1.18.0.
39. J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve, *Radiology* **143**, 29 (1982).
40. E. R. DeLong, D. M. DeLong and D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* **44**, 837 (1988).
41. J. A. Naslund, P. P. Gonsalves, O. Gruebner, S. R. Pendse, S. L. Smith, A. Sharma and G. Raviola, Digital innovations for global mental health: opportunities for data science, task sharing, and early intervention, *Current treatment options in psychiatry* **6**, 337 (2019).
42. K. L. Fortuna, A. L. Myers, J. Ferron, A. Kadakia, C. Bianco, M. L. Bruce and S. J. Bartels, Assessing a digital peer support self-management intervention for adults with serious mental illness: feasibility, acceptability, and preliminary effectiveness, *Journal of Mental Health* **31**, 812 (2021).

A Prospective Comparison of Large Language Models for Early Prediction of Sepsis[1]

Supreeth P. Shashikumar* and Shamim Nemati
Division of Biomedical Informatics, University of California San Diego
La Jolla, California, USA
Email: spshashikumar, snemati@health.ucsd.edu

We present a comparative study on the performance of two popular open-source large language models for early prediction of sepsis: Llama-3 8B and Mixtral 8x7B. The primary goal was to determine whether a smaller model could achieve comparable predictive accuracy to a significantly larger model in the context of sepsis prediction using clinical data.

Our proposed LLM-based sepsis prediction system, COMPOSER-LLM, enhances the previously published COMPOSER model, which utilizes structured EHR data to generate hourly sepsis risk scores. The new system incorporates an LLM-based approach to extract sepsis-related clinical signs and symptoms from unstructured clinical notes. For scores falling within high-uncertainty prediction regions, particularly those near the decision threshold, the system uses the LLM to draw additional clinical context from patient notes; thereby enhancing the model's predictive accuracy in challenging diagnostic scenarios.

A total of 2,074 patient encounters admitted to the Emergency Department at two hospitals within the University of California San Diego Health system were used for model evaluation in this study. Our findings reveal that the Llama-3 8B model based system (COMPOSER-LLM$_{Llama}$) achieved a sensitivity of 70.3%, positive predictive value (PPV) of 32.5%, F-1 score of 44.4% and false alarms per patient hour (FAPH) of 0.0194, closely matching the performance of the larger Mixtral 8x7B model based system (COMPOSER-LLM$_{mixtral}$) which achieved a sensitivity of 72.1%, PPV of 31.9%, F-1 score of 44.2% and FAPH of 0.020. When prospectively evaluated, COMPOSER-LLM$_{Llama}$ demonstrated similar performance to the COMPOSER-LLM$_{mixtral}$ pipeline, with a sensitivity of 68.7%, PPV of 36.6%, F-1 score of 47.7% and FAPH of 0.019 vs. sensitivity of 70.5%, PPV of 36.3%, F-1 score of 47.9% and FAPH of 0.020. This result indicates that, for extraction of clinical signs and symptoms from unstructured clinical notes to enable early prediction of sepsis, the Llama-3 generation of smaller language models can perform as effectively and more efficiently than larger models. This finding has significant implications for healthcare settings with limited resources.

Keywords: Large language model, Unstructured clinical notes, Clinical decision support systems

[1*] This work is supported by the National Institutes of Health (#R01LM013998, #R35GM143121, #R42AI177108). The opinions or assertions contained herein are the private ones of the author and are not to be construed as official or reflecting the views of the NIH or any other agency of the US Government.

1. Introduction

Sepsis is a life-threatening condition that arises when the body's response to infection causes systemic inflammation, leading to tissue damage and organ failure. It is a major cause of mortality and morbidity worldwide, accounting for a significant portion of hospital deaths[1–3]. Early detection and timely intervention are critical to improving patient outcomes, as delayed treatment can lead to severe complications and increased mortality[4–6]. Recent advancements in artificial intelligence (AI) have enabled the development of predictive models that utilize electronic health record (EHR) data to identify early signs of sepsis[7]. These AI-driven models can analyze vast amounts of structured data, such as laboratory results and vital signs, to predict sepsis risk and prompt early clinical intervention. Despite the success of these models, they often overlook the rich contextual information embedded in unstructured clinical notes, which can provide additional insights into a patient's condition.

Large language models (LLMs) have emerged as powerful tools for processing and interpreting unstructured text data, making them valuable assets in predictive analytics for healthcare[8]. LLMs, such as GPT-3, Claude and their variants, are pre-trained on extensive text corpora and fine-tuned for specific tasks. In healthcare, LLMs have shown promise in tasks ranging from generating clinical notes and summarizing patient histories to identifying clinical entities and predicting patient outcomes[9–12]. The integration of LLMs with traditional AI models has the potential to improve predictions by incorporating nuanced information from unstructured data, thereby providing a more comprehensive view of a patient's health status.

However, the deployment of large LLMs in clinical settings presents significant challenges. Large models, such as the Mixtral 8x7B model with 47 billion parameters, require substantial computational resources for training and inference, which can be prohibitive in resource-constrained environments. The motivation for this study was to explore the feasibility of using a new generation smaller LLM, specifically the Llama-3 model with 8 billion parameters, to achieve comparable performance for extraction of clinical signs and symptoms from unstructured clinical notes. By reducing the model size, we aim to address the issues of computational efficiency, scalability, and cost, while maintaining or even improving predictive accuracy.

2. Methods

2.1. *Data*

This study utilized de-identified data from the electronic health records (EHR) of patient encounters in the Emergency Department (ED) at two University of California San Diego (UCSD) Health hospitals, using FHIR and HL7v2 standards. Patients were identified as having sepsis according to the Sepsis-3 international consensus definition for sepsis[2]. The onset time of sepsis was established by following previously published methodology, using evidence of organ dysfunction and suspicion of clinical infection[13–15]. Patients aged 18 and older were monitored throughout their stay until either their first episode of sepsis, transition to comfort care, or transfer out of the ED. To ensure a sufficient quantity of predictor data, we focused on sequential hourly predictions of sepsis starting two hours after ED triage. While the previously established COMPOSER model[14,16] used a decision threshold of 0.6, the COMPOSER-LLM model adopted a lower threshold of 0.5 to enhance sensitivity. To mitigate the potential increase in false alarms, the model incorporated additional contextual information from clinical notes for predictions in the high-uncertainty range of 0.5-0.75. To assess the impact of lowering the decision threshold and to explore the advantages of using an LLM for uncertain predictions, all patients with at least one COMPOSER risk score above 0.5 were included for further analysis. Exclusions were made for patients identified as having sepsis before the prediction start-time or those lacking heart rate or blood pressure measurements prior to this time. Predictions were considered if the following criteria were met: 1) At least one vital sign and lab measurement within the past 24 hours; 2) No antibiotics received; and 3) Availability of an “ED provider note” or “H&P note.”

The *retrospective cohort* included ED patient encounters from October 1, 2023 to December 31, 2023. A total of 1320 ED encounters (16.3% septic) met the inclusion criteria for the *retrospective cohort*. Additionally, the COMPOSER-LLM pipeline was prospectively deployed in silent mode for real-time sepsis prediction in the two EDs within the UCSD Health system starting from May 1, 2024. The prospective data collected during the time period of May 1 - June 15 2024 will be referred to as *prospective cohort*. A total of 754 ED encounters (18.4% septic) met the inclusion criteria for the *prospective cohort*. Table T1 in the appendix shows baseline characteristics and summary characteristics for the *retrospective* and *prospective* cohorts.

This investigation was conducted according to University of California San Diego IRB approved protocol #805726 with a waiver of informed consent.

2.2. *COMPOSER-LLM*

The schematic diagram of the entire COMPOSER-LLM pipeline is shown in Figure 1. Starting from the time of ED admission, COMPOSER[14,16] generated a sepsis risk score at an hourly

resolution. We direct the reader to Shashikumar et al.[16] for more details regarding the input features (structured data) of COMPOSER. If the risk score exceeded a primary decision threshold (θ_1=0.75), an alert was fired. Risk scores closer to a secondary decision threshold (θ_2=0.50) were often associated with false alarms. Consequently, for risk scores within the high-uncertainty region ($\theta_1 \geq$ risk score $> \theta_2$), an LLM-based sepsis likelihood tool was utilized to enhance diagnostic accuracy. Specifically, if the likelihood score surpassed a predetermined likelihood-based decision threshold (α) and the LLM indicated a 'suspicion of bacterial infection,' an alert was fired.

2.2.1. *Sepsis likelihood tool*

The sepsis likelihood tool was designed to improve diagnostic accuracy by confirming the presence of sepsis-related clinical signs and symptoms documented in clinical notes. It first utilized a large language model (LLM) to extract these signs or symptoms from the notes. The extracted symptoms were then processed through a likelihood calculator to assess the probability of sepsis. This calculated likelihood was subsequently used to confirm diagnosis of sepsis.

A Bayesian likelihood calculator was used to compute the likelihood of sepsis based on the clinical signs or symptoms identified by the LLM. The posterior probability of sepsis given a set of clinical signs or symptoms ($\{CS_i\}, i \in 1....M$), $P(D|CS)$, was calculated as follows: $P(D|CS) = \frac{P(D)\cdot P(CS|D)}{P(CS)}$. Where, $CS_i = 1$ corresponded to the scenario under which the clinical signs or symptom CS_i was identified to be present by the LLM pipeline. The set of sepsis-related clinical signs or symptoms used in this study ($\{CS_i\}, i \in 1....9$) were as follows: *fever, hypotension, tachypnea, tachycardia, altered mental status, elevated inflammatory markers, positive blood culture, suspicion of bacterial infection, and organ dysfunction syndrome.*

The likelihood values for each of the clinical signs or symptoms conditioned on sepsis have been tabulated in Table T2 of Appendix.

2.2.2. *LLM-based clinical sign or symptom extractor:*

The LLM-based clinical signs or symptoms extraction pipeline was designed to accept a prompt and clinical notes (all notes generated from admission to the time of prediction) as input and generate a JSON-formatted text output. Given that clinical notes can sometimes exceed the predefined input length (context size) of the LLM, we employed the retrieval augmented generation (RAG) technique[17] to extract smaller, relevant text chunks (context) for the queried clinical sign or symptom. These extracted text chunks were then appended to the input prompt for the LLM. The prompt used in our analysis was as follows:

"You are an ED doctor. Your task is to identify the following abnormal clinical signs and symptoms: {clinical sign or symptom}. Think step-by-step and provide your response in the following JSON format: {<clinical sign or symptom> : ["Yes or No", "Concise justification?"]} Medical note: {RAG context}."

To minimize hallucinations and to maintain consistency in text generation, the temperature parameter of the LLM was set to 0.3. Additionally, for each clinical sign or symptom, the LLM pipeline was run three times and the majority outcome (clinical sign or symptom present or not) across the multiple runs was used for downstream tasks.

Recent advancements, including enhanced training data, advanced architecture (such as group query attention), improved tokenization, and refined training techniques, have enabled newer generations of smaller parameter LLMs (such as Llama-3 8B[18]) to match or even surpass the performance of older, larger models (such as Llama 2 70B[19] and Mixtral 8x7B[20]). In this study, we used the open-source Llama-3 8B and Mixtral 8x7B models to investigate whether the newer, smaller LLM (Llama-3 8B) can achieve comparable effectiveness in extracting clinical signs and symptoms from unstructured clinical notes to the much larger Mixtral 8x7B model. Specifically, the Mixtral 8x7B, developed by Mistral AI, is a sparse mixture of experts LLM with 46.7 billion total parameters, referred to as COMPOSER-LLM$_{Mixtral}$ in this study. The Llama-3 8B, the latest state-of-the-art LLM developed by Meta with 8 billion parameters, is referred to as COMPOSER-LLM$_{Llama}$ in our analysis.

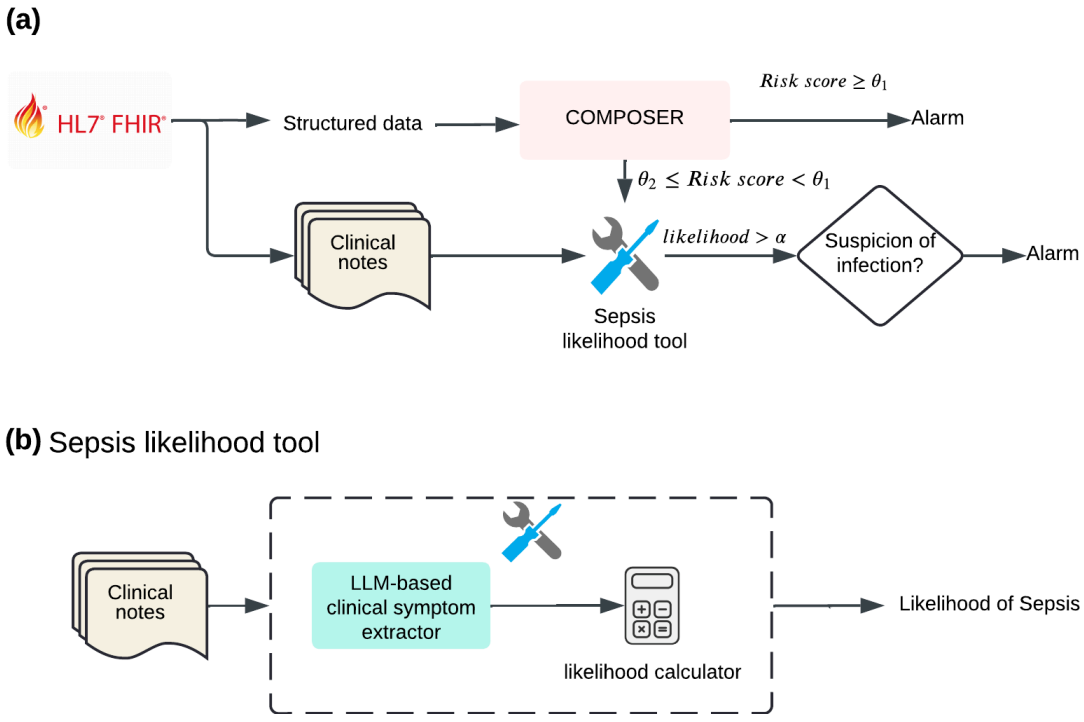


Figure 1. Schematic Diagram of the COMPOSER-LLM pipeline.

2.3. *Experimental setup and evaluation*

For all continuous variables, we have reported medians ([25th–75th percentile]). For binary variables, we have reported percentages. Differences between the septic and non-septic cohort were assessed with Wilcoxon rank sum tests on continuous variables and Pearson's chi-squared tests on categorical variables and significance was assessed at a p-value of 0.05. Sensitivity (SEN), positive predictive value (PPV), and F-1 score at a fixed decision threshold have been reported at the encounter level. SEN, PPV and F-1 score were reported under an end-user clinical response policy in which alarms fired up to 48 hours prior to onset of sepsis were considered as true alarms, and the model was silenced for six hours after an alarm was fired. Additionally, we have reported false alarms per patient hour (FAPH) which can be used to calculate the expected number of false alarms per unit of time in a typical care unit (e.g., a FAPH of 0.025 translates to roughly 1 alarm every 2 h in a 20-bed care unit). The FAPH was calculated by dividing the total number of false alarms by the total number of data points (sum of hourly time points across all patients) in a given cohort.

The COMPOSER model was implemented in TensorFlow. The LLM-based clinical signs or symptom extraction pipeline was implemented using the LangChain framework in Python. The LLM pipeline was run on AWS multi-GPU EC2 instance with NVIDIA A10G GPUs: g5.12xlarge ec2 instance type (cost of $5.672 per hour) for Mixtral 8x7B, g5.2xlarge ec2 instance type (cost of $1.212 per hour) for Llama-3 8B.

2.4. *Prospective deployment:*

COMPOSER-LLM$_{Mixtral}$ and COMPOSER-LLM$_{Llama}$ model were prospectively deployed in silent-mode on a cloud-based platform, as previously described by Boussina et al.[14]. Prospective validation studies are essential in clinical applications of LLMs as retrospective performance may not accurately reflect real-world performance due to factors such as incomplete or missing clinical notes. The real-time platform extracted data at an hourly resolution of all the active patients (across the two Emergency Departments within UCSD Health system) using FHIR APIs with OAuth 2.0 authentication, and passed the input feature set to the COMPOSER-LLM inference engine. The inference engine consisted of COMPOSER microservice and Sepsis likelihood tool microservice hosted within separate EC2 instances. The sepsis risk scores generated by the COMPOSER-LLM pipeline were then written to a flowsheet within the EHR using an HL7v2 outbound message. The flowsheet then triggered a nurse-facing Best Practice Advisory (BPA) that alerted the caregiver that the patient was at risk of developing severe sepsis. As the models were deployed in silent mode, the BPA was not shown to the end-user. The COMPOSER-LLM pipeline was deployed for real-time prediction of sepsis across the two EDs within the UCSD Health system starting from May 1, 2024.

3. Results

Table 1: Comparison of model performance.

		COMPOSER	COMPOSER-LLM$_{Mixtral}$ (Mixtral 8x7B)	COMPOSER-LLM$_{Llama}$ (Llama-3 8B)
Retrospective cohort	**Sensitivity**	72.9%	72.1%	70.3%
	PPV	22.6%	31.9%	32.5%
	F1-Score	34.5%	44.2%	44.4%
	FAPH	0.037	0.021	0.0194
Prospective cohort	**Sensitivity**	70.8%	70.5%	68.7%
	PPV	25.1%	36.3%	36.6%
	F1-Score	37.1%	47.9%	47.7%
	FAPH	0.034	0.020	0.019

The standalone COMPOSER model achieved a sensitivity of 72.9%, positive predictive value (PPV) of 22.6%, F-1 score of 34.5%, and FAPH of 0.037 on the retrospective cohort. In comparison, COMPOSER-LLM$_{Mixtral}$ demonstrated improved performance with a sensitivity of 72.1%, PPV of 31.9%, F-1 score of 44.2%, and FAPH of 0.021. Similarly, COMPOSER-LLM$_{Llama}$ achieved a sensitivity of 70.3%, PPV of 32.5%, F-1 score of 44.1%, and FAPH of 0.0194. Additionally, outputs generated by both the LLMs for the same patient are shown in Table 2.

Within the *prospective cohort*, the performance of COMPOSER-LLM$_{Llama}$ was found to be similar to COMPOSER-LLM$_{Mixtral}$. Specifically, COMPOSER-LLM$_{Llama}$ achieved a sensitivity of 68.7%, PPV of 36.6%, F-1 score of 47.7%, and FAPH of 0.0190 whereas COMPOSER-LLM$_{Mixtral}$ achieved a sensitivity of 70.5%, PPV of 36.3%, F-1 score of 47.9% and FAPH of 0.020.

Table 2: Outputs generated by the LLM for various clinical signs and symptoms for a patient

	COMPOSER-LLM$_{Mixtral}$ (Mixtral 8x7B)	COMPOSER-LLM$_{Llama}$ (Llama-3 8B)
Fever	['Yes', 'Temperature recorded as high as 102.7 degF (4 days ago), and 101degF (today)']	['Yes', 'Temperature of 102.7 degF is greater than or equal to 100.4 degF.']
Hypotension	['No', 'Systolic BP > 100 mmHg, normotensive']	['No', 'BP 122/94 is not less than 90/60.']
Tachypnea	['No', 'Respiratory rate 16-20, within normal limits']	['No', 'Resp rate of 20 is not greater than 20.']
Tachycardia	['No', 'Pulse rate 77-83, within normal limits']	['No', 'Pulse of 83 is not greater than 90.']

Altered mental status	['No', "No evidence of altered mental status in the patient's medical note."']	['No', 'No mention of altered mental status in the medical note.']
Elevated inflammatory markers	['Yes', 'Justification: The CBC with diff shows an elevated white blood cell count (16.1) and increased Imm Gran % (1), indicating possible infection or inflammation.']	['Yes', 'Elevated WBC (16.1) and Anion Gap (16) indicating possible inflammatory response.']
Positive blood culture	['No', 'No information provided about blood culture results in the medical note.']	['No', 'No mention of positive blood culture in the medical note.']
Suspicion of bacterial infection	['Yes', 'Elevated WBC, fevers, nausea, vomiting, worsening right flank and right-sided abdominal pain, and recent instillations into nephrostomy tube suggest possibility of infection.']	['Yes', 'Rising WBC of 16.1, worsening renal insufficiency, and fevers, along with a history of ESBL Kleb from a urine culture 3 days ago, suggest suspicion of bacterial infection.']
Organ dysfunction	['Yes', 'Abnormal renal function with history of CKD, abnormal vital signs (hypotension and tachycardia), abnormal white cell count (leukocytosis)']	['No', "The patient's physical exam and diagnostic testing results do not indicate any organ dysfunction. The patient's vital signs are within normal limits, and the physical exam is notable for moderate to significant tenderness in the right side of the abdomen, but no other abnormalities. "]

4. Discussion

The findings from this study highlight the potential of using the new generation of smaller open-source LLMs for enhancing the early sepsis prediction. The results demonstrated that the Llama-3 8B model (COMPOSER-LLM$_{Llama}$) achieved comparable performance to the larger Mixtral 8x7B model (COMPOSER-LLM$_{Mixtral}$), including sensitivity, positive predictive value (PPV), and F-1 score, with slightly fewer false alarms per patient hour (FAPH). When prospectively evaluated, the COMPOSER-LLM$_{Llama}$ pipeline showed similar performance to the COMPOSER-LLM$_{Mixtral}$ pipeline. These outcomes suggest that, for extraction of clinical signs and symptoms from unstructured clinical notes, the Llama-3 generation of smaller language models can perform as effectively and more efficiently than larger models, providing a more efficient and cost-effective solution for real-time clinical decision support systems.

The new generation of smaller LLMs, such as the Llama-3 8B, possess several advantageous properties over older, larger models like the Mixtral 8x7B. These smaller models have been optimized with improved training data, advanced architectural techniques, and enhanced tokenization methods. Despite their reduced parameter size, these advancements allow smaller models to perform at par or even surpass the performance of older, more extensive models[21]. One of the most significant advantages of smaller LLMs is their lower computational resource

requirement, making them more accessible and scalable for deployment in resource-constrained environments. The reduction in computational overhead (ec2 instance cost of $5.672 per hour for Mixtral 8x7B vs $1.212 per hour for Llama-3 8B) also translates into lower operational costs and faster inference times.

However, this study has several limitations. The sepsis likelihood tool was triggered only after the availability of certain clinical notes ("ED provider note" or "H&P note"), potentially delaying alert generation among patients in the uncertainty interval of 0.5-0.75. However, during the prospective deployment of COMPOSER-LLM, the tool was triggered even if a note was incomplete, as the contextual information within these notes still provided valuable insights. Future research could investigate using LLM-based queries to extract essential patient and provider information, such as suspicion of infection, and explore real-time capture of provider notes through speech recognition and transcription to address issues with missing or incomplete notes. Additionally, while the models were tested on data from two hospitals within a single health system, the generalizability of these findings to other institutions with different patient populations or clinical practices may be limited. Finally, future prospective studies (such as randomized clinical trials) are needed to assess the impact of COMPOSER-LLM on patient care and outcomes.

5. Conclusion

This study demonstrated that a new generation smaller LLM, the Llama-3 8B model (with 8 billion parameters), performed as effectively and more efficiently than an older generation larger LLM, the Mixtral 8x7B model (with 47 billion parameters), for extraction of clinical signs and symptoms from unstructured clinical notes to enable early prediction of sepsis. The results advocate for the potential of smaller models in healthcare, offering a more resource-efficient alternative without compromising accuracy.

Acknowledgments

S.N. is funded by the National Institutes of Health (#R01LM013998, #R35GM143121, #R42AI177108). The opinions or assertions contained herein are the private ones of the author and are not to be construed as official or reflecting the views of the NIH or any other agency of the US Government.

Appendix

Table T1. Patient characteristics of the retrospective and prospective cohorts

	Retrospective cohort		Prospective cohort	
	Septic	***Non-Septic***	***Septic***	***Non-Septic***
# Encounters (%)	215 (16.3%)	1,105	139 (18.4%)	615
Age (in years), median [IQR]	64.7 [51.9 – 75.1]	59.1 [44.9 – 71.2]*	65.1 [50.1 – 75.8]	58.1 [43.3 – 69.6] *
Gender (Male), %	55.8%	50.4%	58.2%	47.7%
Race				
White, %	45.5%	43.9%	50.4%	45.6%
African American, %	11.6%	9.9%	7.2%	7.9%
Asian, %	7.9%	6.5%	4.3%	6.4%
ED Length of Stay (in hours), median [IQR]	24.6 [10.9 – 49.4]	11.8 [6.5 – 29.1] *	22.1 [10.5 – 45.8]	11.9 [6.9 – 30.8] *
CCI, median [IQR]	2 [0 -4]	1 [0 – 3]	2 [0 – 5]	1 [0 -2]
SOFA, median [IQR]	3 [1 – 5]	1 [0 – 2]*	3 [1 – 5]	1 [0 – 2] *
In-hospital mortality, %	8.8%	1.5% *	8.1%	1.2% *
Time from ED triage to onset of sepsis (in hours), median [IQR]	3.2 [1.1 – 8.1]	N/A	2.3 [1.2 – 10.2]	N/A

* p-value<0.05
CCI = Charlson Comorbidity Index
SOFA = Sequential Organ Failure Assessment score

Table T2. Likelihood values for each of the clinical symptoms conditioned on sepsis

Clinical symptoms	Probability value
Fever	0.9
Hypotension	0.05
Tachypnea	0.7
Tachycardia	0.05
Altered mental status	0.05
Elevated inflammatory markers	0.5
Positive blood culture	0.9
Suspicion of bacterial infection	0.75
Organ dysfunction	0.45

References

1. Rudd, K. E. *et al.* Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet* **395**, 200–211 (2020).
2. Singer, M. *et al.* The third international consensus definitions for sepsis and septic shock (Sepsis-3). *J. Am. Med. Assoc.* **315**, 801–810 (2016).
3. Rhee, C. *et al.* Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *J. Am. Med. Assoc.* **318**, 1241–1249 (2017).
4. Ferrer, R. *et al.* Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Crit. Care Med.* **42**, 1749–1755 (2014).
5. Liu, V. X. *et al.* The Timing of Early Antibiotics and Hospital Mortality in Sepsis. *Am. J. Respir. Crit. Care Med.* **196**, 856–863 (2017).
6. Peltan, I. D. *et al.* ED Door-to-Antibiotic Time and Long-term Mortality in Sepsis. *Chest* **155**, 938–946 (2019).
7. Islam, K. R. *et al.* Machine learning-based early prediction of sepsis using electronic health records: a systematic review. *J. Clin. Med.* **12**, 5658 (2023).
8. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
9. Guevara, M. *et al.* Large language models to identify social determinants of health in electronic health records. *NPJ Digit. Med.* **7**, 6 (2024).
10. Wang, H., Gao, C., Dantona, C., Hull, B. & Sun, J. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *Npj Digit. Med.* **7**, 16 (2024).
11. Tai-Seale, M. *et al.* AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. *JAMA Netw. Open* **7**, e246565–e246565 (2024).
12. Ayers, J. W. *et al.* Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
13. Seymour, C. W. *et al.* Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (Sepsis-3). *J. Am. Med. Assoc.* **315**, 762–774 (2016).
14. Boussina, A. *et al.* Impact of a deep learning sepsis prediction model on quality of care and survival. *Npj Digit. Med.* **7**, 14 (2024).
15. Amrollahi, F. *et al.* Inclusion of social determinants of health improves sepsis readmission prediction models. *J. Am. Med. Inform. Assoc.* **29**, 1263–1270 (2022).
16. Shashikumar, S. P., Wardi, G., Malhotra, A. & Nemati, S. Artificial intelligence sepsis prediction algorithm learns to say "I don't know". *NPJ Digit. Med.* **4**, 134 (2021).
17. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
18. The Llama-3 Herd of Models | Research - AI at Meta. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.
19. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *ArXiv Prepr. ArXiv230709288* (2023).

20. Jiang, A. Q. *et al.* Mixtral of Experts. Preprint at http://arxiv.org/abs/2401.04088 (2024).
21. Hassid, M., Remez, T., Gehring, J., Schwartz, R. & Adi, Y. The Larger the Better? Improved LLM Code-Generation via Budget Reallocation. *ArXiv Prepr. ArXiv240400725* (2024).

Using Large Language Models for Efficient Cancer Registry Coding in the Real Hospital Setting: A Feasibility Study

Chen-Kai Wang*
Department of Computer Science, National Yang Ming Chiao Tung University
Hsinchu, 300093, Taiwan, ROC
Advanced Technology Laboratory, Chunghwa Telecom Laboratories
Taoyuan, 326402, Taiwan, ROC
Email: dennisckwang@gmail.com

Cheng-Rong Ke*
Intelligent System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology
Kaohsiung, 80778, Taiwan, ROC
Email: F111154134@nkust.edu.tw

Ming-Siang Huang
Intelligent System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology
Kaohsiung, 80778, Taiwan, ROC
Email: elephant52381@gmail.com

Inn-Wen Chong
Division of Chest Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University
Kaohsiung, 80708, Taiwan, ROC
Department of Biological Science and Technology, National Yang Ming Chiao Tung University
Hsinchu, 30010, Taiwan, ROC
Email: chong@cc.kmu.edu.tw

Yi-Hsin Yang
National Institute of Cancer Research, National Health Research Institutes
Tainan, 70456, Taiwan, ROC
Email: yhyang@nhri.edu.tw

* C.-K. Wang and C.-R. Ke contributed equally to this work.

† Corresponding author

Vincent S. Tseng†
Department of Computer Science, National Yang Ming Chiao Tung University
Hsinchu, 300093, Taiwan, ROC
Email: vtseng@cs.nycu.edu.tw

Hong-Jie Dai†
Intelligent System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology
Kaohsiung, 80778, Taiwan, ROC
Email: hjdai@nkust.edu.tw

The primary challenge in reporting cancer cases lies in the labor-intensive and time-consuming process of manually reviewing numerous reports. Current methods predominantly rely on rule-based approaches or custom-supervised learning models, which predict diagnostic codes based on a single pathology report per patient. Although these methods show promising evaluation results, their biased outcomes in controlled settings may hinder adaption to real-world reporting workflows. In this feasibility study, we focused on lung cancer as a test case and developed an agentic retrieval-augmented generation (RAG) system to evaluate the potential of publicly available large language models (LLMs) for cancer registry coding. Our findings demonstrate that: (1) directly applying publicly available LLMs without fine-tuning is feasible for cancer registry coding; and (2) prompt engineering can significantly enhance the capability of pre-trained LLMs in cancer registry coding. The off-the-shelf LLM, combined with our proposed system architecture and basic prompts, achieved a macro-averaged F-score of 0.637 when evaluated on testing data consisting of patients' medical reports spanning 1.5 years since their first visit. By employing chain of thought (CoT) reasoning and our proposed coding item grouping, the system outperformed the baseline by 0.187 in terms of the macro-averaged F-score. These findings demonstrate the great potential of leveraging LLMs with prompt engineering for cancer registry coding. Our system could offer cancer registrars a promising reference tool to enhance their daily workflow, improving efficiency and accuracy in cancer case reporting.

Keywords: Natural Language Processing; Large Language Models; Electronic Health Record; Cancer registry; Patient Journey.

1. Introduction

Lung cancer stands as the foremost cause of cancer-related deaths among individuals aged 50 years and older, surpassing breast, colorectal, and prostate cancers combined in 2020, as reported by the Global Cancer Observatory, an initiative of the International Agency for Research on Cancer (Ferlay et al., 2020). In the United States, it is projected that 611,720 people will succumb to cancer of all types in 2024, equating to approximately 1,680 deaths per day (Siegel et al., 2024). Similarly, lung cancer has persistently held the top position as Taiwan's leading cause of cancer-specific mortality over the years. The survival rates for patients with lung cancer remain persistently low, often due to late-stage diagnosis that precludes complete surgical resection, thereby reducing long-term survival prospects.

The Taiwan Cancer Registry (TCR), established in 1979 by the Taiwan Society of Cancer Registry, aims to comprehensively measure cancer incidence, morbidity, survival, and mortality among individuals with cancer in Taiwan (Chiang et al., 2015). However, the current method of reporting cancer cases involves labor-intensive and time-consuming manual review of extensive reports, including pathology and radiology reports. Dai et al. (2024) conducted a study at a hospital in southern Taiwan, finding that it takes approximately 30 minutes to process a single case in the

reporting processing. A significant challenge contributing to the time-intensive nature of the process is the large volume and diverse nature of reports associated with each patient. Registrars are required to review and understand a wide array of medical reports, such as pathology reports, radiology reports, and discharge summaries. These reports often cover a span of approximately 1.5 years per patient. One proposed solution to address this challenge involves leveraging artificial intelligence (AI) techniques to automatically parse and extract information from cancer pathology reports. However, these reports are commonly presented in unstructured formats, posing difficulties for machine interpretation due to varying writing styles among different hospitals. Current methodologies predominantly rely on specialized rule-based systems (Coden et al., 2009), machine learning models (Alawad et al., 2020; Dubey et al., 2019; Yoon et al., 2019) or the hybrid of neural symbolic system (Dai, Yang, et al., 2021). Most of these presented works (Alawad et al., 2020; Dubey et al., 2019; Yoon et al., 2019) evaluated their approaches based solely on a single pathology report per patient. This approach may lead to biased results and could struggle to adapt to the real reporting process.

Recently, large language models (LLMs) have emerged as an effective method for extracting information from medical reports (Thirunavukarasu et al., 2023). Due to their large number of parameters and extensive pre-trained on diverse text corpora, LLMs have demonstrated impressive performance across numerous natural language processing (NLP) tasks, including zero-shot and few-shot scenarios (Brown et al., 2020; Nori et al., 2023). Although LLMs have achieved remarkable success in various applications, they still face significant limitations, particularly in domain-specific or knowledge-intensive tasks. These limitations include difficulties with processing long context lengths (Wang et al., 2024) and the potential for generating “hallucinations” when dealing with queries outside their training data or requiring up-to-date information (Zhang et al., 2023). On the other hand, retrieval augmented generation (RAG) is an innovative method for tailoring LLMs to tasks in specific domains (Lewis et al., 2020). The core idea behind RAG is to leverage a vast collection of documents to enhance the capabilities of generative models, thereby improving efficiency in handling complex tasks that require integrated knowledge (Zakka et al., 2024). Unlike traditional LLMs, RAG functions like a search engine by retrieving relevant text data from external knowledge bases through semantic similarity calculations in response to queries. By referencing external knowledge and segmenting large documents into smaller chunks, RAG effectively reduces the problem of generating factually incorrect content and improves the handling of long context data (Kandpal et al., 2023).

In an effort to streamline the data curation process over the various reports of a patient journey while upholding high standards of accuracy, we explore the feasibility of employing LLMs alongside agentic RAG to autonomously extract cancer registry coding items pertaining to lung cancer from various types of clinical reports detailing a patient’s medical journey. This methodology mirrors the responsibilities of a cancer registrar in a real setting, involving the analysis of unstructured reports to identify pertinent data elements essential for cancer registry purposes and their conversion into standardized codes.

Our contributions can be summarized as follows:

(1) We develop an agentic RAG system to facilitate the cancer registry coding process in a real hospital setting. Specifically, we assess the feasibility of directly applying openly available

LLM models without any fine-tuning, utilizing sophisticated crafted prompts through the prompt engineering process.

(2) We empirically show that off-the-shelf LLMs can achieve promising performance on certain cancer registry coding tasks based on the proposed system architecture and the compiled prompts. For example, Mistral-7B (Jiang et al., 2023) can achieve a macro-averaged F-score (F) of 0.637 when evaluated on the test data used in the previous study (Dai et al., 2024).

(3) The LLM, employing strategies such as chain-of-thought (CoT) (Wei et al., 2022) and the proposed coding item grouping, performs better by a large margin than those without these features. When evaluated on the test data, the enhanced strategy outperforms the baseline model without CoT by 0.187 in terms of macro-averaged F-score.

(4) The proposed system can provide a reference text to facilitate the interpretation of the generated outcomes. We conducted an analysis of the presented errors with a detailed discussion for future direction. Through the analysis, we believe that by further validating the generated output with the original reports to reduce the potential hallucinations observed in the presented study, the system could offer cancer registrars a promising reference tool to enhance their daily workflow.

2. Methods

To facilitate the coding process over the large and diverse reports associated with each cancer patient, we propose adapting the agentic RAG system. This system incorporates openly available LLM models along with sophisticatedly designed prompts through the prompt engineering process. In this section, we will first outline the dataset used and the target coding items. Then, we will provide an extensive overview of the proposed agentic RAG system. Subsequently, we will detail the design process and methods for our prompts. Finally, we will describe the evaluation metrics employed to assess the performance of our proposed system.

2.1. *Datasets*

In collaboration with a hospital in southern Taiwan, we collected cancer registry records of lung cancer patients linked with corresponding medical reports in our previous work (Dai et al., 2024). In the compiled dataset, we removed records unrelated to lung cancer based on primary site information, along with patients who had fewer than two reports or only one type of report. This resulted in a final dataset comprising 30 coding item records for 1,629 patients. The dataset was further divided into training and testing sets, comprising 1,287 and 342 patients, respectively. Each patient is associated with an average of 14.6 medical records. Despite Mandarin Chinese being Taiwan's official language, all medical reports were documented mainly in English or a mixture of Chinese and English. The dataset was used for the evaluation of the proposed agentic RAG system for automatic cancer registry coding. For this pilot study, we selected eight coding items to develop our LLM-based cancer registry coding assistant system. These items include pathological TNM classifications (TNM), histology types (H), behavior types (B), primary site (PS), laterality (L), and grades (G).

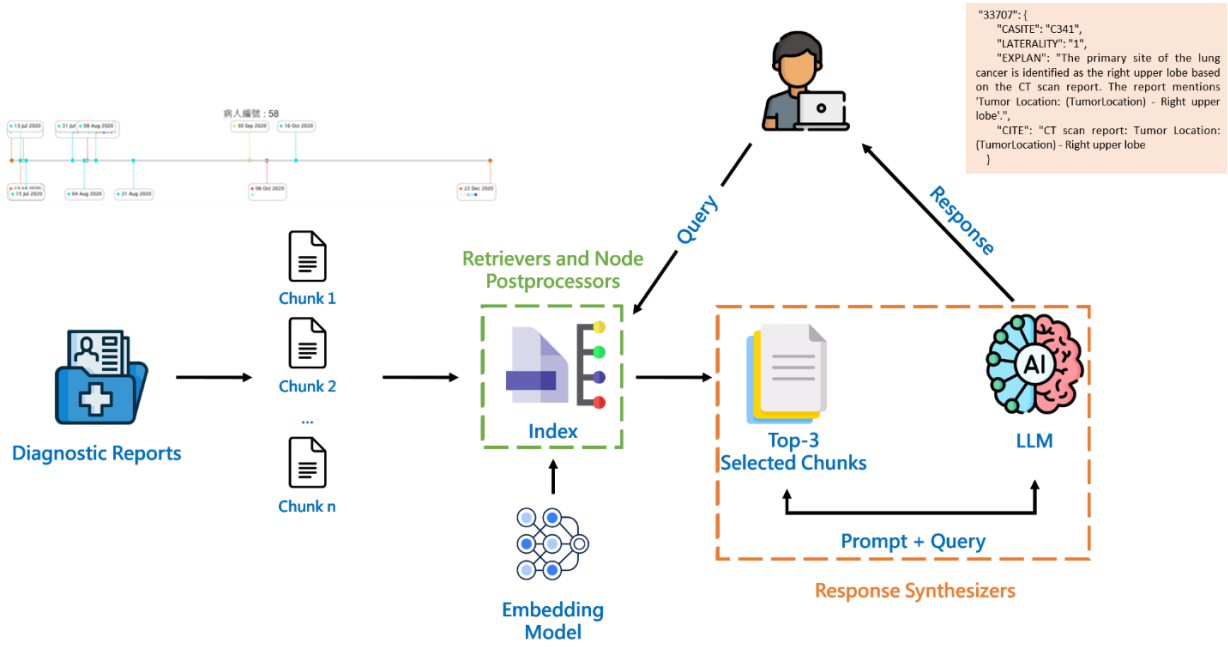


Fig. 1. Workflow of the proposed agentic RAG system.

2.2. *Proposed Agentic RAG System*

We applied RAG to process free-text medical reports collected over approximately 1.5 years to generate recommended cancer registry coding outcomes. Our system employs advanced embedding models to index and retrieve text chunks related to the specific coding prompt from medical reports using retrievers. These chunks are then filtered with post-processors to enhance accuracy before generating standardized cancer registry codes with an LLM. The system functions in two main stages: Top-*k* embedding-based retrieval and LLM-based code generation.

In the first stage, the same embedding model used for indexing the chunks of medical records is used to embed the given prompts for retrieving the most pertinent text chunks from medical reports for each patient. These chunks are refined using keyword-based post-processors to ensure they are among the top three most relevant for the coding task. In the second stage, the refined text chunks are combined with the prompt and query. The LLM then processes this integrated information, learning from patterns in provided examples, analyzing the input, and generating accurate and standardized cancer registry codes.

Figure 1 illustrates the system workflow of the proposed agentic RAG method. The detail workflow operates as follows: Initially, various medical reports for each patient are collected and formatted into JSON lines. These documents are then segmented into smaller, manageable text chunks. Each chunk undergoes processing through an embedding model, transforming it into a vector representation. When a query prompt for a coding task is received, it is similarly converted into a vector representation to facilitate a search within the vector database. This search identifies the most relevant text chunks, which are then combined with the query prompt to create a refined request. This refined request is sent to the LLM for processing, which subsequently provides a comprehensive response. For the underlying LLMs employed by the agents of the proposed RAG

system, we experimented with Mistral-7B and LLaMA3-8B (Touvron et al., 2023). Both Mistral and LlaMA3 are renowned for their balance between computational efficiency and performance across diverse NLP tasks. Therefore, for our implementation of the proposed RAG system, we selected Mistral-7B as the base model and compared its performance with LlaMA3-8B.

2.3. *Prompt Engineering for Cancer Registry Coding*

A prompt is a text-based and task-specific instruction given to a language model to guide its output without altering its parameters. The language model processes the prompt and generates a response based on the provided instructions and context (Marvin et al., 2023). Typically, a prompt may include instructions, input data, context, and an output indicator. According to the information provided, prompts can be categorized into four levels (Heston & Khun, 2023). Level four, known as CoT, breaks down the instruction into step-by-step solutions, offering language models a more structured way to handle the prompt for improved accuracy. Prompt engineering has emerged as a crucial technique for crafting effective prompts. It is an iterative process aimed at refining defined prompts to enhance the capabilities of pre-trained LLMs. In this subsection, we describe the crafted level four prompts through an iterative prompt engineering process.

First, we set the goal to design the initial prompts for the eight coding tasks. We precisely specified the definitions of the coding task along with the desired output formats. In our initial implementation, we used the long-form coding manual of TCR (revision of the 2018v.6) to include detailed explanations and coding guidelines for each coding item. The first and second rows of Table 1 show examples of the PS coding item.

Furthermore, to achieve a more automated and controllable process, we designed output format prompts to instruct the LLM on how to format its output. As shown in the third row of Table 1, we specified that the LLM should generate its response in JSON format to facilitate the extraction of the conclusions. The output JSON object contains three keys: “explain”, “cite”, and the names of the target coding items. The target coding item name key holds the final coding result suggested by the LLM. If the LLM cannot determine the result based on the given report, the values for this key is instructed to assign “NA”. The “explain” key holds the explanation provided by the LLM for the reason why the coding results are suggested. The “cite” key includes the relevant paragraphs from the documents referenced by the LLM to support the coding results. We developed a simple parser based on regular expressions to convert the decoded text response from an LLM into a JSON structured format. If the response for a report cannot be parsed, “NA” is assigned for that report.

During the refining phase, we evaluated the performance of each prompt on the coding tasks using the training set. Biomedical expert MS Huang (listed as the second author) carefully analyzed the models’ responses to identify any errors or areas where the response fell short. Based on the error analysis and the potential solutions observed, we adjusted the prompt content to get a more precise response. This process was repeated until satisfactory performance was achieved. We then evaluate the developed prompts on the test set for performance comparison.

During the iterative process, we observed that certain coding items are often considered together in the actual cancer registration process. Therefore, we treated these related items as a coding item set and integrated their instructions into a single prompt during the design phase. For example, PS and L are often addressed together. This integration helps streamline the process and ensures that

Table 1. Example of the level 3 structured prompt defined for the "primary site" coding item. The ellipsis indicates the placeholder for the prompt string for other coding items belonging to the same group.

Prompt Component	Example
Coding item set definition	Your task as an assistant is to identify and confirm **the primary site […]** of lung cancer. **The primary site refers to specific regions within the respiratory system.** **[…]**
Coding rules for an individual item	It is essential to use only the information provided in the document at hand, considering its date and the pertinent organs or tissues examined. Choose from the following standard codes: - **Primary site codes: C339: Trachea, C340: Main bronchus, C341: Upper lobe, lung, …** […]
Output format (including **examples**)	Your response must be a valid JSON object containing the following keys: **-'primary site': A string containing the code for the primary site.** **[…]** Ensure your response is limited to the provided options for **primary site […]**. For instance, if the pathologic diagnosis specifies '**Lung; upper lobe; left**', this indicates that **the primary site […]** are located in the '**upper lobe**' of the 'left' lung, according to the provided options your JSON response should be: { "explain": "[Insert your explanation here based on the document]", "cite": "[Insert the relevant passages extracted from the document used for your decision]", **"primary site": "C341"**, **[…]**

related items are coded consistently and accurately. Table 2 shows the pre-defined coding item set. The grouped coding items are instructed within the same prompt.

Another significant improvement during the iterative process to the above basic prompt was the introduction of CoT reasoning for coding items like TNM and G. This method involves decomposing the coding task into intermediate steps and solving each step before arriving at the final answer (Wei et al., 2022). For example, consider the coding item G. Initially, we provided detailed coding rules in the prompt, such as:

Table 2. The pre-defined related item groups.

Coding item group type	Coding item
Grouped	- Pathological TNM classification (TNM) - Primary site and Laterality (PS and L) - Histology and Behavior (H and B)
Isolated	Pathological grades (G)

> *…Exclude any data from metastatic sites or recurrent tumors. If an excisional biopsy was conducted at the primary site and subsequent tumor resection shows no residual tumor, use the pathological grade/differentiation from the excisional biopsy. For patients who underwent neoadjuvant treatment before surgery, record the grade/differentiation based on post-surgical tumor tissue pathology. …*

We revised these rules by breaking down the coding task of G into three steps resulting a level four prompt:

1. Identify relevant reports: First, we requested the LLM to identify pathology reports that include surgical procedures from all available medical reports using a list of predefined common surgical terms.
2. Define reference range: Next, we instructed the model to produce the coding result for G based solely on the pathology reports identified in the first step. Coding definition rules similar to the initial detailed definitions shown above were also applied in this step.
3. Point out other key points: Finally, we instruct the model to improve its accuracy by considering the dates of the reports and the specific organs or tissues examined, followed by applying the exact "coding rules" for G.

2.4. *Evaluation Metrics*

We evaluate the performance of the proposed agentic RAG system using the commonly used metrics for evaluating information extraction results: precision (P), recall (R), and F_1-measure (F). P and R are also known as positive predictive value and sensitivity, respectively. The F-score is the weighted harmonic mean of P and R. The formulae for the three metrics are defined as follows:

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

$$F = \frac{2\times P\times R}{P+R}$$

In these formulas, TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively, for each coding item. Specifically, if the model outputs "NA" for a coding item for a patient's entire report set, it is counted as one FN for that patient.

Table 3. Performance comparison of the proposed systems across eight coding items. The highest F-scores for each type are highlighted in bold.

	Mistral-7B			LLaMA3-8B			Neural-symbolic	MT-CNN	HAN
Coding Item	P	R	F	P	R	F	F	F	F
T	0.707	0.915	0.798	0.844	0.972	0.904	**0.905**	0.730	0.763
N	0.845	0.955	0.897	0.860	0.976	0.914	**0.928**	0.830	0.904
M	0.433	0.898	0.584	0.400	0.917	0.557	**0.930**	0.799	0.822
PS	0.877	0.914	0.895	0.894	0.987	**0.938**	0.884	0.750	0.710
L	0.911	0.917	0.914	0.926	0.987	**0.956**	0.948	0.910	0.951
H	0.724	0.710	0.717	0.721	0.964	0.825	**0.871**	0.700	0.760
B	0.942	0.855	0.897	1.000	0.977	0.988	0.934	**0.994**	**0.994**
G	0.815	0.975	0.888	0.883	0.970	0.925	0.932	0.797	**0.939**

3. Results

3.1. *Performance Comparison of the Proposed Agentic RAG System*

To illustrate the effectiveness of the proposed system, we compared it with the previously developed neural symbolic hybrid system (Dai et al., 2024). and two baseline models, as shown in Table 3. For the hierarchical attention network (HAN) model (Gao et al., 2018), we followed the binary relevance transformation method (Dai, Su, et al., 2021) to formulate the coding task for each coding item as a multiclass classification task, training the corresponding number of HAN-based classifiers. For the multi-task convolutional neural network (MT-CNN) model (Alawad et al., 2020), BioWordVec (Zhang et al., 2019) was used to represent tokens, and a single model was trained to generate all eight cancer registry items.

For the proposed RAG systems, LLaMA3-8B clearly outperformed Mistral-7B in almost all coding items under the same configuration and prompt design. LLaMA3-8B also outperformed MT-CNN and HAN in five and six coding items, respectively. Notably, LLaMA3-8B also performed comparably to the neural symbolic system developed in our previous work, achieving the best F-scores in coding items such as PS and L. These promising results demonstrate the feasibility of using LLM models without any fine-tuning for cancer registry coding tasks in the real hospital setting.

3.2. *Ablation Study Results on Different Prompt Engineering Techniques*

To further evaluate the effectiveness of the exectuted prompt engineering process for downstream task performance. We execute our ablation study on four cases: (1) full prompt: the complete prompt with all components shown in Tables 1 and 2 and CoT; (2) a level 3-G prompt: a prompt without CoT; (3) a level 3-I prompt: a prompt without CoT and all coding item groups shown in Table 2 are isolated; and (4) a level 2 prompt: a level 3-I prompt without adding the context. In our implementation, the context refers to the part of “Coding rules for an individual item”. The results are shown in Table 4.

The results from the level three prompt demonstrate the potential of LLMs in performing cancer registry coding tasks from medical reports. This finding is particularly inspiring as it highlights the broader potential of leveraging off-the-shelf LLMs for processing medical text without sophisticated

Table 4. The ablation study results on different prompt engineering techniques

Technique	Macro-P	Macro-R	Macro-F
Full prompt (level 4)	0.782	0.892	0.824
w/o CoT (level 3-G)	0.595	0.733	0.637
w/o CoT & Group (level 3-I)	0.571	0.699	0.609
w/o Coding Rules (level 2)	0.000	0.000	0.000

prompt engineering. Specifically, we observed that the performance for the coding items L, B, and PS is satisfactory, with F-scores over 0.85. However, the performance of the proposed RAG system with grouped prompts on the TN and G coding items is less satisfactory, with F-scores lower than 0.6. Additionally, using isolated prompts alone, the proposed system struggles with additional coding items including TNM and G, showing even lower F-scores (F-score <0.6). By comparing the results of the level three prompt with the full prompt, we found that the inclusion of CoT reasoning significantly boosts the macro-averaged F-score from 0.637 to 0.824. This highlights the effectiveness of the employed prompt engineering process. Additionally, the level two prompt failed to extract any coding items, demonstrating the lack of practical cancer registry coding knowledge in the current off-the-shelf LLMs.

4. Discussion

4.1. *Error Analysis*

Benefiting from the development of pre-trained LLMs, the proposed RAG system can rapidly support most of the cancer registry coding item extraction tasks without further fine-tuning steps. However, from our results, we also observe that the system may occasionally produce conclusions contrary to the facts, even when clear clues are present in the reference texts. These “hallucinations” indicate that the system’s performance has room for improvement. In this section, we outline common error profiles derived from the overall design and present corresponding examples along with potential solutions for future work.

Reference Data Flaws: A single patient may have several to dozens of reports at different times and for different examination items during their treatment period. Using all reports can avoid missing critical information but also introduces computational burdens and noise that may interfere with the decision of the coding results. Therefore, in the retrieval phase for evidential chunks, we only retrieve the top three chunks to narrow the inference space. However, this approach has a double-edged sword effect, which may lead to inappropriate reference chunk citations. Such errors arise when the provided chunks do not offer clear and appropriate clues, leading the model to either refrain from responding or generate hallucinations not mentioned in the original text. Based on our analysis of the presented system errors, it is evident that the current implementation sometimes suffers from the dilemma of similar information retrieved from the top-3 reference data. To address this issue, a post-retrieval process mechanism could be introduced to enhance the diversity among candidate chunks. Balancing data coverage would be helpful for this shortcoming.

Inconsistency with Facts: There are instances where the model produces outputs deviating from the retrieved facts, even when the medical reports already provide a clear basis for concluding the coding results. Despite the defined prompts guiding and restricting the model's behavior, situations that exceed these controls still occur. This type of hallucination, where the model lacks fidelity to the source facts, has also been noted in recent research on LLMs (Tonmoy et al., 2024). Both the initial one-shot prompting and the current self-consistency CoT (Wang et al.) approaches may not be robust enough to assist the model in recalibrating its responses. Future work could explore techniques like Re-Reading (RE2), which enhances understanding by processing questions twice to better focus on the input (Xu et al., 2024), and Self-Reflective Retrieval-Augmented Generation (Self-RAG), which improves both quality and factual accuracy through retrieval and self-reflection (Asai et al., 2023). These methods could help the model produce more consistently and progressively refined outputs.

Knowledge Boundary Limitations: In the process of diagnosing cancer, different examination methods may yield varying results. Summarizing multiple possibilities and ultimately providing a final answer is challenging for both professionals and support systems. For instance, when identifying cancer histology, conclusions derived from surgical pathology are generally more reliable than those obtained from specimens, gross examinations, or microscopic examinations. We noticed that the current applied LLMs are limited by inherent knowledge gaps and may lack the capability to accurately assess the strength of evidence across reports, leading to a higher likelihood of errors.

4.2. *Prompting Engineering for Cancer Registry Coding*

The extraction of target information from clinical texts using LLMs heavily depends on effective prompt design. Due to the multifunctional capabilities of pre-trained models, prompts can be crafted in various ways. This flexibility is particularly useful when considering the professional nature of the input texts and the need for post-processing the output data.

In this study, the aim was to extract specific cancer registry codes from medical reports. The prompt design included a detailed instruction section, coding definitions, and examples, with the output required in a specific JSON format. This comprehensive approach, although necessary for accuracy, resulted in longer prompts. Different studies adopt varying prompt strategies. For instance, Hyeon Seok's work (Choi et al., 2023), which involved extracting cancer features from breast ultrasound and surgery reports, utilized simpler prompts without strict format requirements, as the outputs underwent manual validation. This streamlined approach achieved an accuracy of 87.7%. On the other hand, Huang et al. (2024) study, similar to ours, used detailed prompts for extracting data from public cancer data repositories, requiring output in a JSON format. Their structured prompt design, supported by thorough data preprocessing, achieved an F_1-score of 88%.

These examples demonstrate that while detailed prompts can enhance accuracy, they must be balanced with the need for efficiency and simplicity. A well-designed prompt, aligned with clean data sources and logical objectives, can significantly improve system performance, showcasing the importance of thoughtful prompt construction in utilizing LLMs effectively.

5. Conclusion

Cancer registry tasks involve referencing numerous clinical imaging and diagnostic reports to abstract patient information according to the AJCC-defined codes. These tasks are typically performed by certified clinical personnel with specialized cancer knowledge. The development of cancer registry support systems has the potential to reduce clinical workload and improve healthcare quality. Unlike traditional machine learning models, LLMs can utilize knowledge-guided prompts to predict field codes, making them valuable tools for supporting clinical tasks. In this study, we utilized the Mistral-7B and LLaMA3-8B pre-trained models and designed prompts for eight cancer registry items, including PS, L, H, B, G, and TNM. We observed that providing context and coding rules in a single prompt led to weaker performance due to insufficient reference report extraction. Incorporating CoT prompts, which provide step-by-step guidance toward the final coding output, significantly improved system performance. Additionally, we found that without specific cancer registry rules, the model's outputs became inconsistent and unreliable.

Overall, our findings indicate that LLMs can achieve promising results in lung cancer registry coding tasks even without the need for fine-tuning. Specifically, LLMs demonstrate impressive performance and efficiently utilize auxiliary data for task completion without specific training examples. This underscores their potential as invaluable tools for automating and optimizing cancer data management processes.

Appendix A. Prompt for Grouped Primary Site and Laterality in the Proposed RAG System

Your task as an assistant is to identify and confirm the primary site and laterality of lung cancer. The primary site refers to specific regions within the respiratory system. The laterality refers to whether the cancer originates from a paired organ and is applicable only to primary tumors. It is essential to use only the information provided in the document at hand, considering its date and the pertinent organs or tissues examined. Choose from the following standard codes for lung cancer sites and laterality:

Primary site codes:
- C339: Trachea
- C340: Main bronchus
- C341: Upper lobe, lung
- C342: Middle lobe, lung
- C343: Lower lobe, lung
- C348: Overlapping lesion of lung
- C349: Lung NOS (Not Otherwise Specified)

Laterality codes:
- 1: Primary origin of the cancer is on the right side.
- 2: Primary origin of the cancer is on the left side.
- 3: Unilateral involvement only, but origin unclear whether from left or right side.
- 4: Bilateral involvement with unclear side of origin, and medical records describe a single primary.

Your response must be a valid JSON object containing the following keys:
-'primary site': A string containing the code for the primary site.
-'laterality': A string containing the code for laterality.

Ensure your response is limited to the provided options for primary site and laterality. For instance, if the tissue in the report is labeled as 'Lung; NOS' and the pathologic diagnosis specifies 'Lung; upper lobe; left', this indicates that the primary site and laterality are located in the 'upper lobe' of the 'left' lung, according to the provided options your JSON response should be:

```
{
"explain": "[Insert your explanation here based on the document]",
"cite": "[Insert the relevant passages extracted from the document used for your decision]",
"primary site": "C341",
"laterality": "2"
}
```

Appendix B. Implementation Details for the Proposed RAG System

For the proposed RAG system, we utilize the LlamaIndex (Liu, 2022) framework. The developed system is deployed on a machine equipped with PyTorch libraries and CUDO12.0 along with an Intel i7-13700 processor, 64GB of RAM, and an NVIDIA GeForce RTX 4090 24GB VRAM (video RAM) graphics card. We employ M3-Embedding (Chen et al., 2024) as our embedding model for encoding a patient's every medical report during the indexing stage. For the retrieval module, we set the number of top K candidate chunks to three. In our configuration settings, we set the temperature to 0 and the seed to 42.

It is worth noting that loading models for inference demands a substantial amount of GPU memory. A general rule of thumb is that every billion parameters require 3 GB of graphics double data rate (GDDR) 6 VRAM for the default precision of parameter values (Lin et al., 2024). Due to the limitations of our machine hardware specifications, we quantize the employed LLMs to fixed-point 4 (FP4) for inference, which recasts these model weights into lower precision data types. This method slightly reduces performance but significantly lowers the memory requirement to a quarter of the original.

Appendix C. Definition of 30 Lung Cancer Coding Items in the Dataset for This Study

Coding Type	Description
AJCC Edition	The version and chapters of the AJCC (American Joint Committee on Cancer) cancer staging manual used to determine the cancer stage of the case.
Behavior Code	The morphological code (M-code) in the pathological diagnosis. The 5th code in the M-code is the behavior code. The first four digits of M-code indicate the specific histological term. The fifth digit is the behavior code, which indicates whether a tumor is malignant, benign, in situ, or uncertain.
Clinical Other Staging Group	The classification standards of the selected "Other Staging Systems" (defined below) chosen for staging cancer cases.
Clinical Stage Descriptor	The prefix or suffix used in conjunction with clinical TNM fields. The prefix/suffix denotes special circumstances that may affect the staging and analysis of the data and is based on the clinical T, N, and M categories prior to treatment.

Date of First Microscopic Confirmation	The earliest date when the case's cancer was confirmed by microscopy.
Date of First Surgical Procedure	The earliest date of surgery for cancer performed at any medical institution.
Date of Initial Diagnosis	The earliest date the cancer was diagnosed by a physician.
Date of Surgical Diagnostic and Staging Procedure	The date of the surgical treatment performed for diagnosis or staging at any medical institution.
Diagnostic Confirmation	The most accurate basis of diagnosis at the reporting hospital or an external hospital for the case.
Grade Clinical	The grading/differentiation of the solid tumor before the first treatment. Grading/differentiation refers to the degree of similarity between the tumor and normal tissues. Well differentiated (Grade I) is most similar to normal tissue; undifferentiated (Grade IV) is most dissimilar from normal tissue.
Grade Pathological	The grading/differentiation of the solid tumor after surgery at the primary site. Grading/differentiation refers to the degree of similarity between the tumor and normal tissues. Well differentiated (Grade I) is most similar to normal tissue; undifferentiated (Grade IV) is most dissimilar from normal tissue.
Histology	The structure of the primary tumor cells under the microscope.
Laterality	The specification of whether the cancer originates from one side of a pair of organs or the body. It is a only applicable to the primary tumor site.
Lymph vessels or Vascular Invasion	The code is recorded based on the pathological report of the primary site to indicate the presence or absence of invasion into lymph vessels or blood vessels.
Nodes Examined	The total number of regional lymph nodes examined by a pathologist.
Nodes Positive	The total number of positive regional lymph nodes examined by a pathologist.
Other Staging System	The selection of alternative staging criteria if the AJCC Cancer Staging System is not utilized.
Pathologic M	The presence of distant metastases of the primary tumor.
Pathologic N	The regional lymph nodes involvement of the tumor. The item is encoded based on all clinical evaluations done prior to definitive surgery, plus all information through completion of definitive surgeries in the first course of treatment in the absence of disease progression or within 4 months of diagnosis, whichever is longer.

Pathologic Stage Descriptor	The prefix or suffix used in conjunction with pathologic TNM fields. The prefix/suffix denotes special circumstances that may affect the staging and analysis of the data and is based on the pathologic T, N, and M categories after completion of surgical treatment.
Pathologic T	The size of the primary tumor and its invasion into adjacent tissues. The item is encoded based on all clinical evaluations done prior to definitive surgery, plus all information through completion of definitive surgeries in the first course of treatment in the absence of disease progression or within 4 months of diagnosis, whichever is longer.
Perineural Invasion	The presence of neural invasion as noted in the pathological report of the primary site in the medical records.
Primary Site	The primary site of the cancer.
Scope of Regional Lymph Node Surgery	The extent of regional lymph nodes removed, sectioned, or aspirated during the primary site surgery or another separate surgery at the reporting hospital.
SSF 2 SSF 5 SSF 6 SSF 7	Cancer site-specific factors (SSF) related to prognosis and treatment decisions. SSF2: Visceral pleural Invasion (VPI)/elastic layer value set. SSF5: Sampling or dissection of mediastinal lymph nodes (N2 Nodes) value set. SSF6: EGFR (epidermal growth factor receptor) gene mutation value set. SSF7: ALK (Anaplastic lymphoma kinase) gene translocation value set.
Surgical Margins	The final status of the surgical margins after the primary tumor is removed.
Surgical Margins Date	The closest distance of tumor cells to the surgical margins in the pathological report after the primary tumor is removed.

References

Alawad, M., Gao, S., Qiu, J. X., Yoon, H. J., Blair Christian, J., Penberthy, L., Mumphrey, B., Wu, X.-C., Coyle, L., & Tourassi, G. (2020). Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *Journal of the American Medical Informatics Association*, *27*(1), 89-98.

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Chiang, C.-J., You, S.-L., Chen, C.-J., Yang, Y.-W., Lo, W.-C., & Lai, M.-S. (2015). Quality assessment and improvement of nationwide cancer registration system in Taiwan: a review. *Japanese journal of clinical oncology*, *45*(3), 291-296.

Choi, H. S., Song, J. Y., Shin, K. H., Chang, J. H., & Jang, B.-S. (2023). Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiation Oncology Journal*, *41*(3), 209.

Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K., Cooper, J., Guan, W., & De Groen, P. C. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of biomedical informatics*, *42*(5), 937-949.

Dai, H.-J., Chen, C.-C., Mir, T. H., Wang, T.-Y., Wang, C.-K., Chang, Y.-C., Yu, S.-J., Shen, Y.-W., Huang, C.-J., & Tsai, C.-H. (2024). Integrating predictive coding and a user-centric interface for enhanced auditing and quality in cancer registry data. *Computational and Structural Biotechnology Journal*, *24*, 322-333.

Dai, H.-J., Su, C.-H., Lee, Y.-Q., Zhang, Y.-C., Wang, C.-K., Kuo, C.-J., & Wu, C.-S. (2021). Deep learning-based natural language processing for screening psychiatric patients. *Frontiers in psychiatry*, *11*, 533949.

Dai, H.-J., Yang, Y.-H., Wang, T.-H., Lin, Y.-J., Lu, P.-J., Wu, C.-Y., Chang, Y.-C., Lee, Y.-Q., Zhang, Y.-C., & Hsu, Y.-C. (2021). Cancer registry coding via hybrid neural symbolic systems in the cross-hospital setting. *IEEE Access*, *9*, 112081-112096.

Dubey, A. K., Hinkle, J., Christian, J. B., & Tourassi, G. (2019). Extraction of tumor site from cancer pathology reports using deep filters. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics,

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2020). Global cancer observatory: cancer today. International Agency for Research on Cancer. *Lyon, France*.

Gao, S., Young, M. T., Qiu, J. X., Yoon, H.-J., Christian, J. B., Fearn, P. A., Tourassi, G. D., & Ramanthan, A. (2018). Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association*, *25*(3), 321-330.

Heston, T. F., & Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, *2*(3), 198-205.

Huang, J., Yang, D. M., Rong, R., Nezafati, K., Treager, C., Chi, Z., Wang, S., Cheng, X., Guo, Y., & Klesse, L. J. (2024). A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine*, *7*(1), 106.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., & Saulnier, L. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. International Conference on Machine Learning,

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., & Han, S. (2024). AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. *Proceedings of Machine Learning and Systems*, *6*, 87-100.

Liu, J. (2022). *LlamaIndex*. https://github.com/jerryjliu/llama_index

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. International conference on data intelligence and cognitive informatics,

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. *CA: a cancer journal for clinicians*, *74*(1).

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, *29*(8), 1930-1940.

Tonmoy, S., Zaman, S., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., & Azhar, F. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, X., Salmani, M., Omidi, P., Ren, X., Rezagholizadeh, M., & Eshaghi, A. (2024). Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. Self-consistency improves chain of thought reasoning in language models. arXiv 2022. *arXiv preprint arXiv:2203.11171*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Xu, X., Tao, C., Shen, T., Xu, C., Xu, H., Long, G., & Lou, J.-g. (2024). Re-Reading Improves Reasoning in Large Language Models.

Yoon, H.-J., Gounley, J., Gao, S., Alawad, M., Ramanathan, A., & Tourassi, G. (2019). Model-based hyperparameter optimization of convolutional neural networks for information extraction from cancer pathology reports on HPC. 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI),

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., & Ashley, E. (2024). Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, *1*(2), AIoa2300068.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, *6*(1), 52.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., & Chen, Y. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Automated Evaluation of Antibiotic Prescribing Guideline Concordance in Pediatric Sinusitis Clinical Notes

Davy Weissenbacher†, PhD
Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA
E-mail: davy.weissenbacher@cshs.org

Lauren Dutcher†, MD, MSCE
Division of Infectious Diseases, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
E-mail: LDutcher@pennmedicine.upenn.edu

Mickael Boustany, MD
Division of Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Leigh Cressman, MA
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
E-mail: crel@pennmedicine.upenn.edu

Karen O'Connor, MS
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
E-mail: karoc@pennmedicine.upenn.edu

Keith W. Hamilton, MD
Division of Infectious Diseases, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
E-mail: Keith.Hamilton@pennmedicine.upenn.edu

†Both authors contributed equally to this work

Jeffrey Gerber, MD, PhD
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine,
Philadelphia, PA, USA
E-mail: GERBERJ@chop.edu

Robert Grundmeier, MD
Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia,
Philadelphia, PA, USA
E-mail: GRUNDMEIER@chop.edu

Graciela Gonzalez-Hernandez, PhD
Department of Computational Biomedicine, Cedars-Sinai Medical Center,
Los Angeles, CA, USA
E-mail: Graciela.GonzalezHernandez@csmc.edu

Background: Ensuring antibiotics are prescribed only when necessary is crucial for maintaining their effectiveness and is a key focus of public health initiatives worldwide. In cases of sinusitis, among the most common reasons for antibiotic prescriptions in children, healthcare providers must distinguish between bacterial and viral causes based on clinical signs and symptoms. However, due to the overlap between symptoms of acute sinusitis and viral upper respiratory infections, antibiotics are often over-prescribed.
Objectives: Currently, there are no electronic health record (EHR)-based methods, such as lab tests or ICD-10 codes, to retroactively assess the appropriateness of prescriptions for sinusitis, making manual chart reviews the only available method for evaluation, which is time-intensive and not feasible at a large scale. In this study, we propose using natural language processing to automate this assessment.
Methods: We developed, trained, and evaluated generative models to classify the appropriateness of antibiotic prescriptions in 300 clinical notes from pediatric patients with sinusitis seen at a primary care practice in the Children's Hospital of Philadelphia network. We utilized standard prompt engineering techniques, including few-shot learning and chain-of-thought prompting, to refine an initial prompt. Additionally, we employed Parameter-Efficient Fine-Tuning to train a medium-sized generative model Llama 3 70B-instruct.
Results: While parameter-efficient fine-tuning did not enhance performance, the combination of few-shot learning and chain-of-thought prompting proved beneficial. Our best results were achieved using the largest generative model publicly available to date, the Llama 3.1 405B-instruct. On our evaluation set, the model correctly identified 94.7% of the 152 notes where antibiotic prescription was appropriate and 66.2% of the 83 notes where it was not appropriate. However, 15 notes that were insufficiently, vaguely, or ambiguously documented by physicians posed a challenge to our model, as none were accurately classified.
Conclusion: Our generative model demonstrated good performance in the challenging task of chart review. This level of performance may be sufficient for deploying the model within the EHR, where it can assist physicians in real-time to prescribe antibiotics in concordance with the guidelines, or for monitoring antibiotic stewardship on a large scale.

Keywords: Antibiotic Stewardship, Classification, Large Language Models, Generative Systems

1. Introduction

Antibiotic stewardship programs (ASPs) aim to optimize the use of antibiotics for specific conditions and to combat the growing threat of antimicrobial resistance.[1] Inappropriate prescribing of antibiotics not only contributes to a global health crisis but also exposes patients, particularly pediatric patients, to unnecessary side effects and disrupts their healthy microbiota.[2] Ensuring that antibiotics are prescribed adequately —only when necessary and with the correct dosage and duration— is essential for maintaining their efficacy and is a key focus in public health and research efforts at national and international levels.

Most antibiotic prescribing takes place in the ambulatory setting, and approximately 30% of all outpatient antibiotic prescriptions are unnecessary; a majority of unnecessary outpatient prescribing is for acute upper respiratory tract infections.[3,4] In particular, sinusitis which is among the most common reasons for ambulatory antibiotic prescribing in children.[3] The symptoms of acute sinusitis often overlap significantly with those of uncomplicated viral upper respiratory tract infections. As a result, antibiotics are often over-prescribed for sinusitis, despite guidelines recommending more conservative use.[5,6]

The Centers for Disease Control and Prevention (CDC) Core Elements of Outpatient Antibiotic Stewardship recommend tracking and reporting ambulatory antibiotic prescribing.[7] Some metrics using data from the electronic health record (EHR) have been developed in order to measure unnecessary and guideline-discordant prescribing.[8] Several studies have created classification models to assess appropriate antibiotic prescribing by linking patient diagnoses to tier-based rules where the antibiotic prescription is always, sometimes, or never appropriate depending on the diagnosis.[3,9,10] Others have focused on metrics for specific conditions, such as acute bronchitis, or have addressed antibiotic selection or duration of therapy.[11–13] These metrics have successfully been used in feedback for clinicians and practices and in assessing the impact of stewardship programs on prescribing.

However, while these metrics and classification schemes perform reasonably well, they have primarily only used structured data from the EHR, and have not been able to use information from unstructured text present in clinical notes. This creates a significant gap for conditions in which the assessment of appropriateness using an electronically-based metric from structured data is not feasible. For example, in acute sinusitis, healthcare providers must distinguish bacterial from viral sinusitis based on clinical signs and symptoms alone, and antibiotic prescribing is only considered guideline-concordant for bacterial sinusitis. As such, there are no lab tests or ICD-10 codes (structured data) that can be used to retroactively measure prescribing appropriateness in the absence of time-intensive manual chart review of clinical notes. While audits of patient charts have elicited important findings for the field of antibiotic stewardship, there are limitations to manual review.[9,14,15] Retrospective manual review of charts is labor intensive and time consuming, therefore only small samples of charts can be reviewed, limiting the potential applications in large scale antibiotic stewardship interventions.

This paper explores the significance of antibiotic stewardship for pediatric sinusitis and presents a generative system, utilizing a Large Language Model (LLM) approach, to automate the analysis of unstructured notes from pediatric primary care practices to determine justified vs unjustified prescription of antibiotics given a case presentation, seeking to enable a large-

scale study that aims to improve prescribing practices.

2. Materials and Methods

We represented the task of evaluating the guideline concordance of antibiotic prescribing in clinical notes as a decision task. That is, given a note in which a patient was diagnosed with sinusitis and prescribed antibiotics, our system should predict whether the prescription was 1) appropriate, 2) not appropriate, or 3) insufficient or ambiguous, in cases where the note does not contain enough information to assess the appropriateness of the prescription.

2.1. *Data collection*

We identified all pediatric (younger than 18) clinical encounter notes by ICD-10 code from outpatient billed encounters at one of 32 primary care practices in the Children's Hospital of Philadelphia (CHOP) network from July 1, 2017 through June 30, 2021 using the following criteria: 1) visits with either a J01 (acute sinusitis) or J32 (chronic sinusitis) code and 2) a prescription of an oral antibiotic (excluding antibiotics that would never be prescribed for sinusitis). The following patients were excluded: 1) patients with a confounding chronic medical condition identified by an ICD-10 code;[16] 2) patients with an ICD-10 code for another infection that would warrant an antibiotic prescription at the same visit. Only primary care visits were included; emergency department and urgent care visits were excluded. Only office visit notes from healthcare providers were included.

A total of 10,311 patients met the inclusion criteria 6,377 (61.9%) for acute sinusitis, and 3,934 (38.2%) for chronic sinusitis, seen by 310 providers. The median number of encounters per provider was 12 (3 – 48). To develop, train, and evaluate our classifier, we selected 300 encounter notes at random. Our intent was to reflect the natural distribution of the notes where the system will be deployed, so we did not oversample or undersample any specific group or provider. This resulted in 190 (63.3 %) encounter notes for acute sinusitis and 110 (36.7%) for chronic sinusitis, seen by 132 providers. The median number of encounters per provider was 50 (21.5 – 92).

We split our annotated dataset into three sets, the first two of which were selected from 80 percent of the providers: a training set with 200 notes (117 notes with appropriate prescriptions, 69 not appropriate, and 14 with insufficient or ambiguous documentation), a development set with 50 notes (32 appropriate, 16 not appropriate and 2 insufficient). For the third set (the test set), we selected 50 notes from the remaining 20% of the providers (35 appropriate, 14 not appropriate, 1 insufficient), in order to be able to test the system on how it adapts to notes from new (unseen) providers.

2.2. *Annotation*

We derived a set of criteria by adapting the recommendations of two clinical practice guidelines[17,18] to define the appropriateness of antibiotic prescribing to the patients we selected. Table 1 summarizes our criteria. If a patient met at least one criterion, our annotators labeled the note as appropriate. If there was clear evidence in the note that none of the criteria were

met, the annotators labeled the note inappropriate; otherwise, if it was not possible for the annotator to decide if the criteria were met or not in a note, the note was labeled insufficient. Such cases usually include incomplete, ambiguous or vague documentation. The phrase "*patient had congestion for over a week*" is an example of an ambiguous documentation. If the congestion lasted for 8 or 9 days, criterion 1 in Table 1 would not be met and this would be labeled 'not justified'. However, if the symptom lasted 10 days or longer, then criterion 1 would be satisfied and this would be labeled 'justified'. The phrase "*Fever x 3 days*" is an example of incomplete documentation because it does not specify the exact temperature. Note that our definition focuses solely on the act of prescribing antibiotics and excludes considerations related to the appropriateness of the specific antibiotic prescribed, as well as its dosage and duration.

Table 1: Clinical guidelines used to assess the appropriateness of an antibiotic prescription for patients diagnosed with sinusitis. If the clinical note provided sufficient evidence to meet at least one of the three established criteria, the prescription was annotated as appropriate.

Antibiotics appropriateness
1. Persistent illness: nasal discharge (of any quality), daytime cough, or sinus pain/pressure lasting for $\geq$ 10 days without improvement
2. Severe onset, i.e., concurrent fever (temperature $\geq$ 39°C/102.2°F) and purulent nasal discharge or sinus pain/pressure for at least 3 consecutive days
3. Worsening course, i.e., worsening or new onset of nasal discharge, daytime cough, sinus pain/pressure, or fever after initial improvement

One pediatric physician annotated the 300 notes of our corpus as *appropriate*, *inappropriate*, or *insufficient*. A second pediatric physician is currently annotating 50 notes of our corpus to compute the inter-annotator agreement. To guide this assessment, an annotation guide was developed, using input from a primary care pediatrician, two infectious diseases specialists, and one pediatric infectious diseases specialist. This annotation guide was developed iteratively using practice notes from the same practices with the goal of improving reproducibility as much as possible.

2.3. *Generative models*

Our task presents a significant challenge for conventional natural language processing (NLP) systems, which typically rely on a pipeline approach.[19–21] In such systems, a task is divided into several 'simpler' subtasks, each performed sequentially by independent modules. To complete our task, an NLP pipeline would first require an information extraction module to identify key symptoms in the clinical notes —congestion, cough, sinus pain/discomfort, and fever— as reported by the patient during the encounter. Next, a classification module would detect mentions of symptom severity and assign appropriate labels to each symptom. A third module would normalize the extracted information by identifying and representing the progression of symptoms. Finally, a logical validation module would verify whether the extracted information aligns with the criteria outlined in Table 1, ultimately generating the final decision.

This pipeline approach has several limitations that often result in reduced performance[22] and limited adoption in the medical field. Each module operates based on a set of rules, which can either be manually crafted or automatically learned from training data. Both approaches demand significant human effort. In the medical domain, writing rules requires expertise in both computer science and medicine, and these rules are often difficult to maintain over time.[23] An alternative is to learn the rules directly from annotated examples,[24] but standard machine learning algorithms typically need thousands of examples to achieve acceptable performance, a resource-intensive and costly process. This often leaves modules only partially trained, leading to suboptimal results.[25] Even when the rules are well defined, they rarely account for all possible cases, and module performance is almost never flawless.[24] Since a pipeline approach processes data sequentially through imperfect modules, errors from earlier stages propagate through the system, compounding in later stages and significantly limiting overall performance.[19,26] Moreover, conventional NLP modules —such as classifiers, sequence labelers, or normalizers— are typically designed to output only their labels and confidence scores, without providing explanations for their decisions. This lack of interpretability forces experts to rely on ad-hoc algorithms producing only partial and incomplete explanation of the module behavior.[27] This issue was particularly pronounced with transformer-based encoders like BERT,[28,29] the standard NLP architecture before the recent advancements with large language models-based generative models, which was often qualified as a black box system and not well adopted by medical professionals who doubted their decision.

As an alternative to conventional NLP systems, we propose using state-of-the-art generative systems powered by large language models (LLMs). In recent years, generative systems have become the leading approach in NLP as evidenced by the widespread success of chatGPT.[30] Generative systems feature interfaces that allow users to submit prompts in natural language, an intuitive interface to perform a task.[31] These prompts typically include an instruction specifying the desired action, along with optional data needed to perform the task. Generative systems leverage semi-supervised training to transfer general knowledge acquired from extensive text corpora, enabling them to generate appropriate responses and execute instructions for tasks they were not explicitly trained on. This eliminates the need to retrain the system for each specific task, a requirement often necessary in conventional NLP systems.

In this study, we applied a generative system to address the specific challenge of antibiotic stewardship, a task for which no established benchmarks exist. In accordance with common practices for deploying generative systems in clinical settings,[32] we utilized prompt engineering with few-shot learning and chain-of-thought reasoning. Instead of adopting more advanced and resource-intensive techniques —such as full fine-tuning on large clinical datasets,[33] knowledge injection via retrieval-augmented generation,[34] or self-correction through multi-agent interactions— we chose to evaluate the system's inherent capabilities,[35,36] reserving these enhancements for future research.

2.4. *Classification with Generative systems*

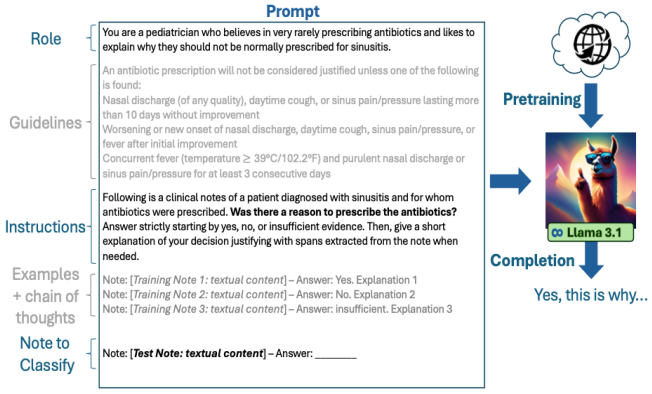
We performed our classification using generative systems from the Llama 3 family,[37] which is one of the largest freely available sets of models offering competitive performance compared

to proprietary alternatives. We progressively refined an initial simple prompt by following few-shot learning and chain-of-thought techniques to enhance the models' performance on our task. Additionally, we fine-tuned a Llama-3-70B-Instruct model using a parameter-efficient fine-tuning (PEFT) approach, namely LoRA,[38] to specialize the model for our specific task.

Initial prompt. Figure 1 outlines the various components of the prompt we designed to instruct our model on how to classify the appropriateness of antibiotic prescription in clinical notes. We began our experiments with an initial straightforward prompt that defined the role the generative model should assume, followed by a brief paragraph specifying the instructions for the task. This paragraph included the following key components:

(1) The role specifying the function the model should adopt when generating response, in our case a pediatrician.
(2) The context which describes the notes; specifically, the input note is a clinical note of a patient diagnosed with sinusitis who received antibiotics.
(3) The question the model should answer.
(4) The format in which we wanted the model to present its response.
(5) The text of the note to be classified
(6) The keyword *Answer:* to initiate the model's completion according to our specified format.

The authors, during an interactive session, tried multiple initial prompts and evaluated the Llama 3 70B-instruct model's results on the development set. At the end of the interactive session, we selected the initial prompt illustrated in Figure 1 Left.



Prompt with conditional guidelines

Role: You are a pediatrician who believes in very rarely prescribing antibiotics and likes to explain why they should not be normally prescribed for sinusitis.
Definitions: A patient has fever if the patient has a body temperature above 102.2°F or 39°C. Nasal discharge quality can be a. clear and watery, b. thick and white, c. yellow, green, brown or grey, d. bloody, e. foul-smelling, f. thick and stringy, g. purulent. If a patient has nasal passages congested, then the patient has nasal discharges.
Conditional guidelines: As a pediatrician you would consider that the prescription of an antibiotic was not justified unless one or more of the following rule is satisfied:
Rule 1. If the patient had nasal discharge of any quality for 10 or more days without improvement, then the prescription of the antibiotics was justified.
Rule 2. If the patient was coughing during the day for 10 or more days without improvement, then the prescription of the antibiotics was justified.
Rule 3. If the patient experienced discomfort or pain in the areas around the sinus for 10 or more days without improvement, then the prescription of the antibiotics was justified.
Rule 4. If the patient had nasal discharge of any quality, was recovering, but then experienced an increase of its severity or its reappearance, then the prescription of the antibiotics was justified.
Rule 5. If the patient was coughing during the day, was recovering, but then experienced an increase of its severity or its reappearance, then the prescription of the antibiotics was justified.
Rule 6. If the patient experienced discomfort or pain in the areas around the sinus, was recovering, but then experienced an increase in the pain/discomfort severity or its reappearance, then the prescription of the antibiotics was justified.
Rule 7. If the patient had fever and, on the same time, had purulent nasal discharge for 3 or more consecutive days, then the prescription of the antibiotics was justified.
Rule 8. If the patient had fever and, on the same time, experienced discomfort or pain in the areas around the sinus, for 3 or more consecutive days, then the prescription of the antibiotics was justified.
Instructions: Following is are clinical notes of patients diagnosed with sinusitis and for whom antibiotics were prescribed. By default, assume that the prescription of antibiotics was not appropriate unless you find evidence in the note indicating that at least one of the preceding rules was satisfied. Answer strictly starting by Yes, No Insufficient. Then, give a short explanation of your decision justifying with spans extracted from the note when needed.
Examples - (note excerpt, answer, explanation, quotes) :
Note: {note 1 excerpt} Answer: Yes. Explanation: The patient was coughing for two weeks which is more than 10 days therefore the prescription of antibiotics was justified (Rule 2. is satisfied) Quote: "cough x 2 weeks"
Note: {note 2 excerpt} Answer: No. Explanation: The patient had nasal passages congested for only 5 days which is less than 10 days (Rule 1. is not satisfied). The patient experienced pain in the areas around the sinus pain but no duration was specified (Rule 3. is not satisfied). Quote: "Thick yellow green congestion for about 5 days"
Note: {Note 3 excerpt} Answer: Insufficient. Explanation: The patient was coughing during the day between 7 to 10 days. This duration is vague. If the patient was coughing for less than 10 days then the prescription of the antibiotics was not justified (Rule 2. is not satisfied), on the contrary, if the patient was coughing for 10 days then the prescription of the antibiotics was justified (Rule 2. is satisfied) Quote: "Cough? 7-10 days"
Note to classify:
Note: {textual content of the note to classify} Answer:

Fig. 1: Left: Iterative construction of a prompt to classify antibiotic prescription appropriateness using a Llama 3 generative model. Right: Our prompt with conditional guidelines.

Guidelines. We first extended this initial prompt by inserting the clinical guidelines that our annotators followed when labeling the notes in our corpus. The generative models that are publicly available were pretrained on large corpora from the internet, which contain few, if any, professional medical documents[39,40] and may not have encountered or memorized our specific guidelines. By including these guidelines directly in the prompt, we ensured that the model had direct access to the criteria defining the task. The guidelines in Table 1 were written for medical professionals. To make them more accessible and easier for our generative model to interpret, *LD*, Assistant Professor in Medicine, simplified the language used in the guidelines.

Few-shot learning. Much like humans, generative models can benefit from seeing a few examples before attempting a task, a concept known as few-shot learning. We implemented this approach by including the text of three notes from our training set in the prompt, each accompanied by their appropriateness labels. Although complex conditions could be used to select these examples -such as choosing notes with close semantic similarity to the one being classified or those that annotators found challenging[41]- we opted to select the examples randomly. We chose this approach for simplicity and left the exploration of more sophisticated selection strategies for future work.

Chain-of-thought prompting. Together with few-shot learning, we also employed chain-of-thought prompting.[42] After each label of our training examples, we included a brief explanation of the label, along with the relevant quotes that demonstrated the extracted span from the example note supporting the explanation. *LD* provided these explanations, highlighting which criteria from Table 1 were met, missing, or challenging to verify based solely on the note's text. Requesting explanations along with quotes forces the model to ground its responses within the text of the notes, thereby reducing hallucinations. Despite the large context window of 8,192 tokens, the Llama-3-70B-Instruct model still has a limited prompt capacity, which restricted us to including no more than three example notes. In a supplementary experiment, to include additional examples, we did not input the entire text of the training notes. Instead, we truncated the notes, only incorporating the sentences containing the relevant quoted phrases.

We hypothesized that chain-of-thought reasoning is an important component for improving the performance of a generative model, and conducted additional experiments by reformulating our initial prompt and its components. While the description of the model's role remained unchanged, we revised the context and question to predispose the model to answer 'not appropriate' by default unless it identified evidence in the notes that satisfied a criterion from our guidelines. We also introduced simple definitions for *'fever,' 'nasal discharge,' and 'nasal congestion'* before presenting the guidelines, and we rephrased the guidelines as a set of eight conditional rules. Additionally, we revised the explanations for all ten examples in the prompt to explicitly indicate which rules were met or unmet (Line 9 in Table 2). We provide the exact prompt used in these experiments in Figure 1) Right.

Parameter Efficient Fine-tuning with LoRA. Although generative models achieve state-of-the-art performance on general NLP tasks, they may benefit from being fine-tuned to perform more specific and challenging tasks. A standard method for training generative models is full fine-tuning, a supervised training process. In this process, the model is presented with instructions and corresponding data required to perform a task. It generates a response, which is then

automatically compared to the expected gold-standard answer from the training examples. If the model generates the expected answer, no adjustments are made. However, if it deviates, all weights of its underlying neural network are updated to increase the likelihood of generating the correct response. While fine-tuning can enhance the model's performance on specific tasks, updating all weights of a very large neural network is computationally intensive and requires a significant number of expensive GPUs, which were not available for our experiments. LoRA is a heuristic proposed by Yu et al.[38] to update only a small portion of the weights in the neural network. We trained the Llama 3 70B-instruct model using the implementation of LoRA from the litGPT 0.40 library.[43] We employed the default learning parameters provided in litgpt, which included the cross-entropy loss function and the AdamW optimizer instantiated with a a learning rate equal to 1e-3. The model was trained using bfloat16 precision, with the low-rank adaptation (LoRA) matrix rank set to 32. Training was conducted across 4 A100 GPUs for 20 epochs, with a batch size of 4. We retained the model checkpoint that achieved the highest performance on our development set as the final trained model.

Larger language model. It has been demonstrated that increasing the size of generative models not only improve their performance on known tasks but also unlocks new capabilities exclusive to the largest models.[40] For instance chain of thoughts, sufficiently large models can mimic the logical steps humans follow when solving problems, and by learning to explain their reasoning, they improve their performance. Considering the potential benefits of larger foundational models, we also evaluated the Llama 3.1 405B-instruct quantized (int4) model, which was released shortly before our submission deadline.

Evaluation. We evaluated the performance of the Llama 3 70B-instruct model using our initial prompt on the development set, then assessed its performance as we sequentially added each component designed to enhance the prompt —namely, guidelines, a few examples, explanations for the labels, and finally, fine-tuning on our training corpus. We conducted all experiments with a temperature of 0.001, top-p of 0.01, and top-k of 1, to ensure deterministic responses by consistently selecting the most likely token when generating its answers. Due to time constraints and the slow processing speed of the Llama 3.1 405B-instruct model, approximately 30 minutes to classify a single note, we were unable to rerun all experiments to find the best prompt settings for this model. Instead, we evaluated this model with the best-performing settings from the Llama 3 70B-instruct model on the development set. We conducted all experiments with the default temperature of the Llama 3.1 405B-instruct model set to 0.6, top-p to 0.9, and top-k to 50, allowing for more variety in its responses. Because of time limitations, we did not run additional experiments with the temperature settings adjusted to ensure deterministic responses. After identifying the best model and settings, we performed a final evaluation on the test set. Our test set consists of only 50 notes, making it relatively small. Since our results indicate that the best-performing model used few-shot prompting and was not fine-tuned on our training examples, those examples remained unused. To assess how well our system scales, we reassigned the training examples and evaluated the model on the training set. We define the evaluation set as the combined set of all examples from both the training and test sets. Since there were very few notes labeled as insufficient in our gold standard, most errors involved the model confusing notes with appropriate (guideline-

concordant) prescriptions with those that were not appropriate (not guideline-concordant) and vice versa. Therefore, we chose to report all results by only providing the percentage of notes in each class that were correctly labeled by the generative model and did not report the more standard F1-scores.

3. Results and Discussion

We present our results in Table 2, highlighting best performance on the test set. The system correctly identified 10 out of 14 (71.4%) notes labeled as not appropriate and 32 out of 35 (91.4%) notes labeled as appropriate (line 9). This performance was achieved by providing the model with instructions and logical guidelines to perform the task, without training on the training set. The classifier demonstrated good correctness on a complex task typically performed by trained physicians when only given a few examples and clear explanations indicating whether the rules of the guidelines were satisfied or not. We also evaluated the model's performance on the entire training set to provide a more comprehensive assessment (line 9). On this larger dataset, the system maintained comparable performance, with improved detection of notes labeled as appropriate, correctly identifying 112 out of 117 (95.7%), while its detection of unlabeled notes was slightly lower, identifying 45 out of 69 (65.2%).

The table shows that all modifications made to the initial prompt (line 1.) led to incremental improvements in the model's classification performance. The table offers several interesting insights. Firstly, it is surprising that truncating the text of the example notes did not lead to a performance drop (line 4. *vs.* line 5.). This suggests that most of the text in a note is not utilized by the model for understanding the examples and can be omitted without losing essential information. Secondly, it is worth noting that LoRA, the parameter-efficient technique we employed using our training set, did not enhance the model's performance (line 6. vs. line 7.).This unexpected result requires additional experiments for further explanation. Thirdly, our findings align with recent trends in the NLP community, which indicate that generative models based on larger language models perform better than their smaller counterparts. This is evident in Table 2, where the Llama 3.1 405B-instruct model, at the time of writing, the largest model freely available to the community, outperformed the Llama 3 70B-instruct model. Lastly, since the model was not trained on our training set, it was not biased toward recognizing the style of certain providers over others. It demonstrated robustness to variations in providers' styles and achieved comparable performance on the test set as it did on the development set.

Several prior studies have used NLP and/or LLMs in infectious diseases to aid in the diagnosis and treatment of infections, such as through the review of radiology reports or in infection surveillance.[44,45] To our knowledge, however, we present the first use of LLMs in the assessment of antibiotic prescribing appropriateness using clinician notes. While these methods require further refinement and validation in larger cohorts, use of LLMs can complement previously-developed EHR-based stewardship metrics that use structured data elements, and thus improve the ability to assess prescribing practices.[3]

The methods presented here have the potential for broad application. Sinusitis is one of the most common infectious diagnoses in the outpatient setting for both adults and children, with

Table 2: Classification results on the evaluation sets for various prompts and large language models. For each line, the top percentage represents the proportion of notes labeled as appropriate correctly retrieved by the model, while the bottom percentage indicates the proportion of notes labeled as not appropriate retrieved. We omitted the percentages of notes labeled as insufficient because none of the models were able to retrieve any in this category.

	Development	**Test**	**Train**
Llama 3 70B-instruct model			
1. Role + Instructions	0.0 (0/32) 1.0 (16/16)	—	—
2. Role + Instructions + Guidelines	0.0 (0/32) 1.0 (16/16)	—	—
3. Role + Instructions + Guidelines + 3 Examples (Full text) w/o explanations	9.4 (3/32) 87.5 (14/16)	—	—
4. Role + Instructions + Guidelines + 3 Examples (Full text) & explanations	87.5 (28/32) 18.8 (3/16)	—	—
5. Role + Instructions + Guidelines + 3 Examples (Excerpt text) & explanations	53.1 (17/32) 62.5 (10/16)	—	—
6. Role + Instructions + Guidelines + 10 Examples (Excerpt text) & explanations	90.6 (29/32) 31.2 (5/16)	—	—
7. Role + Instructions + Guidelines + 10 Examples (Excerpt text) & explanations + LoRA fine-tuning	90.6 (29/32) 31.2 (5/16)	—	—
Llama 3.1 405B-instruct model			
8. Role + Instructions + Guidelines + 10 Examples (Excerpt text) & explanations	93.8 (30/32) 68.8 (11/16)	91.4 (32/35) 64.3 (9/14)	—
9. Role + Instructions + conditional Guidelines + 10 Examples (Excerpt text) & explanations	**90.6** (29/32) **93.8** (15/16)	**91.4** (32/35) **71.4** (10/14)	95.7 (112/117) 65.2 (45/69)

most encounters occurring in primary care, urgent care, and the emergency department.[3] If deployed across these settings, use of LLMs to assess prescribing appropriateness for sinusitis has the potential to impact the care of millions of people. In practice, we envision that this tool could be used in several ways. First, it could be used in provider-based feedback interventions where prescribers receive feedback on their prescribing appropriateness retrospectively at regular intervals (e.g. monthly), similar to prior work that utilized structured EHR-based metrics.[11,12,46] Additionally, this approach could be an important tool in tracking guideline concordant prescribing over time at clinic or health system level, as recommended by the CDC.[7] Finally, this also has the potential to be deployed to aid in real-time decision support during clinic visits, though modifications may need to be made given that not all notes are completed during the visit.

3.1. *Error analysis*

We analyzed the errors made by the best classifier, the Llama 3.1 405B-instruct model (line 8. in Table 2), on the examples in the test set. The model misclassified a total of 8 notes. The most frequent errors were False Positives (FPs), where the notes were labeled as inappropriate

for antibiotic prescription, but the classifier predicted them as appropriate. There were 5 such misclassified notes. Upon re-examining the notes, *LD* reviewed the explanations provided by the model and determined whether they were valid. It was found that 2 FPs occurred in notes that could have been labeled as insufficient due to ambiguous temperature documentation. For the remaining 3 FPs, *LD* confirmed the errors made by the system. One error resulted from the incorrect resolution of a deictic time reference; another from a misinterpretation of the term 'worsening' (in the phrase 'acutely worsening symptoms overnight', where 'worsening' refers to an increase in the severity of symptoms, not the progression pattern where the patient initially feels sick, then slightly better, and then worse); and the final FP was due to the system's hallucination, incorrectly stating that a temperature of 102°F is higher than 102.2°F.

The model had more success classifying the notes in which antibiotics were prescribed appropriately. There were only 3 False Negatives (FNs), as these notes clearly mentioned the onset and duration of symptoms. One FN occurred due to the under-specification of the definition of fever in criterion 3 in Table 1; unlike criterion 2, the exact temperature defining a fever is not specified. As a result, there was a disagreement between the annotator and the system regarding the resolution of this criterion in the note. The last two FNs were made on notes that were ambiguous and could have been labeled insufficient.

Finally, we analyzed the errors made by the best-performing classifier (line 9.) on the 15 notes labeled as *insufficient* in the evaluation set. All misclassifications involved ambiguous symptom duration phrases, such as "*congestion for over a week*", "*cough 7-10 days*", or "*nasal discharge about 1.5 weeks*". In 7 instances, the model correctly identified the temporal expressions that were vague but failed to recognize the ambiguity and inaccurately assign either a shorter or longer duration. In 8 other cases, the model explicitly flagged the expressions as ambiguous but it still opted for an incorrect duration inference. Given that all errors stemmed from ambiguity in symptom duration —often involving similar phrasing— we could incorporate additional examples into the prompt to help the model better recognize those phrases and correctly class insufficient documentation.

3.2. *Limitations and future work*

The largest model, Llama 3.1 405B-instruct, demonstrated good performance on our task. It was able to follow the logic of our guidelines and provide reasonable explanations for its decisions without explicit training. Although the task is challenging, it only requires the system to identify four common symptoms, assess their severity, and understand their duration or progression patterns. As evidenced by our performance with general generative models, the necessary knowledge to perform the task was available in their training data from the internet. However, most clinical NLP tasks will require specialized knowledge available only in clinical notes and ontologies. Researchers will need to continue pretraining or fine-tuning these models to integrate this domain-specific knowledge. As the size of generative models continues to grow, these training tasks become increasingly challenging for standard institutions such as hospitals and universities, which may lack the necessary hardware for the required computations.[47]

Note that our evaluation has several limitations. First, all notes were sampled from a single clinical institution. We are currently annotating 281 notes from primary care visits

for adult sinusitis at one of the University of Pennsylvania Health System's practices. To assess the robustness of our system in a different clinical setting, we plan to apply it to these newly annotated notes. Additionally, future evaluations should test the accuracy of our methods in other clinical environments, such as urgent care and emergency departments, and across institutions that use different EHR systems. Second, our cohort was identified using ICD-10 codes, which have suboptimal sensitivity and specificity for infectious diagnoses.[48] Moreover, we only included visits where an antibiotic was prescribed. It is possible that some visits for sinusitis, where an antibiotic was justified but not prescribed (guideline discordant), were missed. However, given that the majority of patients with a sinusitis diagnosis receive antibiotics, this scenario is likely infrequent.[46]

Ambiguous and vague documentation in the notes continues to pose a challenge for our best model, as none of the insufficiently documented notes were correctly classified. With larger language models now supporting input prompts of up to 16,000 tokens, we plan to include more examples of vague and ambiguous notes, along with explanations, to help the model recognize and classify these cases appropriately. Despite forcing the models to justify their decisions and anchor their answers within the input texts, we still found instances of hallucination. Integrating 'debates' among several generative LLM-based models has been proposed as an effective solution to detect and reduce hallucinations.[35,49] Our approach could easily be extended from a single generative model performing classification to a deliberative panel finding consensus for each debated note. We leave the deployment and evaluation of this approach to future work.

4. Conclusion

To address the challenge of over-prescribing antibiotics for sinusitis in children, this study proposes using natural language processing to automate the assessment of prescription appropriateness, overcoming the limitations of time-consuming manual chart reviews. We developed, trained, and evaluated generative models to classify the appropriateness of antibiotic prescriptions in 300 clinical notes from pediatric patients with sinusitis at the Children's Hospital of Philadelphia primary care network. Although Parameter-Efficient Fine-Tuning did not improve performance, the combination of few-shot learning and chain-of-thought prompting proved beneficial. Our best results were achieved using the largest generative model available at the time, the Llama 3.1 405B-instruct. On our evaluation set, the model correctly identified 144 (94.7%) of the 152 notes where the antibiotic prescription was appropriate and 55 (66.2%) of the 83 notes where it was not. Without training, our generative model demonstrated good performance in this complex task, suggesting it could be effectively deployed within the EHR to assist physicians in real-time to prevent over-prescribing as well as in monitoring antibiotic prescribing on a large scale. The clinical notes annotated for this study are Protected Health Information and not publicly available at this point. We have shared the code for access at `https://bitbucket.org/hlpgonzalezlab/naps/`.

Acknowledgements

The research reported in this publication was supported by a Centers for Disease Control and Prevention (CDC) Cooperative Agreement Funding Opportunity Announcement (FOA) U54 CK000610, Epicenters for the Prevention of Healthcare-Associated Infections.

References

1. T. F. Barlam, S. E. Cosgrove, L. M. Abbo *et al.*, Implementing an Antibiotic Stewardship Program: Guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America, *Clinical Infectious Diseases* **62**, e51 (04 2016).
2. L. McDonnell, A. Gilkes, M. Ashworth, V. Rowland, T. H. Harries, D. Armstrong and P. White, Association between antibiotics and gut microbiome dysbiosis in children: systematic review and meta-analysis, *Gut Microbes* **13**, p. 1870402 (2021).
3. K. E. Fleming-Dutra, A. L. Hersh, D. J. Shapiro *et al.*, Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011, *JAMA* **315**, 1864 (05 2016).
4. K. J. Suda, L. A. Hicks, R. M. Roberts, R. J. Hunkler, L. M. Matusiak and G. T. Schumock, Antibiotic Expenditures by Medication, Class, and Healthcare Setting in the United States, 2010–2015, *Clinical Infectious Diseases* **66**, 185 (08 2017).
5. A. W. Chow, M. S. Benninger, I. Brook, J. L. Brozek, E. J. C. Goldstein, L. A. Hicks, G. A. Pankey, M. Seleznick, G. Volturo, E. R. Wald and J. File, Thomas M., IDSA Clinical Practice Guideline for Acute Bacterial Rhinosinusitis in Children and Adults, *Clinical Infectious Diseases* **54**, e72 (04 2012).
6. E. R. Wald, K. E. Applegate, C. Bordley, D. H. Darrow, M. P. Glode, S. M. Marcy, C. E. Nelson, R. M. Rosenfeld, N. Shaikh, M. J. Smith, P. V. Williams and S. T. Weinberg, Clinical Practice Guideline for the Diagnosis and Management of Acute Bacterial Sinusitis in Children Aged 1 to 18 Years, *Pediatrics* **132**, e262 (07 2013).
7. Core elements of outpatient antibiotic stewardship `https://www.cdc.gov/antibiotic-use/hcp/core-elements/outpatient-antibiotic-stewardship.html`, Accessed July 30, 2024.
8. Measurement and evaluation approaches to improve outpatient antibiotic prescribing in health systems `https://www.cdc.gov/antibiotic-use/pdfs/Measurement-Evaluation-Improve-Outpatient-508.pdf`, Accessed July 30, 2024.
9. K. O. Degnan, V. Cluzet, M. Z. David, L. Dutcher, L. Cressman, E. Lautenbach and K. W. Hamilton, Development and validation of antibiotic stewardship metrics for outpatient respiratory tract diagnoses and association of provider characteristics with inappropriate prescribing, *Infection Control 38; Hospital Epidemiology* **43**, p. 56–63 (2022).
10. K.-P. Chua, M. A. Fischer and J. A. Linder, Appropriateness of outpatient antibiotic prescribing among privately insured US patients: ICD-10-CM based cross sectional study, **364**, p. k5092.
11. D. Meeker, J. A. Linder, C. R. Fox, M. W. Friedberg, S. D. Persell, N. J. Goldstein, T. K. Knight, J. W. Hay and J. N. Doctor, Effect of Behavioral Interventions on Inappropriate Antibiotic Prescribing Among Primary Care Practices: A Randomized Clinical Trial, *JAMA* **315**, 562 (02 2016).
12. J. S. Gerber, P. A. Prasad, A. G. Fiks, A. R. Localio, R. W. Grundmeier, L. M. Bell, R. C. Wasserman, R. Keren and T. E. Zaoutis, Effect of an Outpatient Antimicrobial Stewardship Intervention on Broad-Spectrum Antibiotic Prescribing by Primary Care Pediatricians: A Randomized Trial, *JAMA* **309**, 2345 (2013).
13. L. Dutcher, K. Degnan, A. B. Adu-Gyamfi, E. Lautenbach, L. Cressman, M. Z. David, V. Cluzet, J. E. Szymczak, D. A. Pegues, W. Bilker, P. Tolomeo, f. t. C. f. D. C. Hamilton, Keith W and P. C. P. E. Program, Improving Outpatient Antibiotic Prescribing for Respiratory Tract Infections in

Primary Care: A Stepped-Wedge Cluster Randomized Trial, *Clinical Infectious Diseases* **74**, 947 (07 2021).
14. K. N. Truitt, T. Brown, J. Y. Lee and J. A. Linder, Appropriateness of antibiotic prescribing for acute sinusitis in primary care: A cross-sectional study, **72**, 311.
15. D. J. Livorsi, C. M. Linn, B. Alexander, B. H. Heintz, T. A. Tubbs and E. N. Perencevich, The value of electronically extracted data for auditing outpatient antimicrobial prescribing, **39**, 64.
16. C. Feudtner, J. A. Feinstein, W. Zhong, M. Hall and D. Dai, Pediatric complex chronic conditions classification system version 2: updated for icd-10 and complex medical technology dependence and transplantation, *BMC Pediatrics* **14**, p. 199 (2014).
17. A. Chow, M. Benninger, I. Brook *et al.*, Idsa clinical practice guideline for acute bacterial rhinosinusitis in children and adults, *Clin Infect Dis* **54**, e72 (2012).
18. E. R. Wald, K. E. Applegate, C. Bordley *et al.*, Clinical Practice Guideline for the Diagnosis and Management of Acute Bacterial Sinusitis in Children Aged 1 to 18 Years, *Pediatrics* **132**, e262 (07 2013).
19. H. Cunningham, V. Tablan, A. Roberts and K. Bontcheva, Getting more out of biomedical documents with gate's full lifecycle open source text analytics, *PLOS Computational Biology* **9**, 1 (02 2013).
20. C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky, The Stanford CoreNLP natural language processing toolkit, in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014.
21. D. Ferrucci, A. Lally, K. Verspoor and E. Nyberg, Unstructured information management architecture (UIMA) version 1.0 OASIS Standard (mar, 2009).
22. D. Weissenbacher, Influence des annotations imparfaites sur les systèmes de traitement automatique des langues, un cadre applicatif: la résolution de l'anaphore pronominale, PhD thesis, Université Paris-Nord - Paris XIII2008.
23. T. Poibeau, *Extraction Automatique D'information* (Hermes, 2003).
24. A. Magge, D. Weissenbacher, A. Sarker, M. Scotch and G. Gonzalez-Hernandez, Deep neural networks and distant supervision for geographic location mention extraction, *Bioinformatics* **34**, i565 (2018).
25. W. H. Clark and A. J. Michaels, Training from zero: Forecasting of radio frequency machine learning data quantity, *Telecom* **5**, 632 (2024).
26. D. Weissenbacher and Y. Sasaki, Which factors contributes to resolving coreference chains with bayesian networks?, in *14th International Conference on Intelligent Text Processing and Computational Linguistics*, 2013.
27. A. Thampi, *Interpretable AI: : Building explainable machine learning systems* (Manning, 2022).
28. J. Vig, A multiscale visualization of attention in the transformer model, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Association for Computational Linguistics, July 2019).
29. A. Rogers, O. Kovaleva and A. Rumshisky, A Primer in BERTology: What We Know About How BERT Works, *Transactions of the Association for Computational Linguistics* **8**, 842 (2021).
30. L. Ouyang, J. Wu, X. Jiang *et al.*, Training language models to follow instructions with human feedback, in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Curran Associates, Inc., 2022).
31. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* **21** (jan 2020).
32. R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting and N. Liu, Large language models in health care: Development, applications, and challenges, *Health Care Science* **2**, 255 (2023).

33. X. Yang, A. Chen, N. PourNejatian *et al.*, A large language model for electronic health records, *npj Digit. Med* **5** (2022).
34. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel and D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20 (Curran Associates Inc., Red Hook, NY, USA, 2020).
35. H. Wang, X. Du, W. Yu, Q. Chen, K. Zhu, Z. Chu, L. Yan and Y. Guan, Apollo's oracle: Retrieval-augmented reasoning in multi-agent debates (2023).
36. T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest and X. Zhang, Large language model based multi-agents: A survey of progress and challenges, in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24* , ed. K. Larson (International Joint Conferences on Artificial Intelligence Organization, 8 2024). Survey Track.
37. Llama Team, AI @ Meta, The llama 3 herd of models `https://ai.meta.com/research/publications/the-llama-3-herd-of-models/`, Access July 28, 2024.
38. Y. Yu, C.-H. H. Yang, J. Kolehmainen, P. G. Shivakumar, Y. Gu, S. Ryu, R. Ren, Q. Luo, A. Gourav, I.-F. Chen, Y.-C. Liu, T. Dinh, A. Gandhe, D. Filimonov, S. Ghosh, A. Stolcke, A. Rastrow and I. Bulyko, Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition, *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* , 1 (2023).
39. B. Jimenez Gutierrez, N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun and Y. Su, Thinking about gpt-3 in-context learning for biomedical ie? think again, in *Findings of the Association for Computational Linguistics: EMNLP 2022*, (Association for Computational Linguistics, 2022).
40. M. Moor, O. Banerjee, Z. Abad, H. Krumholz, J. Leskovec, E. Topol and P. Rajpurkar, Foundation models for generalist medical artificial intelligence, *Nature* **616** (2023).
41. D. Weissenbacher, X. Zhao, J. R. C. Priestley, K. M. Szigety, S. F. Schmidt, K. O'Connor, I. M. Campbell and G. Gonzalez-Hernandez, Biocreative viii – task 3: Genetic phenotype normalization from dysmorphology physical examinations (2023).
42. J. Wei, X. Wang, D. Schuurmans *et al.*, Chain-of-thought prompting elicits reasoning in large language models (2023).
43. Lightning AI, Litgpt `https://github.com/Lightning-AI/litgpt`, Accessed October 1, 2024.
44. G. Jones, J. Amoah, E. Y. Klein, H. Leeman, A. Smith, S. Levin, A. M. Milstone, K. Dzintars, S. E. Cosgrove and V. Fabre, Development of an Electronic Algorithm to Identify in Real Time Adults Hospitalized With Suspected Community-Acquired Pneumonia, *Open Forum Infectious Diseases* **8**, p. ofab291 (2021).
45. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes, *Annals of Internal Medicine* **156**, 11 (2012), PMID: 22213490.
46. A. A. Vazquez Deida, D. J. Bizune, C. Kim, J. M. Sahrmann, G. V. Sanchez, A. L. Hersh, A. M. Butler, L. A. Hicks and S. Kabbani, Opportunities to Improve Antibiotic Prescribing for Adults With Acute Sinusitis, United States, 2016–2020, *Open Forum Infectious Diseases* **11**, p. ofae420 (2024).
47. C. Peng, X. Yang, A. Chen *et al.*, A study of generative large language model for medical research and healthcare, *npj Digital Medicine* **6** (2023).
48. D. J. Livorsi, C. M. Linn, B. Alexander, B. H. Heintz, T. A. Tubbs and E. N. Perencevich, The value of electronically extracted data for auditing outpatient antimicrobial prescribing, *Infection Control 38; Hospital Epidemiology* **39**, p. 64–70 (2018).
49. T. Xiangru, Z. Anni, Z. Zhuosheng, L. Ziming, Z. Yilun, Z. Xingyao, C. Arman and G. Mark, Medagents: Large language models as collaborators for zero-shot medical reasoning (2024).

Implications of An Evolving Regulatory Landscape on the Development of AI and ML in Medicine

Nicole Rincon
Penn State Law, University Park, PA 16802, USA
Email: nar5442@psu.edu

Sara Gerke
University of Illinois Urbana-Champaign College of Law, Champaign, IL 61820, USA
Email: gerke@illinois.edu

Jennifer K. Wagner
School of Engineering Design and Innovation; Department of Anthropology; Department of Biomedical Engineering; Institute for Computational and Data Science; Huck Institutes of the Life Sciences; Rock Ethics Institute, Pennsylvania State University, University Park, PA 16802 USA; and Penn State Law, University Park, PA 16802 USA
Email: jkw131@psu.edu

The rapid advancement of artificial intelligence and machine learning (AI/ML) technologies in healthcare presents significant opportunities for enhancing patient care through innovative diagnostic tools, monitoring systems, and personalized treatment plans. However, these innovative advancements might result in regulatory challenges given recent Supreme Court decisions that impact the authority of regulatory agencies like the Food and Drug Administration (FDA). This paper explores the implications of regulatory uncertainty for the healthcare industry related to balancing innovation in biotechnology and biocomputing with ensuring regulatory uniformity and patient safety. We examine key Supreme Court cases, including *Loper Bright Enterprises v. Raimondo*, *Relentless, Inc. v. Department of Commerce*, and *Corner Post, Inc. v. Board of Governors of the Federal Reserve System*, and their impact on the *Chevron* doctrine. We also discuss other relevant cases to highlight shifts in judicial approaches to agency deference and regulatory authority that might affect how science is handled in regulatory spaces, including how biocomputing and other health sciences are governed, how scientific facts are applied in policymaking, and how scientific expertise guides decision making. Through a detailed analysis, we assess the potential impact of regulatory uncertainty in healthcare. Additionally, we provide recommendations for the medical community on navigating these challenges.

Keywords: Artificial Intelligence, Medicine, Regulation, Science Policy, SCOTUS, ELSI

1. Introduction

The use of artificial technologies (AI) and machine learning (ML) in healthcare has been continuously expanding,[1] paving the way for more personalized, preventative, and innovative treatments that improve patient outcomes in the future.[2] Its implementation will likely be widespread in various areas of medicine and can be especially beneficial in prevention, therapeutics, and diagnostics.[2] The relevance of AI/ML advancements in medicine is evident, as seen in the rise of new innovative treatments and diagnostics, such as the use of AI algorithms for diabetic retinopathy screening,[3] smart sensors that assist with more accurately estimating the probability of heart attacks,[4] and imagining systems that use algorithms for diagnostics information for skin cancer in patients.[5] In addition, the development of Generative AI (GenAI) tools, which leverage ML, has been skyrocketing since the launch of OpenAI's ChatGPT in 2022. GenAI can create images, videos, and text, and it is expected to revolutionize healthcare, including providing a more patient-tailored approach.[6] While exciting, these new developments come alongside various ethical and legal issues, many of which, due to their novelty, have yet to be addressed.[7] This is especially true regarding the regulatory paths for these new developing technologies.[8-9]

Upon the emergence of new AI/ML tools in healthcare, there have been clear efforts by the U.S. Food and Drug Administration (FDA), the regulatory agency for various medical products, including medical devices, to create new pathways and expectations for how to bring such tools to market.[10-13] Recent Supreme Court decisions, such as *Loper Bright Enterprises v. Raimondo*[14] and *Relentless, Inc. v. Department of Commerce*,[15] have further complicated the process of determining more concrete paths and expectations for these newly developed technologies. These rulings led to the overruling of the *Chevron* doctrine, which has, in turn, diminished agency deference and set the stage for more litigation and more stringent criteria for regulatory actions. These decisions might have significant implications for the FDA and its current approach to AI/ML technologies in healthcare. Awareness of recent case law can allow those in the medical community to better anticipate expectations surrounding the development and implementation of new technologies.

This Article first discusses the current regulatory framework for medical AI/ML and some of the FDA's recent initiatives. It then discusses the *Chevron* doctrine and relevant Supreme Court decisions before the doctrine's overruling. Next, it analyzes the two Supreme Court cases that led to *Chevron*'s fall, *Loper Bright Enterprises v. Raimondo* and *Relentless, Inc. v. Department of Commerce*. It also discusses a subsequent and relevant Supreme Court case, namely *Corner Post*.[16] Finally, this Article analyzes how recent judicial decisions might affect regulatory practices and explores strategies for those looking to develop AI/ML in healthcare. In particular, we show that these uncertainties are likely to result in an increase in litigation, along with a need for more explicit rulemaking to constrain interpretations. However, given the speed at which medical AI/ML has advanced, it is important to keep up to date with recent case law and regulatory developments in the field.

2. The FDA and the Regulation of Medical AI/ML

The FDA protects public health by assuring the safety, efficacy, and security of various medical products, including biological products and medical devices.[17] It draws its authority from various statutes, primarily the Federal Food, Drug, and Cosmetic Act (FDCA).[18] Its authority permits the

FDA to create legally binding regulations in cases where Congress delegated such power to the agency. In addition, the FDA also publishes guidance documents that reflect the FDA's thinking on specific topics and developments. While acting through guidance documents (rather than regulations) seemed to be the FDA's preferred means in recent years, these are not legally binding. Still, guidance documents play an essential role in practice because they reflect the agency's current interpretation of the respective topics.

The FDCA covers various products, such as food, drugs, medical devices, and cosmetics. The law, initially enacted in 1938, has been amended over 100 times to address new advancements and emerging challenges. Amendments relevant to the medical field specifically, are the Kefauver-Harris Drug Amendments of 1962[19](which required drug manufacturers to provide proof of efficacy and safety before a drug could be marketed and approved by the FDA) and the Medical Device Amendment of 1976[20] (which established a risk-based classification system for medical devices and expanded the regulatory authority of the FDA over medical devices). Ultimately, the FDCA regulates three main pathways for medical devices to obtain marketing authorization: premarket notification or 510(k),[21] premarket approval or PMA,[22] and De Novo Classification request.[23] The FDCA does not, however, allow the FDA to develop new pathways for medical devices unilaterally. This limitation on authority can present challenges as new technologies like AI/ML develop that do not fall under classic categories and might require more innovative regulatory pathways.

Additionally, the FDA operates within the Administrative Procedure Act (APA) framework, which essentially governs the processes by which federal agencies may develop and issue binding regulations.[24] The purpose of the APA, enacted in 1946, was to promote transparency, public participation, and accountability in the regulatory process. Under the APA, the FDA must follow specific procedures when creating its regulations, such as publishing notices of proposed rulemaking,[25] allowing for public comment,[26] and providing detailed explanations of the final rules. The procedures and limitations set forth by APA are to ensure that the FDA's regulatory acts are both well-informed and open to comments. In contrast, unlike formal actions, the FDA can publish guidance documents much easier and faster because they are legally *not* binding in nature but still have a considerable impact in practice.

The FDA also regulates medical AI/ML-enabled products so long as they are classified as medical devices under FDCA Section 201(h)(1).[27] Recently, the FDA has attempted to advance the regulation of medical Al/ML due to its potential for transforming healthcare delivery and improving patient outcomes. The FDA has been proactively attempting to address the regulatory challenges posed by new AI/ML-enabled products through various guidance documents, publications, regulations, plans, and programs. For example, already back in 2019, the FDA published a discussion paper suggesting a new framework for changes to AI/ML-based Software as a Medical Device (SAMD).[10] This was followed by an Action Plan on AI/ML-based SaMD issued in 2021.[11] Just in March of 2024, the FDA also released a paper as a complement to its Action Plan outlining its intent to incorporate collaboration between its different departments to better protect public health while still encouraging innovation through AI/ML.[12,13] While unable to issue new authorization pathways on its own, the FDA may still determine how best to combine the pathways it is authorized to use to fit the needs of new and developing products.

Moreover, there have been more legislative actions attempting to proactively address the potential challenges raised by new medical technology such as AI/ML. For instance, the 21st Century Cures Act,[28] signed into law on December 13, 2016, encouraged the innovation and development of medical devices that could be more effective in treating or diagnosing. However, the 21st Century Cures Act also explicitly excluded several categories of software functions from the FDCA definition of a "device," which also includes certain clinical decision support software functions.[29] The FDA has heavily interpreted the meaning of such statute through guidance documents to state its current thinking on that topic.[30-32]

3. *Chevron* and Relevant Supreme Court Decisions Before Its Overruling

3.1 Chevron U.S.A, Inc. v. Natural Resources Defense Council, Inc. (1984)

In 1984 the Supreme Court decided *Chevron U.S.A., Inc. v. Natural Resources Defense Council, Inc.*,[33] establishing a key principle in administrative law known as the *Chevron* doctrine. The case involved a challenge to an Environmental Protection Agency (EPA) regulation, with litigants claiming the EPA regulation was inconsistent with the Clean Air Act. The Court upheld the EPA's interpretation, establishing judicial deference to agency interpretations of ambiguous statutes. The *Chevron* doctrine involved a two-step process to resolve a challenge to an agency's actions. First, the court would determine whether Congress had already directly addressed the matter in question. In other words, the court would decide whether the statute contains ambiguity or is silent on the matter, thereby necessitating any interpretation. If Congressional intent was clear and unambiguous, that intent was to be followed without deviation by the agency. If, however, silence or ambiguity necessitated interpretation of the statute on the matter in question, the court's second step would be to decide whether the agency's interpretation was based on a permissible construction of the statute. If the interpretation of delegated authority was reasonable, the agency's expertise was entitled to deference by the court.[33 at 843-844] Essentially, the *Chevron* doctrine was the primary means by which courts would determine whether an administrative agency was acting on or exceeding its congressionally delegated authority.

What became known as "*Chevron* deference" acknowledged that federal agencies possess relevant expertise (such as scientific or technical expertise) critical for proper interpretation and implementation of statutes that the agency is charged by Congress with administering and that courts often lack this expertise.[33 at 865] The *Chevron* doctrine reflected and respected a delicate balance of power between the branches of government: its two-step process ensured that Congress established the initial statutory framework for the federal policy, that administrative agencies implemented the statutory framework using the legislative instructions when available and filling in gaps only when it is permissible to do so; and that judicial oversight would ensure such agency interpretations were necessary and reasonable. The *Chevron* doctrine was heavily relied upon for 40 years, although pushback and critiques (particularly from those seeking to limit the power of federal agencies) caused judicial application of *Chevron* deference to evolve over time.

3.2 West Virginia v. Environmental Protection Agency (2022)

In 2022 the Supreme Court decided *West Virginia v. EPA*.[34] West Virginia, backed by several other states and industry groups, challenged the EPA's authority to regulate greenhouse gas emissions via the Clean Power Plan, arguing the EPA overstepped its authority under the Clean Air Act.[34 at 715]

The Supreme Court agreed and held the agency could not make such significant changes to the nation's policies unilaterally.[34 at 733] This ruling foreshadowed more recent decisions further limiting agency power.[35] The court sidestepped *Chevron* precedent in this case. Instead, the court diverted attention to a new doctrine it introduced for the first time: the "Major Questions Doctrine." The Major Questions Doctrine enabled the court to assume broad powers to invalidate agency actions, positing that agencies cannot act on matters of "economic and political significance" unless there is explicit authority from Congress for the agency to take actions of such importance. Considering the FDA's scope of authority for regulating health technologies and their sizable economic consequences, this doctrine has the potential for courts to ultimately require more explicit congressional authorization for *any* regulatory actions of importance.

3.3. Securities and Exchange Commission v. Jarkesy (2024)

In June 2024 the Supreme Court decided *Securities and Exchange Commission (SEC) v. Jarkesy*,[36] making it more difficult for federal agencies to impose monetary penalties using administrative processes overseen by administrative law judges (ALJs). The SEC had brought an enforcement action against George Jarkesy for allegedly misleading investors regarding the hedge funds he managed, and the SEC ALJ found Jarkesy liable and imposed sanctions. Jarkesy successfully challenged the decision before the U.S. Court of Appeals for the Fifth Circuit, arguing that the agency's use of an administrative process (i.e., the ALJ adjudication) violated his constitutional right to a jury trial under the Seventh Amendment.[36 at 660,678] The Supreme Court agreed. In reaching its decision, the court examined the "Public Rights Doctrine,"[36 at 660] a doctrine that acknowledges Congress may delegate adjudicative authority to a federal agency in some situations without infringing upon the Seventh Amendment right to a jury trial. Here, under the facts of this case, the court determined that actions regarding fraud and civil penalties are traditionally handled by courts and cannot be assigned by Congress to an agency.

The case has ramifications beyond the SEC. All agencies, including the FDA, might need to reevaluate the types of cases enforced under its administrative adjudication process. Agencies enforcing health fraud and abuse laws could be on particularly shaky ground now.[37] The use of ALJs, subject to more stringent constitutional constraints under the Roberts Court, might make agencies more hesitant to rely upon them. The use of ALJs has been an essential component of federal agencies' regulatory and enforcement powers. These administrative proceedings, initially created to streamline the regulatory process, might now require more detailed consideration to ensure compliance and avoid litigation.

The FDA's administrative proceedings, such as hearings and product seizures for noncompliance,[38] could be scrutinized as overstepping the constraints set by *Jarkesy* and the Supreme Court's interpretation of the Public Rights Doctrine. The FDA's processes for enforcing compliance with its regulations, including issuing fines or sanctions, might be challenged in a similar fashion to that of *Jarkesy*. Admittedly, though, the FDA's enforcement actions have predominantly included informal ones such as warning letters. Thus, *Jarkesy* might even prompt the FDA to continue its path to use those.[39]

4. *Chevron*'s Fall and the *Corner Post* Case

4.1 Loper Bright Enterprises v. Raimondo and *Relentless, Inc. v. Department of Commerce (2024)*

The groundbreaking 5-4 Supreme Court decision on June 28th, 2024, in *Loper Bright Enterprises v Raimondo* and *Relentless v. Dep't of Commerce* (two companion cases collectively referred to as "Loper Bright"[14]) expressly overruled *Chevron*, after 40 years of it functioning—as Justice Kagan described in the dissenting opinion joined by Justices Sotomayor and Jackson—"as a cornerstone of administrative law."[14 (dissent at 112)] The case involved challenges to the National Marine Fisheries Service's interpretation of statutory language for its fishing regulations because the agency's interpretation required clear congressional authorization, which they did not have. The Court, calling *Chevron* deference "fundamentally misguided"[14 at 14] and describing it as incompatible with the APA, held specifically that courts "under the APA may not defer to agency interpretation of the law simply because a statute is ambiguous."[14 at 62]

The Majority explained their decision by asserting not only that "...agencies have no special competence in resolving statutory ambiguities" but also that "[c]ourts do."[14 at 10] Rather than viewing subject matter expertise (including scientific and technical expertise) relevant to an agency's scope of authority as reason to defer to the agency interpretation so long as it is reasonable, the Court took a dismissive and even hostile[41] view to agencies, explaining that such deference would be an "abdication" of the court's responsibilities to use "the traditional tools of statutory construction"[14 at 46] to "determine the best reading of the statute and resolve the ambiguity."[14 at 44-45]

The Court noted that the information provided by litigating parties and others through amici briefs would offer sufficient perspective even on technical details and that a court could consider agency interpretations (e.g., as one persuasive but not conclusive interpretation [See 41,42]). The Court also claimed that technical expertise could still be considered under *Skidmore* deference.[14 at 11] However, regarding deference under *Skidmore,* a court is not required to follow an agency's interpretation of a statute but rather has the choice of determining the amount of deference, if any, to give an agency when considering "the thoroughness evident in its [the agency's] consideration, the validity of its reasoning, its consistency with earlier and later procurements, and all of those factors which give it power to persuade."[42 at 11] While the Court in *Loper Bright* emphasized that *Skidmore* deference still permits courts to consider agency interpretations, it does so only if those interpretations are persuasive—a standard that is not governed by a strict rule but left to the courts' discretion on a case-by-case basis. This level of deference provides significantly less binding authority for federal agencies, shifting more interpretative power to the courts while also limiting agencies' ability to make decisions with the same level of control that they had previously under *Chevron*.

As noted in the dissent, *Chevron* had "formed the backdrop against which Congress, courts, and agencies—as well as regulated parties and the public—all have operated for decades. It has been applied in thousands of judicial decisions. It has become part of the warp and woof of modern government, supporting regulatory efforts of all kinds—to name a few, keeping air and water clean, food and drugs safe, and financial markets honest."[14 (dissent at 113)] Its overturning, therefore, could result in regulatory uncertainty, and federal agencies will need to relocate the boundaries of agency

authority (even boundaries that previously seemed clear and established). Additionally, the decision to overturn *Chevron* is likely to slow the efficiency of courts and agencies due to an anticipated flood of litigation.[40]

4.2 Corner Post, Inc. v. Board of Governors of the Federal Reserve System

Just days after overturning *Chevron* in *Loper Bright*, the Supreme Court decided *Corner Post, Inc. v. Board of Governors of the Federal Reserve System*[16] and further molded the regulatory landscape by extending the period during which agency regulations can be challenged. The case involved a challenge from a truck stop business to certain regulations imposed by the Federal Reserve Board regarding limits on fees to merchants for debit card purchases. The regulations were promulgated in 2011, and the truck stop (which was established in 2018) brought the challenge in 2021, a decade after the regulations were issued.

The lower courts dismissed the lawsuit due to the expiration of the APA's 6-year period within which suits against agencies can be filed,[43] but the Supreme Court reversed this decision. The Court's holding rests on conceptualizing the APA's period as a "statute of limitations"[44] (which would begin to run when a specific plaintiff has been injured by the agency action) rather than a "statute of repose"[45] (which would begin to run as soon as an agency's action occurred, such as issuance of final regulations). The dissent highlighted that this understanding contradicts statutes within the administrative law context[16 (dissent at 76,78)] and warned that, under this interpretation, potential litigation against agency regulations would never end.[16 (dissent at 83)] Rather than an agency having some confidence that regulations issued a decade ago would be safe from legal challenges, this decision ultimately renders them vulnerable indefinitely. This decision opens the door for challenges to older regulations, resulting in a significant extension of the timeframe during which regulations can be contested than was previously understood.

5. Impact of the Recent Supreme Court Decisions on Medical AI/ML

Legal scholars have started unpacking the broad impacts of these recent Supreme Court decisions for healthcare generally[37,41] and medical AI/ML specifically.[35,47] Under the leadership of Chief Justice Roberts, the Supreme Court has issued several decisions that are cause for concern for federal oversight of medical AI/ML. Previous works[35] have highlighted cases (e.g., *TransUnion v. Ramirez*[48] and *Dobbs v. Jackson Women's Health Org*.[49]) that frustrate attempts by the Federal Trade Commission (FTC) and other policymakers to promote algorithmic fairness and responsible data practices in the context of digital health. This trend of the Supreme Court issuing "industry friendly" decisions[37] that make it *more* challenging for federal agencies to engage in effective, adaptive governance has continued with the cases summarized here (i.e., *Jarkesy*, *Loper Bright*, and *Corner Post*). Even prior to the 2024 cases, scholars had remarked that the Supreme Court was, in effect, making America "ungovernable"[50] by weakening the powers of every governmental branch *except its own*.[51] Some have gone so far as to dub the Roberts Court as the "anti-innovation Court,"[52] as its holdings would suppress policy innovations many think are essential to impelling responsible conduct in a rapidly changing, AI-enabled world. The increased polarization of public opinion regarding the Supreme Court is likely expected to continue as policymakers debate President Biden's proposed court reforms.[54]

The recent Supreme Court decisions have strongly signaled a deregulation era is upon us;[54] however, the impact of these decisions directly on the FDA's oversight of medical AI/ML remains

to be seen. While it is unlikely that the agency will be immune from these headaches of legal uncertainty, the agency's habit of choosing to govern mainly by informal, non-binding guidance documents (as opposed to formal rulemaking) means that *Loper Bright* and *Corner Post* are not directly applicable to them. Governance by guidance, a subject garnering its own distinct criticisms,[55] means in practical terms that there might rarely be regulatory interpretations of statutes ripe for legal challenges against the FDA.

The FDA has so far mainly published discussion papers and non-binding guidance documents related to AI (as highlighted in Section 2 above).[56] These informal actions are intended to offer agency flexibility, allowing the FDA to quickly change its current thinking on a topic to keep pace with innovation. Guidance documents can be released through a more direct and efficient process than proposed rules.[25] These AI-related documents over the last few years have certainly provided helpful information to manufacturers and other stakeholders on the new regulatory challenges raised by AI, the FDA's thinking and initiatives on this topic, and likely expectations during premarket reviews of AI/ML-based medical devices.

Even if *Loper Bright* is not directly applicable to guidance documents, the holding still has important indirect implications for the regulation of medical AI/ML. In particular, the holding might encourage the FDA to continue using primarily informal actions like guidance documents rather than creating legally binding regulations, which are more time-consuming and costly for the agency.[47] On the flip side, with the Supreme Court's holding in *Loper Bright*, this approach will likely receive even more skepticism from courts in the future and might be seen as a potential bypass of formal rulemaking.[47]

But even when creating legally binding regulations regarding AI/ML, the FDA would need to ensure that such regulations would be explicitly defined and authorized by Congress, as the agency's flexibility to interpret ambiguous statutes has been reduced.[47] When ambiguities arise, courts can now take the primary role in deciding outcomes. This could be difficult given the in-depth scientific and technological details specific to the areas of AI/ML. The courts can now rule in these highly specialized areas and attempt to piece together the relevant information to make decisions where Congress has not provided unambiguous language (which is often the case). Additionally, decisions on ambiguities by courts might carry the weight of stare decisis. In the jurisdictions bound by these holdings, agencies will not be allowed to enforce regulations that are inconsistent with the interpretation provided by those courts. Moreover, the courts' need for a comprehensive understanding in anticipation of these rulings—each likely to be in different industries with different facts—will likely slow both the courts and the implementation of new regulations while legal challenges await adjudication.

Moreover, *Corner Post* poses serious implications for the FDA, FTC, and other federal agencies, with the indefinite timeframe for challenges of agency regulations under the APA, complicating even those areas of regulatory compliance that might have been considered settled, well-established frameworks. Agencies might need to anticipate extended periods of litigation and AI/ML developers might experience ongoing regulatory uncertainty, which could impact market strategies, encourage litigious strategies, and delay innovations. Regulatory uncertainty has the potential to create spaces in which corporations might, for example, be tempted to engage in cost-cutting measures to increase profits but also result in compromised safety and efficacy.[37] These conditions also might enable

some biotech/biomedical corporations to engage in questionable data practices and increased litigious practices to dominate the market, stifle competition, limit access to alternatives, and increase prices to consumers,[37] necessitating more antitrust measures by the FTC.

Congress might also begin to face pressure to revise statutes to make them less ambiguous to ensure their intent is reflected correctly and will likely need to draft and pass more detailed legislation for future issues. This might very well be the case for legislation specific to AI/ML technologies—including those developed for medicine,[47] in light of AI in clinical applications presumptively considered "rights impacting" and "safety impacting"[57] and could result in waiting for specific and comprehensive laws governing these technologies.

With existing legal mechanisms available for use in the federal government's oversight of medical AI/ML in question and new legislation slow to pass, it is incumbent upon the AI/ML community to be more proactive in their commitments to responsible development and use of these technologies. Innovative, integrated policy research is needed, with one clear example being the opportunity for AI/ML developers to pilot the CAITE model.[58] Moreover, the integrated, holistic training of the medical AI/ML workforce is needed so that researchers and clinical practitioners can anticipate the evolving expectations and constraints in regulations across the AI/ML product (or system) lifecycle. This need for interdisciplinary training—such as that offered within Penn State's Law, Policy, and Engineering initiative; Arizona State University's School for the Future of Innovation in Society; or MIT's Institute for Data, Systems, and Society, for example—is growing more urgent with the tremendous challenges and opportunities on the horizon for digital twins, as noted recently by the National Academies of Science, Engineering, and Medicine.[59]

6. Conclusion

This paper has examined the regulatory challenges and implications of recent Supreme Court decisions on the FDA's oversight of AI/ML technologies in healthcare. The dual challenge of promoting innovation while safeguarding public health underscores the importance of a balanced and nuanced regulatory framework. Ethical practices, risk management, and proactive compliance might become essential in navigating these uncertainties and ensuring the successful integration of AI/ML technologies into clinical practice. A collaborative approach involving regulators, industry stakeholders, and the biomedical community might become necessary to develop effective strategies for balancing innovation with patient safety and public health protection.

Acknowledgments

This legal research was supported by Grant No. 1R21EB035474-01 awarded by the National Institutes of Health Office of the Director (NIH OD) and the National Institute of Biomedical Imaging and Bioengineering (NIBIB). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work is undertaken as part of the "Bioethical, Legal, and Anthropological Study of Technologies (BLAST)," for which an IRB exemption determination has been obtained from the Penn State University Institutional Review Board (STUDY00023241) on August 4, 2023. Outside the submitted work, SG also reports grants from the European Union (Grant Agreement no. 101057099 and no. 101057321). SG is also a Research Fellow at the University of Copenhagen, Faculty of Law, supported by a Novo Nordisk

Foundation Grant for a scientifically independent International Collaborative Bioscience Innovation & Law Programme (Inter-CeBIL programme - grant no. NNF23SA0087056).

References

1. World Health Organization, *WHO Outlines Considerations for Regulation of Artificial Intelligence for Health* (Oct. 19, 2023), https://www.who.int/news/item/19-10-2023-whooutlines-considerations-for-regulation-of-artificial-intelligence-for-health (Last accessed July 31, 2024).
2. Bajwa J, Munir U, Nori A, Williams B. *Artificial Intelligence in Healthcare: Transforming the Practice of Medicine. Future Healthc J.* 2021 Jul;8(2):e188-e194. doi: 10.7861/fhj.2021-0095. PMID: 34286183; PMCID: PMC8285156.
3. Lim JI, Regillo CD, Sadda SR, Ipp E, Bhaskaranand M, Ramachandra C, Solanki K. *Artificial Intelligence Detection of Diabetic Retinopathy: Subgroup Comparison of the EyeArt System with Ophthalmologists' Dilated Examinations*. Ophthalmol Sci. 2022 Sep 30;3(1):100228. doi: 10.1016/j.xops.2022.100228. PMID: 36345378; PMCID: PMC9636573.
4. Hannan, A., Cheema, S. M., & Pires, I. M. (2024). *Machine Learning-Based Smart Wearable System for Cardiac Arrest Monitoring Using Hybrid Computing. Biomedical Signal Processing and Control*, 87(Part B), 105519. https://doi.org/10.1016/j.bspc.2023.105519 (Last accessed July 31, 2024).
5. Melarkode N, Srinivasan K, Qaisar SM, Plawiak P. *AI-Powered Diagnosis of Skin Cancer: A Contemporary Review, Open Challenges and Future Research Directions. Cancers (Basel)*. 2023 Feb 13;15(4):1183. doi: 10.3390/cancers15041183. PMID: 36831525; PMCID: PMC9953963.
6. AlphaSense. (June 19,2024). *Generative AI in Healthcare: Use Cases, Benefits, and Drawbacks*. Retrieved July 27, 2024, from https://www.alpha-sense.com/blog/trends/generative-ai-healthcare/ (Last accessed July 31, 2024).
7. Gerke S, Minssen T, Cohen G. *Ethical and Legal Challenges of Artificial IntelligenceDriven Healthcare. Artificial Intelligence in Healthcare*. 2020:295–336. doi: 10.1016/B978-0-12-818438-7.00012-5. Epub 2020 Jun 26. PMCID: PMC7332220.
8. Gerke S. *Health AI for Good Rather Than Evil? The Need for a New Regulatory Framework for AI-Based Medical Devices* (2021). Yale Journal of Health Policy, Law, and Ethics, Vol. 20, Issue 2, 433 (2021), Available at SSRN: https://ssrn.com/abstract=4070947 (Last accessed July 31, 2024).
9. Gerke S., *'Nutrition Facts Labels' for Artificial Intelligence/Machine Learning-Based Medical Devices—The Urgent Need for Labeling Standards* (February 1, 2023). 91 Geo. Wash. L. Rev. 79, (2023). Available at SSRN: https://ssrn.com/abstract=4404252.
10. U.S. Food & Drug Admin., *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback* (Apr. 2, 2019), https://www.fda.gov/media/122535/download (Last accessed July 31, 2024).
11. U.S. Food & Drug Admin., *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan* (Jan. 2021), https://www.fda.gov/media/145022/download (Last accessed July 31, 2024).
12. U.S. Food & Drug Admin., *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled*

Device Functions, https://www.fda.gov/media/166704/download (Last accessed July 30, 2024).
13. U.S. Food & Drug Admin., *Artificial Intelligence & Medical Products: How CBER, CDER, CDRH, and OCP Are Working Together*, https://www.fda.gov/media/177030/download (Last accessed July 21, 2024).
14. *Loper Bright Enters. v. Raimondo*, 144 S. Ct. 2244 (2024).
15. *Relentless, Inc. v. Dep't of Commerce*, No. 21-1886 (1st Cir. 2023), cert. granted, No. 221219, 2023 WL 7030273 (U.S. Oct. 13, 2023).
16. *Corner Post, Inc. v. Bd. of Governors of the Fed. Rsrv. Sys.*, 144 S. Ct. 2440 (2024).
17. U.S. Food & Drug Admin., https://www.usa.gov/agencies/food-and-drug-administration (Last accessed July 27, 2024).
18. Federal Food, Drug, and Cosmetic Act, 21 U.S.C. §§ 301-399i.
19. Kefauver-Harris Amendments, Pub. L. No. 87-781, 76 Stat. 780 (1962).
20. Medical Device Amendments of 1976, Pub. L. No. 94-295, 90 Stat. 539.
21. 21 U.S.C. § 360(k) (2023).
22. 21 U.S.C. § 515(b)(2) (2023).
23. 21 U.S.C. § 360c(f)(2) (2023).
24. Administrative Procedure Act, 5 U.S.C. §§ 551-559 (2023).
25. 5 U.S.C. § 553 (2023).
26. 5 U.S.C. § 552 (2023).
27. 21 U.S.C. § 321(h)(1) (2023).
28. 21st Century Cures Act, Pub. L. No. 114-255, 130 Stat. 1033 (2016).
29. 21 U.S.C. § 360j(o)(1)(E) (2023).
30. U.S. Food & Drug Admin., *Changes to Existing Medical Software Policies Resulting from Section 3060 of the 21st Century Cures Act: Guidance for Industry and Food and Drug Administration Staff* (Sept. 27, 2019), https://www.fda.gov/media/109622/download.
31. U.S. Food & Drug Admin., *Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff* (Sept. 28, 2022), https://www.fda.gov/media/109618/download (Last accessed July 31, 2024).
32. Sara Gerke, *"Nutrition Facts Labels" for Artificial Intelligence/Machine Learning-Based Medical Devices—The Urgent Need for Labeling Standards*, 91 Geo. Wash. L. Rev. 79, 99-119 (2023).
33. *Chevron, U.S.A., Inc. v. NRDC, Inc.*, 467 U.S. 837, 104 S. Ct. 2778 (1984)
34. *West Virginia v. EPA*, 597 U.S. 697, 142 S. Ct. 2587 (2022)
35. Wagner JK. *Algorithmic Fairness in the Roberts Court Era*. Pac Symp Biocomput. 2023;28:519-530. PMID: 36541005; PMCID: PMC9782697.
36. *Jarkesy v. SEC*, 144 S. Ct. 677 (2024).
37. Sachs RE, Fuse Brown EC. *Supreme Power - The Loss of Judicial Deference to Health Agencies*. N Engl J Med. 2024 Jul 17. doi: 10.1056/NEJMp2408197. Epub ahead of print. PMID: 39018525.
38. U.S. Food & Drug Admin., *Compliance & Enforcement*, https://www.fda.gov/animalveterinary/compliance-enforcement (Last accessed July 28, 2024).
39. U.S. Food & Drug Admin., *FDA Enforcement Statistics Summary: Fiscal Year 2017*, https://www.fda.gov/media/110196/download (Last accessed July 31, 2024).

40. Crowell & Moring LLP, *The Supreme Court's Double Hammer to Agencies: Loper Bright and Corner Post Set New Precedents for Challenging Federal Agency Action* (July 11, 2024), https://www.crowell.com/en/insights/client-alerts/the-supreme-courts-doublehammer-to-agencies-loper-bright-and-corner-post-set-new-precedents-for-challengingfederal-agency-action (Last accessed July 31, 2024).
41. Parmet, W. E. (2024, July 18). *Loper Bright And The Death Of Deference In The Administration Of Health Policy*. Health Affairs Forefront. DOI: 10.1377/forefront.20240717.933482 (describing the case as "a symbol of the Court's hostility to the administrative state").
42. *Skidmore v. Swift & Co.*, 323 U.S. 134, 65 S. Ct. 161 (1944).
43. 28 U.S.C. § 2401(a) (2023).
44. STATUTE OF LIMITATIONS, Black's Law Dictionary (12th ed. 2024).
45. STATUTE OF REPOSE, Black's Law Dictionary (12th ed. 2024).
46. Joshua Weiss, Sage Schaftel & Mae Babbington, *Health Care Impacts Following Chevron Decision,* Brownstein Client Alert (July 3, 2024), https://www.bhfs.com/insights/alertsarticles/2024/health-care-impacts-following-chevron-decision (Last accessed July 31, 2024).
47. Gerke, S and Simon, D.A., *Chevron's Fall and Its Impact on Medical Artificial Intelligence* (forthcoming).
48. *TransUnion LLC v. Ramirez*, 141. S. Ct. 2190, 210 L. Ed. 2d 568 (2021).
49. *Dobbs v. Jackson Women's Health Org.*, 142 S. Ct. 2228 (2022).
50. L. Heinzerling, *The Supreme Court is Making America Ungovernable*, The Atlantic, July 29, 2022, https://www.theatlantic.com/ideas/archive/2022/07/supreme-court-majorquestions-doctrinecongress/670618/ (Last accessed July 30, 2024).
51. Lemley, Mark A., *The Imperial Supreme Court* (July 28, 2022). Available at SSRN: https://ssrn.com/abstract=4175554 or http://dx.doi.org/10.2139/ssrn.4175554.
52. Beermann, Jack Michael, *The Anti-Innovation Supreme Court: Major Questions, Delegation, Chevron and More* (March 9, 2023). William & Mary Law Review, Forthcoming, Boston Univ. School of Law Research Paper No. 4383132, Available at SSRN: https://ssrn.com/abstract=4383132 or http://dx.doi.org/10.2139/ssrn.4383132.
53. Savage, C. *Weakening Regulatory Agencies Will Be a Key Legacy of the Roberts Court: News Analysis. New York Times* (Online). June 28, 2024. https://www.nytimes.com/2024/06/28/us/politics/supreme-court-regulatory-agencies.html (Last accessed July 30, 2024).
54. The White House. *FACT SHEET: President Biden Announces Bold Plan to Reform the Supreme Court and Ensure No President Is Above the Law*. July 29, 2024. https://www.whitehouse.gov/briefing-room/statements-releases/2024/07/29/fact-sheetpresident-biden-announces-bold-plan-to-reform-the-supreme-court-and-ensure-nopresident-is-above-the-law/ (Last accessed July 31, 2024).
55. K.M. Lewis, *Informal Guidance and the FDA*, 66 Food & Drug LJ 507 (2011)
56. U.S. Food & Drug Admin., *Artificial Intelligence: FDA Publications*, https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/artificial-intelligence-fda-publications (Last accessed Sept. 23, 2024).
57. *OMB Memorandum for the Heads of Executive Departments and Agencies (Young Memo): Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial*

Intelligence. Mar. 28, 2024. Available at https://www.whitehouse.gov/wpcontent/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-RiskManagement-for-Agency-Use-of-Artificial-Intelligence.pdf (Last accessed July 31, 2024).

58. Schmit CD, Doerr MJ, Wagner JK. *Leveraging IP for AI governance*. Science. 2023 Feb 17;379(6633):646-648. doi: 10.1126/science.add2202. Epub 2023 Feb 16. PMID: 36795826.
59. National Academies of Sciences, Engineering, and Medicine. 2024. *Foundational Research Gaps and Future Directions for Digital Twins*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26894 (Last accessed July 31, 2024).

Investigating the Differential Impact of Psychosocial Factors by Patient Characteristics and Demographics on Veteran Suicide Risk Through Machine Learning Extraction of Cross-Modal Interactions*

Joshua Levy†
Department of Computational Biomedicine, Cedars Sinai Medical Center
Los Angeles, CA USA
Email: joshua.levy@cshs.org

Monica Dimambro
White River Junction VA Medical Center
White River Junction, VT USA
Email: monica.dimambro@va.gov

Alos Diallo, Jiang Gui
Dartmouth College Geisel School of Medicine
Hanover, NH USA
Email: alos.b.diallo.gr@dartmouth.edu, jiang.gui@dartmouth.edu

Brian Shiner, Maxwell Levis
White River Junction VA Medical Center
White River Junction, VT USA
Email: brian.shiner@va.gov, maxwelle.levis@va.gov

Accurate prediction of suicide risk is crucial for identifying patients with elevated risk burden, helping ensure these patients receive targeted care. The US Department of Veteran Affairs' suicide prediction model primarily leverages structured electronic health records (EHR) data. This approach largely overlooks unstructured EHR, a data format that could be utilized to enhance predictive accuracy. This study aims to enhance suicide risk models' predictive accuracy by developing a model that incorporates both structured EHR predictors and semantic NLP-derived variables from unstructured EHR. XGBoost models were fit to predict suicide risk– the interactions identified by the model were extracted using SHAP, validated using logistic regression models, added to a ridge regression model, which was subsequently compared to a ridge regression approach without the use of interactions. By introducing a selection parameter, α, to balance the influence of structured (α=1) and unstructured (α=0) data, we found that intermediate α values achieved optimal performance across various risk strata, improved model performance of the ridge regression approach and uncovered significant cross-modal interactions between psychosocial constructs and patient characteristics. These interactions highlight how psychosocial risk factors are influenced by individual patient contexts, potentially informing improved risk prediction methods and personalized interventions. Our findings underscore the importance of incorporating nuanced narrative data into

* This work is supported by Department of Defense grant PR220927 to JL, ML, JG, BS, NIH P30CA023108 support for JL, and by VA Clinical Science Research and Development Career Development Award (CX002630) to ML.
† To whom correspondence should be addressed.

predictive models and set the stage for future research that will expand the use of advanced machine learning techniques, including deep learning, to further refine suicide risk prediction methods.

Keywords: machine learning, suicide risk, clinical notes, electronic health records, Veterans

1. Introduction

Veterans are at an elevated risk of suicide, underscoring the critical need for advanced risk stratification methods within the US Department of Veterans Affairs. The primary tool currently in use is Recovery Engagement and Coordination for Health – Veterans Enhanced Treatment (REACH-VET), an AI-driven model that utilizes structured data to assess and categorize suicide risk [1,2]. This model plays a crucial role in pinpointing Veterans who are at the highest risk and disproportionately contribute to the annual suicide statistics [3].

Recent research efforts have focused on augmenting the REACH-VET model by integrating unstructured data sources, such as clinical notes, to uncover additional predictors of risk [4,5]. Our previous studies have aimed at identifying specific risk groups and stratifying the impact of various textual suicide risk factors within these groups [6–10]. By using REACH-VET to establish baseline risk, we have developed NLP models that employ semantic databases and textual analysis to track risk factors across different risk tiers and determine optimal intervention timing.

Our prior research effectively identified novel NLP-derived variables that complement traditional demographic and structured risk predictors. By matching cases and controls based on their risk percentiles as measured through structured predictors, we sought to control for potential confounding factors. Controlling for confounding factors, however, does not address the possibility of effect modification, an area that remains relatively underexplored in this context. Understanding how risk factors differ by Veteran subgroups is crucial not only for improving predictive accuracy but also for enhancing the explainability of how psychosocial factors relate to suicide risk across diverse groups. Exploring these interactions could lead to interventions that are more tailored and effective, underscoring the importance of this research for future clinical applications.

Classification and regression trees (CART) are particularly useful for examining effect modifiers among predictors through conditional decision splits. Previous studies have demonstrated their utility in revealing complex statistical interactions. Despite the effectiveness of CART in managing interactions, a significant challenge persists due to the overwhelming number of NLP variables compared to the relatively fewer structured predictors and patient-level clinical factors. This imbalance complicates their effective integration into the predictive model. This disparity necessitates innovative approaches to manage and interpret the extensive data generated by NLP techniques within our predictive models in the context of these patient factors.

This manuscript describes our methodology for refining risk prediction models by integrating both structured and unstructured data within a risk-matched Veteran population, aiming to deliver a more intricate comprehension of suicide risk. Our approach not only seeks to provide more pertinent risk assessments tailored to specific subpopulations but also aims to demonstrate how machine learning models can effectively identify effect modifiers of crucial psychosocial variables based on patient characteristics. These modifiers, once validated through conventional statistical regression

methods, have the potential to significantly improve the interpretability and precision of existing models for assessing suicide risk.

2. Methods

2.1. *Patient Selection*

To establish our study group, we integrated data from the VA Corporate Data Warehouse (CDW) Electronic Health Records (EHR) with mortality information from the VA-Department of Defense Mortality Data Repository [11]. This allowed us to pinpoint Veterans who died by suicide and interacted with VHA healthcare services in 2017 or 2018, totaling 2,842 cases. Following established recommended guidance for matched case-control studies that focus on infrequent events, we matched each suicide case with five controls. Assistance from the VA Office of Mental Health and Suicide Prevention was crucial in selecting control subjects who were treated at the same VHA facility and during the same period as the cases. Controls were chosen to match the deceased cases' REACH-VET risk percentile and were alive at the time the cases died (totaling 14,042 controls) [12]. Controls were unique and non-overlapping, such that no cases could share the same controls. We validated the effectiveness of our matching approach by evaluating the standardized mean differences in various demographic and clinical parameters between cases and controls (**Table 1**). In a prior study that analyzed risk trends in a national sample of recent VA suicide deaths [12], we found that patients at varying suicide risk tiers (high, moderate/med, and low), have very different diagnostic, service usage, and demographic patterns. To best develop targeted risk models, we stratified the present study's sample using these risk tiers.

Table 1: Patient Characteristics/Demographics

	Case (N=2842)	Control (N=14042)	p-value
Demographics			
Female	119 (4.2%)	1079 (7.7%)	0.149
Non married	1688 (59.4%)	7861 (56.1%)	0.068
Married	1154 (40.6%)	6163 (43.9%)	0.038
Homeless_prior24m	212 (7.5%)	1189 (8.5%)	
Veteran	2834 (99.7%)	13971 (99.6%)	0.017
Rural	635 (22.3%)	3215 (22.9%)	0.013
Risk Tier			0.007
High	389 (13.7%)	1940 (13.8%)	
Moderate	1436 (50.5%)	7040 (50.2%)	
Low	1017 (35.8%)	5044 (36.0%)	
Race			0.273
Am. Ind. or Asian Pacific	61 (2.1%)	308 (2.2%)	
Black	154 (5.4%)	1638 (11.7%)	
Hispanic	124 (4.4%)	875 (6.2%)	
Unknown	129 (4.5%)	306 (2.2%)	
White	2374 (83.5%)	10897 (77.7%)	
Age			0.008
Mean (SD)	60.5 (18.0)	60.4 (15.7)	
Deployment			
Vietnam	1100 (38.7%)	5862 (41.8%)	0.066
Afghanistan or Iraq	957 (33.7%)	4761 (33.9%)	0.017
Mental Health Diagnosis/ Risk Flag			
Anxiety	1341 (47.2%)	6686 (47.7%)	0.009
Bipolar	545 (19.2%)	2238 (16.0%)	0.085
Conduct	56 (2.0%)	316 (2.3%)	0.02
Depression	1876 (66.0%)	9137 (65.2%)	0.02
Neurocognitive	316 (11.1%)	1671 (11.9%)	0.025
OCD	80 (2.8%)	325 (2.3%)	0.032
PTSD	1060 (37.3%)	5273 (37.6%)	0.005
Personality	389 (13.7%)	1599 (11.4%)	0.070
Sleeping	1331 (46.8%)	7270 (51.8%)	0.100
Substance	1249 (43.9%)	5401 (38.5%)	0.112
Trauma	1442 (50.7%)	7235 (51.6%)	0.016
Combat	731 (25.7%)	2680 (19.1%)	0.159
Military Sexual Trauma	126 (4.4%)	875 (6.2%)	0.080
Number of Inpatient Mental Health Days within 1 Year of Death			
Mean (SD)	17.2 (66.1)	15.6 (64.6)	0.024
Prescriptions			
Opioid Rx_prior12	885 (31.1%)	4338 (30.9%)	0.004
Opioid Rx_prior24	1104 (38.8%)	5686 (40.5%)	0.035
Mood Stabilizer Rx_prior12	1017 (35.8%)	4718 (33.6%)	0.045
Mood Stabilizer Rx_prior24	1178 (41.4%)	5455 (38.9%)	0.052
Antipsychotic Rx_prior12	616 (21.7%)	2364 (16.9%)	0.122
Antipsychotic Rx_prior24	708 (24.9%)	2791 (19.9%)	0.120
Antidepressant Rx_prior12	1573 (55.3%)	7661 (54.6%)	0.014
Antidepressant Rx_prior24	1733 (61.0%)	8401 (59.9%)	0.022

2.2. *Data Collection and Partitioning*

2.2.1. *Clinical Note Retrieval*

We retrieved unstructured EHR notes from the CDW that were recorded within 30 days before each case's death. This timeframe was chosen based on earlier research that highlighted the significance of clinical notes during the period immediately leading up to death by suicide. To prevent the

influence of potential data leakage / endogeneity, we excluded notes from the final two days before death and any notes that referenced death or a high likelihood of death within the five days prior to the suicide. Additionally, we removed patients from our analysis if their records contained more than six times the average number of notes, thus preventing a disproportionate focus on individuals with higher healthcare engagement. This resulted in a dataset of 92,399 notes from 389 cases and 1,940 controls at high risk, 107,532 notes from 1,436 cases and 7,040 controls at moderate risk, and 44,613 notes from 1,017 cases and 5,044 controls at low risk. Model training and interpretation was conducted on the note-level, whereas performance was reported on the patient level (**see 2.4**).

2.3. *Data Preparation*

2.3.1. *Derivation of NLP Variables*

To capture word counts, we first converted all our text to lowercase, removed stop words like "his/hers", "were/would", "and/with", etc. and tokenized our data set into unigrams or bigrams. We used Sentiment Analysis and Cognition Engine (SÉANCE) to analyze sentiment from these tokens, transforming our corpus into 516 semantic variables. SÉANCE is a Python-based software package that is accessible on VA servers and has been found to be comparable to the commonly used Linguistic Inquiry and Word Count (LIWC) software [13,14]. SÉANCE utilizes a variety of established linguistic databases, including SemanticNet [15,16], General Inquirer Database (GID) [17], EmoLex [18,19], Lasswell [20], Valence Aware Dictionary and sEntiment Reasoner (VADER) [21], Hu–Liu [22,23], Harvard IV-4 [17], and the Geneva Affect Label Coder (GALC) [24]. Each database consists of expert-derived dictionary lists and rule-based systems [25], comprising over 250 unique variables, which can be assessed in positive and negative iterations, leading to 516 SÉANCE variables.

2.3.2. *Extraction of Patient Characteristics*

Using data from the Corporate Data Warehouse (CDW), we extracted a comprehensive array of information encompassing demographics, social determinants of health, patterns of service usage, prescription histories, and diagnostic details. This data set included key demographic variables such as age, gender, marital status, and race. Social determinants like homelessness and military service were also considered, providing context to the healthcare challenges these Veterans may face. The service usage patterns captured included the number and types of visits to emergency departments and mental health services, which are critical indicators of health engagement and potential crisis points. Prescription data detailed the use of critical medications such as opioids and antipsychotics, while diagnostic information covered a wide range of mental health conditions from anxiety and depression to PTSD and substance abuse disorders (**Table 1**). We observed a significant disparity in the number of traditional patient characteristics available (n=66) compared to the number of NLP-derived variables (n=516), which include terms and their negations extracted from clinical notes. A complete list of variables included in the model can be found in the Supplementary Material, available at the following URL: https://github.com/jlevy44/NLP_Demographics_VA/tree/main/Data_Dictionaries .

2.3.3. *Training, Validation and Test Patient Cohorts*

For each risk tier, patients were stratified into training, validation, and test sets using an 80%, 10%, and 10% split, respectively. We utilized the *GroupShuffleSplit* function from the scikit-learn

package (Python v3.8) to ensure that all notes and patient characteristics from the same individual were grouped into the same set [26]. This approach prevents any notes from a single patient from being distributed across different sets, thereby avoiding data leakage and ensuring the integrity of test set statistics. Variables were standardized via scaling parameters estimated from the training set.

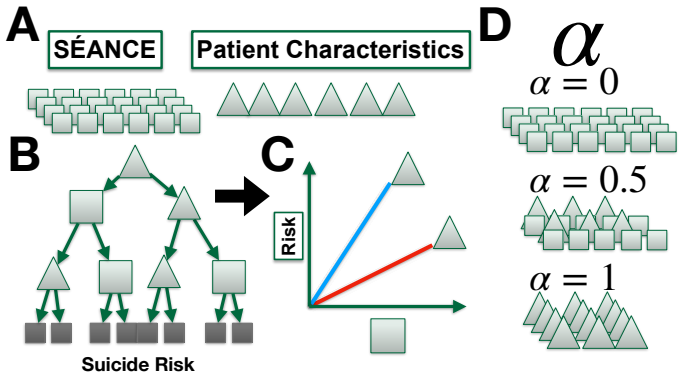


Figure 1: Workflow overview: A) The number of SÉANCE variables dwarfs the number of patient characteristics / demographics, **B)** Cross-modal interactions between SÉANCE and patient characteristics can be identified using a CART approach (e.g., XGBoost) through conditional decision splits between the different sets of variables; **C)** Shows how the relationships between NLP variables (squares) and suicide risk vary across different demographic subgroups (triangles). The lines represent these varying associations, providing simplified interpretations based on the GLM approach. **D)** Selection of SÉANCE and patient characteristics variables controlled through α, intermediate values reflect selection of both variable types, increasing likelihood of detecting cross-modal interactions

2.4. *Selected Machine Learning Models*

All 582 patient characteristics / demographic and SÉANCE variables were modeled simultaneously to predict whether a clinical note corresponded to a patient who had died by suicide. Note-level predicted probabilities ($p = f(\vec{x})$) were averaged across the notes within each patient into a final patient-level score ($\bar{p}$) reflecting the risk of suicide used as the final comparison. We evaluated model performance on both the validation and test sets by calculating the patient-level area under the receiver operating characteristic curve (AUROC) which compared $\bar{p}$ to whether the patient died by suicide. To ensure robustness, we employed a 1000-sample non-parametric bootstrapping to compute 95% confidence intervals for AUROC estimates.

We aimed to evaluate the performance of two machine learning models: 1) Penalized high-dimensional generalized linear models, exemplified by ridge logistic regression [27], which apply an L2-norm penalty to shrink model coefficients (set to 2.5e5 after a coarse hyperparameter search). This method reduces model complexity and prevents overfitting by addressing multicollinearity. 2) Classification and regression trees (CART) [28], as implemented by Extreme Gradient Boosting (XGBoost). XGBoost is an advanced form of gradient boosting that incrementally refines decision trees by concentrating on errors from previous trees [29]. It uses a gradient descent algorithm to meticulously adjust tree parameters, optimizing them based on the error gradient relative to earlier predictions. While the ridge regression model serves as a baseline for performance comparison, the XGBoost model is expected to enhance performance by capturing statistical interactions within and across modalities—specifically among patient characteristics / demographic and SÉANCE variables, as well as interactions between these modalities. XGBoost was specifically chosen for its capability to assign weights to features, directly influencing their selection probabilities during model training. This feature is crucial for effectively balancing the influence of different predictors

(see **2.5**). While LightGBM and BART (Bayesian Additive Regression Trees) offer similar functionalities, they were not selected for specific reasons [30,31]. LightGBM, for instance, only reweights the feature split gain after their initial selection, without altering initial selection probabilities. BART, on the other hand, allows for assignment of priors for variable selection, but computational demands are significantly higher, making it less suitable for our current scope but a potential candidate for future exploration. Ridge regression was selected as representative of generalized linear modeling approaches after initial comparisons to LASSO and ElasticNet [32].

The primary objective of this study is not simply to compare Ridge regression with XGBoost. Rather, our aim is to show that the interactions identified by XGBoost can offer additional valuable information, enhancing the predictive accuracy of Ridge regression and other generalized linear models that are known for their parsimonious interpretations. Consequently, we expect that incorporating these interactions as predictors—a method we have named *Ridge-Int*—will significantly improve the performance of Ridge regression, bridging the gap between complex machine learning and traditional statistical models [33].

2.5. *Key Contribution: Weighting the selection of NLP variables and patient characteristics*

In tree-based models, the selection of variables for inclusion at various levels or nodes typically occurs with uniform probability. This approach can inadvertently lower the probability of selecting variables from smaller sets of variables, such as patient characteristics, compared to larger sets like those from NLP-derived variables. Consequently, this bias in variable selection could hinder the identification of meaningful interactions between patient characteristics / demographic and NLP variables, as the former are less likely to be chosen as nodes or leaves in the model.

To address this imbalance, we hypothesize that strategically weighting the selection of variables from these two distinct sets—patient characteristics / demographic and NLP variables—could be crucial for uncovering optimal interactions between them. In this study, we conduct a sensitivity analysis to explore the impact of different weighting strategies on the detection of interactions (**Figure 1**). Specifically, we investigate three scenarios:

1. **Upweighting Patient Characteristics / Demographics:** We hypothesize that increasing the selection probability of patient characteristics (including demographics) could enhance the identification of interactions within these features.
2. **Upweighting NLP Features:** Conversely, increasing the weight of NLP features is expected to surface more interactions within the NLP data.
3. **Balanced Weighting:** Applying an equal weighting strategy, adjusted for the numerical disparity between the sets (upweighting patient characteristics / demographics proportionally to the number of NLP features), is hypothesized to facilitate the detection of cross-modal interactions, balancing the trade-offs between the two.

To test these hypotheses, we introduce a selection hyperparameter, $\alpha \in [0,1]$, which determines the extent to which one set of predictors is favored over the other. The weighting formula for individual patient characteristics / demograpgics is defined as $\alpha * \frac{n_{SEANCE}}{n_{demographics}} + \epsilon$, where $\frac{n_{SEANCE}}{n_{demographics}}$ represents the ratio of the number of NLP variables to patient characteristics, adjusting for their discrepancy. Conversely, the weight for selecting NLP variables is set as $1-\alpha+ \epsilon$. Thus, an α value of 0 would give priority to NLP variables, highlighting interactions within the NLP data, whereas an α of 1 would prioritize patient characteristics / demographics, enhancing the identification of interactions solely between structured patient characteristics. Here, ϵ is a small constant (ϵ = 1e-7) introduced to

ensure that the probability of selecting variables from either predictor set never reaches zero as required by the XGBoost package. This minimal adjustment allows for the rare but possible selection of variables from the non-prioritized set.

Table 2: XGBoost hyperparameter search grid

Hyperparameter	Values
colsample_bynode	0.25, 0.5, 0.75, 1
colsample_bylevel	0.25, 0.5, 0.75, 1
colsample_bytree	0.5, 0.75, 1
subsample	0.6, 0.8, 1
min_child_weight	1, 3, 5, 7
max_depth	3, 4, 5, 6
gamma	0, 1, 5, 10
reg_alpha	0, 0.1, 1, 10
reg_lambda	0, 0.1, 1, 10
Number of Trees	25, 50, 100

2.6. *Model Fitting, Interaction Extraction and Validation for Experimental Comparisons*

We trained XGBoost models using various values for α, including 0 (favoring SÉANCE variables), 0.1, 0.3, 0.5, 0.7, 0.9, and 1 (favoring patient characteristics). The model fitting process involved a 50-iteration randomized search for optimal hyperparameters (**Table 2**), with early stopping for tree boosting based on validation set performance. This procedure was repeated for all α and risk tiers.

For each value of α, we used the *interactiontransformer* package [33] to select candidate interactions for further analysis via the tree explainer. This assigns each interaction a global SHAP score, which represents the average influence of the interaction across all notes and patients, reflecting its overall contribution to the model's performance [34]. SHAP interaction scores were computed separately for the validation and test sets. To validate candidate interactions, we examined the top 1000 interactions identified by SHAP. For each interaction, we fit unpenalized generalized linear models (GLM, logistic regression) incorporating the interaction term (**Figure 1C**):

$$logit(p_{suicide}) = \beta_0 + \beta_1 feature_1 + \beta_2 feature_2 + \beta_3 feature_1 * feature_2$$

A candidate interaction was confirmed as validated if the p-value for the coefficient of the interaction term, β_3, was less than 0.05 divided by 1000. This stringent criterion reflects the Bonferroni adjustment applied to account for multiple comparisons (1000 candidate interactions), ensuring the robustness of our findings against Type I errors.

To evaluate the effectiveness of SHAP values in identifying and prioritizing key interactions, we used Fisher's exact test to compare the likelihood of GLM-validated variables appearing in the top 100 versus the top 1000 SHAP-ranked interactions. By calculating an odds ratio (OR) and a corresponding p-value as a measure of enrichment in the top 100 set, we quantified the degree to which SHAP values not only identify but also accurately prioritize the most impactful interactions.

We hypothesized that validated interactions would be predominantly found among the highest-ranked interactions by SHAP, indicating the effectiveness of SHAP in identifying the most influential interactions in terms of their contribution to the model's predictive accuracy. This step serves not only to validate the interactions but also to verify the reliability of SHAP's ranking mechanism in prioritizing the most statistically significant and predictive interactions [a].

For the validated interactions, we categorized the nature of each interaction based on its modality: either within modality interactions (such as demographic-demographic or SÉANCE-SÉANCE) or cross-modality (demographic-SÉANCE) interactions. We quantified these categories by calculating their proportions within the overall set of validated interactions, providing insight into the patterns of relationships that significantly contribute to the model.

After categorizing the interactions, we incorporated them into the predictive model. Specifically, we enhanced the Ridge regression model by adding either the validated interactions or the top 50

[a] Further clarification on these calculations can be found in Supplementary materials: "Clarification on Role of Algorithms and Methods", at: https://github.com/jlevy44/NLP_Demographics_VA/blob/main/suppl_material.docx

interactions ranked by p-value—whichever count was greater. This enhanced model, referred to as *Ridge-Int*, was designed to assess the impact of including significant interaction terms on the predictive accuracy. The performance of the *Ridge-Int* model was compared against the baseline Ridge model, which did not include interaction terms. We evaluated the models' effectiveness on both the validation and test sets using the AUROC, with 95% confidence intervals calculated using the previously described bootstrapping method.

Following the validation and integration of interactions into our models, we plotted the AUROCs for XGBoost, Ridge, and *Ridge-Int* against the hyperparameter α, along with the odds ratios for the significance of validated interactions and the proportion of validated interactions that were cross-modal within the validation and test set. We anticipated that the interactions identified by XGBoost would enhance the performance of Ridge regression, and that the performance metrics for both XGBoost and *Ridge-Int* would likely reach a plateau at an intermediate α value between 0 and 1. Similarly, we expected the enrichment of validated interactions and the proportion of cross-modal interactions to saturate at a midpoint α, demonstrating the relevance of cross-modal interactions for enhancing predictiveness. This analysis was stratified and performed across each risk tier, allowing for a nuanced evaluation of how the inclusion of interactions influences model performance within distinct suicide risk tiers.

2.7. *Interpretation of Randomly Selected Interactions*

To deepen our understanding of the interactions between different modalities, we analyzed the statistical interaction models fitted to the data, employing estimated marginal means as a post hoc comparison to elucidate the effects of various psychosocial constructs obtained through NLP on suicide risk [35,36]. These effects were specifically examined as conditioned by patient characteristics, and similarly, how patient characteristics / demographics influence the impact of psychosocial factors on suicide risk (**Figure 1C**). For illustrative clarity, effect estimates for randomly selected interactions were presented in detailed tables and supportive visualizations demonstrating the conditional/stratified effects of these psychosocial constructs by patient characteristics.

3. Results

3.1. *Affirming the Relevance of Cross-modal Interactions*

In our study, we adjusted the selection probability between two predictor sets to investigate their potential trade-offs in influencing model performance. By altering the hyperparameter α, we modulated the selection bias towards either patient characteristics / demographics or SÉANCE variables within the XGBoost model, and then examined the nature and impact of the interactions identified by SHAP. The interactions that were extracted and validated using unpenalized GLMs– prior to their inclusion in *Ridge-Int*– with interaction terms demonstrated statistical validity. Importantly, these interactions, once confirmed through statistical modeling, showed a high enrichment within the top 100 SHAP-ranked interactions, evidenced by significant odds ratios. This result supports the effectiveness of XGBoost and SHAP in pinpointing genuine interactions.

Our analysis revealed that interactions were more prevalently validated at intermediate values of α, suggesting an optimal balance at these levels for extracting meaningful interactions between different types of data (**Figure 2B, Table 3**). The proportion of cross-modal interactions that were validated peaked at these intermediate α values (**Figure 2C, Table 3**). This finding corroborates the

hypothesis that adjusting α allows for fine-tuning of the XGBoost model to effectively balance the contribution of both predictor sets, enhancing the model's capacity to uncover and utilize significant interactions between different data types. It should be noted that the enrichment of validated interactions was especially pertinent for low-risk tier patients, though a lower proportion of these interactions were cross-modal in nature.

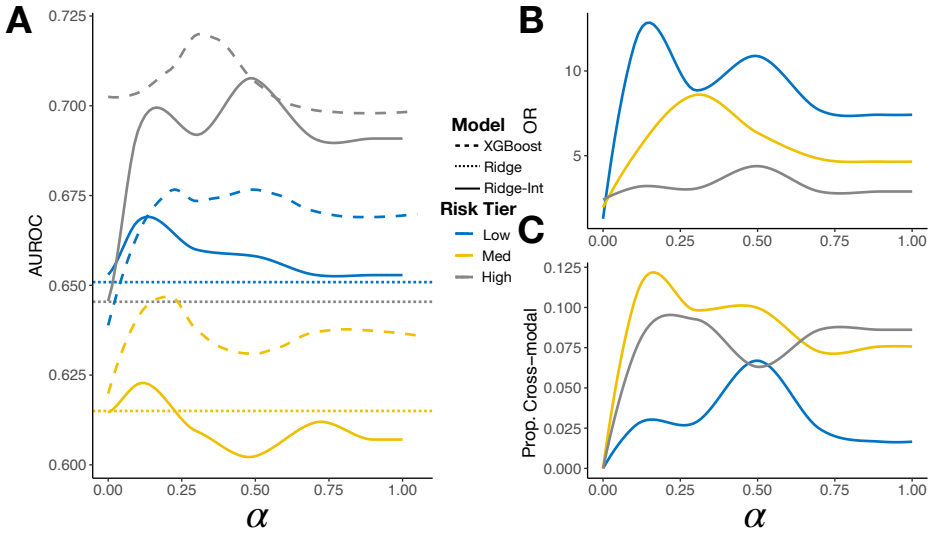


Figure 2: Model Comparison and Interaction Validation/Delineation: A) Test set model performance reported via the AUROC on the patient-level, aggregated across notes, for each model type and risk tier. **B)** Odds Ratio (OR) versus α. OR reflects the enrichment of validated interactions among the top-ranked interactions identified through SHAP, serving as a measure of how well the statistical model validates the interactions identified by SHAP. **C)** The proportion of validated interactions that were identified as cross-modal as a function of α.

3.2. *Model Performance Comparisons*

In our study, we hypothesized that cross-modal interactions would significantly enhance the predictive performance in suicide risk assessment, particularly when analyzing aggregated data across patient notes. Our results confirmed this hypothesis, demonstrating the critical role of these interactions in improving model accuracy. Notably, the XGBoost model, which explicitly accounts for statistical interactions through conditional decision splits, consistently outperformed the traditional Ridge regression model, which does not inherently consider interactions (**Figure 2A, Table 4).** Upon integrating these extracted interactions from the XGBoost model into the Ridge regression frameworks (*Ridge-Int*), we observed marked performance improvement for the low and high risk patients and modest improvement in the moderate risk patients for α=0.1 (**Figure 2A, Table 4)**.

Interestingly, the most pronounced gains were observed when the selection parameter α, which balances the influence of structured patient characteristics / demographics versus SÉANCE features, was set to intermediate values. This suggests that neither purely patient characteristics nor purely SÉANCE features are sufficient on their own; rather, it is their combination and the interactions between them that drive the predictive accuracy of the models. This phenomenon was corroborated by the relative alignment of these optimal α values with where the highest number of crossmodal interactions were identified and incorporated into the statistical modeling of the Ridge regression (**Figure 2A,C, Tables 2,3**).

Table 3: Validation and Analysis of XGBoost-Derived Interactions via SHAP and Subsequent Logistic Regression Modeling. OR indicate the degree to which validated interactions as confirmed through logistic regression modeling are enriched among the top SHAP-ranked interactions, reflecting the effectiveness of the XGBoost in identifying relevant interactions. Additionally, the table lists the proportion of these validated interactions that were characterized as cross-modal, highlighting their potential for bridging distinct data modalities.

Risk	α	OR	p	% Cross-Modal	Risk	α	OR	p	% Cross-Modal	Risk	α	OR	p	% Cross-Modal
Low	0	1.3	0.672	0.0%	**Med**	0	2.0	0.082	0.0%	**High**	0	2.4	0.017	0.0%
	0.1	**11.5**	**<0.001**	2.6%		**0.1**	5.0	<0.001	**10.3%**		0.1	3.1	<0.001	7.1%
	0.3	8.9	<0.001	2.9%		**0.3**	**8.6**	**<0.001**	9.8%		**0.3**	3.1	<0.001	**9.3%**
	0.5	10.9	<0.001	**6.7%**		**0.5**	6.3	<0.001	**10.0%**		**0.5**	**4.4**	**<0.001**	6.3%
	0.7	7.7	<0.001	2.5%		0.7	4.8	<0.001	7.3%		0.7	2.9	<0.001	8.6%
	0.9	7.4	<0.001	1.7%		0.9	4.6	<0.001	7.6%		0.9	2.9	<0.001	8.6%
	1	7.4	<0.001	1.7%		1	4.6	<0.001	7.6%		1	2.9	<0.001	8.6%

Table 4: Test Set Model Performance for XGBoost and Ridge Regression models, comparing performance across low, medium, and high-risk tiers. The AUROC values are presented alongside 95% CIs calculated through 1000-sample non-parametric bootstrapping. For Ridge Regression, 'n/a' indicates the performance of the model without the inclusion of interactions derived from XGBoost, serving as a baseline comparison. The variations in AUROC values across different α (ranging from 0 to 1) illustrate the impact of emphasizing either patient characteristics / demographics or SÉANCE features, or a balanced consideration of both, in predicting suicide risk.

XGBoost

Risk	α	AUROC	2.5%CI	97.5%CI	Risk	α	AUROC	2.5%CI	97.5%CI	Risk	α	AUROC	2.5%CI	97.5%CI
Low	0	0.622	0.565	0.681	**Med**	0	0.602	0.555	0.646	**High**	0	0.715	0.642	0.788
	0.1	**0.678**	**0.624**	**0.735**		**0.1**	**0.658**	**0.61**	**0.705**		0.1	0.687	0.61	0.767
	0.3	0.672	0.618	0.721		0.3	0.637	0.59	0.681		**0.3**	**0.726**	**0.65**	**0.799**
	0.5	**0.679**	**0.622**	**0.734**		0.5	0.629	0.579	0.681		0.5	0.705	0.627	0.781
	0.7	0.669	0.611	0.723		0.7	0.638	0.589	0.686		0.7	0.698	0.618	0.782
	0.9	0.669	0.611	0.723		0.9	0.637	0.587	0.683		0.9	0.698	0.618	0.782
	1	0.669	0.611	0.723		1	0.637	0.587	0.683		1	0.698	0.618	0.782

Ridge Regression

Risk	α	AUROC	2.5%CI	97.5%CI	Risk	α	AUROC	2.5%CI	97.5%CI	Risk	α	AUROC	2.5%CI	97.5%CI
Low	n/a	0.651	0.593	0.704	**Med**	n/a	0.615	0.566	0.666	**High**	n/a	0.645	0.561	0.729
	0	0.653	0.596	0.706		0	0.615	0.565	0.665		0	0.646	0.562	0.729
	0.1	**0.668**	**0.61**	**0.72**		**0.1**	**0.622**	**0.574**	**0.672**		0.1	0.693	0.602	0.778
	0.3	0.66	0.6	0.714		0.3	0.609	0.559	0.659		0.3	0.692	0.603	0.773
	0.5	0.658	0.599	0.715		0.5	0.602	0.552	0.652		**0.5**	**0.707**	**0.621**	**0.784**
	0.7	0.653	0.594	0.708		0.7	0.612	0.562	0.66		0.7	0.691	0.602	0.775
	0.9	0.653	0.594	0.708		0.9	0.607	0.557	0.655		0.9	0.691	0.602	0.775
	1	0.653	0.594	0.708		1	0.607	0.557	0.655		1	0.691	0.602	0.775

Table 5: Randomly Selected Validated Cross-Modal Interactions Across All Risk Tiers. Showcases a sample of cross-modal interactions validated from logistic regression analyses on note-level across different risk tiers, with findings adjusted for 1000 multiple comparisons for each risk tier using Bonferroni correction. Each column corresponds to different α levels (0.3, 0.5, and 0.7 as representative), demonstrating the variability in interaction significance and strength (log(OR)) with changing α values. See supplementary materials for term descriptions.

	α=0.3			α=0.5			α=0.7		
Risk	**Interaction**	**log(OR)**		**Interaction**	**log(OR)**		**Interaction**	**log(OR)**	**p-adj**
High	days_inpatMH_prior_12mo:Notlw_Lasswell	-0.02	7.4e-07	Sleeping:vader_neutral	-0.98	2.6e-02	Calc_age:Know_GI	0.16	1.0e-04
	days_inpatMH_prior_12mo:Coll_GI_neg_3	-0.03	4.0e-14	antipsy_prior24:vader_compound	0.19	1.0e-05	age_55_74:Hu_GI_neg_3	3.59	3.4e-10
	Substance:negative_adjectives_component	0.12	4.3e-02	age_55_74:Posaff_Lasswell	-4.22	1.6e-03	days_inpatMH_prior_12mo:Tool_GI_neg_3	0.02	7.4e-05
	Calc_age:negative_adjectives_component	0.01	7.0e-08	Substance:Work_GI	-4	1.5e-03	age_55_74:Socrel_GI_neg_3	2.74	4.8e-04
	age_55_74:Coll_GI_neg_3	10.47	2.0e-22	Calc_age:Endslw_Lasswell	0.1	7.9e-03	elix_cat:Know_GI	3.24	1.2e-06
Med	elix_cat:fear_and_digust_component	0.74	3.2e-03	elix_cat:Male_GI	-4.36	9.5e-04	elix_cat:Male_GI	-4.36	9.5e-04
	elix_cat:Fear_EmoLex	3.26	6.1e-03	elix_cat:fear_and_digust_component	0.74	3.2e-03	moodst_prior24:Abs_GI	4.22	8.4e-03
	Substance:Powcoop_Lasswell_neg_3	-12.25	2.4e-05	opioid_prior12:Submit_GI_neg_3	-4.52	7.0e-05	di_cat:hu_liu_pos_perc_neg_3	-0.19	1.9e-02
	Nonmarried:hu_liu_pos_perc_neg_3	0.41	7.4e-03	MH_cat:hu_liu_pos_nwords	3.1	2.5e-02	Trauma:vader_positive	2.24	4.6e-04
	Unknown:Pleasur_GI_neg_3	8.63	3.4e-02	MH_cat:Tranlw_Lasswell	2.85	1.8e-03	Unknown:Pleasur_GI	9.05	1.8e-02
Low	elix_cat:fear_and_digust_component	0.74	3.2e-03	elix_cat:Male_GI	-4.36	9.5e-04	elix_cat:Male_GI	-4.36	9.5e-04
	elix_cat:Fear_EmoLex	3.26	6.1e-03	elix_cat:fear_and_digust_component	0.74	3.2e-03	moodst_prior24:Abs_GI	4.22	8.4e-03
	Substance:Powcoop_Lasswell_neg_3	-12.25	2.4e-05	opioid_prior12:Submit_GI_neg_3	-4.52	7.0e-05	di_cat:hu_liu_pos_perc_neg_3	-0.19	1.9e-02
	Nonmarried:hu_liu_pos_perc_neg_3	0.41	7.4e-03	MH_cat:hu_liu_pos_nwords	3.1	2.5e-02	Trauma:vader_positive	2.24	4.6e-04
	Unknown:Pleasur_GI_neg_3	8.63	3.4e-02	MH_cat:Tranlw_Lasswell	2.85	1.8e-03	Unknown:Pleasur_GI	9.05	1.8e-02

3.3. *Select Interpretation of Findings from Cross-modal Workflow*

The XGBoost model successfully identified numerous cross-modal interactions, specifically at intermediate α, of which we selectively analyzed a few at random to elucidate their implications for decision-making and potential therapeutic advancements. For example, in **Table 5**, the results from GLM of validated interaction terms are presented. **Table 6** and **Figure 3** provide detailed breakdowns of four key interactions with further interpretation presented in the Discussion.

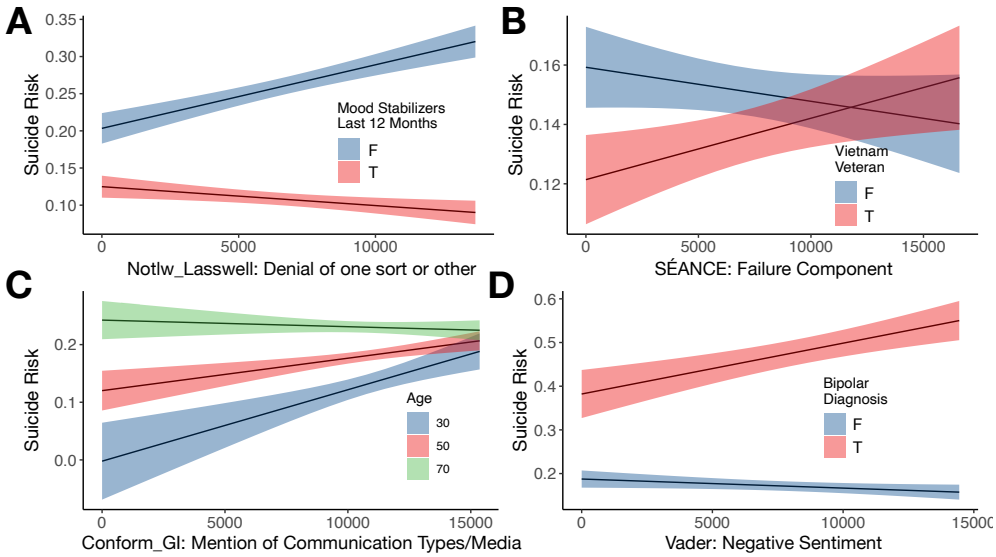


Figure 3: Interpretation of Conditional Effects of Psychosocial Constructs Across Patient Subgroups for Randomly Selected Validated Interactions. A) Suicide risk associated with mood stabilizer use fluctuates based on mention of denial in clinical notes. **B)** Intensified impact of failure mentions in notes on suicide risk among Vietnam Veterans compared to other Veterans. **C)** Varying effects of mentioning communication forms, such as mentions of social media, on suicide risk across age groups, with younger Veterans showing heightened sensitivity. **D)** Increased suicide risk due to negative sentiments among patients with bipolar disorder. Note that interpretations are on the note-level. Risk scale is expressed using the inverse logit link function.

Table 6: Select Interactions and Conditional Effects from Logistic Regression Analysis on Randomly Selected Validated Interactions. Interaction terms are denoted using ":", followed by a conditional effect, denoted by "|", representing the evaluation of the variable's effect under specific conditions set by the modifying variable on the right. Conditional effects are derived using estimated marginal means.

Risk	Term	log(OR)	p-value
High	moodst_prior12:Notlw_Lasswell	-1.11e-05	8.95e-12
	Notlw_Lasswell|moodst_prior12	8.55e-06	7.40e-11
	Notlw_Lasswell|No moodst_prior12	-2.54e-06	7.96e-03
	Vietnam:failure_component	3.21e-06	5.25e-03
	failure_component|Vietnam	-1.15e-06	1.45e-01
	failure_component|Not Vietnam	2.07e-06	1.40e-02
Med	Calc_age:Comform_GI	-3.38e-07	1.87e-04
	Comform_GI|Age=30	1.24e-05	4.05e-05
	Comform_GI|Age=50	5.63e-06	3.25e-04
	Notlw_Lasswell|Age=70	-1.14e-06	4.56e-01
Low	Bipolar:vader_negative	1.37e-05	6.51e-06
	vader_negative|Bipolar	-2.08e-06	4.66e-02
	vader_negative|Not Bipolar	1.16e-05	4.61e-05

4. Discussion

Recent advancements have emphasized the critical role of machine learning and the analysis of unstructured clinical reports in augmenting suicide risk prediction models [37]. These developments

aim to complement existing models that leverage structured predictors already operational within the VA system, which have been further repurposed to categorize risks into defined tiers for population studies [12]. Despite these innovations, the dynamics between predictors derived from structured and unstructured data, and their combined potential to improve suicide risk prediction, remain largely unexplored.

In this study, we aimed to refine suicide risk predictive models to cater specifically to relevant subgroups. Our strategy involved developing models that balanced the inclusion of both structured and unstructured (NLP) predictors. This approach allowed us to delve into the trade-offs and synergies between these predictor types through traditional statistical modeling of the interactions extracted from them. We introduced a predictor set selection parameter, α, to regulate the extent to which predictors from semantic NLP variables (SÉANCE) and structured EHR were utilized.

Our findings revealed that this methodology not only enhanced the accuracy of suicide risk predictions but also illuminated how cross-modal interactions between NLP variables and structured predictors could demonstrate the altered risk associated with various psychosocial constructs based on patient characteristics / demographics and vice versa. The ability to discern these interactions underscores the pivotal role of cross-modal dynamics in improving model performance, validating their importance in complex predictive tasks such as suicide risk assessment.

The implications of our analytical approach are significant– we will now discuss key lessons and insights derived from interpreting the interaction terms (**Figure 3, Tables 5, 6**). A positive interaction effect estimate signifies an elevated suicide risk when one variable increases, conditional on the rise of another variable. Conversely, a negative interaction effect indicates reduced risk under the same conditions. For example, our analysis showed that patients with a substance use disorder who frequently use negative adjectives in their clinical notes are at an increased risk of suicide [38]. In contrast, the presence of negative words has a less pronounced effect on patients without such a disorder. Another notable observation (**Figure 3, Table 6**) is that negative sentiments significantly elevate suicide risk among bipolar patients compared to those who are not bipolar, consistent with prior literature [39]. These instances demonstrate how psychosocial constructs variably affect different patient groups, paving the way for future large-scale studies aimed at identifying novel intervention targets and enhancing preventive strategies in suicide risk management.

This study has several limitations that merit consideration and can inform future work. Firstly, while the predictive modeling results were aggregated across patient notes, the initial predictive modeling and interpretation were conducted at the individual note level. Surprisingly, models trained solely with patient characteristics / demographics (α=1) showed an AUROC greater than expected, given that they were matched based on REACH-VET percentile scores derived from these same characteristics. This outcome suggests two key insights: additional stratification of suicide risk within defined risk tiers can unearth predictive factors not captured by models trained exclusively on structured predictors across the entire population (i.e., effect modification by risk tier); furthermore, the design of XGBoost, which ensures non-zero selection probabilities for variables, allowed the inclusion of a small yet significant set of SÉANCE variables to bolster model predictiveness. We did not compare the usage of TreeSHAP to other interaction extraction approaches [40–42]. Another limitation is that while the structured variables span over a year or more, the NLP variables are derived from observations within the past 30 days. Despite these observed limitations, the fundamental principles and broader findings of our research remain sound and valid. Another limitation is the statistical power to detect interactions, which may have been constrained by the limited sample size of this study. Future work aims to extend this analysis across a broader

temporal and demographic scope at the national level, which should incorporate a more diverse array of characteristics and potentially yield more robust findings above and beyond current suicide risk prediction approaches. Further external validation is limited by the focus on US veterans, and results may not generalize to other populations [43–46]. It is common to train models on historical data and validate them on more recent data, which could have strengthened the validity of our findings.

It should also be noted that previous research has highlighted that increased care utilization significantly influences the REACH-VET scores used for matching and stratification by risk tier [12]. Generally, more comprehensive data on Veterans can contribute to higher inferred suicide risk, whereas patients with less comprehensive records are typically assigned a lower risk. This variability in data completeness across different subpopulations underscores the need for our models to identify associations within these groups, especially since they may be differentially impacted by the extent of their record completeness [47–50].

Looking ahead, we plan to develop machine learning models that are not solely dependent on structured predictors for matching (i.e., randomly matched). This approach will allow us to potentially identify patterns that were previously obscured due to the biases introduced by data completeness. This could lead to more nuanced and effective predictive models that better address the diverse needs of all subgroups within the Veteran population.

The interpretation of findings from SÉANCE terms should be approached with caution [51,52]. SÉANCE encompasses a diverse array of lexical variables, each with different standards and encompassing varying numbers of words. However, these terms face challenges in capturing the nuanced contexts in which these words are used, which can complicate the interpretation of these concepts beyond their mere mention. This limitation is akin to the current challenges faced in sentiment analysis, even with the incorporation of negation terms. Initially, we adopted a semantic database approach as a proof of concept for this method. While we plan to expand our analysis to include a “bag-of-words” approach that captures all words within the corpus, this method also has its limitations as it tends to disregard their context within sentences. Therefore, our future work will focus on employing deep learning techniques to mine for motifs and patterns that can capture more complex and nuanced narratives. This approach will allow us to better contextualize these constructs and understand their differential impacts, informing future interventions more effectively [53–58].

Furthermore, deep learning models offer the flexibility to weigh different forms of information—including social determinants of health—on a patient-by-patient basis. They can also help identify critical timepoints for collecting notes that are most relevant to assessing suicide risk and determining optimal times for intervention. Our earlier work relied on count-based approaches, partly due to the limitations of computing resources available within the VA VINCI computing system. However, as advanced graphics processing units (GPU) systems become more accessible at the VA, we anticipate a shift towards more sophisticated deep learning approaches.

5. Conclusion

In conclusion, this study demonstrates the potential of integrating structured and unstructured data sources to enhance the predictiveness of suicide risk models for Veterans. The nuanced insights gained from cross-modal interactions identified through this comprehensive approach can better appreciate the dynamic interplay between numerical data from electronic health records and rich, psychosocial constructures available in clinical notes. As we move forward, the incorporation of more advanced machine learning techniques, particularly deep learning, promises to further refine our predictive capabilities and offer more targeted, effective interventions and risk prioritization.

References

1. McCarthy JF, Cooper SA, Dent KR, Eagan AE, Matarazzo BB, Hannemann CM, Reger MA, Landes SJ, Trafton JA, Schoenbaum M. Evaluation of the recovery engagement and coordination for health–veterans enhanced treatment suicide risk modeling clinical program in the veterans health administration. JAMA network open. American Medical Association; 2021;4(10):e2129900–e2129900.
2. Kessler RC, Hwang I, Hoffmire CA, McCarthy JF, Petukhova MV, Rosellini AJ, Sampson NA, Schneider AL, Bradley PA, Katz IR, Thompson C, Bossarte RM. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. Int J Methods Psych Res. 2017 Sep;26(3):e1575.
3. McCarthy JF, Bossarte RM, Katz IR, Thompson C, Kemp J, Hannemann CM, Nielson C, Schoenbaum M. Predictive Modeling and Concentration of the Risk of Suicide: Implications for Preventive Interventions in the US Department of Veterans Affairs. Am J Public Health. American Public Health Association; 2015 Sep;105(9):1935–1942.
4. Leonard Westgate C, Shiner B, Thompson P, Watts BV. Evaluation of Veterans' Suicide Risk With the Use of Linguistic Detection Methods. PS. 2015 Oct 1;66(10):1051–1056.
5. Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, Watts B, Flashman L, McAllister T. Predicting the risk of suicide by analyzing the text of clinical notes. PloS one. Public Library of Science San Francisco, USA; 2014;9(1):e85733.
6. Levis M, Westgate CL, Gui J, Watts BV, Shiner B. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychological medicine. Cambridge University Press; 2021;51(8):1382–1391.
7. Levis M, Levy J, Dufort V, Russ CJ, Shiner B. Dynamic suicide topic modelling: Deriving population-specific, psychosocial and time-sensitive suicide risk variables from Electronic Health Record psychotherapy notes. Clin Psychology and Psychoth. 2023 Jul;30(4):795–810.
8. Levis M, Levy J, Dufort V, Gobbel GT, Watts BV, Shiner B. Leveraging unstructured electronic medical record notes to derive population-specific suicide risk models. Psychiatry research. Elsevier; 2022;315:114703.
9. Levis M, Levy J, Dimambro M, Dufort V, Ludmer DJ, Goldberg M, Shiner B. Using natural language processing to evaluate temporal patterns in suicide risk variation among high-risk Veterans. Psychiatry Research. Elsevier; 2024;116097.
10. Levis M, Levy J, Dent KR, Dufort V, Gobbel GT, Watts BV, Shiner B. Leveraging natural language processing to improve electronic health record suicide risk prediction for Veterans Health Administration users. The Journal of clinical psychiatry. Physicians Postgraduate Press, Inc.; 2023;84(4):47557.
11. Department of Veterans Affairs, Department of Defense. Joint Department of Veterans Affairs (VA) and Department of Defense (DoD) Mortality Data Repository – National Death Index (NDI) [Internet]. 2017. Available from: https://www.mirecc.va.gov/suicideprevention/documents/VA_DoD-MDR_Flyer.pdf
12. Levis M, Dimambro M, Levy J, Dufort V, Fraade A, Winer M, Shiner B. Characterizing Veteran suicide decedents that were not classified as high-suicide-risk. Psychological Medicine. Cambridge University Press; 2024;1:1–10.
13. Crossley SA, Kyle K, McNamara DS. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. Behav Res. 2017 Jun;49(3):803–821.

14. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annu Rev Psychol. 2003 Feb;54(1):547–577.
15. Angioni M, Demonits R, Deriu M, Tuveri F. Semanticnet: a WordNetbased Tool for the Navigation of Semantic Information. Proceedings, GWC. 2008;21–34.
16. Das A, Bandyopadhyay S. Semanticnet-perception of human pragmatics. Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon [Internet]. 2010 [cited 2024 Jul 31]. p. 2–11. Available from: https://aclanthology.org/W10-3402.pdf
17. Stone PJ, Dunphy DC, Smith MS. The general inquirer: A computer approach to content analysis. MIT press; 1966 [cited 2024 Jul 31]; Available from: https://psycnet.apa.org/record/1967-04539-000
18. Mohammad SM, Turney PD. CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON. Computational Intelligence. 2013 Aug;29(3):436–465.
19. Mohammad S, Turney P. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text [Internet]. 2010 [cited 2024 Jul 31]. p. 26–34. Available from: https://aclanthology.org/W10-0204.pdf
20. Lasswell HD, Namenwirth JZ. The Lasswell value dictionary. New Haven. 1969;
21. Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the international AAAI conference on web and social media [Internet]. 2014 [cited 2024 Jul 31]. p. 216–225. Available from: https://ojs.aaai.org/index.php/icwsm/article/view/14550
22. Hu M, Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining [Internet]. Seattle WA USA: ACM; 2004 [cited 2024 Jul 31]. p. 168–177. Available from: https://dl.acm.org/doi/10.1145/1014052.1014073
23. Hu M, Liu B. Mining opinion features in customer reviews. AAAI [Internet]. 2004 [cited 2024 Jul 31]. p. 755–760. Available from: https://cdn.aaai.org/AAAI/2004/AAAI04-119.pdf
24. Scherer KR. What are emotions? And how can they be measured? Social Science Information. SAGE Publications Ltd; 2005 Dec 1;44(4):695–729.
25. Urbanowicz RJ, Moore JH. Learning Classifier Systems: A Complete Introduction, Review, and Roadmap. Journal of Artificial Evolution and Applications. 2009 Sep 22;2009:1–25.
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825–2830. PMID: 34682092
27. Hoerl AE, Kennard RW. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics. 1970 Feb;12(1):69–82.
28. Loh WY. Classification and Regression Trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011 Jan 1;1:14–23.
29. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016 [cited 2019 Nov 26]. p. 785–794. Available from: http://doi.acm.org/10.1145/2939672.2939785

30. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems [Internet]. 2017 [cited 2024 Jul 31];30. Available from: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
31. Tan YV, Roy J. Bayesian additive regression trees and the General BART model. Statistics in Medicine. 2019;38(25):5048–5069. PMID: 31460678
32. Ranstam J, Cook JA. LASSO regression. British Journal of Surgery. 2018 Sep 1;105(10):1348.
33. Levy JJ, O'Malley AJ. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. BMC Med Res Methodol. 2020 Jun 29;20(1):171. PMCID: PMC7325087
34. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020 Jan;2(1):56–67. PMID: 32607472
35. Lenth RV, Buerkner P, Giné-Vázquez I, Herve M, Jung M, Love J, Miguez F, Riebl H, Singmann H. emmeans: Estimated Marginal Means, aka Least-Squares Means [Internet]. 2023 [cited 2023 Mar 1]. Available from: https://CRAN.R-project.org/package=emmeans
36. Searle SR, Speed FM, Milliken GA. Population marginal means in the linear model: an alternative to least squares means. The American Statistician. Taylor & Francis; 1980;34(4):216–221.
37. Riblet NB, Matsunaga S, Lee Y, Young-Xu Y, Shiner B, Schnurr PP, Levis M, Watts BV. Tools to detect risk of death by suicide: A systematic review and meta-analysis. The Journal of clinical psychiatry. Physicians Postgraduate Press, Inc.; 2022;84(1):43891.
38. Harlow LL, Newcomb MD, Bentler PM. Depression, self-derogation, substance use, and suicide ideation: Lack of purpose in life as a mediational factor. Journal of Clinical Psychology. 1986;42(1):5–21.
39. Stange JP, Hamilton JL, Burke TA, Kleiman EM, O'Garro-Moore JK, Seligman ND, Abramson LY, Alloy LB. Negative cognitive styles synergistically predict suicidal ideation in bipolar spectrum disorders: A 3-year prospective study. Psychiatry Research. 2015 Mar 30;226(1):162–168.
40. Agrawal R, Trippe B, Huggins J, Broderick T. The Kernel Interaction Trick: Fast Bayesian Discovery of Pairwise Interactions in High Dimensions. Proceedings of the 36th International Conference on Machine Learning [Internet]. 2019 [cited 2024 Sep 29]. Available from: https://proceedings.mlr.press/v97/agrawal19a.html
41. Gui J, Andrew AS, Andrews P, Nelson HM, Kelsey KT, Karagas MR, Moore JH. A Robust Multifactor Dimensionality Reduction Method for Detecting Gene–Gene Interactions with Application to the Genetic Analysis of Bladder Cancer Susceptibility. Annals of Human Genetics. 2011;75(1):20–28.
42. Tsang M, Cheng D, Liu Y. Detecting Statistical Interactions from Neural Network Weights. ArXiv [Internet]. 2017 May 14 [cited 2024 Sep 29]; Available from: https://www.semanticscholar.org/paper/Detecting-Statistical-Interactions-from-Neural-Tsang-Cheng/5f85a8eaa7a1a1686f5a2bf721c63e337f03d8eb
43. Nock MK, Millner AJ, Ross EL, Kennedy CJ, Al-Suwaidi M, Barak-Corren Y, Castro VM, Castro-Ramirez F, Lauricella T, Murman N. Prediction of suicide attempts using clinician

assessment, patient self-report, and electronic health records. JAMA network open. American Medical Association; 2022;5(1):e2144373–e2144373.
44. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Scientific reports. Nature Publishing Group UK London; 2018;8(1):7426.
45. Atmakuru A, Shahini A, Chakraborty S, Seoni S, Salvi M, Hafeez-Baig A, Rashid S, San Tan R, Barua PD, Molinari F. Artificial Intelligence-based Suicide Prevention and Prediction: A Systematic Review (2019-2023). Information Fusion. Elsevier; 2024;102673.
46. Bayramli I, Castro V, Barak-Corren Y, Madsen EM, Nock MK, Smoller JW, Reis BY. Temporally informed random forests for suicide risk prediction. Journal of the American Medical Informatics Association. Oxford University Press; 2022;29(1):62–71.
47. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, Nock MK, Smoller JW, Reis BY. Predicting Suicidal Behavior From Longitudinal Electronic Health Records. AJP. 2017 Feb 1;174(2):154–162.
48. Bostwick JM, Pabbati C, Geske JR, McKean AJ. Suicide Attempt as a Risk Factor for Completed Suicide: Even More Lethal Than We Knew. AJP. 2016 Nov 1;173(11):1094–1100.
49. Tanguturi Y, Bodic M, Taub A, Homel P, Jacob T. Suicide risk assessment by residents: Deficiencies of documentation. Academic Psychiatry. Springer; 2017;41:513–519.
50. Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, Walsh CG, Brent DA. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA open. Oxford University Press; 2021;4(1):ooab011.
51. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence. Nature Publishing Group UK London; 2019;1(5):206–215.
52. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019 Dec;17(1):195.
53. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May;521(7553):436–444. PMID: 36774395
54. Sawhney R, Joshi H, Gandhi S, Shah R. A time-aware transformer based model for suicide ideation detection on social media. Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) [Internet]. 2020 [cited 2024 Jul 31]. p. 7685–7697. Available from: https://aclanthology.org/2020.emnlp-main.619/
55. Hsieh TY, Wang S, Sun Y, Honavar V. Explainable Multivariate Time Series Classification: A Deep Neural Network Which Learns to Attend to Important Variables As Well As Time Intervals. Proceedings of the 14th ACM International Conference on Web Search and Data Mining [Internet]. Virtual Event Israel: ACM; 2021 [cited 2024 Jul 31]. p. 607–615. Available from: https://dl.acm.org/doi/10.1145/3437963.3441815
56. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, Melnikov A, Kliushkina N, Araya C, Yan S, Reblitz-Richardson O. Captum: A unified and generic model interpretability library for PyTorch. arXiv:200907896 [cs, stat] [Internet]. 2020 Sep 16 [cited 2021 Feb 11]; Available from: http://arxiv.org/abs/2009.07896
57. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances

in Neural Information Processing Systems 30 [Internet]. Curran Associates, Inc.; 2017 [cited 2019 Jun 9]. p. 4765–4774. Available from: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf PMID: 31050537
58. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [Internet]. arXiv; 2022 [cited 2024 Jul 31]. Available from: http://arxiv.org/abs/2203.05794

ReXamine-Global: A Framework for Uncovering Inconsistencies in Radiology Report Generation Metrics

Oishi Banerjee[1*] Agustina Saenz[1*] Kay Wu[1*] Warren Clements[2,†] Adil Zia[2,†]
Dominic Buensalido[2,†] Helen Kavnoudias[2,†] Alain S. Abi-Ghanem[3,†] Nour El Ghawi[3,†]
Cibele Luna[4,†] Patricia Castillo[5,†] Khaled Al-Surimi[6,†] Rayyan A. Daghistani[7,†] Yuh-Min Chen[8,†]
Heng-sheng Chao[8,†] Lars Heiliger[9,†] Moon Kim[9,†] Johannes Haubold[10,†]
Frederic Jonske[11,†] Pranav Rajpurkar[1]

[*]Equal Contribution.
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Department of Radiology, Alfred Health, Melbourne, Victoria, Australia
[3]Department of Diagnostic Radiology, American University of Beirut, Beirut, Lebanon
[4]Department of Radiology, University of Miami Miller School of Medicine, Miami, Florida, USA
[5]University of Miami / Jackson Memorial Hospital, Miami, Florida, USA
[6]Department of Healthcare Management, University of Doha for Science and Technology, Doha, Qatar
[7]Department of Medical Imaging, King Abdulaziz Medical City, Riyadh, Saudi Arabia
[8]Department of Chest Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, Republic of China
[9]Institute for AI in Medicine, University Hospital Essen, Essen, North Rhine-Westphalia, Germany
[10]Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, North Rhine-Westphalia, Germany
[11]Department of Medical Machine Learning, Institute of AI in Medicine, University Medicine Essen, Essen, North Rhine-Westphalia, Germany
[†]MAIDA Initiative Partners

Oishi Banerjee: oishi_banerjee@g.harvard.edu
Agustina Saenz: ads006@mail.harvard.edu
Kay Wu: kay.wu@medportal.ca

Given the rapidly expanding capabilities of generative AI models for radiology, there is a need for robust metrics that can accurately measure the quality of AI-generated radiology reports across diverse hospitals. We develop ReXamine-Global, a LLM-powered, multi-site framework that tests metrics across different writing styles and patient populations, exposing gaps in their generalization. First, our method tests whether a metric is undesirably sensitive to reporting style, providing different scores depending on whether AI-generated reports are stylistically similar to ground-truth reports or not. Second, our method measures whether a metric reliably agrees with experts, or whether metric and expert scores of AI-generated report quality diverge for some sites. Using 240 reports from 6 hospitals around the world, we apply ReXamine-Global to 7 established report evaluation metrics and uncover serious gaps in their generalizability. Developers can apply ReXamine-Global when designing new report evaluation metrics, ensuring their robustness across sites. Additionally, our analysis of existing metrics can guide users of those metrics towards evaluation procedures that work reliably at their sites of interest.

Keywords: radiology report generation; metrics; evaluation; generalization

1. Introduction

The capabilities of AI are rapidly expanding in the field of radiology, with recent generative AI models comprehensively interpreting all aspects of radiology images and describing them in sophisticated text reports [1, 2, 3, 4]. To compare models and efficiently track progress in this space, developers rely heavily on automatic metrics that can efficiently score AI-generated radiology reports, measuring the accuracy of their content. These metrics measure the similarity between AI-generated candidate reports and ground-truth, radiologist-written reports; a candidate is assumed to be high-quality when metrics show it is similar to the corresponding ground-truth report. However, there are concerns that scores from commonly used metrics may not accurately evaluate the content of AI-generated reports, thus providing a misleading impression of model performance [5]. Furthermore, automatic metrics have historically been used to evaluate models trained on and tested against reports from a handful of single-institution datasets [6, 7], and it is unclear whether they generalize well across diverse reports from external sites.

In our work, we developed ReXamine-Global, a method for testing potential metrics across different writing styles and patient populations and exposing gaps in their generalizability. Using ground-truth reports from diverse hospitals, our method tests whether metrics are prone to two possible failure modes. First, we test whether metrics are undesirably sensitive to reporting style. Specifically, we explore whether they provide different scores depending on whether AI-generated reports are stylistically similar to ground-truth reports (e.g. during internal validation, when the model is tested against a familiar distribution) or not (as might occur during external validation, when model is tested against an unfamiliar distribution). Second, we check whether metric scores correlate with expert scores, with the expectation that an ideal metric would rank candidate reports exactly as an expert would. Using reports from 6 hospitals in different countries, we applied ReXamine-Global to test the generalizability of 7 established metrics for evaluating AI-generated radiology reports, revealing flaws in existing metrics.

Our work makes two primary contributions:

(1) We introduced ReXamine-Global, a new method for testing how report evaluation metrics generalize across diverse writing styles and patient populations. When creating new report evaluation metrics, developers can apply our method to determine whether metrics are overly sensitive to report-writing style or otherwise prone to poor generalization.

(2) By applying ReXamine-Global to 7 existing metrics, we uncovered gaps in the generalizability

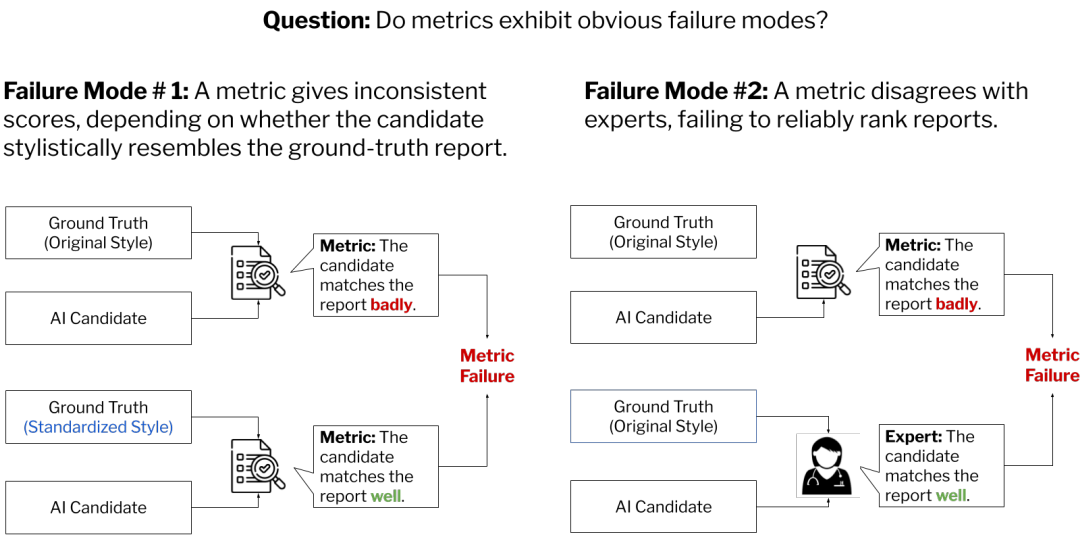


Fig. 1. ReXamine-Global tests how metrics generalize when used across distributions, with the goal of uncovering two failure modes. First, we test whether automatic metrics are undesirably sensitive to clinically irrelevant differences in report style, providing different scores depending on whether candidates are stylistically similar to the ground truths. Next, we test whether metrics disagree with expert scores, providing unreliable judgments at some sites. A successful metric would avoid both failure modes.

of many popular metrics, with a GPT-4-based metric outperforming all other approaches. These insights can help users of existing metrics design more reliable evaluation procedures for their sites of interest.

2. Methods

The ReXamine-Global Framework

We proposed a LLM-powered framework for testing how a report evaluation metric performs across different writing styles and patient populations:

(1) **Multi-site data collection:** Gather a diverse dataset of ground-truth reports from multiple hospitals, representing a range of patient populations and writing styles.

(2) **Standardization of ground-truth texts:** Use a large language model (LLM) to rewrite the original ground-truth reports in a standardized style, while preserving the original content.

(3) **Generation of error-containing 'candidate' texts:** Use a LLM to insert errors into standardized ground-truth reports. This step produces 'candidate' reports, representing outputs from an imperfect radiology report generation model.

(4) **Application of metric:** Use the metric to compare two pairs of reports: 1.) each candidate vs. its original ground-truth report (a stylistically different pair) and 2.) each candidate vs. its standardized ground-truth report (a stylistically similar pair).

(5) **Expert evaluation:** Engage clinical experts to manually evaluate the candidate reports, comparing them against ground-truth reports and counting the number of errors.

(6) **Assessment of metric consistency across styles:** Test whether, for any site, the metric produces significantly different scores for "candidate-original" pairs and "candidate-standardized" pairs. Ideally, a metric would always give a candidate the same score, regardless of whether it is being compared against the original or standardized ground-truth report.

Country	Example Reports	
Australia	ECMO catheter via inferior vena cava, tip in mid right atrium. Nasogastric tube in stomach. Left internal jugular central line tip in left brachiocephalic SVC junction. ETT 1 cm above carina. Left lower lobe collapse/consolidation. No pneumothorax or pleural effusion.	ETT and pacemaker position. ETT tip 4 cm from carina. Increased density in left hemithorax consistent with pleural fluid collection. No consolidation seen.
Germany	Rightly inserted endotracheal tube. Gastric tube subphrenically blanked out. Right transjugular CVC and sheath with tip projection to superior vena cava. New delineable sternal cerclages. Delineable clip material after mitral valve replacement. Progressive ateal confluent shading in left lung inferior field, mixed picture of pleural effusion and decreased ventilation. Increasing inferior ventilation in right lung subfield. Minor congestion signs. No pneumothorax.	Heart and mediastinum widened in supine position. Patchy shadowing bipulmonary, likely due to congestion, concomitant atypical infiltrates cannot be excluded by projection radiography. Clinical correlation required. No major pleural effusion. No pneumothorax delineable in supine position. Properly inserted endotracheal tube. Transjugular CVC on right side with tip projection to superior vena cava. Gastric tube ending in projection onto left upper abdomen.
Lebanon	Mild pulmonary edema. Cardiomegaly with cardiothoracic index of 0.57. No large pleural effusion or detectable pneumothorax. Single lead pacemaker with intact lead terminating in right ventricle topography. Chest wall intact.	Increase in left basal pleural effusion with overlying haziness likely related to basal atelectasis. Right basal atelectatic bands. Right lung otherwise clear. No detectable right pleural effusion. Cardiac silhouette is in size.
Saudi Arabia	Enlarged cardiac/pericardiac silhouette. Prominent central pulmonary vasculatures and bronchovascular markings suggest pulmonary congestion. Bilateral lower lung more of linear opacities may reflect atelectatic changes although infectious process not entirely excluded.	Left upper lobe atelectatic band otherwise unremarkable study.
Taiwan	Elevated right hemidiaphragm, tracheal deviated to Rt side. Right lung volume reduction is considered. Consolidation over right upper lung field, tumor growth cannot be r/o. R/o bullae over right lung apex	Consolidation over right hemithorax, cause to be determined. Lung consolidation change and/or pleural effusion cannot be r/o. Trachea slightly deviated to Rt side.
United States	IMPRESSION: Lines, tubes, etc: None. Cardiomediastinal silhouette: Within normal limits. Mediastinum midline. Lungs: Questionable subtle patchy right lower lung zone opacity which could represent an infectious process in the appropriate clinical setting, although limited due to overlying breast tissue summation. Pleura: Bilateral costophrenic angles sharp. No pneumothorax. Mild biapical pleural thickening/scarring. Bones/soft tissues: Unremarkable.	IMPRESSION: Intact median sternotomy wires. Scattered surgical clips projecting over heart. Cardiac silhouette top normal in size. Trachea and mediastinum midline. Mild tortuosity of descending thoracic aorta. Greater than expected density of midline lower mediastinum, could reflect hiatal hernia, other lower mediastinal pathology not entirely excluded. No significant edema. No airspace consolidation. Mild asymmetric elevation of right hemidiaphragm. No appreciable pleural effusion or pneumothorax, though lung apex clipped from field-of-view. No aggressive osseous lesion.

Table 1. Our dataset represents hospitals in 6 different countries, with reports that vary widely in content, terminology and organization. For example, the reports from Germany were automatically translated to English, resulting in atypical wording choices (e.g. "delineable", "ateal"). Reports from Taiwan heavily featured abbreviations (e.g. "Rt" for "right"), while reports from the United States were longer than average, frequently containing several subsections. Variations such as these can pose a challenge for automatic metrics.

(7) **Assessment of metric agreement with expert scores:** Test whether, for any site, the metric's scores fail to agree with expert scores. Ideally, metrics and experts will agree about which reports are the highest- and lowest-quality at every site, regardless of ground-truth style.

Using this framework, we assessed 7 existing automatic metrics for report evaluation.

Dataset

To apply ReXamine-Global, we sampled reports from a private dataset containing chest X-ray reports from around the world, with a focus on emergency departments and intensive care units. We selected

reports that were either originally written in or translated into English. We included data from 6 hospitals in 6 different countries: United States, Saudi Arabia, Taiwan, Australia, Germany, and Lebanon. We randomly sampled 40 reports from each hospital, resulting in a total dataset of 240 reports. We refer to these reports as "original ground-truth reports."

These radiology reports represent different patient populations as well as different writing styles, with marked differences in terminology, syntax and organization. For example, the reports from Germany were automatically translated to English, leaving artifacts that can prove challenging for automatic metrics. We give examples of these diverse reports in Table 1, which shows two examples from each site.

2.1. *Generation of Candidate Radiology Reports Using GPT-4*

After choosing 240 cases, we created 240 candidate reports, representing AI generations requiring evaluation. Our aim was to simulate outputs from an advanced but still flawed report generation model trained on MIMIC-CXR, a dataset widely used in the field [8]. We used GPT-4 to produce a candidate report based on each radiologist-written ground-truth report, using a two-step process described further in Appendix A:

(1) **Standardizing Style:** Initially, GPT-4 was tasked with rewriting the 'Findings' and 'Impression' sections of an original ground-truth report, using an example from MIMIC-CXR as a style guide. This step produced reports that preserved the original content but were written in a standardized, MIMIC-based style. A clinical expert checked 10 randomly sampled reports to ensure that this step did not meaningfully change report content. We refer to these reports as "standardized ground-truth reports."

(2) **Introducing Errors:** In the subsequent step, GPT-4 was instructed to deliberately introduce a few errors into the paraphrased report, thereby producing the final candidate report. We suggested several possible types of errors, such as the addition of a new finding, omission of an existing finding, or modification of the size or severity of a finding (Figure 2).

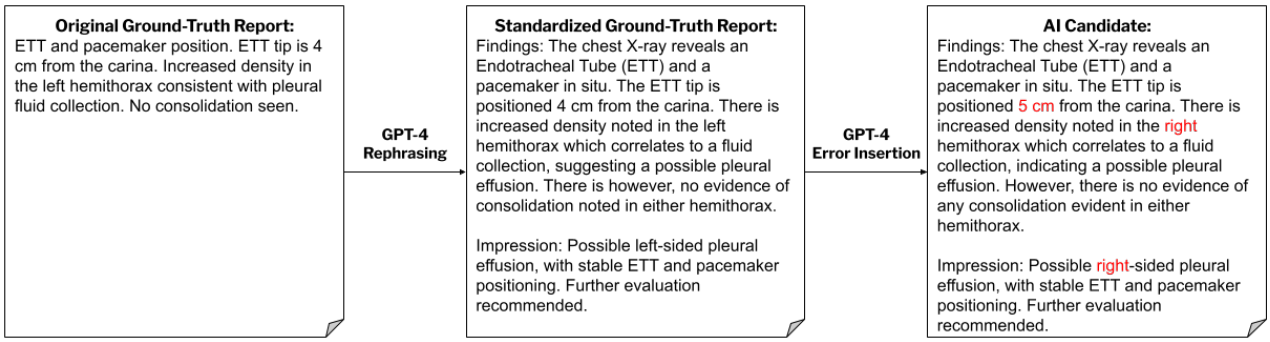


Fig. 2. Using GPT-4, we first standardized the style of the ground-truth reports and then introduced errors to create AI candidates. For details on our prompts, please see Appendix A.

2.2. *Automatic Metrics*

We examined seven existing automatic metrics used to judge the quality of AI-generated radiology reports. We included two general-purpose metrics that are not specialized for medical text: BLEU-2, which counts overlapping substrings in the ground-truth text and AI-generated text [9], and BERTScore, which computes the similarity of embeddings produced by passing each text through

a general-purpose BERT model [10] . Additionally, we considered clinical metrics such as CheXbert vector similarity, which compares the similarity of embeddings produced by passing each text through a specialized medical BERT model [11], and RadGraph-F1, which uses a specialized medical model to extract a graph of medical entities and relations from each text and measures the similarity of the graphs [12]. Additionally, we studied two versions of the RadCliQ metric, recently proposed specifically for evaluating AI-generated reports [5]. RadCliQ-v0 and RadCliQ-v1 both use a machine learning model to take in values from other metrics, such as BERTScore and CheXbert vector similarity, and then produce a composite score based on these input values. Finally, we considered FineRadScore, a recently proposed method that uses LLMs to perform a line-by-line comparison of ground-truth and candidate reports [13]. In our implementation of FineRadScore, we used GPT-4 to identify lines requiring corrections and treated the total number of problematic lines as the final score, which we refer to FineRadScore-GPT-4. We use implementations of these metrics from previously established repositories [14, 15].

2.3. *Expert Evaluation*

To obtain gold-standard measurements of candidate report quality, we conducted a manual evaluation engaging both an internal medicine attending physician and a radiology resident. The evaluation protocol was based on a scoring system adapted from the American College of Radiology [16] and from prior research studies [5], designed to assess the clinical significance of discrepancies in report interpretations. Errors were classified into seven independent categories: False prediction of finding; Omission of finding; Incorrect location of finding; Incorrect position of finding; Incorrect severity of finding. Mention of comparison that is not present in the reference impression; Omission of comparison describing a change from a previous study. We counted the total number of errors found in each report to produce our final expert score, so lower-quality candidates receive higher scores. For this study, each reviewer was assigned 120 unique reports, with an additional 10 reports each to assess inter-rater agreement.

2.4. *Experiments*

We used our 7 automatic metrics and expert evaluation to compare two types of report pairs: (1) the original ground-truth report vs. the AI candidate report, and (2) the standardized ground-truth report vs. the AI candidate. We assessed how automatic metrics performed on these comparisons using two approaches. First, we tested whether AI candidates received different scores when compared against the standardized ground-truth report rather than the original ground-truth report; we assume an ideal metric would be robust against clinically irrelevant stylistic variations and therefore give the same scores in both experiments. Second, we tested whether metric scores agreed with expert scores, as an ideal metric would provide the same ranking of a site's reports as experts do. These two approaches allowed us to compare how metrics behave when assessing reports with different styles (original ground truth vs. AI candidate) and reports with similar styles (standardized ground truth vs. AI candidate), as the standardized ground truth and AI candidate reports share a common GPT-4-generated style.

To facilitate interpretation of our results, we standardized the directionality of all automatic and human evaluation metrics, so that a higher score consistently indicates worse performance from the report generation model. Originally, higher scores for BLEU-2, BERTScore, CheXbert vector similarity, and RadGraph-F1 indicated better performance, while lower scores for RadCliQ and FineRadScore-GPT-4 indicated better performance. To align all metrics so a higher score indicates worse performance, we multiplied the scores of BLEU-2, BERTScore, CheXbert vector similarity, and RadGraph-F1 by -1. This standardization makes it easier to compare our results across different evaluation metrics.

We employed two main statistical approaches to study the behavior of automatic metrics across different countries and ground-truth styles. First, we conducted paired t-tests to determine whether automatic metrics provide different scores depending on whether original or standardized ground-truth reports are used. These tests were performed independently for each country to account for potential regional variations. To address the issue of multiple comparisons in our t-test analyses, we applied a Bonferroni correction to control the familywise error rate. The significance level α was set at 0.05, and the Bonferroni-corrected threshold was calculated as α/n, where n is the total number of paired t-tests conducted (number of metrics $\times$ number of countries = 42). Second, we calculated Spearman's rank correlation coefficients (ρ) to quantify the agreement between automatic metrics and human evaluations for each country. This analysis was performed separately when using original and standardized ground-truth reports, allowing us to assess how well our automatic metrics aligned with human judgments across different ground-truth styles and geographical regions.

3. Results

3.1. *Effect of Stylistic Differences on Metric Scores*

We found that stylistic differences significantly impacted scores from all metrics, with the exception of FineRadScore-GPT-4. Across all non-GPT metrics and countries, paired t-tests revealed significant differences in scores depending on whether original or standardized ground-truth reports were used (Bonferroni-corrected $p < 0.05$) (Table 2). BERTScore showed the highest mean t-statistics across all countries (mean t-stat = -29.72, range: -17.24 to -37.09), indicating a substantial and consistent difference in scores between the two report styles. FineRadScore-GPT-4 exhibited the smallest t-statistics (mean t-stat = -1.07, range: -1.50 to -0.42) and was the only metric that did not show significant differences for any country after Bonferroni correction. All t-statistics were negative, indicating that comparisons between standardized ground truth reports and AI candidates consistently yielded lower scores (i.e. indicating higher-quality AI candidates) compared to comparisons between original ground-truth reports and AI candidates. In other words, metrics rated the AI model as better-performing when the ground truth stylistically resembled the AI candidate. More details on the distribution of metric and expert scores can be found in Appendix B.

Metric	Mean t-stat	Min t-stat	Max t-stat	Significant Countries
BLEU-2 [9]	-27.23	-31.01	-20.60	6
BERTScore [10]	-29.72	-37.09	-17.24	6
CheXbert Similarity [11]	-6.29	-8.15	-3.97	6
RadCliQ-v0 [5]	-20.50	-30.08	-11.20	6
RadCliQ-v1 [5]	-22.23	-32.37	-12.77	6
RadGraph-F1 [12]	-13.66	-19.18	-9.65	6
FineRadScore-GPT-4 [13]	**-1.07**	**-1.50**	**-0.42**	**0**

Table 2. Negative t-statistics indicate that standardized ground truth-AI candidate pairs (similar styles) consistently received lower scores than original ground truth-AI candidate pairs (different styles). The magnitude of the t-statistic reflects the strength of this difference. The "Mean" value gives the average t-statistic across all 6 countries, while the "Min" and "Max" t-stat values show the lowest and highest values seen across the 6 countries. The "Significant Countries" column indicates the number of countries (out of 6) where the metric showed a significant difference between ground truth-AI candidate and standardized ground truth-AI candidate pairs after Bonferroni correction. FineRadScore-GPT-4 is the only metric whose scores were not significantly affected by the ground-truth style.

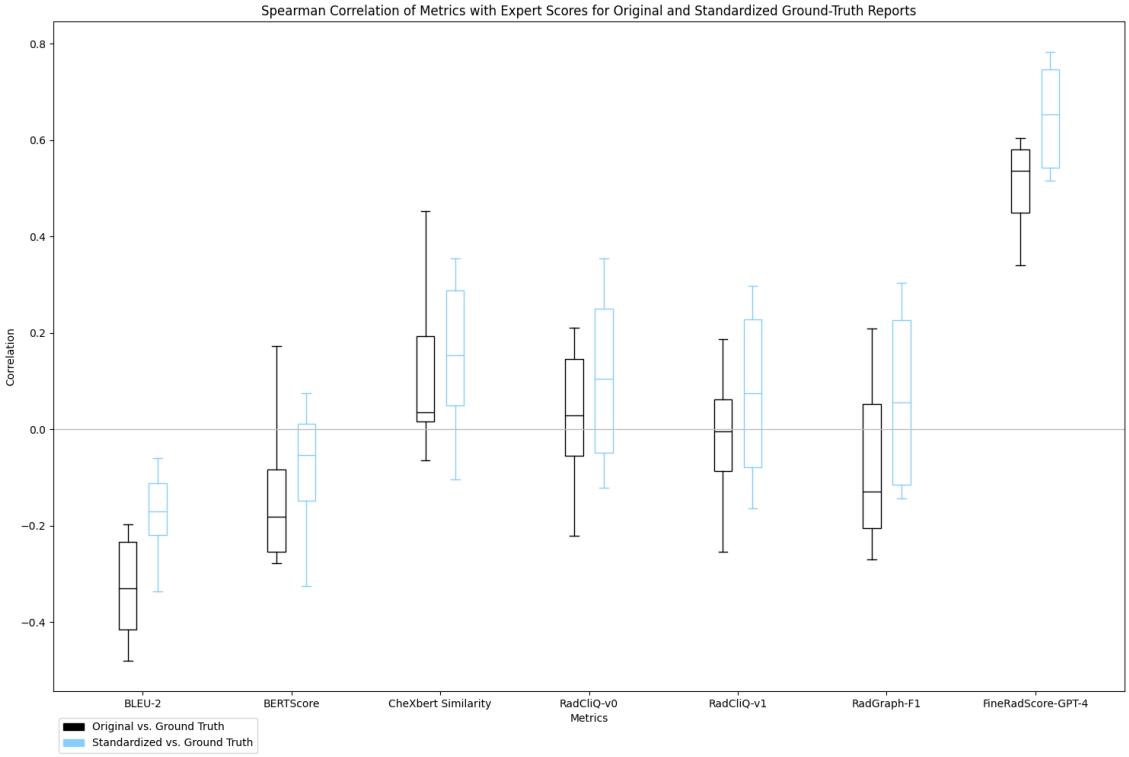


Fig. 3. Except for FineRadScore-GPT-4, no metric achieved positive Spearman correlations with expert scores at every site, indicating poor generalization. Correlations for original ground-truth reports are shown in the black box plots (left). Correlations for standardized ground-truth reports are shown in blue box plots (right). Metrics typically achieved higher performance with standardized ground-truth reports than original ground-truth reports. For detailed numerical results, see the table in Appendix C.

3.2. *Correlation of Automatic Metrics with Expert Scores on Stylistically Diverse Reports*

When comparing original ground-truth reports against stylistically different candidates, metrics frequently failed to align with experts (Figure 3). FineRadScore-GPT-4, the only metric using a LLM, offered the best performance, with coefficients ranging from ($\rho = 0.34$ to 0.60). Despite achieving positive correlations at some sites, each of the other metrics had negative coefficients for at least one site. BLEU-2 showed especially poor performance, with Spearman's rank correlation coefficients (ρ) ranging from $\rho = -0.20$ to -0.48.

3.3. *Correlation of Automatic Metrics with Expert Scores on Stylistically Standardized Reports*

After standardizing ground-truth reports to resemble the style of the candidates, metrics generally showed better agreement with experts (Figure 3). For example, FineRadScore-GPT-4's coefficients rose across all sites, now ranging from $\rho = 0.52$ to 0.78. Despite similar increases, every other metric still had a negative coefficient for at least one site, suggesting that metrics can fail to generalize even after standardization. Notably, BLEU-2's correlation coefficients remained consistently negative even after standardization, ranging from $\rho = -0.34$ to -0.06.

4. Discussion

ReXamine-Global, which tests report evaluation metrics across diverse distributions, successfully revealed critical gaps in metric generalizability. By applying ReXamine-Global to 7 existing metrics, we found that most automatic metrics are undesirably sensitive to stylistic differences, giving significantly different scores depending on the style of the ground-truth report. The only exception was FineRadScore-GPT-4, which used a powerful LLM to evaluate reports [13]. Furthermore, we observed that automatic metrics of all kinds demonstrated, at best, moderate correlation with expert opinions when using original ground-truth reports. Metrics generally attained better correlations when comparing candidates against standardized ground-truth reports, opening the possibility that preprocessing candidates and ground-truth reports to make them stylistically similar can improve evaluation procedures. Importantly, we observed that metric behavior sometimes varied across hospitals; for example, CheXbert Similarity's correlations when comparing candidates and original ground-truth reports ranged from -0.065 to 0.45. This finding shows the importance of including data from a range of diverse hospitals.

The clear variability in metric performance across sites highlights important directions for future work. ReXamine-Global automatically identifies extreme failure cases, surfacing candidate-report pairs that could benefit from metric-specific, qualitative analysis to reveal concrete mechanisms behind metric failure. We provide an example of such a qualitative analysis in Appendix D, manually reviewing reports to identify specific scenarios where BLEU-2 and RadGraph-F1 perform poorly. Furthermore, ReXamine-Global can guide the development of more robust report evaluation metrics, capable of generalizing effectively across diverse healthcare settings. We also hope our work can warn users about the risks of naively applying metrics to new distributions and help them choose high-performing metrics for their specific sites of interest.

4.1. *Limitations*

While we utilized GPT-4 to generate standardized ground-truth and candidate reports, candidate reports generated by other models may elicit different behavior from metrics, so a metric that performs well on ReXamine-Global may generalize poorly to some other distribution of generated reports. In addition, our manual evaluation scoring system did not encompass all possible error categories, potentially overlooking some types of inaccuracies, and our evaluation was conducted by only two physicians, which significantly limits the breadth and diversity of expert assessment. We also assumed that the same number of errors is present regardless of whether the candidate is compared against the original ground truth or the standardized ground truth, though it is possible that errors were occasionally added or removed by GPT-4 during standardization. These constraints may have introduced bias and reduced the robustness of our manual evaluation results. Ideally, each candidate-report pair would be reviewed by multiple physicians from diverse specialties, with a third reviewer to resolve discrepancies. This approach would provide a more comprehensive and reliable assessment of report quality and error identification. A larger pool of reviewers would also make it possible to conduct inter-rater reliability analyses, which could confirm the reliability of manual evaluation.

5. Institutional Review Board (IRB)

All data was obtained with approval from Institutional Review Board (IRB) and Data Use Agreement (DUA) protocols.

6. Data Contribution

Authors who are MAIDA Initiative Partners made substantial contributions to data collection for this study. More information about the MAIDA initiative can be found at

https://www.rajpurkarlab.hms.harvard.edu/maida.

7. Acknowledgments

We thank Xiaoman Zhang for figure design assistance, Wendy Erselius for managing data collection and Jonathan Dreyfuss for statistical analysis support. This work was supported by Biswas Family Foundation's Transformative Computational Biology Grant in Collaboration with the Milken Institute.

8. Appendices

Appendix A. GPT-4 Instructions

We gave GPT-4 the following instructions when standardizing the style of our original ground-truth reports:

```
Pretend you are a radiologist and format the content of these notes in a
polished findings and impressions section. Your findings section may be
long or short. Your impression should only have 1-3 lines. If you are
unsure about an abbreviation, term, or other odd phrasing, make your best
guess. Match the style of this radiology report:

Report:
Findings: Single frontal view of the chest demonstrates a right
Port-A-Cath in unchanged position, terminating at the cavoatrial junction.
Median sternotomy wires are present, along with surgical clips in the left
upper quadrant.  The heart is mildly enlarged, but stable compared with
prior examinations, with redemonstration of calcified mediastinal lymph
nodes. A rounded opacity in the lower left lung likely correlates to a
calcified granuloma as seen on CT of the chest from ___.  There is no
evidence of pneumonia, pleural effusion, pneumothorax or overt pulmonary
edema.  The lung volumes are low, accentuating bibasilar atelectasis.  No
subdiaphragmatic free air is present.

Impression: No subdiaphragmatic free air or other acute cardiopulmonary
process.
```

After standardizing the style of our reports, we used the following instructions to introduce errors, producing the final candidate:

```
Please write a report using the above report as a template. Perturb the
content of a few existing lines. Here are some examples of how a line
could be changed:
- If the report says X condition is present, state that X condition is
absent.
- If the report rules out X condition, state that X condition is present.
- Change the location, size, severity, or implications of a condition.

Only perturb a few lines. Keep the other lines exactly the same. Your
report should still sound fluent, like a radiologist wrote it.
```

Appendix B. Distribution of Metric and Expert Scores

This table gives more details on metric and expert scores per country. On average, metrics gave lower scores when comparing AI candidates to standardized ground-truth reports, rather than to original ground-truth reports.

Metric	Ground Truth	Australia	Lebanon	Taiwan	Saudi Arabia	United States	Germany
BLEU-2	Original	−0.23 ± 0.10	−0.25 ± 0.09	−0.17 ± 0.06	−0.13 ± 0.07	−0.24 ± 0.07	−0.20 ± 0.06
	Standardized	−0.70 ± 0.13	−0.69 ± 0.11	−0.72 ± 0.11	−0.70 ± 0.12	−0.74 ± 0.13	−0.69 ± 0.13
BERTScore	Original	−0.47 ± 0.09	−0.52 ± 0.08	−0.41 ± 0.08	−0.43 ± 0.15	−0.49 ± 0.08	−0.44 ± 0.06
	Standardized	−0.87 ± 0.07	−0.86 ± 0.06	−0.86 ± 0.06	−0.87 ± 0.06	−0.87 ± 0.08	−0.85 ± 0.08
CheXbert Similarity	Original	−0.69 ± 0.19	−0.64 ± 0.14	−0.70 ± 0.19	−0.57 ± 0.24	−0.66 ± 0.18	−0.65 ± 0.19
	Standardized	−0.83 ± 0.17	−0.78 ± 0.15	−0.83 ± 0.14	−0.78 ± 0.16	−0.78 ± 0.19	−0.74 ± 0.18
RadCliQ-v0	Original	2.31 ± 0.65	2.09 ± 0.46	2.45 ± 0.52	2.64 ± 0.96	2.29 ± 0.61	2.55 ± 0.50
	Standardized	0.83 ± 0.57	0.88 ± 0.48	0.77 ± 0.42	0.83 ± 0.57	0.83 ± 0.64	1.01 ± 0.47
RadCliQ-v1	Original	0.47 ± 0.41	0.30 ± 0.30	0.57 ± 0.32	0.70 ± 0.59	0.45 ± 0.39	0.64 ± 0.31
	Standardized	−0.61 ± 0.39	−0.59 ± 0.32	−0.66 ± 0.27	−0.63 ± 0.39	−0.65 ± 0.43	−0.51 ± 0.32
RadGraph-F1	Original	−0.41 ± 0.12	−0.52 ± 0.10	−0.40 ± 0.11	−0.39 ± 0.17	−0.44 ± 0.13	−0.36 ± 0.11
	Standardized	−0.65 ± 0.13	−0.68 ± 0.11	−0.69 ± 0.09	−0.69 ± 0.17	−0.71 ± 0.13	−0.66 ± 0.12
FineRadScore-GPT-4	Original	4.15 ± 1.00	3.73 ± 1.34	4.80 ± 1.51	3.60 ± 1.58	4.88 ± 1.68	4.60 ± 1.61
	Standardized	3.92 ± 1.35	3.65 ± 1.10	4.47 ± 1.47	3.33 ± 1.42	4.58 ± 1.52	4.35 ± 1.44
Expert Errors	Both	3.48 ± 1.71	3.15 ± 1.31	3.60 ± 1.45	2.38 ± 1.31	4.05 ± 1.50	3.65 ± 1.44

Table 3: Means and standard deviations of metrics and expert scores.

Appendix C. Full Correlation Results

This table gives detailed results about how metric scores were correlated with expert scores, across sites and ground-truth report styles.

Metric	Ground Truth	Australia	Lebanon	Taiwan	Saudi Arabia	United States	Germany
BLEU-2	Original	-0.48	-0.44	-0.35	-0.20	-0.31	-0.21
	Standardized	-0.10	-0.34	-0.06	-0.20	-0.23	-0.15
BERTScore	Original	-0.26	-0.28	-0.07	-0.25	-0.11	0.17
	Standardized	0.07	-0.33	-0.02	0.02	-0.17	-0.08
CheXbert Similarity	Original	0.24	-0.06	0.45	0.03	0.01	0.04
	Standardized	0.36	-0.10	0.30	0.24	0.06	0.04
RadCliQ-v0	Original	0.06	-0.22	0.17	-0.00	-0.07	0.21
	Standardized	0.35	-0.12	0.25	0.25	-0.05	-0.04
RadCliQ-v1	Original	0.00	-0.25	0.08	-0.01	-0.11	0.19
	Standardized	0.30	-0.16	0.24	0.20	-0.09	-0.05
RadGraph-F1	Original	-0.06	-0.27	-0.21	0.09	-0.20	0.21
	Standardized	0.17	-0.14	0.30	0.24	-0.13	-0.06
FineRadScore-GPT-4	Original	0.56	0.59	0.51	0.60	0.43	0.34
	Standardized	0.78	0.76	0.62	0.52	0.52	0.69

Table 4: Spearman correlations between metric and expert scores.

Appendix D. Failure Modes of BLEU-2 and RadGraph-F1

By examining extreme failure cases where metrics gave particularly incorrect scores, we were able to identify some qualitative reasons for metric failure. We focused on BLEU-2 and RadGraph-F1, which represent the overlap between short phrases and clinically relevant keywords respectively, and are more readily interpretable than other embedding-based approaches. To understand how BLEU-2 can fail, we studied reports from Australia, where BLEU-2 achieved its lowest correlation of -.48. To understand how RadGraph-F1 can fail, we studied reports from Lebanon, where RadGraph-F1 achieved its lowest correlation of -.27.

In Table 5, we looked at cases where metrics marked reports as being high-quality, though experts found many errors in them. We observed that these errors tended to affect only one or two words in the report (e.g. "left" vs. "right", "4.5" vs. "5"). As a result, these errors barely reduced the overall scores from these metrics, which directly measure word or keyword overlap. Embedding or LLM-based approaches may have been more capable of recognizing that these particular mismatches had an outsized impact on the meaning of the report, even though they affected fairly few words.

Original Ground-Truth Report	**AI Candidate**	**Metric Score**	**Expert Score**
The ETT is **4.5** cm above the carina. A catheter enters via **right** internal jugular approach, looped in the lower portion of the right atrium and has its tip in the right ventricle. This position is unusual and clinical correlation suggested. Mediastinal and **left** basal chest drains noted. Nasogastric tube has its tip in the body of the stomach. The lungs are **mildly** congested. There is minor **left** basal atelectasis. Minor blunting of the bilateral costophrenic angles is in keeping with small pleural effusions. **No pneumothorax**.	Report: Findings: Assessment of the chest X-ray reveals an endotracheal tube positioned **5** cm above the carina. A catheter is noted to enter through the **left** internal jugular vein looping into the lower right atrium and ending in the right ventricle, an unusual placement. Mediastinal and **right** basal chest drains are evident. Nasogastric tube placement is accurate with its tip located in the body of the stomach. Radiographic evidence of **moderate** bilateral lung congestion is observed along with minor presence of **right** basal atelectasis. There is minor blunting of the bilateral costophrenic angles suggestive of small pleural effusions. A **small pneumothorax** is identified. Impression: Unusual catheter positioning, presence of minor right basal atelectasis, small bilateral pleural effusions, and **identified pneumothorax**. Clinical correlation recommended.	BLEU-2: -.42	6
No pleural effusion or detectable pneumothorax. The cardiac silhouette is **enlarged** with a cardiothoracic ratio of **0.6**. A coronary stent is noted There is an impression of a rounded area of shadowing projecting over the mid aspect of the cardiac silhouette on the PA projection and retrocardiac space on the lateral projection, that may be related to a hiatal hernia, **less** likely a pulmonary opacity. The rest of the lungs are clear. No overt pulmonary oedema. No displaced rib fracture. There is **exaggerated** kyphosis of the thoracic spine with mild central loss of height of few of the visualised vertebral bodies mainly in the **mid** dorsal spine. Metallic hardware is noted along the **left** humeral head and proximal shaft. Comparison to previous studies if available is suggested	Report: Findings: The X-Ray shows no discernible pneumothorax or pleural effusion. The cardiac silhouette is **normal-sized** with a cardiothoracic ratio of **0.4**, and a coronary stent is present. A pronounced area of shadowing is observed over the mid aspect of the cardiac silhouette on the PA projection and retrocardiac space on the lateral projection. This may be related to a hiatal hernia or **more** likely, a pulmonary opacity. No signs of pulmonary oedema or rib fractures are detected. The rest of the lungs appear clear. A **mild** kyphosis of the thoracic spine is noted with significant loss of height in several of the visualised vertebral bodies, mainly in the **upper** dorsal spine. Metallic evidence is seen along the **right** humeral head and proximal shaft. Impression: Normal cardiac size and probable pulmonary opacity noted. Suggest comparison with previous studies for further evaluation.	RadGraph-F1:-.62	5

Table 5. These candidates received low metric scores despite containing many errors. Errors are highlighted in bold. Scores for these metrics have been multiplied by -1, so higher scores indicate worse performance.

In Table 6, we examined cases where metrics marked reports as being low-quality, though experts found few errors. Here, we found that stylistic differences in how normal findings were described substantially impacted BLEU-2's performance. In the first example, the predicted report was much

longer than the ground-truth report, in part because it was more explicit in ruling out abnormalities ("no additional abnormalities", "no other acute cardiopulmonary conditions"). This discrepancy reflects realistic variation among radiologists. Some radiologists may prefer concise reports which imply that unmentioned abnormalities are absent, while other radiologists write longer reports that clearly rule out many abnormalities. Unfortunately, BLEU-2 performed poorly when facing this stylistic diversity and found little overlap due to the substantial difference in wording, even though the ground truth and candidate conveyed nearly identical content.

RadGraph-F1 failed due to even subtler differences in writing style. The RadGraph pipeline extracts keywords and labels them based on their function in the report, and the RadGraph-F1 score quantifies how often the labeled keywords from the ground-truth and candidate reports match. While the names of abnormalities and anatomical locations stayed largely consistent between the ground truth and AI candidate, other differences in wording led to markedly different sets of keywords (e.g. "within", "acceptable", and "range" vs. "normal"), causing RadGraph-F1 to indicate that these reports do not match well. Additionally, we observed discrepancies in how the RadGraph model labeled keywords across stylistically different reports. For example, "infiltrates" was labeled as being "located at 'pneumonic'" in the ground-truth report, yet it was labeled as "modifying 'pneumonic'" in the AI candidate, an inconsistency that further harmed RadGraph-F1's performance.

Original Ground-Truth Report	**AI Candidate**	**Metric Score**	**Expert Score**
Bilateral **lower** hilar peribronchial thickening in this setting is likely to represent bronchitis.	Report: Findings: The scrutinized chest X-ray exhibits bilateral **mid** hilar peribronchial thickening. The pattern of these findings is frequently seen in instances of bronchitis. No additional abnormalities or significant changes in comparison to preceding examinations are identified. Impression: The noted markers are suggestive of bronchitis. No other acute cardiopulmonary conditions have been detected.	BLEU-2: -.09	1
Impression **No consolidation or definite pneumonic infiltrates**. No pneumothorax or pleural effusion. The cardiomediastinal silhouette is normal. The visualised bones are unremarkable.	Report: Findings: Single frontal view of the chest shows **evidence of consolidation and immediate pneumonic infiltrates**. There are no signs of pneumothorax or pleural effusion. The cardiomediastinal silhouette falls within the acceptable range. Evaluation of the visible bones does not reveal any conspicuous anomalies. Impression: **Chest X-ray exhibits acute cardiopulmonary changes** and still does not present any bone abnormalities.	RadGraph-F1:-.35	1

Table 6. These candidates received high metric scores despite containing almost no errors. Errors are highlighted in bold. Scores for these metrics have been multiplied by -1, so higher scores indicate worse performance.

References

[1] Hong-Yu Zhou et al. *A Generalist Learner for Multifaceted Medical Image Interpretation.* 2024. arXiv: 2405.07988 [cs.CV]. URL: https://arxiv.org/abs/2405.07988.

[2] Stephanie L Hyland et al. "MAIRA-1: A Specialised Large Multimodal Model for Radiology Report Generation". In: *arXiv preprint arXiv:2311.13668* (2023). URL: http://arxiv.org/abs/2311.13668.

[3] Tim Tanida et al. "Interactive and Explainable Region-Guided Radiology Report Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2023. URL: http://openaccess.thecvf.com/content/

REFERENCES

CVPR2023 / html / Tanida _ Interactive _ and _ Explainable _ Region - Guided _ Radiology_Report_Generation_CVPR_2023_paper.html.

[4] Tao Tu et al. "Towards Generalist Biomedical AI". In: *arXiv preprint arXiv:2307.14334* (2023). URL: http://arxiv.org/abs/2307.14334.

[5] Feiyang Yu et al. "Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation". In: *Patterns* (2023). URL: https://doi.org/10.1016/j.patter.2023.100802.

[6] Alistair E W Johnson et al. "MIMIC-CXR, a de-Identified Publicly Available Database of Chest Radiographs with Free-Text Reports". In: *Scientific Data* 6.1 (2019), p. 317.

[7] Dina Demner-Fushman et al. "Preparing a Collection of Radiology Examinations for Distribution and Retrieval". In: *Journal of the American Medical Informatics Association: JAMIA* 23.2 (2016), pp. 304–310.

[8] Alistair E. W. Johnson et al. *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. 2019. arXiv: 1901.07042 [cs.CV]. URL: https://arxiv.org/abs/1901.07042.

[9] Kishore Papineni et al. "Bleu: A Method for Automatic Evaluation of Machine Translation". In: *Annual Meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[10] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *arXiv preprint arXiv:1904.09675* (2019). URL: http://arxiv.org/abs/1904.09675.

[11] Akshay Smit et al. *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT*. 2020. arXiv: 2004.09167 [cs.CL]. URL: https://arxiv.org/abs/2004.09167.

[12] Saahil Jain et al. *RadGraph: Extracting Clinical Entities and Relations from Radiology Reports*. 2021. arXiv: 2106.14463 [cs.CL]. URL: https://arxiv.org/abs/2106.14463.

[13] Alyssa Huang et al. *FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores*. 2024. arXiv: 2405.20613 [cs.CL]. URL: https://arxiv.org/abs/2405.20613.

[14] Feiyang Yu. *Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation*. https://github.com/rajpurkarlab/CXR-Report-Metric. 2023.

[15] Alyssa Huang. *FineRadScore*. https://github.com/rajpurkarlab/FineRadScore. 2024.

[16] Shlomit Goldberg-Stein et al. "ACR RADPEER Committee White Paper with 2016 Updates: Revised Scoring System, New Classifications, Self-Review, and Subspecialized Reports". In: *Journal of the American College of Radiology: JACR* 14.8 (2017), pp. 1080–1086.

Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions

Guangzhi Xiong[1,*], Qiao Jin[2,*], Xiao Wang[3], Minjia Zhang[3], Zhiyong Lu[2,†], Aidong Zhang[1,†]

[1]*Department of Computer Science, University of Virginia, VA 22904, USA*
[2]*National Library of Medicine, National Institutes of Health, MD 20892, USA*
[3]*Department of Computer Science, University of Illinois Urbana-Champaign, IL 61801, USA*
[*]*Equal contribution.* [†]*Co-correspondence.*
E-mail: hhu4zu@virginia.edu, qiao.jin@nih.gov, xiaow4@illinois.edu, minjiaz@illinois.edu, zhiyong.lu@nih.gov, aidong@virginia.edu

The emergent abilities of large language models (LLMs) have demonstrated great potential in solving medical questions. They can possess considerable medical knowledge, but may still hallucinate and are inflexible in the knowledge updates. While Retrieval-Augmented Generation (RAG) has been proposed to enhance the medical question-answering capabilities of LLMs with external knowledge bases, it may still fail in complex cases where multiple rounds of information-seeking are required. To address such an issue, we propose iterative RAG for medicine (*i*-MedRAG), where LLMs can iteratively ask follow-up queries based on previous information-seeking attempts. In each iteration of *i*-MedRAG, the follow-up queries will be answered by a vanilla RAG system and they will be further used to guide the query generation in the next iteration. Our experiments show the improved performance of various LLMs brought by *i*-MedRAG compared with vanilla RAG on complex questions from clinical vignettes in the United States Medical Licensing Examination (USMLE), as well as various knowledge tests in the Massive Multitask Language Understanding (MMLU) dataset. Notably, our zero-shot *i*-MedRAG outperforms all existing prompt engineering and fine-tuning methods on GPT-3.5, achieving an accuracy of 69.68% on the MedQA dataset. In addition, we characterize the scaling properties of *i*-MedRAG with different iterations of follow-up queries and different numbers of queries per iteration. Our case studies show that *i*-MedRAG can flexibly ask follow-up queries to form reasoning chains, providing an in-depth analysis of medical questions. To the best of our knowledge, this is the first-of-its-kind study on incorporating follow-up queries into medical RAG.

Keywords: Large Language Models; Retrieval-Augmented Generation; Medical Question Answering; AI for Healthcare.

1. Introduction

Generative artificial intelligence (AI) technologies such as large language models (LLMs) have brought a wide variety of opportunities for biomedical applications.[1–4] For example, they have shown great potential for answering biomedical questions,[5–9] summarizing medical documents,[10–12] and matching patients to clinical trials.[13–16] However, LLMs often generate plausible-sounding but inaccurate content, an issue commonly known as "hallucination" in the literature.[17] They also possess outdated knowledge obtained from a fixed set of training data.[18] Retrieval-augmented generation (RAG) provides a lightweight post-training solution to these issues by providing LLMs with relevant documents retrieved from up-to-date and trustworthy sources.[19,20]

While there have been several medical applications of RAG, such as Almanac,[21] Clinfo.ai,[22] and MedRAG,[23] their RAG component is mainly beneficial to questions that have direct answers in a single document, such as those in the PubMedQA[24] and BioASQ[25] datasets. However, only marginal improvements are seen with RAG for questions that require multiple rounds of clinical reasoning like MedQA,[26] a dataset curated from medical license examinations. For example, to recommend a treatment for a patient with certain symptoms, a system needs to first infer the potential diagnosis from the symptoms and then find a suitable treatment for the diagnosis. Nevertheless, only one round of retrieval is conducted in the vanilla RAG architecture, prohibiting multiple rounds of information seeking that are required in complex clinical reasoning.

In this work, we propose *i*-MedRAG, a simple and effective framework for incorporating follow-up queries into RAG. Specifically, we prompt LLMs to iteratively generate follow-up queries to search for additional information from external medical corpora. The queries and the corresponding answers generated with RAG will be used to augment the answer generation of the original question. Empirical results demonstrate the effectiveness of *i*-MedRAG on both open- and close-source LLMs, which show improved performance on the United States Medical Licensing Examination (USMLE) subset of MedQA and medical questions from the Massive Multitask Language Understanding (MMLU) dataset. Our further analysis of the number of iterations and number of queries per iteration used in *i*-MedRAG reflects how its performance scales with different settings. Additionally, we present several case studies of *i*-MedRAG, showing how it overcomes the limitations in vanilla RAG to find the correct answers.

In summary, our contributions are three-fold:

- We introduce *i*-MedRAG, a novel RAG architecture that incorporates follow-up queries to solve complex reasoning tasks.
- We have conducted comprehensive experiments on medical question answering, and the results demonstrate that *i*-MedRAG not only outperforms vanilla RAG approaches but also surpasses all other prompt engineering approaches on MedQA with GPT-3.5, setting a new state-of-the-art performance of 69.68%.
- We also provide analyses to further characterize *i*-MedRAG, showing how its performance varies with the scaling of follow-up queries.

2. Related Work

2.1. *Retrieval-Augmented Generation for Medicine*

Retrieval-augmented generation (RAG) has been widely adopted in medicine. Here, we discuss several representative approaches. Almanac[21] is a system that augments LLMs with curated resources for medical guidelines and treatment recommendations, which shows improvements over the standard LLMs in six manually assessed metrics. Similarly, Low *et al.*[27] demonstrate the improvements of RAG-based systems for real-world clinical queries with manual evaluation. Clinfo.ai[22] is an open-source web application that answers clinical questions based on retrieved scientific literature from PubMed articles. Xiong *et al.*[23] conduct a benchmarking study with the MedRAG toolkit, and show the benefits of RAG in several medical multi-choice question answering datasets. There are also various biomedical literature search products[28] that use RAG to summarize the retrieved articles,[29] such as OpenEvidence[a] and ChatRWD[b]. However, most of the RAG studies in medicine use the vanilla architecture with only one round of retrieval. There have been several attempts to use iterative data refinement for LLM training[30] or RAG[31–33] in the general domain. Nevertheless, similar ideas have not yet been explored in the medical domain. To the best of our knowledge, our study presents the first approach and evaluations on incorporating follow-up queries in RAG for medicine.

2.2. *Medical Question Answering*

Question answering tasks such as MedQA,[26] PubMedQA,[24] MedMCQA,[34] BioASQ,[25] and Massive Multitask Language Understanding (MMLU)[35] are commonly used to benchmark the medical knowledge and reasoning capabilities of LLMs.[36] Most of these datasets focus on single-hop questions such as "what is the most common symptom of hypertension?", while only MedQA questions are longer patient vignettes where both medical knowledge and multi-step reasoning are required. As such, there have been many studies working on improving the GPT-3.5 performance on MedQA with prompt engineering. Figure 1 shows the comparison among different representative prompt engineering approaches on MedQA, including chain-of-thought (CoT) prompting,[37] self-consistency (SC) prompting,[38] multi-agent communication with MedAgents,[39] and RAG-based approaches such as Knowledge Solver (KSL),[40] LLMs Augmented with Medical Textbooks (LLM-AMT),[41] and MedRAG.[23] Much fewer studies focus on prompt engineering with GPT-4 on MedQA,[7,42] probably because the raw GPT-4 error rate[43] is close to the noise rate in MedQA annotations.[44] In this study, we focus on the zero-shot setting as it reflects realistic clinical scenarios. While not requiring any instances for training or few-shot learning, our approach surpasses all previous methods with GPT-3.5 on the MedQA dataset.

3. Methods

Figure 2 shows the overview of our *i*-MedRAG and its comparison to the vanilla Retrieval-Augmented Generation (RAG). Different from RAG, our *i*-MedRAG modifies its pipeline by

[a] `https://www.openevidence.com/`

[b] `https://www.atroposhealth.com/chatrwd`

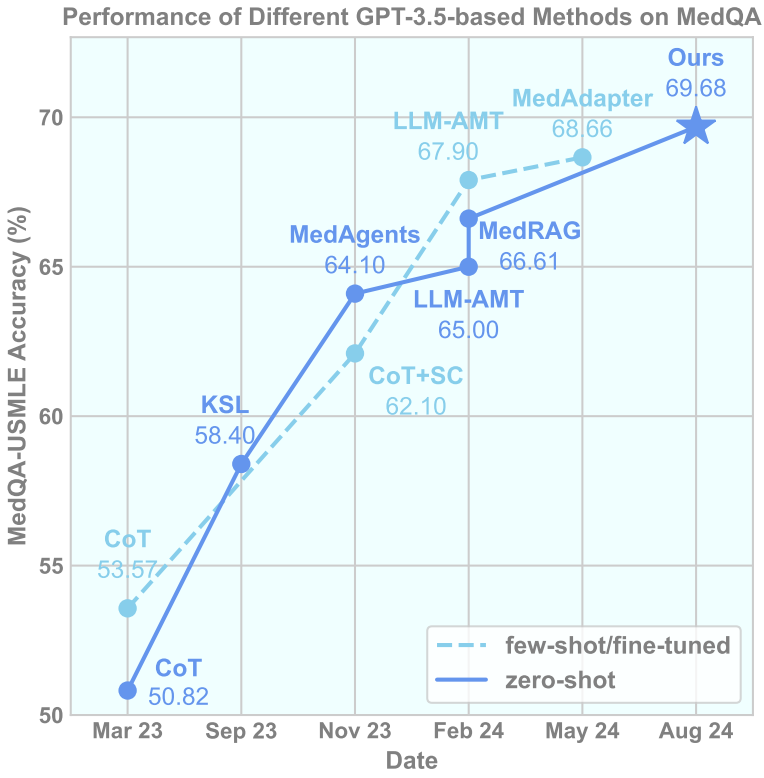


Fig. 1. Comparison of various methods proposed to improve GPT-3.5 performance on MedQA. Our zero-shot i-MedRAG outperforms all previous prompt engineering and fine-tuning methods.

replacing the information retrieval step (Figure 2 left) with our proposed iterative question-answering step (Figure 2 middle and right). The settings of RAG are described in Section 3.1 and the pipeline of our new i-MedRAG is discussed in Section 3.2. The details of the iterative question answering are described in Section 3.3.

3.1. *Retrieval-Augmented Generation*

In the zero-shot setting of medical question answering, the task of LLM $\mathcal{M}$ is trying to find the correct answer $\mathcal{A}$ given the question $\mathcal{Q}$ only. The ideal answer prediction $\tilde{\mathcal{A}}$ can be provided by

$$\tilde{\mathcal{A}} = \arg\max_{\mathcal{A}} \mathbb{P}_{\mathcal{M}}(\mathcal{A} \mid \mathcal{Q}, \text{inst.}), \tag{1}$$

where the "inst." is the task instruction the user provides that instructs the model to perform the task. As medical questions are knowledge-intensive,[36] it benefits from accessing large-scale external corpora to search for useful information.[21–23] A typical method to combine LLM reasoning with external corpora is RAG, which first retrieves relevant documents from the corpus for the given medical question and enters the retrieved documents along with the question into LLM to augment its answer generation. Formally, the RAG pipeline can be

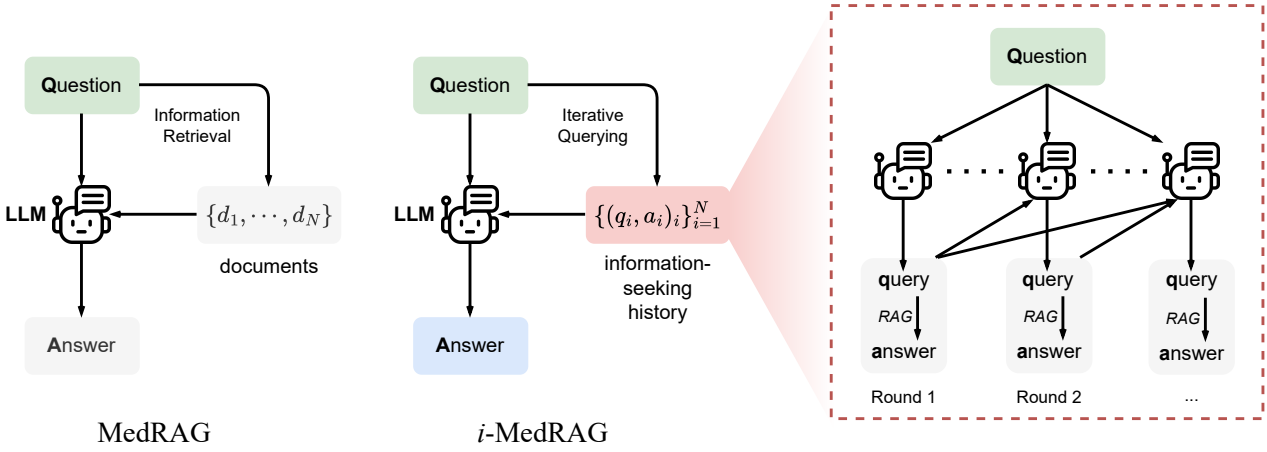


Fig. 2. Overview of i-MedRAG and its comparison to RAG (MedRAG). Left: the pipeline of Retrieval-Augmented Generation (RAG). Middle: the pipeline of our proposed i-MedRAG. Right: the iterative generation of question-specific medical query-answer (QA) pairs by asking follow-up queries.

described as

$$\tilde{\mathcal{A}} = RAG(\mathcal{Q}; \mathcal{M}, \mathcal{R}, \mathcal{D}) = \arg\max_{\mathcal{A}} \mathbb{P}_{\mathcal{M}}(\mathcal{A} \mid \mathcal{Q}, \text{inst.}, \{d_i\}_{i=1}^N), \tag{2}$$

where $\{d_i\}_{i=1}^N$ are the question-specific retrieved documents given by

$$\{d_i\}_{i=1}^N = \mathcal{R}(\mathcal{Q}; \mathcal{D}). \tag{3}$$

Here R is the text retriever and $\mathcal{D}$ is the corpus with a collection of documents.

3.2. *Iterative Retrieval-Augmented Generation*

While RAG exhibits promising performance in medical question answering,[23] it may be unable to handle certain complex medical questions in real-world cases. As text retrievers are typically trained to find relevant documents based on text similarity or lexicon overlap, they cannot break down a complex question and search for relevant information in a step-by-step manner. Thus, the inflexible retrieval step (Formula 3) in RAG may fail to analyze medical questions and find useful information to augment the answer generation, especially in complex clinical cases, where multiple rounds of information-seeking are required.

To address the issues mentioned, we propose to incorporate flexible information retrieval by prompting LLMs to iteratively generate follow-up queries based on the given medical question and previous information-seeking history. Moreover, as the context lengths of LLMs are limited, it can be impractical and infeasible to include all retrieved documents in the LLM context. Therefore, we prompt LLMs to directly answer the raised queries with relevant information and use such query-answer pairs as the information-seeking history. The pipeline of our proposed system can be formulated as

$$\tilde{\mathcal{A}} = i\text{-}MedRAG(\mathcal{Q}; \mathcal{M}, \mathcal{R}, \mathcal{D}) = \arg\max_{\mathcal{A}} \mathbb{P}_{\mathcal{M}}(\mathcal{A} \mid \mathcal{Q}, \text{inst.}, \{(q_i, a_i)\}_{i=1}^N), \tag{4}$$

where $\{(q_i, a_i)\}_{i=1}^{N}$ are the queries and the corresponding answers generated by LLMs with the help of RAG. The iterative process of query and answer generation will be detailed in Section 3.3.

3.3. *Iterative Generation of Follow-up Questions*

While the retrieved documents in RAG are determined by the question and the retrieval system, we propose to incorporate the reasoning capabilities of LLMs in i-MedRAG by prompting them to dynamically generate helpful queries in a step-by-step manner. Specifically, the LLM will be encouraged to generate n different queries to help find useful additional information for m iterations. In all iterations except for the first one, the model will be given the information-seeking history to generate context-specific follow-up queries. The queries $q_{i1}, \cdots, q_{in}$ generated in the i-th iteration can be formulated as

$$q_{i1}, \cdots, q_{in} = \begin{cases} \underset{q_{i1}, \cdots, q_{in}}{\arg\max}\, \mathbb{P}_{\mathcal{M}}(q_{i1}, \cdots, q_{in} \mid \mathcal{Q}, \text{inst.}'), & \text{if } i = 1, \\ \underset{q_{ic1}, \cdots, q_{in}}{\arg\max}\, \mathbb{P}_{\mathcal{M}}(q_{i1}, \cdots, q_{in} \mid \mathcal{Q}, \text{inst.}', \{(q_{jk}, a_{jk})\}_{j=1,\cdots,i-1}^{k=1,\cdots,n}), & \text{if } i > 1. \end{cases} \tag{5}$$

Different from the "inst." in Formula 2, the "inst.′" here is a modified instruction which focuses on generating follow-up queries instead of answering the medical question. For each query generation step, we prompt the LLM to analyze the existing information first and then generate new queries for additional knowledge. The step-by-step "reason-then-query" pipeline helps LLMs break down complex medical questions and find useful information from the external corpus. The answer to each generated query is given by a RAG system mentioned in Formula 2. This enables the system to leverage existing literature to provide grounded answers for generated queries.

The overall algorithm of i-MedRAG is presented in Algorithm 1.

4. Experiments

4.1. *Evaluation settings*

To evaluate the performance of our proposed i-MedRAG on knowledge-intensive medical question-answering tasks and compare it with other approaches, we select MedQA[26] as the testbed, which contains medical questions collected from United States Medical Licensing Examination (USMLE). With complex clinical cases in the dataset, MedQA reflects the difficulty of decision-making in real-world clinical medicine. The approaches for comparison are prompt engineering or fine-tuning methods that try to improve the performance of GPT-3.5 on MedQA, including chain-of-thought (CoT) prompting,[43] self consistency (SC), knowledge solver (KSL),[40] medical agents (MedAgents),[39] LLMs augmented with medical textbooks (LLM-AMT),[41] medical retrieval-augmented generation (MedRAG),[23] and LLMs with test-time adaptations (MedAdapter).[45]

Additionally, we evaluate the generalizability of our i-MedRAG with more LLMs and medical datasets. Llama-3.1-8B is selected as the representative of open-source models, which has a context window of 128k tokens. We also include MMLU-Med, a set of six medical tasks (anatomy, clinical knowledge, professional medicine, human genetics, college medicine,

Algorithm 1 The algorithm of i-MedRAG for medical question answering

Input medical question $\mathcal{Q}$, large language model $\mathcal{M}$, text retriever $\mathcal{R}$, medical corpus $\mathcal{D}$, query instruction "inst.′", answer instruction "inst.", hyperparameters m, n, N

Output answer prediction $\tilde{\mathcal{A}}$

Initialize the information-seeking history $\mathcal{H}$ = emptylist()
for i in $1, 2, \cdots, m$ **do**
    if $i = 1$ **then**
        generate n new queries $q_{i1}, \cdots, q_{in}$ using $\mathcal{M}$ given $\mathcal{Q}$
    else if $i > 1$ **then**
        generate n new queries $q_{i1}, \cdots, q_{in}$ using $\mathcal{M}$ given $\mathcal{Q}$ and $\mathcal{H}$
    end if
    for j in $1, 2, \cdots, n$ **do**
        retrieve N relevant documents $d_{ij}^1, \cdots, d_{ij}^N$ using $\mathcal{R}$ and $\mathcal{D}$ given q_{ij}
        generate the answer a_{ij} using $\mathcal{M}$ given q_{ij} and $d_{ij}^1, \cdots, d_{ij}^N$
        add the query-answer pair (q_{ij}, a_{ij}) to the list $\mathcal{H}$
    end for
end for
generate the predicted answer $\tilde{\mathcal{A}}$ using $\mathcal{M}$ given $\mathcal{Q}$ and $\mathcal{H}$
return $\tilde{\mathcal{A}}$

college biology) from Massive Multitask Language Understanding (MMLU), following previous studies.[8,23] MMLU-Med serves as a testbed to show the performance of i-MedRAG on a variety of different medical tasks.

Both MedQA and MMLU-Med are composed of multi-choice questions, whose evaluation metric is the accuracy of predicted answers chosen from given options. For the retrieval part in i-MedRAG, we select the Textbooks[26] and Statpearls[c] corpora introduced in MedRAG,[23] which are shown effective on medical examination questions. MedCPT[46] is chosen as the text retriever, which has been trained on domain-specific literature. For other baselines compared, the official settings described in their papers are used.

4.2. *Main results*

Table 1 shows the comparison results of i-MedRAG and other baseline approaches on MedQA using GPT-3.5. Official scores reported in previous research are used for a fair comparison. While methods with few-shot learning or model fine-tuning tend to perform better than the ones in a zero-shot setting, our i-MedRAG set a state-of-the-art performance of GPT-3.5 on MedQA without any training samples or parameter tuning. Among zero-shot approaches, i-MedRAG (69.68%) has a significant performance improvement ($p < 0.05$) compared to the previous best record achieved by MedRAG (66.61%).

The results of generalizing i-MedRAG to more LLMs and data are presented in Table 2. We compare i-MedRAG with our implemented CoT and MedRAG to see if i-MedRAG

[c] `https://www.statpearls.com/`

Table 1. Performance of GPT-3.5 with different prompt engineering / fine-tuning methods on MedQA. The "External Knowledge" column denotes if the method augments LLM generation with information retrieval of external knowledge.

Method	External Knowledge	Setting	Accuracy (%)
Chain of Thought[43]	No	zero-shot	50.82
Knowledge Solver[40]	Yes	zero-shot	58.40
Chain of Thought + Self Consistency[39]	No	zero-shot	61.30
MedAgents[39]	No	zero-shot	64.10
LLMs Augmented with Medical Textbook[41]	Yes	zero-shot	65.00
MedRAG[23]	Yes	zero-shot	66.61
Chain of Thought[43]	No	five-shot	53.57
Chain of Thought + Self Consistency[39]	No	five-shot	62.10
LLMs Augmented with Medical Textbook[41]	Yes	fine-tuned	67.90
MedAdapter[45]	No	fine-tuned	68.66
***i*-MedRAG (ours)**	**Yes**	**zero-shot**	**69.68**

can bring a consistent improvement of LLM performance in medical question answering. For all experiments with *i*-MedRAG, we tune the hyperparameters on a validation set of 100 samples and then report its scores on the test set. Similar to the results on GPT-3.5, the open-source Llama-3.1-8B also shows improved performance on MedQA with the help of *i*-MedRAG. While Llama-3.1-8B shows a close performance to GPT-3.5 in CoT and MedRAG settings, its performance is significantly improved with *i*-MedRAG, achieving an accuracy of 75.02%. The improved performance of GPT-3.5 and Llama-3.1-8B on MMLU-Med also demonstrates the generalizability of *i*-MedRAG to more medical data. As medical questions in MMLU-Med are less complex than the USMLE questions in MedQA, follow-up queries may not be necessary to find relevant information for the given question. Thus, it can be observed that the improvement by *i*-MedRAG compared to MedRAG is less significant in MMLU-Med than in MedQA.

Table 2. Performance of *i*-MedRAG on different LLMs and datasets. "Acc." denotes the accuracy. "Δ" shows the relative performance improvement compared with CoT.

Model	**Method**	**MedQA-USMLE**		**MMLU-Med**		**Average**	
		Acc.	Δ	Acc.	Δ	Acc.	Δ
GPT-3.5-Turbo	CoT	65.04	+0.00%	72.91	+0.00%	68.98	+0.00%
GPT-3.5-Turbo	MedRAG	66.61	+2.41%	75.48	+3.52%	71.05	+3.00%
GPT-3.5-Turbo	*i*-MedRAG	**69.68**	**+7.13%**	**75.85**	**+4.03%**	**72.77**	**+5.49%**
Llama-3.1-8B	CoT	64.73	+0.00%	77.23	+0.00%	70.98	+0.00%
Llama-3.1-8B	MedRAG	66.54	+2.80%	78.05	+1.06%	72.30	+1.86%
Llama-3.1-8B	*i*-MedRAG	**73.61**	**+13.72%**	**78.42**	**+1.54%**	**76.02**	**+7.10%**

4.3. *Scaling with iterations and queries*

As we described in Section 3.3, the number of iterations to ask follow-up queries and the number of queries generated in each iteration are the two critical hyperparameters in our proposed iterative generation of follow-up queries. To explore how different selections of the hyperparameter values affect the model performance, we run i-MedRAG with different settings and compare their results. We test both GPT-3.5 and Llama-3.1-8B on MedQA and MMLU-Med to examine if there are model-specific or task-specific patterns.

Figure 3 shows the model performance with different hyperparameter settings. Generally, MedQA and MMLU-Med show distinct patterns in performance change with the increasing number of iterations. While the performance of both GPT-3.5 and Llama-3.1-8B on MedQA tends to improve with more iterations of follow-up queries, their performance on MMLU-Med converges or starts to drop with just one or two iterations, corresponding to the different complexities of these two tasks.

From the results on MedQA, it is also empirically shown that the number of generated queries per iteration determines the rate of performance improvement and convergence over multiple iterations. LLMs with more queries generated per iteration tend to have a larger improvement in accuracy but also converge more quickly. Such a result is intuitively reasonable as more information can be collected each iteration with more generated queries.

4.4. *Case studies*

Table 3 shows the predictions of GPT-3.5 on a MedQA question with different prompt engineering approaches. The question asks about the mechanism of the drug for transitional cell carcinoma of the bladder which causes hearing loss. To solve the problem, it is important to find the exact drug and then figure out how it causes the mentioned symptoms. However, the CoT result shows that GPT-3.5 does not inherently contain sufficient medical knowledge to solve this problem. Instead of inferring the described drug, GPT-3.5 with CoT directly hallucinates a wrong option as the answer. While free radicals are relevant to hearing loss, their connection to the disease of the patient is unclear and not discussed. Compared to CoT which solely relies on the internal knowledge of LLMs, MedRAG provides an opportunity for LLMs to augment their answer generation with external medical knowledge. Nevertheless, the model output shows that the MedRAG system fails to retrieve useful information about the drug from medical corpora. Given the complex problem description, it is difficult for text retrievers to find the asked mechanism without knowing the drug.

With iteratively generated follow-up queries, our i-MedRAG manages to identify the described drug and find information about its mechanism. From Table 3, it can be observed that GPT-3.5 starts with a general query about the asked mechanism. However, similar to the case in MedRAG, the RAG system fails to provide useful information about the query. With the information-seeking history, GPT-3.5 updates its actions with follow-up queries with respect to side effects especially hearing loss. With the updated queries, it manages to identify "cisplatin" as the drug which is not explicitly mentioned in the question. A query about the mechanism of action of cisplatin is further proposed to search for information about the answer to the original question. With several iterations of adaptive question answering, GPT-3.5

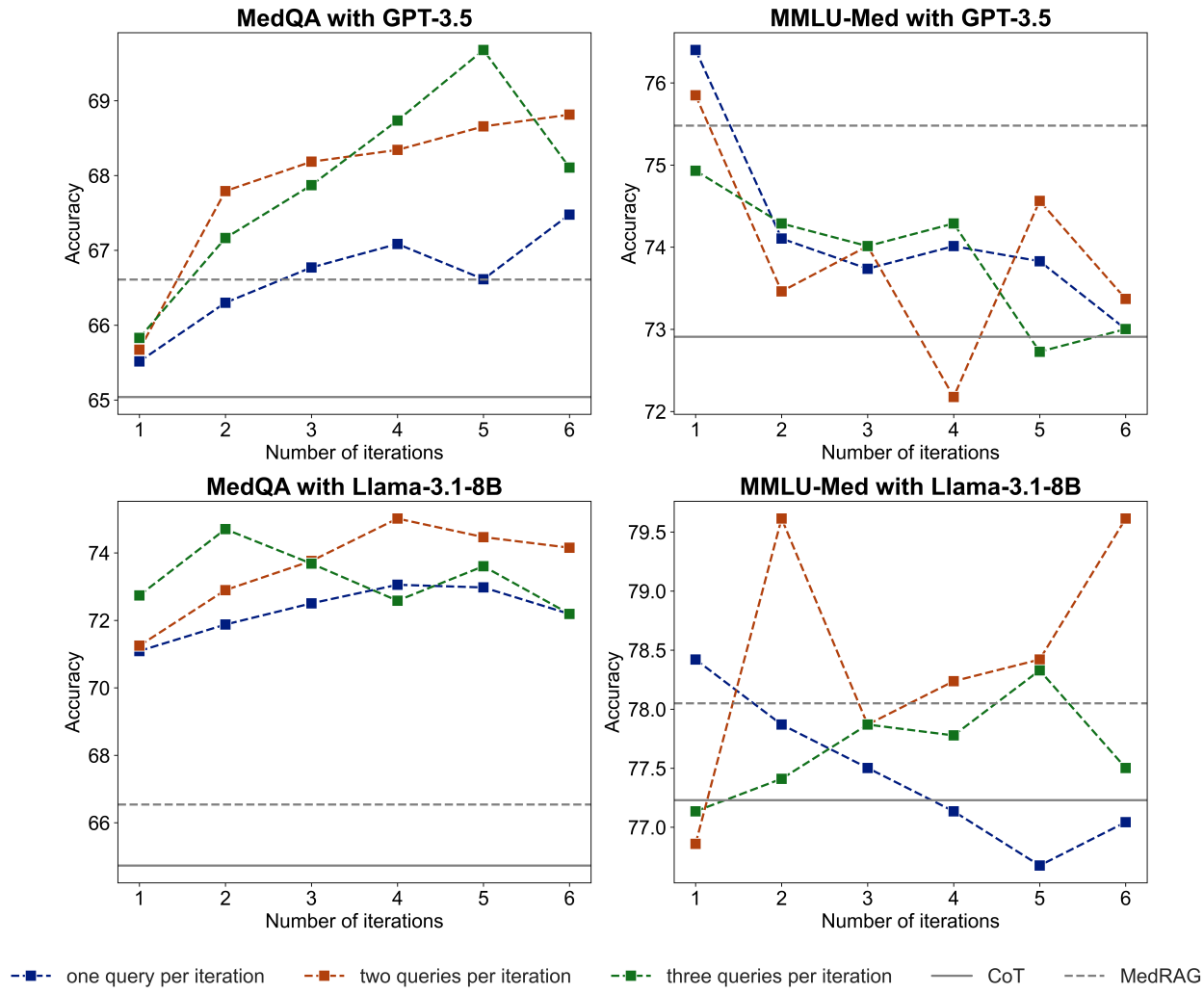


Fig. 3. i-MedRAG performance on MedQA and MMLU-Med with different hyperparameters.

successfully finds the correct answer for the given clinical medical question.

Table 4 shows another case from the MMLU-Med dataset by Llama-3.1-8B. The case is selected from the "professional medicine" subtask of MMLU-Med, which contains complex clinical cases similar to those in MedQA. In the question, LLMs are asked to choose the next step in diagnosis given the described symptoms, which requires the initial analysis of potential diseases. Similar to GPT-3.5, the open-source Llama-3.1-8B may hallucinate wrong answers with just CoT or MedRAG, as shown by the highlighted incorrect rationales in Table 4. With iterative follow-up queries, i-MedRAG enables the LLM to find out the correct answer by making specific queries given existing information. For example, Llama-3.1-8B asks about the symptoms of *Clostridium difficile* infection (CDI) and diverticulitis in its first iteration which helps it to identify diverticulitis as the potential disease for the diagnosis. In the second iteration, the model queries about the CT scan for diverticulitis, where the answer provides key information that helps it identify the correct next step.

Table 3. An example of how *i*-MedRAG helps GPT-3.5 to find the correct answer. The incorrect rationale and useful evidence are highlighted.

MedQA Question
0001: A 67-year-old man with transitional cell carcinoma of the bladder comes to the physician because of a 2-day history of ringing sensation in his ear. He received this first course of neoadjuvant chemotherapy 1 week ago. Pure tone audiometry shows a sensorineural hearing loss of 45 dB. The expected beneficial effect of the drug that caused this patient's symptoms is most likely due to which of the following actions? **A.** Inhibition of proteasome, **B.** Hyperstabilization of microtubules, **C.** Generation of free radicals, **D.** Cross-linking of DNA
CoT
Analysis: The patient is experiencing a ringing sensation in his ear, which is a symptom of tinnitus. [...] Among the given options, the most likely mechanism of action that could cause sensorineural hearing loss is the generation of free radicals. Free radicals can cause oxidative damage to the hair cells and auditory nerve, leading to hearing loss. Therefore, the answer choice is C. Generation of free radicals. **Answer**: C ✗
MedRAG
Retrieved documents: [...] **Analysis**: [...] To determine the mechanism of action of the drug causing the symptoms, we need to identify the drug used for neoadjuvant chemotherapy. However, the drug is not mentioned in the question stem or the provided documents. Without this information, it is not possible to definitively determine the mechanism of action. Therefore, the answer cannot be determined based on the given information. **Answer**: Cannot be determined. ✗
***i*-MedRAG**
Query (Iteration 1): What is the mechanism of action of the neoadjuvant chemotherapy drug used in transitional cell carcinoma of the bladder? **Retrieved documents**: [...] ***Answer***: Based on the provided documents, the mechanism of action of the neoadjuvant chemotherapy drug used in transitional cell carcinoma of the bladder is not explicitly mentioned [...] ***Query (Iteration 2)***: What are the potential side effects of neoadjuvant chemotherapy drugs used in transitional cell carcinoma of the bladder? **Retrieved documents**: [...] ***Answer***: The potential side effects [...] include myelosuppression, dysuria, urinary frequency, contact dermatitis, and systemic illness associated with granulomatous infections in multiple sites. ***Query (Iteration 3)***: What is the specific neoadjuvant chemotherapy drug [...] that can cause sensorineural hearing loss? **Retrieved documents**: [...] ***Answer***: The specific neoadjuvant chemotherapy drug used in transitional cell carcinoma of the bladder that can cause sensorineural hearing loss is **cisplatin**. ***Query (Iteration 4)***: What is the mechanism of action of **cisplatin**? **Retrieved documents**: [...] ***Answer***: The mechanism of action of cisplatin is achieved through the covalent binding of platinum to the purine bases guanine and adenine in DNA. This covalent binding leads to intra-strand and inter-strand crosslinks causing subsequent strand breaks [...] **Final analysis**: [...] The drug that is most likely responsible for the patient's symptoms is cisplatin, which is known to cause sensorineural hearing loss. The mechanism of action of cisplatin involves cross-linking of DNA, leading to subsequent strand breaks and cell death. Therefore, the correct answer is D. **Answer**: D ✔

Table 4. An example of how *i*-MedRAG helps Llama-3.1-8B to find the correct answer. The incorrect rationale and useful evidence are highlighted.

MMLU Question
professional_medicine-020: A 52-year-old woman comes to the emergency department because of a 1-week history of low-grade fever and increasing abdominal cramps that are exacerbated by bowel movements [...] Her temperature is 38.1°C (100.6°F), pulse is 75/min, and blood pressure is 150/80 mm Hg [...] there is tenderness to palpation of the left lower quadrant with guarding but no rebound [...] leukocyte count is 15,000/mm3 [...] Which of the following is the most appropriate next step in diagnosis? **A.** Test of the stool for Clostridium difficile toxin **B.** Endoscopic retrograde cholangiopancreatography **C.** Pelvic ultrasonography **D.** CT scan of the abdomen with contrast
CoT
Analysis: [...] the differential diagnosis includes diverticulitis, appendicitis, and bowel obstruction. However, the absence of bloody stools and the patient's response to antibiotics makes diverticulitis less likely. Therefore, the most appropriate next step in diagnosis is to test for Clostridium difficile toxin in the stool. **Answer**: A ✗
MedRAG
Retrieved documents: [...] **Analysis**: [...] Given the patient's symptoms and history, the most likely diagnosis is a pelvic abscess or a complication related to the previous surgery. The most appropriate next step in diagnosis would be to investigate the pelvic area for any potential complications or abscesses. **Answer**: C ✗
***i*-MedRAG**
Query 1 (Iteration 1): What are the common symptoms and diagnostic criteria for Clostridium difficile infection (CDI)? **Retrieved documents**: [...] **Answer**: [...] ***Query 2 (Iteration 1)***: What are the typical findings on physical examination and laboratory tests for diverticulitis? **Retrieved documents**: [...] **Answer**: Based on the provided documents, the typical findings on physical examination and laboratory tests for diverticulitis are: *Physical Examination*: Tenderness to palpation over the area of inflammation (almost always present) [...] Fever is almost always present, but hypotension and shock are uncommon. *Laboratory Tests*: Leukocytosis (elevated white blood cell count) [...] ***Query 1 (Iteration 2)***: What are the typical imaging findings for diverticulitis on a CT scan of the abdomen with contrast? **Retrieved documents**: [...] **Answer**: [...] These findings are consistent with the diagnosis of diverticulitis, and are often used to differentiate it from other conditions that may present with similar symptoms. ***Query 2 (Iteration 2)***: What are the indications for surgical intervention in a patient with diverticulitis? **Retrieved documents**: [...] **Answer**: [...] **Final Analysis**: [...] The patient's laboratory results show a leukocyte count of 15,000/mm^3, indicating leukocytosis [...] The patient's symptoms of abdominal cramps, fever, and leukocytosis, combined with tenderness to palpation of the left lower quadrant, are consistent with diverticulitis [...] the most appropriate next step in diagnosis is to perform a CT scan of the abdomen with contrast to evaluate for diverticulitis and rule out other causes of her symptoms. **Answer**: D ✔

5. Discussion

Overall, our proposed *i*-MedRAG effectively improves the performance of LLMs on complex medical questions by prompting them to iteratively ask follow-up queries. The experimental results show that our approach is better than previously proposed prompt engineering and fine-tuning methods, and is generalizable to various LLMs and medical question-answering datasets. Nevertheless, our approach has certain limitations which need to be discussed. It is also worthwhile to discuss the future work of this study to analyze how it can be further improved to facilitate real-world medical assistance.

5.1. *Limitations*

The first limitation of *i*-MedRAG is its high cost. While generating more follow-up queries tends to provide LLMs with more comprehensive and focused information about the given medical question, the cost also grows linearly with the number of queries generated. The time cost can be further increased if more documents are used to help answer the generated queries with RAG. While the cost is comparable to approaches using multiple LLM agents[39] or self consistency[47] which also prompt LLMs multiple times for each question, it is much more costly than baseline prompting methods such as CoT.[37]

Another limitation is the selection of hyperparameter values for optimal performance. As shown in Figure 3, different LLMs can have different hyperparameter settings for their optimal performance. Even for the same LLM, its optimal hyperparameters can vary based on the medical questions being processed. Thus, it is non-trivial to find the optimal hyperparameters of *i*-MedRAG for a new medical task, which may be inefficient for real-world deployments.

5.2. *Future work*

Given the limitations of *i*-MedRAG, we consider several potential future directions that could further improve the performance of retrieval-augmented generation for medicine. The first direction is the automation of hyperparameter selection in *i*-MedRAG. To reduce the laborious process of hyperparameter selection, one may use an LLM agent to dynamically determine how many follow-up queries should be asked each iteration. This can improve the efficiency and flexibility of the hyperparameter selection process. Another future direction is to improve the performance of *i*-MedRAG with few-shot demonstrations. While few-shot CoT prompting is demonstrated to perform better than the zero-shot counterpart,[43] it is not easy to adapt such strategies to *i*-MedRAG as the reasoning process can be dynamically affected by the use of external corpora and retrievers. Investigating how *i*-MedRAG can benefit from one or few-shot samples could be a potential direction to further enhance its performance on medical question answering. More quantitative analysis can also be performed to examine the error types of *i*-MedRAG compared to existing methods.

6. Acknowledgments

Guangzhi Xiong and Aidong Zhang are supported by NIH grant 1R01LM014012 and NSF grant 2333740. Qiao Jin and Zhiyong Lu are supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

References

1. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan and D. S. W. Ting, Large language models in medicine, *Nature medicine* **29**, 1930 (2023).
2. H. Zhou, B. Gu, X. Zou, Y. Li, S. S. Chen, P. Zhou, J. Liu, Y. Hua, C. Mao, X. Wu *et al.*, A survey of large language models in medicine: Progress, application, and challenge, *arXiv preprint arXiv:2311.05112* (2023).
3. J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou and R. Daneshjou, Large language models in medicine: the potentials and pitfalls: a narrative review, *Annals of Internal Medicine* **177**, 210 (2024).
4. S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D. C. Comeau *et al.*, Opportunities and challenges for chatgpt and large language models in biomedicine and health, *Briefings in Bioinformatics* **25**, p. bbad493 (2024).
5. R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon and T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, *Briefings in bioinformatics* **23**, p. bbac409 (2022).
6. V. Liévin, C. E. Hother, A. G. Motzfeldt and O. Winther, Can large language models reason about medical questions?, *Patterns* **5** (2024).
7. H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu *et al.*, Can generalist foundation models outcompete special-purpose tuning? case study in medicine, *arXiv preprint arXiv:2311.16452* (2023).
8. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023).
9. E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin *et al.*, Biomedlm: A 2.7 b parameter language model trained on biomedical text, *arXiv preprint arXiv:2403.18421* (2024).
10. C. Shaib, M. Li, S. Joseph, I. Marshall, J. J. Li and B. C. Wallace, Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success), in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023.
11. L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau *et al.*, Evaluating large language models on medical evidence summarization, *npj Digital Medicine* **6**, p. 158 (2023).
12. D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová *et al.*, Adapted large language models can outperform medical experts in clinical text summarization, *Nature Medicine* , 1 (2024).
13. Q. Jin, Z. Wang, C. S. Floudas, F. Chen, C. Gong, D. Bracken-Clarke, E. Xue, Y. Yang, J. Sun and Z. Lu, Matching patients to clinical trials with large language models, *ArXiv* (2023).
14. C. Wong, S. Zhang, Y. Gu, C. Moung, J. Abel, N. Usuyama, R. Weerasinghe, B. Piening, T. Naumann, C. Bifulco *et al.*, Scaling clinical trial matching using large language models: A case study in oncology, in *Machine Learning for Healthcare Conference*, 2023.
15. M. Wornow, A. Lozano, D. Dash, J. Jindal, K. W. Mahaffey and N. H. Shah, Zero-shot clinical trial patient matching with llms, *arXiv preprint arXiv:2402.05125* (2024).
16. S. Zhuang, B. Koopman and G. Zuccon, Team ielab at trec clinical trial track 2023: Enhancing clinical trial retrieval with neural rankers and large language models, *arXiv preprint arXiv:2401.01566* (2024).
17. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto and P. Fung, Survey of hallucination in natural language generation, *ACM Computing Surveys* **55**, 1 (2023).
18. T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu and G. Haffari, Continual learning for large language models: A survey, *arXiv preprint arXiv:2402.01364* (2024).

19. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in Neural Information Processing Systems* **33**, 9459 (2020).
20. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun and H. Wang, Retrieval-augmented generation for large language models: A survey, *arXiv preprint arXiv:2312.10997* (2023).
21. C. Zakka, R. Shad, A. Chaurasia, A. R. Dalal, J. L. Kim, M. Moor, R. Fong, C. Phillips, K. Alexander, E. Ashley *et al.*, Almanac—retrieval-augmented language models for clinical medicine, *NEJM AI* **1**, p. AIoa2300068 (2024).
22. A. Lozano, S. L. Fleming, C.-C. Chiang and N. Shah, Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature, in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, 2023.
23. G. Xiong, Q. Jin, Z. Lu and A. Zhang, Benchmarking retrieval-augmented generation for medicine, *arXiv preprint arXiv:2402.13178* (2024).
24. Q. Jin, B. Dhingra, Z. Liu, W. Cohen and X. Lu, Pubmedqa: A dataset for biomedical research question answering, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.
25. G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos *et al.*, An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, *BMC bioinformatics* **16**, 1 (2015).
26. D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang and P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, *Applied Sciences* **11**, p. 6421 (2021).
27. Y. S. Low, M. L. Jackson, R. J. Hyde, R. E. Brown, N. M. Sanghavi, J. D. Baldwin, C. W. Pike, J. Muralidharan, G. Hui, N. Alexander *et al.*, Answering real-world clinical questions using large language model based systems, *arXiv preprint arXiv:2407.00541* (2024).
28. Q. Jin, R. Leaman and Z. Lu, Pubmed and beyond: biomedical literature search in the age of artificial intelligence, *EBioMedicine* **100** (2024).
29. Q. Jin, R. Leaman and Z. Lu, Retrieve, summarize, and verify: how will chatgpt affect information seeking from the medical literature?, *Journal of the American Society of Nephrology* **34**, 1302 (2023).
30. X. Li, P. Yu, C. Zhou, T. Schick, O. Levy, L. Zettlemoyer, J. E. Weston and M. Lewis, Self-alignment with instruction backtranslation, in *The Twelfth International Conference on Learning Representations*, 2024.
31. Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan and W. Chen, Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy, in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.
32. H. Trivedi, N. Balasubramanian, T. Khot and A. Sabharwal, Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.
33. Z. Jiang, M. Sun, L. Liang and Z. Zhang, Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach, *arXiv preprint arXiv:2407.13101* (2024).
34. A. Pal, L. K. Umapathi and M. Sankarasubbu, Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, in *Conference on Health, Inference, and Learning*, 2022.
35. D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song and J. Steinhardt, Measuring massive multitask language understanding, *arXiv preprint arXiv:2009.03300* (2020).

36. Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu and S. Yu, Biomedical question answering: a survey of approaches and challenges, *ACM Computing Surveys (CSUR)* **55**, 1 (2022).
37. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* **35**, 24824 (2022).
38. X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery and D. Zhou, Self-consistency improves chain of thought reasoning in language models, *arXiv preprint arXiv:2203.11171* (2022).
39. X. Tang, A. Zou, Z. Zhang, Y. Zhao, X. Zhang, A. Cohan and M. Gerstein, Medagents: Large language models as collaborators for zero-shot medical reasoning, *arXiv preprint arXiv:2311.10537* (2023).
40. C. Feng, X. Zhang and Z. Fei, Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs, *arXiv preprint arXiv:2309.03118* (2023).
41. Y. Wang, X. Ma and W. Chen, Augmenting black-box llms with medical textbooks for clinical question answering, *arXiv preprint arXiv:2309.02233* (2023).
42. T. Savage, A. Nayak, R. Gallo, E. Rangan and J. H. Chen, Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine, *NPJ Digital Medicine* **7**, p. 20 (2024).
43. H. Nori, N. King, S. M. McKinney, D. Carignan and E. Horvitz, Capabilities of gpt-4 on medical challenge problems, *arXiv preprint arXiv:2303.13375* (2023).
44. K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi *et al.*, Capabilities of gemini models in medicine, *arXiv preprint arXiv:2404.18416* (2024).
45. W. Shi, R. Xu, Y. Zhuang, Y. Yu, H. Wu, C. Yang and M. D. Wang, Medadapter: Efficient test-time adaptation of large language models towards medical reasoning, *arXiv preprint arXiv:2405.03000* (2024).
46. Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur and Z. Lu, Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval, *Bioinformatics* **39**, p. btad651 (2023).
47. X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery and D. Zhou, Self-consistency improves chain of thought reasoning in language models, in *The Eleventh International Conference on Learning Representations*,

ClinValAI: A framework for developing Cloud-based infrastructures for the External Clinical Validation of AI in Medical Imaging

Ojas A. Ramwala[1], Kathryn P. Lowry[2], Daniel S. Hippe[3], Matthew P.N. Unrath[4], Matthew J. Nyflot[2,5], Sean D. Mooney[6], Christoph I. Lee[2,7,8†]

[1]*Department of Biomedical Informatics and Medical Education, University of Washington School of Medicine, Seattle, Washington, 98195, USA*
[2]*Department of Radiology, University of Washington School of Medicine, Seattle, Washington, 98195, USA*
[3]*Clinical Research Division, Fred Hutchinson Cancer Center, Seattle, WA, 98109, USA*
[4]*Pariveda Solutions, Seattle, Washington, 98101, USA*
[5]*Department of Radiation Oncology, University of Washington School of Medicine, Seattle, 98195, USA*
[6]*Center for Information Technology, National Institutes of Health, Bethesda, Maryland, 20892, USA*
[7]*Department of Health Systems and Population Health, University of Washington School of Public Health, Seattle, Washington, 98195, USA*
[8]*Northwest Screening and Cancer Outcomes Research Enterprise, University of Washington, Seattle, Washington, 98195, USA*
[†]*Email: stophlee@uw.edu*

Artificial Intelligence (AI) algorithms showcase the potential to steer a paradigm shift in clinical medicine, especially medical imaging. Concerns associated with model generalizability and biases necessitate rigorous external validation of AI algorithms prior to their adoption into clinical workflows. To address the barriers associated with patient privacy, intellectual property, and diverse model requirements, we introduce ClinValAI, a framework for establishing robust cloud-based infrastructures to clinically validate AI algorithms in medical imaging. By featuring dedicated workflows for data ingestion, algorithm scoring, and output processing, we propose an easily customizable method to assess AI models and investigate biases. Our novel orchestration mechanism facilitates utilizing the complete potential of the cloud computing environment. ClinValAI's input auditing and standardization mechanisms ensure that inputs consistent with model prerequisites are provided to the algorithm for a streamlined validation. The scoring workflow comprises multiple steps to facilitate consistent inferencing and systematic troubleshooting. The output processing workflow helps identify and analyze samples with missing results and aggregates final outputs for downstream analysis. We demonstrate the usability of our work by evaluating a state-of-the-art breast cancer risk prediction algorithm on a large and diverse dataset of 2D screening mammograms. We perform comprehensive statistical analysis to study model calibration and evaluate performance on important factors, including breast density, age, and race, to identify latent biases. ClinValAI provides a holistic framework to validate medical imaging models and has the potential to advance the development of generalizable AI models in clinical medicine and promote health equity.

Keywords: Artificial Intelligence; Bias; Breast Cancer; Clinical Validation; Cloud Infrastructures; Generalizability; Medical Imaging

1. Introduction

Artificial Intelligence (AI) algorithms have demonstrated encouraging results in the field of biomedical signal[1,2] and image[3–11] processing, electronic health record (EHR) analysis[12], and clinical text processing[13] to provide improved diagnostic outcomes, early intervention strategies, and well-tailored patient-specific management options. The performance of AI algorithms has been on par with radiologists[14] and even better in a few scenarios[15].

However, deep learning models are susceptible to generalizability challenges[16]. Diagnostic AI models have demonstrated deteriorated performance during independent evaluation on datasets reflecting real-world healthcare settings, especially for specific subpopulations[17]. The adoption of such algorithms can have critical implications for patients' safety. Thus, large-scale independent external validation of AI models is imperative before adopting them into clinical workflows.

Nevertheless, there are several barriers to robust evaluation. Since AI vendors are protective of their intellectual property, they may be unwilling to provide their algorithms to health institutions for validation, especially prior to their purchase. Per HIPAA guidelines, medical centers cannot share patient data with commercial organizations without their consent since it contains protected health information. Moreover, different AI algorithms have varying storage and computing requirements. Planning and budgeting for resources to cater to such varying needs can cause substantial financial and cognitive burdens on health systems evaluating multiple AI tools on-premises for clinical adoption. Outsourcing clinical validation work to third-party services can be costly and involve legal and operational complications while sharing access to clinical data.

To address the limited technical guidance on developing methods that can aid in monitoring the performance of AI in clinical medicine[18], we propose ClinValAI – an open-source cloud-agnostic unified framework for establishing robust infrastructures to validate AI algorithms. We customize its functionalities for the clinical validation of AI models for medical imaging applications. Our work aims to enable medical institutions to rigorously evaluate models prior to their integration into clinical workflows. By leveraging our framework, healthcare institutions can screen data from large populations to accurately assess model generalizability and investigate latent biases.

To demonstrate the capabilities of our framework, we used our ClinValAI-based cloud infrastructure to perform large-scale clinical validation of Mirai[19], a state-of-the-art open-source mammography-based AI algorithm for breast cancer risk prediction. We comprehensively evaluate its generalizability on a large and diverse dataset of 26,449 2D screening mammography exams from 14,291 patients, demonstrating the reliability of our work in monitoring AI models and assessing algorithmic bias in healthcare settings. Our framework has the potential to improve a clinical institution's AI model selection process to enhance patient care for their target population.

2. Methods

The clinical validation of AI algorithms can be performed via on-premises as well as cloud-based infrastructures. While medical institutions traditionally trust on-premises setups with their patient data, challenges, including upfront resource investments, scalability issues, and maintenance overhead, can obstruct validation efforts. In contrast, the configurability of cloud-based storage and computational environments, their cost-effective setup and maintenance, built-in network security

and information recovery services, and rapid acquisition render cloud infrastructures an appealing choice for deploying and rigorously validating AI models on large datasets. ClinValAI can be leveraged to establish innovative, effective, and secure cloud-based validation infrastructures. Figure 1 details our conceptual framework for externally validating AI models in clinical medicine.

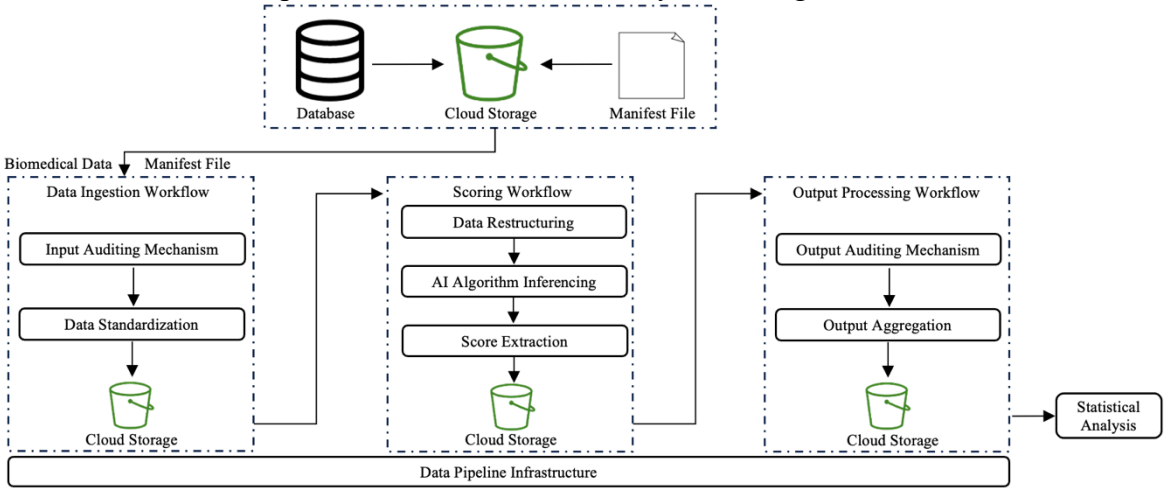


Fig. 1. Conceptual overview of the ClinValAI framework.

2.1. *Preserving Patient Data Privacy*

Patient privacy and information security concerns constrain biomedical data sharing and stymie AI algorithm development and validation efforts[20]. ClinValAI leverages the "Model to Data" (MTD) paradigm[21,22] to validate AI models on private biomedical data. Cloud infrastructure and containerization techniques form the foundation of the MTD framework. Rather than providing direct data access to the vendors, the Dockerized models are uploaded to the cloud host as containers encapsulating the AI algorithms, their dependencies, and other configuration settings required for successfully testing the models on the data stored in the cloud. To address intellectual property concerns, ClinValAI supports license files for Docker images, allowing AI vendors to control access to their AI models. Thus, ClinValAI enables health institutions to preserve patient data within a firewall and run models on medical imaging exams without providing vendors direct data access.

2.2. *Data Pipeline Infrastructure*

ClinValAI features multiple computational pipelines for biomedical data processing and clinical validation of AI algorithms through a combination of series and parallel jobs.

2.2.1. *Workflow Representation*

To comprehensively express the workflow design, we leverage the Workflow Description Language (WDL)[23] due to its comprehensibility and cross-platform interoperability. WDL enables defining pipelines to process and analyze data. WDL necessitates an engine to execute its functionalities. Our proposed framework utilizes miniWDL[24], a WDL execution engine for biomedical applications that functions as a job orchestrator for executing multiple data processing workflows in a parallel

fashion, depending on the available memory and computing resources. The customizability of ClinValAI's workflow representation method bolsters its utilization for the clinical validation of AI.

2.2.2. *Job Scheduling and Batch Processing Orchestration Mechanism*

Our framework is equipped with tools that provision compute instances and communicate with the miniWDL engine and a container job scheduling mechanism to automate infrastructure deployment (Figure 2). It can be further modified for more granular control over those pipelines. After the workflow submission, the WDL script is uploaded to cloud storage, and the job scheduling mechanism is invoked to run a miniWDL container, known as the "head" job container. ClinValAI implements data processing pipelines through the miniWDL engine operating on this container. The head job pulls the WDL script from the cloud storage and, per its instructions, directs the scheduling mechanism to spin up "task" job Docker containers that execute individual components of the workflow. ClinValAI enables the head job containers to spin up multiple sets of task job containers to achieve the parallel execution of computational steps.

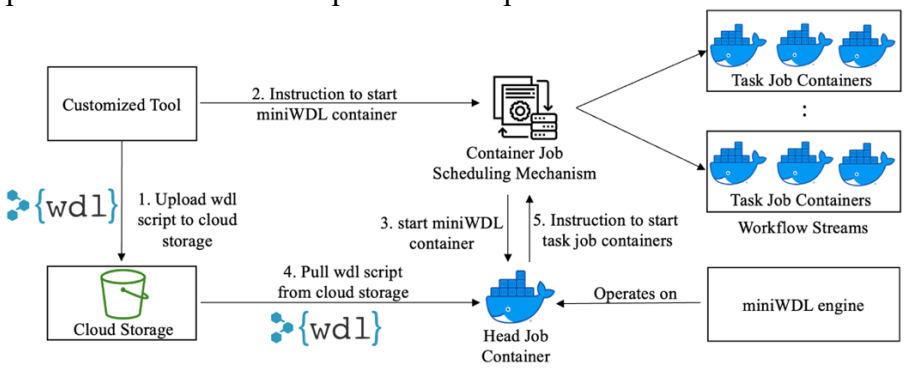


Fig. 2: ClinValAI's job scheduling and batch processing orchestration mechanism.

For our case study on clinical validation of the Mirai algorithm for breast cancer risk prediction on screening mammograms, our ClinValAI-based cloud infrastructure ingests a *set* of compressed files, each representing a *batch* comprising multiple sub-folders corresponding to patients' mammography *exams*. ClinValAI creates multiple execution streams for each set; all exams in one batch are processed serially by leveraging numerous task containers running sequentially. Exams in one batch are scored independently of other batches in a parallel fashion. Thus, ClinValAI's data pipeline enables leveraging the full potential of the cloud computing environment.

In addition to validating AI models using their Docker images, our framework supports customizing Linux Docker images to establish optimized workflows. Rather than dynamically pulling scripts from cloud storage at run-time, ClinValAI facilitates configuring the Docker images at build time. This approach avoids inadvertent version updates in the sequence of instructions during run-time, which could produce inconsistent results. While fetching scripts from cloud storage during run-time is more convenient, baking them into the Docker image enhances reliability.

2.3. *Data Ingestion Workflow*

ClinValAI's data ingestion workflow (Figure 3) is the first of the three stages in the framework. It comprises an input auditing and a data standardization mechanism.

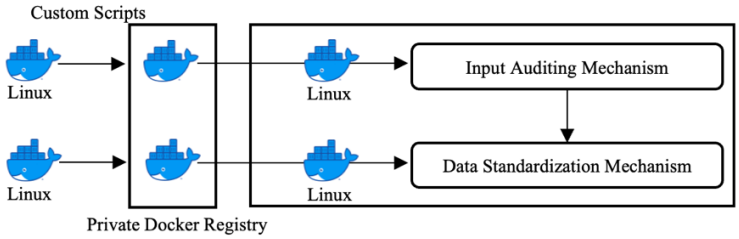


Fig. 3. ClinValAI's data ingestion workflow ensures that inputs are consistent with model prerequisites.

2.3.1. *Input Auditing Mechanism*

ClinValAI's input auditing mechanism performs the vital task of verifying if the data can be processed and are aligned with the model's prerequisites before initiating the scoring process. This can help ensure a sample size that preserves statistical power for meaningful analysis. Through a configured Docker image, it compares the uploaded data with a manifest file and algorithm-specific requirements to verify that the dataset is complete with all the required information.

To validate Mirai, a manifest file comprising the accession numbers, data modality, the corresponding number of images in each exam, image laterality and projection, file sizes, etc., is created. The auditing logic checks for corrupted files, DICOMs with missing pixel array data, and unsupported manufacturing devices and monitors if the image metadata contains all the information required by the algorithm. For example, AI models for mammography interpretation may not be able to process images if view/projection (Cranio-Caudal (CC) or Medio-Lateral Oblique (MLO)) or laterality (left or right breast) information is missing from DICOMs. ClinValAI thoroughly analyzes the data to identify such inconsistencies and features a comprehensive input auditing mechanism to ensure a seamless external validation study.

2.3.2. *Data Standardization*

Standardizing inputs before initiating AI inferencing is necessary if there is variation from multiple data sources or if a data source requires enrichment before algorithmic processing can take place. ClinValAI's data standardization mechanism analyzes the findings of the input auditing logic and provides the feature of customizing the associated Linux Docker image to achieve data standardization and ensure the quality of the study data.

For our validation study of Mirai's performance, if a set of DICOMs is corrupted or missing pixel array information, the standardization mechanism does not pass them through the scoring workflow. Similarly, it removes images that do not match the study criteria – for example, deleting all the non-mammography images to ensure that only the acceptable modalities are included.

One of the important aspects of ClinValAI's data standardization mechanism is its ability to impute missing information. For example, if an image does not have laterality or projection information in the DICOM headers, the framework populates the DICOM metadata using the details from the manifest files. Moreover, if the required data is not available in the manifest file, it parses other descriptive DICOM headers to look for specific information for imputation. For example, AI algorithms for mammography interpretation expect laterality information in one of the *ImageLaterality*, *Laterality*, or *FrameLaterality* headers and projection information in the

ViewPosition header. If these tags are missing, ClinValAI's data standardization mechanism analyzes other subjective headers like *SeriesDescription* to systematically impute laterality and projection information into their respective tags. Thus, ClinValAI can be customized to facilitate effective data cleaning and preprocessing, information imputation, and data standardization.

2.4. *Scoring Workflow*

ClinValAI's scoring workflow (Figure 4) is the second stage in the framework. It comprises a data restructuring, an algorithm inferencing, and a score extraction mechanism.

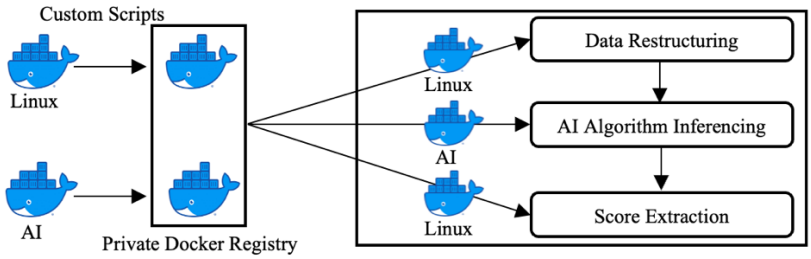


Fig. 4. ClinValAI's scoring workflow ensures consistent inferencing and systematic troubleshooting.

2.4.1. *Data Restructuring*

Various health institutions can organize patient data and medical images in different formats, requiring datasets to be systematized by patient ID, accession numbers, or date of collection. Different AI models can have their specific input structuring requirements. For example, a model may require all images from a patient to be in a single folder, while another may need additional sub-folders based on exam ID or modality. Different models may need varying numbers of images per exam – for instance, Mirai needs four standard 2D mammograms (CC and MLO views of the left and right breast), whereas some models can function even with unilateral exams. Some models can raise errors if inputs contain multiple images of the same view and laterality combination, while others can successfully score them. Moreover, some models can process 2D and 3D images simultaneously, while others can leverage separate Docker images depending on shape and modality. ClinValAI supports extensive data restructuring by enabling the customization of Docker images to account for model-specific variations by holistically analyzing the DICOM metadata and pixel array information, thereby establishing consistency between input and model criteria.

2.4.2. *AI Algorithm Inferencing*

ClinValAI enables effective customization of AI algorithms' Docker images to facilitate accurate scoring of exams. The Docker file is specified with the required environment variables and necessary scoring scripts, and the updated Docker image is used to spin up the AI model's Docker container to execute algorithmic processing. Information about the computational requirements of the AI algorithms can be utilized to identify the appropriate compute instances to be specified in our framework. To work with asynchronous inferencing workflows, our framework also features a polling mechanism depending on the inference time of each algorithm to ensure that the compute instances are not stalled due to inconsistent data, node failures, or other issues. Furthermore, our framework provides the flexibility of incorporating additional steps, such as drafting a list of input

studies to be processed or creating corresponding output folders for storing final results, depending on the models' prerequisites. Similarly, using our framework, we reconfigure Mirai's open-source Docker image via programmatic steps to streamline its inferencing workflow. Thus, by facilitating multiple customization features, ClinValAI enables robust validation of AI algorithms.

2.4.3. *Score Extraction*

After the completion of the scoring process, the model's generated files need to be processed to retrieve specific outputs of interest, such as image-, exam-, or patient-level scores. Different AI algorithms have different ways of representing outputs. ClinValAI enables customizing the Linux Docker image to follow the modes and steps to extract scores from diverse formats – from flat files like comma-separated values (CSV) documents to highly nested DICOM Structured Reports (SRs) and JavaScript Object Notation (JSON) objects. Similarly, ClinValAI also facilitates the storage of supplementary files, such as annotations in processed images or heat maps, and associated model explanations, if available, to facilitate improved interpretation for radiologists. Moreover, this step also records and organizes logs specific to the algorithm and workflow. Thus, ClinValAI facilitates systematic troubleshooting, effective scoring, and rigorous clinical validation of AI algorithms.

2.5. *Output Processing Workflow*

ClinValAI's output processing workflow (Figure 5) is the third and final stage in the framework. It comprises an output auditing and an output aggregation mechanism.

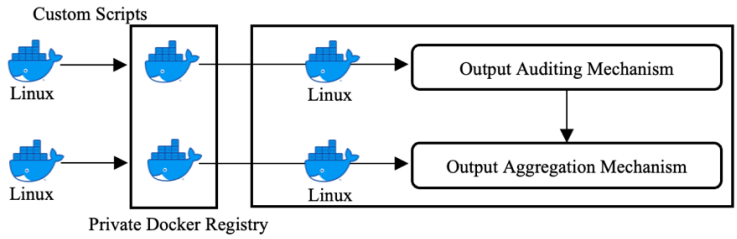


Fig. 5. ClinValAI's output processing workflow helps identify and analyze samples with missing results and aggregates final outputs for downstream analysis.

2.5.1. *Output Auditing Mechanism*

Once the scoring workflow has been executed, ClinValAI performs the essential task of verifying if results have been produced for all the inputs and if the generated files comply with the algorithm's expected outputs. Moreover, the framework facilitates examining if the required numeric values of interest, inference reports, and supplementary files can be extracted from the resulting outputs. ClinValAI identifies samples with missing output data, irretrievable scores, and corrupted output files to enable analysis of samples to be re-scored. If no outputs are generated for a patient's exam, infrastructure-specific logs can be inspected to check for issues related to compute instances or customization of the Docker images. If scores cannot be extracted from the model's output for an exam, algorithm-specific logs can be analyzed to check for inconsistencies and errors. Overall, ClinValAI facilitates a holistic output auditing mechanism for the streamlined validation of models.

2.5.2. *Output Aggregation Mechanism*

Outputs from individual workflows are hierarchically stored based on set number, batch number, and exam ID. Analyzing the complete dataset in the distributed format of cloud storage can be cumbersome. Before statistical analysis can be performed, ClinValAI systematically aggregates relevant details by appending all results to a relational database. After the completion of scoring workflows for all standardized batches of exams, the pipeline connects to the database and hierarchically uploads data from the audited results, supplementary files, and logs by inserting rows for every set, batch, and exam as demonstrated by the entity relationship diagram (Figure 6). During statistical analysis, this database is pulled to analyze findings. ExamIDs and Study UIDs (Unique Identifiers) are used to cross-reference the AI algorithm's results and the attributes of interest.

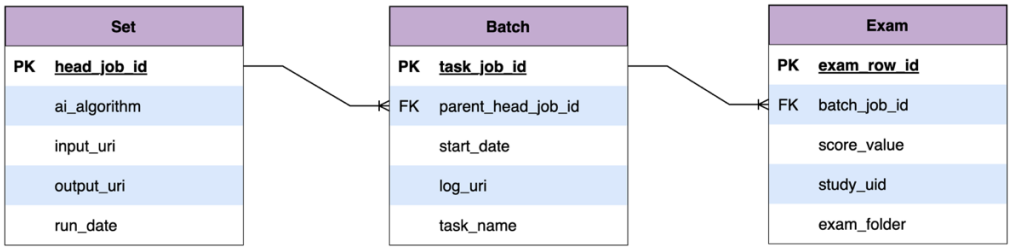


Fig. 6. Entity Relationship Diagram for the aggregated output data. Abbreviations: URI=Uniform Resource Identifier; PK=Primary Key; FK=Foreign Key

Thus, ClinValAI features multiple customizable workflows to establish optimized infrastructures for robust clinical validation of AI algorithms for medical imaging applications.

3. Evaluation and Results

To evaluate the utility of our framework, we used our ClinValAI-based cloud infrastructure to perform a rigorous external validation of Mirai, a state-of-the-art deep learning algorithm that predicts future breast cancer risk across five years by processing the four standard views of a 2D digital mammogram – Cranio-Caudal and Medio-Lateral Oblique views of the left and right breast.

3.1. *Patient Cohort*

All mammography screening exams from 2010-2014 performed across four imaging facilities in the University of Washington (UW) Medicine health system were reviewed for eligibility. Exams of women with age < 40 or ≥ 80 years, a personal history of breast cancer, or the presence of breast implants were excluded. Cancer outcomes at year 5 after every exam were collected via linkage to the Washington State cancer registry, which captures all breast cancers diagnosed within the state of Washington through December 31st, 2020, allowing for robust ground truth for all screening exams. Information on breast density and patient demographics, including age at the time of imaging and race, were obtained from the University of Washington Medicine electronic medical records. ClinValAI excluded exams with insufficient 2D screening images and processed 26,449 exams from 14,291 patients to generate Mirai scores. A total of 543 exams (2.1%) were followed by a breast cancer diagnosis within five years (88 in year 1, 92 in year 2, 112 in year 3, 119 in year 4, and 132 in year 5). Table 1 shows the patient characteristics. BI-RADS[25] categories 'heterogeneously dense'

and ‘extremely dense’ correspond to dense breasts, and ‘almost entirely fatty’ and ‘scattered fibroglandular’ correspond to non-dense breasts.

Table 1. Patient characteristics at each exam.

		Breast Cancer within 5 years	
	All	Yes	No
Variable	(n = 26,449)	(n = 543)	(n = 25,906)
Age			
40-49	7,014 (26.5%)	114 (21.0%)	6,900 (26.6%)
50-59	9,431 (35.7%)	151 (27.8%)	9,280 (35.8%)
60-69	7,082 (26.8%)	171 (31.5%)	6,911 (26.7%)
70-79	2,922 (11.0%)	107 (19.7%)	2,815 (10.9%)
Race			
White	20,365 (82.6%)	460 (87.1%)	19,905 (82.5%)
Black	1,649 (6.7%)	31 (5.9%)	1,618 (6.7%)
Asian	2,394 (9.7%)	33 (6.2%)	2,361 (9.8%)
Other	241 (1.0%)	4 (0.8%)	237 (1.0%)
Unknown	1,800	15	1,785
Breast density			
Not dense	11,659 (44.1%)	216 (39.8%)	11,443 (44.2%)
Dense	14,786 (55.9%)	327 (60.2%)	14,459 (55.8%)
Unknown	4	0	4

Values are number (%).

3.2. *Statistical Analysis*

A mammography exam was used as the unit of analysis. Nonindependence of multiple exams from the same women was accounted for in calculations of 95% confidence intervals (CIs) and p-values by using generalized estimating equations (GEE) or the nonparametric bootstrap, clustered by woman[26]. The Mirai algorithm provides cumulative risk predictions for years 1-5 following the index examination. The outcome used for evaluating the performance of Mirai was the presence/absence of a cancer diagnosis at each timeframe. The discrimination performance of Mirai was evaluated using receiver operating characteristic (ROC) curves, the area under the ROC curve (AUC), and Uno’s concordance index (c-index) as an overall summary over the 5-year timeframe[27]. The calibration of Mirai was evaluated using calibration plots and corresponding summaries of overall calibration (calibration-in-the-large) and the calibration slope[28]. To help distinguish between breast cancer detection vs. risk prediction performance, we performed the analyses using all available exams and then repeated the analyses after excluding exams that had a breast cancer diagnosis within six months. All statistical analyses were conducted using R (version 4.3, R Foundation for Statistical Computing, Vienna, Austria). All hypothesis tests were two-sided, with statistical significance defined as $p < 0.05$.

3.3. *Discrimination Performance*

AUCs ranged from 0.81 (95% CI: 0.75-0.86) for 1-year cancer outcomes with the 1-year Mirai scores to 0.70 (95% CI: 0.0.67-0.72) for 5-year cancer outcomes with the 5-year Mirai scores when including all examinations (Figure 7, Table 2). The c-index was 0.70 (95% CI: 0.67-0.72). After excluding 70 exams with a cancer diagnosis within six months, the AUC was 0.72 (95% CI: 0.56-

0.84) at 1 year and 0.68 (95% CI: 0.65-71) at 5 years, while the c-index was 0.68 (95% CI: 0.65-0.70). These values were more similar to previously reported results in other cohorts[19,29] after applying the same type of exclusion (Table 2)[29], though they were still on the lower end of the range.

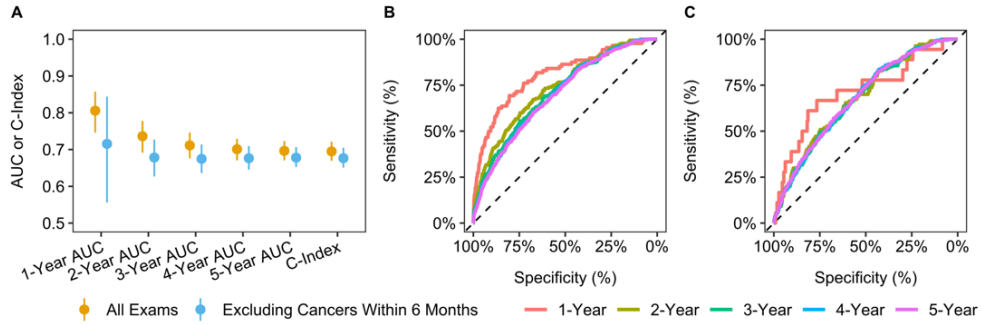


Fig. 7. Discrimination performance of Mirai. Panel A: ROCAUC values over time and the overall c-index. The orange values are based on all exams, and the blue values are after excluding cancers within six months. Error bars represent 95% CIs. Panel B: ROC curves at different time points based on all exams. Panel C: ROC curves at different time points after excluding cancers within six months.

Table 2. Discrimination performance of Mirai in the University of Washington and 7 previously reported cohorts.

	1-Year AUC (95% CI)	5-Year AUC (95% CI)	C-index (95% CI)
All Exams			
University of Washington, USA	**0.81 (0.75-0.86)**	**0.70 (0.67-0.72)**	**0.70 (0.67-0.72)**
MGH, USA[19]	0.84 (0.80-0.87)	0.76 (0.73-0.79)	0.75 (0.72-0.78)
Novant, USA[29]	0.78 (0.73-0.84)	0.75 (0.70-0.80)	0.75 (0.70-0.80)
Emory, USA[29]	0.83 (0.81-0.86)	0.76 (0.74-0.79)	0.77 (0.75-0.79)
Maccabi-Assuta, Israel[29]	0.86 (0.81-0.91)	0.75 (0.71-0.79)	0.77 (0.73-0.81)
Karolinska, Sweden[19]	0.90 (0.89-0.92)	0.78 (0.76-0.80)	0.81 (0.79-0.82)
CGMH, Taiwan[19]	0.90 (0.87-0.93)	0.79 (0.75-0.82)	0.79 (0.76-0.83)
Barretos, Brazil[29]	0.89 (0.86-0.93)	0.82 (0.78-0.86)	0.84 (0.81-0.88)
Excluding Cancers within 6 Months			
University of Washington, USA	**0.72 (0.56-0.84)**	**0.68 (0.65-0.71)**	**0.68 (0.65-0.70)**
MGH, USA[19]	0.71 (0.60-0.84)	0.71 (0.68-0.75)	0.69 (0.66-0.73)
Novant, USA[29]	N/A	0.72 (0.66-0.79)	0.72 (0.66-0.79)
Emory, USA[29]	0.74 (0.66-0.84)	0.71 (0.68-0.74)	0.69 (0.66-0.72)
Maccabi-Assuta, Israel[29]	N/A	0.68 (0.62-0.74)	0.70 (0.64-0.76)
Karolinska, Sweden[19]	N/A	0.71 (0.69-0.73)	0.71 (0.69-0.74)
CGMH, Taiwan[19]	0.84 (0.72-0.99)	0.70 (0.66-0.75)	0.70 (0.66-0.75)
Barretos, Brazil[29]	0.87 (0.80-0.94)	0.75 (0.70-0.80)	0.78 (0.74-0.83)

MGH = Massachusetts General Hospital; CGMH = Chang Gung Memorial Hospital.

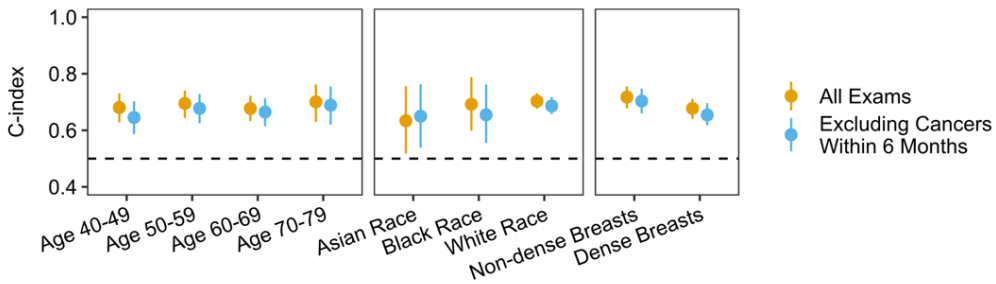


Fig. 8. Discrimination performance of Mirai within subgroups. Error bars: 95% CIs. Dashed line: AUC = 0.50.

Discrimination, as measured by the overall c-index, was also examined within subgroups defined by age, race, and breast density, as shown in Figure 8. There were no statistically significant differences in the c-index between subgroups (unadjusted $p > 0.094$ for each comparison).

3.4. *Calibration Performance*

Calibration plots for Mirai risk predictions versus observed at different timeframes are shown in Figure 9. The corresponding metrics of overall calibration (observed risk minus mean predicted risk) and the calibration slope are shown in Table 3. When all exams are included, the metrics indicated significantly overestimated risk in years 1-2 (overall calibration: -0.15% to -0.10%, $p < 0.014$ for both), but that Mirai was overall reasonably well calibrated for the later years, where the 95% CIs for overall calibration included zero (no difference between observed and predicted risk on average) and the 95% CIs for the calibration slope included 1 (predictions were not more or less extreme [farther from the mean] than observed on average).

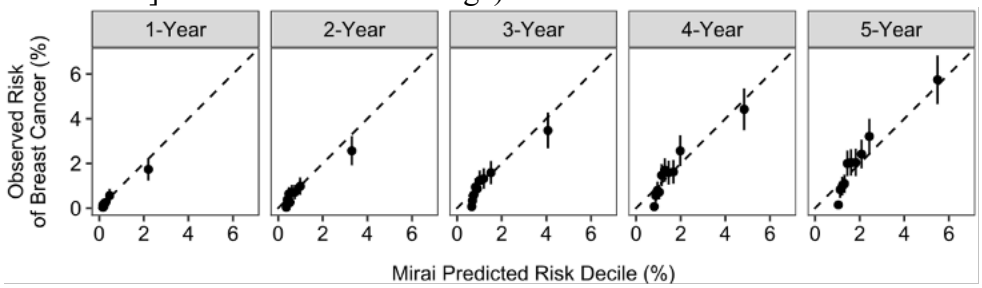


Fig. 9. Calibration plots of Mirai with all exams included, with predicted risks grouped into deciles of approximately equal size. Error bars represent 95% CIs. The dashed line corresponds to perfect calibration (intercept = 0, slope = 1).

Table 3. Calibration statistics for Mirai.

	All Exams				After Excluding Cancers within 6 Months			
	Overall Calibration*		Calibration Slope†		Overall Calibration*		Calibration Slope†	
Timeframe	Estimate (%)	(95% CI)	Estimate	(95% CI)	Estimate (%)	(95% CI)	Estimate	(95% CI)
1-year risk	-0.10	(-0.17, -0.03)	0.83	(0.52, 1.22)	-0.36	(-0.39, -0.32)	0.05	(0.00, 0.12)
2-year risk	-0.15	(-0.25, -0.04)	0.76	(0.51, 1.06)	-0.40	(-0.48, -0.32)	0.21	(0.08, 0.36)
3-year risk	-0.13	(-0.29, 0.03)	0.89	(0.60, 1.20)	-0.38	(-0.53, -0.26)	0.40	(0.20, 0.63)
4-year risk	-0.05	(-0.25, 0.14)	0.90	(0.63, 1.20)	-0.31	(-0.49, -0.12)	0.50	(0.27, 0.76)
5-year risk	0.09	(-0.15, 0.33)	1.03	(0.75, 1.32)	-0.16	(-0.38, 0.07)	0.66	(0.44, 0.91)

*Observed risk minus mean predicted risk; a value > 0 indicates the prediction under-estimated risk on average, and a value < 0 indicates the prediction over-estimated risk.

†A well-calibrated model has a calibration slope of 1; slope > 1 indicates that high predictions tended to underestimate risk (not high enough) and low predictions tended to overestimate risk (not low enough); slope < 1 indicates predictions tended to be more extreme than observed (high values too high and low values too low).

When exams with cancer diagnoses within six months were excluded, the calibration metrics substantially worsened (Table 3). Overall, Mirai significantly overestimated risk, more so at earlier timeframes (overall calibration: -0.40% to -0.36% in years 1-2 and -0.31% to -0.16% in years 4-5), and the calibration slopes were significantly less than 1 at all timeframes (calibration slopes: 0.05 to 0.66, $p < 0.012$ across years).

Thus, ClinValAI enabled the establishment of an effective cloud infrastructure to successfully perform the clinical validation of Mirai on a large and diverse dataset to study its generalizability.

4. Discussion

We introduce ClinValAI to promote the external clinical validation of AI algorithms on medical imaging exams, thereby providing the opportunity to reliably understand their real-world performance in healthcare settings and their impact on patient care, health, and safety. Our framework can be leveraged to evaluate the generalizability of deep learning models on healthcare data from diverse demographics to analyze the differences in performance across various sub-populations and identify biases. ClinValAI can facilitate the detection of models' failure modes and enable an understanding of AI's potential to function as a standalone tool for diagnostic applications.

An important consideration while using our work is the requirement to specify necessary programmatic steps while configuring Docker images to execute individual mechanisms. However, ClinValAI's multiple customizable features enhance its usability for validating AI models.

Our presented analysis is limited to just one deep learning algorithm. As a next step, we plan to leverage ClinValAI to perform a rigorous external validation study of Mirai and three commercial AI algorithms for breast cancer risk prediction on a large and diverse dataset of $\geq$ 40,000 mammograms from seven registries affiliated with the Breast Cancer Surveillance Consortium (BCSC). Utilizing our framework for this study will enable the evaluation of model performance at the woman, exam, and tumor levels, facilitating a comprehensive assessment of the generalizability of AI models. While we showcase ClinValAI's usability for medical imaging models, our work can be extended to validate AI models for various biomedical data modalities.

Finally, ClinValAI is equipped to provide opportunities to periodically retest performance. Vendors can analyze performance based on detailed findings from the results communicated by our framework. This encourages the development of explainable models to better reason performance, thereby enhancing the potential of receiving clinicians' trust. The streamlined feedback mechanism can support targeted algorithm fine-tuning efforts. This can foster enhanced academic-industry partnerships. The continuous monitoring feature enables analyzing variations in model performance vis-à-vis data drift and model drift. Overall, ClinValAI can pave the way for studying the capabilities of AI algorithms in optimizing clinical workflows and reducing the burden on the medical fraternity. ClinValAI's codebase and scripts for statistical analysis can be accessed here: https://github.com/OjasRamwala/ClinValAI.

5. Conclusion

The rise in commercial AI algorithms in clinical medicine and the associated generalizability concerns make rigorous validation indispensable to the clinical translation of AI tools. ClinValAI addresses critical challenges associated with external validation efforts and provides an easily customizable and cloud-agnostic framework to build scalable infrastructures to audit and monitor AI algorithms. By enabling large-scale external validation efforts on data from diverse cohorts, our work has the potential to foster health equity and overcome health disparities by promoting the development of robust, interpretable, and generalizable AI algorithms for healthcare applications.

Acknowledgments

This work was funded in part by the National Cancer Institute (grants P01CA154292, R01CA262023, R37CA240403, and R37CA292399), the American Cancer Society (grant 21-078-01-CPSH), the University of Washington Institute of Medical Data Science Pilot Award, an Amazon Web Services Health Equity Award, and the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

References

1. Parmar, S. K., Ramwala, O. A. & Paunwala, C. N. Performance Evaluation of SVM with Non-Linear Kernels for EEG-based Dyslexia Detection. in *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)* 1–6 (2021). doi:10.1109/R10-HTC53172.2021.9641696.
2. Ramwala, O. A., Paunwala, C. N. & Paunwala, M. C. GRU-Based Parameter-Efficient Epileptic Seizure Detection. in *Biomedical Signal and Image Processing with Artificial Intelligence* (eds. Paunwala, C. et al.) 73–86 (Springer International Publishing, Cham, 2023). doi:10.1007/978-3-031-15816-2_4.
3. Fatemi, M. *et al.* Inferring spatial transcriptomics markers from whole slide images to characterize metastasis-related spatial heterogeneity of colorectal tumors: A pilot study. *J. Pathol. Inform.* **14**, 100308 (2023).
4. Schopf, C. M. *et al.* Artificial Intelligence-Driven Mammography-Based Future Breast Cancer Risk Prediction: A Systematic Review. *J. Am. Coll. Radiol.* (2023).
5. Ramwala, O. A., Dhakecha, S. A., Ganjoo, A., Visiya, D. & Sarvaiya, J. N. Leveraging Adversarial Training for Efficient Retinal Vessel Segmentation. in *2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* 1–6 (2021). doi:10.1109/ECAI52376.2021.9515093.
6. COVID-19 Diagnosis from Chest Radiography Images using Deep Residual Network | IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9225521.
7. Mulchandani, H. *et al.* Tonsillitis based Early Diagnosis of COVID-19 for Mass-Screening using One-Shot Learning Framework. in *2020 IEEE 17th India Council International Conference (INDICON)* 1–6 (2020). doi:10.1109/INDICON49873.2020.9342371.
8. Novel Multi-Modal Throat Inflammation and Chest Radiography based Early-Diagnosis and Mass-Screening of COVID-19. *Open Biomed. Eng. J.* **15**, 226–234 (2021).
9. Dalal, P. *et al.* Throat Inflammation Based Mass Screening of Covid-19 on Embedded Platform. in *Soft Computing and its Engineering Applications* (eds. Patel, K. K., Garg, D., Patel, A. & Lingras, P.) 277–288 (Springer, Singapore, 2021). doi:10.1007/978-981-16-0708-0_23.
10. Levy, J. *et al.* Artificial Intelligence, Bioinformatics, and Pathology: Emerging Trends Part I—an Introduction to Machine Learning Technologies. *Adv. Mol. Pathol.* **5**, e1–e24 (2022).
11. Levy, J. *et al.* Artificial Intelligence, Bioinformatics, and Pathology: Emerging Trends Part II––Current Applications in Anatomic and Molecular Pathology. *Adv. Mol. Pathol.* **5**, e25–e52 (2022).
12. Yang, X. *et al.* A large language model for electronic health records. *Npj Digit. Med.* **5**, 1–9 (2022).
13. Wu, S. *et al.* Deep learning in clinical natural language processing: a methodical review. *J. Am. Med. Inform. Assoc. JAMIA* **27**, 457–470 (2019).

14. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms | Breast Cancer | JAMA Network Open | JAMA Network. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2761795.
15. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
16. Maleki, F. *et al.* Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. *Radiol. Artif. Intell.* **5**, e220028 (2022).
17. Hsu, W. *et al.* External Validation of an Ensemble Model for Automated Mammography Interpretation by Artificial Intelligence. *JAMA Netw. Open* **5**, e2242343 (2022).
18. Ramwala, O. A. *et al.* Establishing a Validation Infrastructure for Imaging-Based Artificial Intelligence Algorithms Before Clinical Implementation. *J. Am. Coll. Radiol.* (2024) doi:10.1016/j.jacr.2024.04.027.
19. Yala, A. *et al.* Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* **13**, eaba4373 (2021).
20. Mooney, S. J. & Pejaver, V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annu. Rev. Public Health* **39**, 95–112 (2018).
21. Alternative models for sharing confidential biomedical data | Nature Biotechnology. https://www.nature.com/articles/nbt.4128.
22. Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction | Journal of the American Medical Informatics Association | Oxford Academic. https://academic.oup.com/jamia/article/27/9/1393/5868591.
23. Voss, K., Gentry, J. & Auwera, G. V. D. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. (2017) doi:10.7490/F1000RESEARCH.1114631.1.
24. miniwdl — miniwdl documentation. https://miniwdl.readthedocs.io/en/latest/.
25. D'Orsi, C. J., Sickles, E. A., Mendelson, E. B. & Morris, E. A. *2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. (American College of Radiology, 2014).
26. Huang, F. L. Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters. *Educ. Psychol. Meas.* **78**, 297–318 (2018).
27. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117 (2011).
28. Crowson, C. S., Atkinson, E. J. & Therneau, T. M. Assessing Calibration of Prognostic Risk Scores. *Stat. Methods Med. Res.* **25**, 1692–1706 (2016).
29. Yala, A. *et al.* Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **40**, 1732–1740 (2022).

PGxQA: A Resource for Evaluating LLM Performance for Pharmacogenomic QA Tasks[†]

Karl Keat[1]*, Rasika Venkatesh[1]*, Yidi Huang[1]*, Rachit Kumar[1], Sony Tuteja[2], Katrin Sangkuhl[3], Binglan Li[3], Li Gong[3], Michelle Whirl-Carrillo[3], Teri E. Klein[3,4,5], Marylyn D. Ritchie[6,7,8]**, Dokyoon Kim[7,8]**
[1]*Genomics and Computational Biology Graduate Program,* [2]*Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[3]*Department of Biomedical Data Science,* [4]*Department of Medicine (BMIR),* [5]*Department of Genetics, Stanford University, Stanford, CA USA*
[6]*Department of Genetics,* [7]*Institute for Biomedical Informatics,* [8]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA*
*, **The Authors contributed equally to this work
***Emails: marylyn@pennmedicine.upenn.edu, dokyoon.kim@pennmedicine.upenn.edu*

Pharmacogenetics represents one of the most promising areas of precision medicine, with several guidelines for genetics-guided treatment ready for clinical use. Despite this, implementation has been slow, with few health systems incorporating the technology into their standard of care. One major barrier to uptake is the lack of education and awareness of pharmacogenetics among clinicians and patients. The introduction of large language models (LLMs) like GPT-4 has raised the possibility of medical chatbots that deliver timely information to clinicians, patients, and researchers with a simple interface. Although state-of-the-art LLMs have shown impressive performance at advanced tasks like medical licensing exams, in practice they still often provide false information, which is particularly hazardous in a clinical context. To quantify the extent of this issue, we developed a series of automated and expert-scored tests to evaluate the performance of chatbots in answering pharmacogenetics questions from the perspective of clinicians, patients, and researchers. We applied this benchmark to state-of-the-art LLMs and found that newer models like GPT-4o greatly outperform their predecessors, but still fall short of the standards required for clinical use. Our benchmark will be a valuable public resource for subsequent developments in this space as we work towards better clinical AI for pharmacogenetics.

Keywords: Pharmacogenetics; Pharmacogenomics, Large Language Models, Artificial Intelligence, Clinical Informatics.

1. Introduction

1.1. *Pharmacogenetics*

Pharmacogenetics (PGx) is the study of the role of genetics on an individual's response to medication, with the aim of bringing tools to the clinic that can utilize a patient's genetic information to improve medication safety and efficacy. Genetic variations that lead to changes in the activity or availability of drug metabolizing enzymes (DMEs), receptors, channels, and other proteins involved in pharmacodynamics and pharmacokinetics can contribute strongly to interindividual variability in drug response, resulting in an increased risk of adverse drug reactions (ADRs) and nonresponse

[†]KK is funded through National Human Genome Research Institute (NHGRI) F31HG013246. RK is funded through NHGRI T32HG000046. This work was also supported by the following grants from the National Institutes of Health (NIH): U24HG010615, U24HG013077, UL1TR001878, and K23HL143161.

phenotypes.[1] By identifying genetic markers that influence drug response, PGx enables healthcare providers to predict which patients are more likely to experience adverse reactions or treatment failure. This knowledge allows for more individually tailored medication regimens, optimizing therapeutic outcomes while minimizing the risk of side effects.[2] The overarching goal of PGx is promoting personalized medicine, such that patients receive the right drug and the right dose, at the right time. In doing so, the field aims to improve patient outcomes, enhance medication safety, and reduce healthcare costs associated with ineffective or harmful treatments.

Despite the availability of numerous well-characterized, clinically actionable PGx guidelines for widely used medications, the clinical implementation of PGx has been slow. Very few medical centers and clinics routinely use this technology. This gap is due to various factors such as a lack of awareness and education among healthcare providers, the constantly evolving body of PGx guidelines, and technical challenges in integrating PGx data into electronic health records (EHRs).[3] The cost of PGx testing and variable insurance coverage can also pose significant financial barriers, while regulatory and legal concerns may also impact the extent of implementation of PGx testing in hospital systems.[4] Lack of domain expertise and education among healthcare providers, patients, and researchers in particular poses a critical barrier to the implementation of PGx-guided therapies in clinical settings as this leads to difficulty understanding and interpreting test results, in addition to limited research conducted regarding the clinical impact of such technologies.[5]

1.2. *Existing PGx Resources and Limitations*

Given that there are many causes for interindividual variability in treatment response as well as a need for guidance in interpreting PGx screening results, multiple independent bodies of experts have published research and guidelines to inform PGx-guided treatment. The Clinical Pharmacogenetics Implementation Consortium (CPIC) is one such group that has generated a set of specific drug recommendations to guide prescribing practices in the presence of genetic test results. CPIC has established 43 evidence-based clinical guidelines for 151 commonly prescribed medications. These recommendations were created based on a large body of evidence showing the impact of known PGx alleles in altering drug metabolism or response. Level A refers to gene-drug pairs where genetic information "should be used" for prescribing decisions and alternative therapies or dosing are highly likely to be effective and safe. At least one moderate or strong action (change in prescribing) is recommended for Level A pairs. Level B refers to pairs where genetic information "could be used" to change prescribing because alternative therapies/dosing are extremely likely to be as effective and as safe as non-genetically based dosing. Other international committees with their own sets of guidelines include The Dutch Pharmacogenetics Working Group (DPWG), and the French National Network (Réseau) of Pharmacogenetics (RNPGx).[6] The Pharmacogenomics Knowledge Base (PharmGKB), is a resource that aims to comprehensively aggregate, curate, and characterize PGx knowledge including the literature and guidelines from these distinct sources.[7]

While these resources are highly comprehensive, most require a moderate to high degree of domain knowledge to understand and interpret the provided information. Clinicians and patients, in particular, need PGx expertise to understand reports and utilize them to inform treatment decisions. Clinicians typically receive limited PGx training and therefore rely heavily on these resources for guidance.[5,8–12] Moreover, differences among guidance sources and the rapid pace of new discoveries and guidelines create potential for misunderstandings and confusion. While PharmGKB curates,

aggregates, and presents guidance across sources, clinicians, patients, and researchers may prefer an interface that allows them to query and access targeted information using natural language instead of menus and tables.

1.3. *Opportunities for Large Language Models to Guide PGx*

Large language models (LLMs) represent a major advance in artificial intelligence, allowing for the creation of seemingly intelligent chatbots which can interpret questions and assist with various tasks. LLMs have shown promise in a variety of natural language tasks, including those in medicine. For example, chatbots using LLMs can accurately answer patient queries in a conversational manner preferred by patients. GPT-4 has also achieved human-level accuracy on the United States Medical Licensing Exam (USMLE), outperforming the minimum passing threshold on short answer and multiple-choice questions.[13] LLMs have been proposed for integration into clinical workflows to handle administrative tasks, which include managing appointment scheduling by patient request, answering routine inquiries about medication or treatment plans, and assisting in the preparation of medical records.[14,15] Additionally, LLMs can support clinical decision-making by providing real-time information retrieval and analysis, potentially reducing the cognitive load on healthcare professionals and improving patient outcomes.[16] For these reasons, advances in LLMs have created an exciting opportunity to build chatbots to assist with complex medical specialties like PGx, providing a powerful and intuitive interface to access pharmacogenetic knowledge.

Despite the promise of LLMs in medicine, there are significant issues that must be addressed before widespread clinical integration. These models are limited to the information they were trained on and can produce fabricated responses with an authoritative and confident tone when lacking information. There are numerous examples of this phenomenon across disciplines, but this poses a particularly large barrier to use in healthcare, where real time patient decisions rely on the presence of accurate information and mistakes can cost lives.[17–19] Moreover, LLMs are costly to update and retrain as new information becomes available.[20–22] This poses a challenge in fields where clinical guidelines are routinely updated, such as in PGx, and even current state-of-the-art LLMs had their training data capped several months before the latest CPIC guideline release. Despite these risks, LLMs are already being employed by clinicians, patients, and researchers to answer medical questions and their performance must be studied in order to understand their limitations.[23]

1.4. *Prior work on LLMs for PGx*

PGx is a specialized area of medicine with limited and variable levels of coverage in the US medical and pharmacy curriculum.[5,10–12] Despite this, PGx has a wide impact on several specialties due to the variety of drugs with actionable guidelines. Therefore, leveraging LLMs in this field has the potential to significantly enhance clinical practice and patient care. For instance, Murugan et al., used GPT-4 and retrieval-augmented generation (RAG) to build PGx4Statins, a PGx chatbot for answering questions about statin therapy guidelines.[24] However, the limitations of LLMs may pose a particular risk in this field, as PGx guidelines are revised and updated irregularly as new evidence becomes available, and inaccurate or outdated advice may result in adverse drug reactions or

treatment nonresponse. As such, any PGx chatbot would need to be thoroughly vetted before clinical implementation is possible.

While the performance of LLMs at answering general medical questions has been demonstrated, there is limited data on how LLMs perform with PGx queries. Prior to now, there have been no comprehensive, publicly available benchmarks to assess the performance of LLM chatbots in answering PGx queries. PGx4Statins was benchmarked manually, requiring a team of scorers to rate LLM responses based on the criteria of accuracy, relevancy, risk management, language clarity, bias neutrality, empathetic sensitivity, citation support, and hallucination limitation on a 1-5 scale. While this likely represents a gold-standard approach for evaluating real-world performance of a PGx clinical chatbot, PGx4Statins was only able to be tested on a small number of questions and for a single drug, demonstrating the limitations of this evaluation strategy.[24] As new chatbots and language models are released, a more scalable solution is needed to comprehensively test the accuracy of these tools, so that we can then prioritize top performers for more rigorous, labor-intensive testing.

To address the absence of evaluation strategies for PGx chatbots, we have developed PGxQA, a resource for evaluating the performance of LLMs in a variety of PGx-related tasks for multiple identified stakeholders: patients, clinicians, and researchers. PGxQA consists of a large corpus of PGx questions generated directly from CPIC data resources, CPIC PGx guidance for Level A drug-gene pairs, or provided by experts in the field. In addition, PGxQA includes tools for higher throughput manual and automated evaluation of accuracy and completeness. PGxQA's question set covers all of the CPIC Level A guidelines across several dimensions, such as translating genotypes into phenotypes, naming the dbSNP ID(s) for variant(s) that define a particular star-allele, and most importantly, translating phenotypes into clinical recommendations. These resources will help promote the responsible development of medical chatbots by allowing us to assess their knowledge of PGx topics, thus lowering barriers to implementation of PGx in the clinic and providing easier access to PGx knowledge for clinicians, patients, and researchers.

2. Methods

2.1. *Automated Question Generation*

To generate a meaningfully large corpus of evaluation questions, a significant proportion of the question bank was generated using custom python scripts to extract relevant information from the 'CPIC Data' database from their GitHub repository and format the information as question-answer pairs.[25] The psycopg2 package was used to load and query CPIC's postgresql database and pandas was used to output tables of questions.[26–28]

Due to a large degree of redundancy in questions and the potential for an over-weighting of pharmacogenes with many defined star alleles in our overall scoring, we implemented a subsetting tool which takes each set of questions and drops redundant questions to maintain roughly even proportions of questions based on which genes they cover and what answer choices they cover. All generated questions are available for download, such that users can run the entire set or generate custom subsets based on their own criteria.

2.2. *LLM Querying*

To query the various studied LLMs, we wrote a set of python scripts to load in our questions and send them to a local or remote LLM server. We defined a universal base prompt for all LLMs to ensure that all LLMs are working with similar basic instructions. We used the 'openai' python package along with an OpenAI API key to remotely query GPT-3.5-turbo, GPT-4-turbo, and OpenAI's latest model as of writing, GPT-4o. We were also able to use the 'openai' python interface to send queries to a locally hosted instance of the open-source LLM Llama 3. Lastly, we used the 'requests' library in python to connect to Google's Generative Language REST API to query Gemini 1.5 Pro, Google's flagship LLM product.[29] We used our python code to query the LLMs with all of the questions in our subsets, outputting tables containing the original question, question metadata, the ground-truth reference answer, the LLM answer, and some automated scoring metrics.

2.3. *Manual Question Generation*

2.3.1. *External Provided Questions*

While the structured information within the CPIC database allows us to cover a large proportion of the potential use cases for a PGx chatbot, we sought out real world sources of PGx questions to represent what information is being sought by actual clinicians, researchers, and patients. We acquired a set of questions sent to PharmGKB scientists from 2020-2024, containing queries about PGx and the PharmGKB scientists' responses. Additionally, we obtained an anonymized set of questions and answers from Penn Medicine's Pharmacogenetics Consult Service, which provided a rich source of clinician-centric questions on PGx testing, results interpretation, and other relevant queries. We manually pruned these datasets to stay within our scope of queries about PGx information retrieval and formatted them into tables as short answer questions for our LLMs.

2.3.2. *Adversarial Questions*

To assess how the models perform when presented with incorrect information, insufficient information, or information outside of the scope of queries regarding PGx, we devised sets of structured adversarial questions. These queries were structured to be nearly identical to the question bank extracted directly from the CPIC database, with the exception of having extraneous or missing information. For these queries, we evaluate whether LLMs answer that sufficient information was not available to answer the question, scoring based on the rate of refusal to respond. We additionally ran the whole set of LLM queries, giving the LLMs the option to refuse to respond, as to compare refusal rates between standard and adversarial queries.

2.4. *Automated LLM Metrics*

To rapidly score the large corpus of questions and reduce reliance on expert labor, we generated a set of automated scoring functions to directly measure or approximate the performance of the LLMs on each specific task.

2.4.1. *Numeric Scoring*

For questions requiring a numeric answer, such as the allele frequency tests, LLMs were instructed to format their response as a number. We then parsed out this number and calculated the mean absolute deviation (defined as the mean of the differences) between the LLM answer and the reference answer for the entire question set.

2.4.2. *Information Retrieval Scoring*

For questions where the task involved returning non-sentence information such as dbSNP IDs, gene symbols, or generic drug names, we instructed the LLMs to return the desired information in a predictable format that can be parsed using regular expressions or by splitting a defined delimiting character like ';'. For question sets where there are multiple values making up the answer (for example to list all of the drugs which have CPIC guidelines linked to a particular gene), performance was measured as precision and recall, where precision is the proportion of values in the LLM answer that are found in the reference answer, and recall is the proportion of values in the reference answer that were correctly included in the LLM answer.

2.4.3. *Multiple Choice Scoring*

For question sets where the questions had a small finite set of possible answers, we constrained them to multiple choice, where the LLM was told to select the correct answer from a provided list of options, facilitating the process of detecting if the LLM answered correctly programmatically. For these queries, the accuracy of the LLM in identifying the correct response was computed as the proportion of answers that were correctly selected.

2.4.4. *Automated Text Similarity Metrics*

In the case of short-answer questions where we wanted the LLMs to answer in one or two sentences, it is nontrivial to directly score the accuracy without human graders with the expertise to evaluate the answers, which presents a scalability issue. To roughly approximate human scoring, we computed automated text similarity metrics between the LLM answer and a human-written reference answer. Specifically, we compute the cosine similarity of the answers under different text embedding models as well as BERTScore using the microsoft/deberta-xlarge-mnli base model. We selected the model that most closely resembled human judgement by comparing the embedding scores' concordance with human-scored answers.[30–34] We then calculated the "win-rate" of the LLM answers by looking at the percentage of answers where the LLM similarity score to the reference answer was higher than the LLM similarity score to a generic discordant answer. For example, if asked to make a clinical recommendation, where the correct answer is to avoid the drug and the discordant answer is to take the drug as normal, the LLM would "win" if its answer has higher similarity to the reference answer than the discordant answer.

2.5. *Human Review of LLM Answers*

2.5.1. *Concordance with Automated Metrics*

To determine which text metric best captures the semantics of PGx recommendations, we manually reviewed a set of 77 short-answer questions and responses from GPT-4o. For each question, we manually annotated whether the LLM answer was closest to the ground truth reference answer, or an alternative response containing a discordant recommendation. Using these human labels as ground truth, we computed the F1 score of each text metric by classifying an example positive if the LLM-reference pair has the highest metric value among all LLM-response pairs.[30–34] We found that BERTScore Precision maximizes agreement with human judgment.

2.5.2. *Subject Matter Expert Reviews*

We recruited 4 PGx experts to perform a granular manual review of a selected subset of short-answer LLM responses. For each question, reviewers were shown a human-written and LLM-generated response in randomized blinded order and asked to rate each answer on a five-point Likert scale along attributes of accuracy (i.e. "This response is clinically/scientifically accurate"), completeness (i.e. "This response contains all of the necessary information to address the question fully"), and safety (i.e. "This answer does not pose any danger to human health or safety). For each question, reviewers were also presented with the relevant CPIC guideline document. Ratings were collected using the open-source Data Annotator for Machine Learning tool[35], which was deployed on an AWS EC2 instance with a public IP address so that expert reviewers from around the country could easily work on the assigned scoring task or quit and return to the task later.

2.6. *Data Analysis and Visualization*

The results of our various scoring approaches were analyzed in a Jupyter notebook with pandas, which is included in the GitHub repository for this project.[27,28,36] All plots were generated using the matplotlib and seaborn python packages.[37,38]

3. Results

3.1. *The PGxQA Question Corpus*

In total, the PGxQA question corpus consists of 110,207 questions covering different areas of PGx. While we subsequently present our own tools for querying and evaluating LLMs using this expansive dataset, we make available the entire set of questions as a resource agnostic of downstream evaluation approach. We detail the question types covered in **Table 1**.

Table 1: Representative examples of PGxQA questions generated from CPIC database or external sources

Question Type	Description	Number of questions	Example Prompt	Expected Response
Allele frequency	Ask for a value indicating the allele frequency of a given allele in a population.	2,548	"What is the average allele frequency of ABCG2 rs2231142 reference (G) in the African American/Afro-Caribbean population? Respond with just a number, rounded to 4 decimal places, with no additional text."	0.9651
Allele definition	Ask for dbSNP IDs for variants that define or are part of a given allele. Note that some alleles consist of multiple SNPs.	901	"What SNPs are in the allele definition for CFTR F1052V? Provide a dbSNP ID (also known as an rsID, starting with rs) when available."	rs150212784
Allele function	Determine how an allele affects the overall function of a gene.	1,111	"What is the allele functionality of CYP2C9 *9? Please select the answer from the following choices: {'Normal function', 'Decreased function', 'Uncertain function', 'No function', 'Unknown function'}, and respond with only your selection."	Normal function
Genes to drugs	Ask for drugs with actionable CPIC guidelines for a given gene. Note that multiple drugs can be listed.	22	"Which drugs have actionable CPIC guidelines for CYP2C19? Please respond with nothing but a list of generic drug names delimited by ';'."	pantoprazole;sertraline;omeprazole;lansoprazole;amitriptyline;citalopram;voriconazole;escitalopram;clopidogrel
Diplotype to phenotype	Ask what the defined pharmacogenetic phenotype is for a given set of alleles in a gene.	101,138	"What is the pharmacogenetic phenotype for CYP2C9 *1/*1? Please select the answer from the following choices: {'Intermediate Metabolizer', 'Normal Metabolizer', 'Poor Metabolizer', 'Indeterminate'}, and respond with only your selection."	Normal Metabolizer
Drugs to genes	Ask what genes a clinician might want to include in a panel given what drug a patient is taking OR what genes have actionable guidelines for certain drugs for an interested researcher. Note that multiple genes can be listed.	79 (each); 158 (total)	Clinician: "I want to give my patient paroxetine. What genes should I include in a pharmacogenetics panel? Please respond with nothing but a list of gene symbols delimited by ';'." Researcher: "What genes have actionable pharmacogenetic guidelines for paroxetine? Please respond with nothing but a list of gene symbols delimited by ';'."	Clinician: CYP2D6 Researcher: CYP2D6
Phenotype to category	Given an individual with a certain allele and a drug, provide a guideline for that phenotype-drug combination if applicable in terms of drug dosing (multiple choice). Note: this is a multiple choice version of "Phenotype to guideline".	2,145	"What would be the clinical guidance for someone who is HLA-B*57:01 negative for HLA-B with regards to taking abacavir? Please respond with just 'Avoid' if the guidance is to avoid the drug or take an alternate drug, 'Alter dose' if the guideline is to raise, lower, or start with a specific dose, or 'Unchanged', if there are no clinical recommendations or there is no deviation from standard care for this phenotype and drug."	Unchanged
Phenotype to guideline	Ask the LLM to, given an individual with a certain allele and a drug, provide a guideline for that allele-drug combination if applicable in a short-answer format (not multiple choice). Note: this is a short answer version of "Phenotype to category".	2,133	"What would be the clinical guidance for someone who is HLA-B*57:01 negative for HLA-B with regards to taking abacavir?"	Use abacavir per standard dosing guidelines
Adversarial questions (refusal)	For the above categories, provide a similar prompt, but with one of the entities (genes, drugs, alleles, etc.) being fabricated or incorrect. A model is expected to refuse to answer.	36	"What SNPs are in the allele definition for QSTG1 reference (C)? Provide a dbSNP ID (also known as an rsID, starting with rs) when available or answer UNKNOWN if unknown."	UNKNOWN
External Questions	Questions provided by one or more external sources, as described in Section 2.3.1. Note that these were all scored manually using expert raters, as described in Section 2.5.2.	15	"My patient underwent a percutaneous coronary intervention (PCI) and I want to prescribe clopidogrel. They had pharmacogenetic testing and are a CYP2C19 rapid metabolizer (*1/*17). Do they need a different dose of clopidogrel from the standard 75 mg daily?"	"Per the current CPIC guidelines, patients who are CYP2C19 poor metabolizers have significantly reduced CYP2C19 activity, and should avoid clopidogrel if possible due to increased risk of adverse cardiac and cerebrovascular events."

3.2. *Automated Performance Metric Results*

3.2.1. *Quantitative or Categorical Responses*

OpenAI's GPT models almost universally performed better than Llama or Gemini on numeric, information retrieval, and multiple-choice query metrics (**Table 2**). In particular, GPT-4o, outperformed or was in second place for nearly every metric. However, overall performance varied widely across question categories, with models performing worse at Allele Definition, Allele Function, Diplotype to Phenotype, and Phenotype to Category questions than the other question categories. Performances of less than 0.5 for most metrics and LLMs indicate that allele-related questions were more likely to lead to incorrect answers, potentially because allele definitions are dependent on contextual information such as genes. This potentially highlights that LLM training data or approaches may not properly encode allele information, particularly if they do not incorporate tabular data like the CPIC allele tables. Additionally, the number of star alleles has grown massively as new variants and combinations of variants are discovered. Limited references to these alleles in scientific literature likely contribute to poor performance, since LLMs primarily draw from natural language and at baseline struggle with tabular data.[39]

In contrast, other categories saw stronger performance such as the "Genes to drugs" or "Drugs to genes" categories, particularly in the average recall of the LLMs in identifying the expected entities. This indicates that entities such as drugs and genes, which have been described in text for much longer, and across a wider variety of sources, may be better encoded within the LLM weights. However, the precision in these categories was lacking for several LLMs, indicating that such LLMs may be prone to so-called "hallucinations" when responding to these questions, or may make claims backed up by inconclusive evidence.

Table 2. Mean scores for each automated question category except for Phenotype to Guideline. The top scoring model for each category is bolded

Question Category	Metric	Llama 3	Gemini Pro 1.5	GPT3.5	GPT4	GPT4o
Allele frequency	Mean Absolute Deviation	0.1178	0.1465	0.1147	0.0601	**0.0561**
Allele definition	Average Precision	0.1443	0.1341	0.1750	**0.2599**	**0.2599**
	Average Recall	**0.2274**	0.1422	0.2107	0.2221	0.2229
Allele function	Accuracy	0.3856	0.3791	0.3333	**0.5033**	0.4771
Genes to drugs	Average Precision	0.2870	0.1364	0.5459	0.4760	**0.6843**
	Average Recall	0.3955	0.1104	**0.6810**	0.6719	0.6300
Diplotype to phenotype	Accuracy	0.3770	0.3455	0.2565	0.3665	**0.4346**
Drugs to genes (clinician)	Average Precision	0.3177	0.1706	0.2169	0.4424	**0.5992**
	Average Recall	0.7679	0.4494	0.7152	0.8481	**0.9367**
Drugs to genes (researcher)	Average Precision	0.4325	0.3430	0.2968	0.5580	**0.8091**
	Average Recall	0.7489	0.5190	0.7278	0.6667	**0.8418**
Phenotype to category	Accuracy	0.4365	0.3538	0.3212	0.4385	**0.5635**
Phenotype to guideline	BERTscore Precision Win rate	0.7056	0.5499	0.7178	**0.7251**	0.7056

3.2.2. *Short Answer Responses*

After comparing each text embedding method to human classification results, the BERTScore Precision metric was the most concordant with human similarity assessments in indicating which of several reference answers the GPT-4o-generated response was the most concordant with (**Figure 1a., Supplementary Table S1**).[30–34] Because this metric seemed the closest to capturing human judgment on a broad scale, we used it as an automated scoring proxy for LLM performance on our short answer "Phenotype to guideline" tests. Based on automated tests, GPT-4-turbo slightly outperformed GPT-3.5-turbo, GPT-4o, and Llama 3 in average win rate as defined in the methods (**Figure 1b.**). However, Gemini-Pro seems to greatly underperform relative to its counterparts, having an average win rate roughly 0.15 lower than the other models, indicating that its answers likely significantly diverged from the other models and from the ground truth reference.

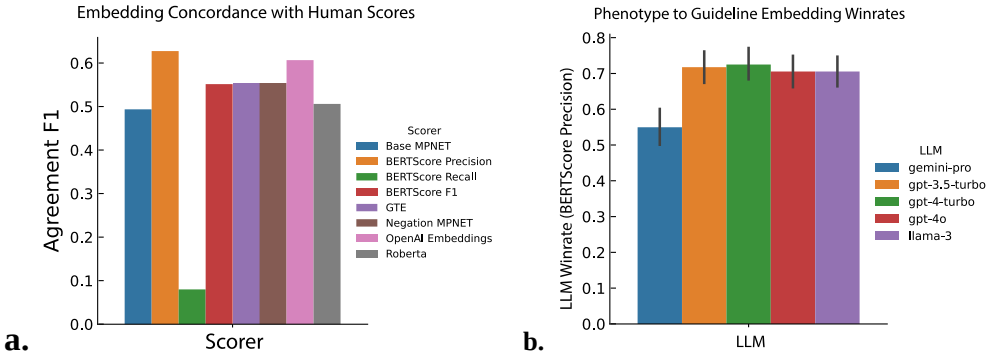


Figure 1: a.) Scorer concordance with human ratings of response similarity as defined by the F1 of the agreement for the Phenotype to Guideline question category for GPT-4o. b.) Model win rates in Phenotype to Guideline Tests.

3.2.3. *Refusal Assessment*

When given the option to refuse to respond, LLMs had highly variable rates of refusal on misspecified and properly specified questions (where misspecified refers to questions where there is not sufficient information to answer, or there exist no clinical guidelines for the requested information). Ideally, a medical chatbot should refuse to answer misspecified questions (a refusal rate of 1 is best) and answer properly specified questions (a refusal rate of 0 is best). Llama, Gemini, and GPT3.5 all refused to answer both types of questions at roughly equal rates. Llama and Gemini tended to refuse very infrequently (<0.2 refusal rate) in either circumstance, while GPT-3.5 refused at roughly equal rates for both circumstances (~0.3 refusal rate) (**Figure 2**). A low refusal rate for misspecified queries might indicate a higher tendency to hallucinate information when given confusing or contradictory queries. In contrast, GPT-4 and GPT-4o showed a higher rate of refusal for misspecified questions (~0.7) compared to properly specified questions (~0.3), indicating that these two models exhibit ability to identify questions with incorrect information as well as a propensity to avoid hallucinations, though there remains significant room for improvement. These results are further broken down in Supplementary Table S2, which shows the refusal rates for different categories.

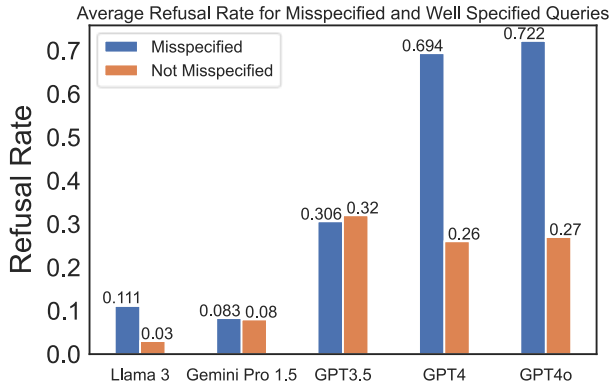


Figure 2: Refusal rates of the different LLMs for misspecified and properly specified question sets.

3.3. *LLM Results with Human Scoring*

3.3.1. *Manual LLM Metrics*

Although the emphasis of this work is on large scale benchmarks that can be employed widely, even in settings where manual expert review would be intractable, it is undeniable that expert reviewers provide invaluable understanding of the nuances and details of PGx which cannot easily be measured by automated scorers and text similarity scores. We recruited 4 PGx experts to manually score a set of GPT-4o responses to 15 short answer questions, and had those same experts score the human-written reference answers. On average, GPT-4o performed lower than the reference answer in all categories, with 'Accuracy' having the largest gap (**Table 3**). While these results reflect that GPT-4o performed well for many questions, there were some answers where it provided highly incorrect or even dangerous responses, such as when it gave incorrect recommendations on tacrolimus PGx in the context of liver transplant.

Table 3. Average Likert scores for Accuracy, Completeness, and Safety of GPT-4o and reference answers as scored by PGx experts

Metric	GPT4o	Reference Answer	Performance Gap
Accuracy (Likert)	3.917	4.917	-1.000
Completeness (Likert)	4.167	4.533	-0.367
Safety (Likert)	4.083	4.850	-0.767

4. Discussion

This work provides a framework and dataset to evaluate LLM-based chatbots in their ability to answer PGx questions derived from gold-standard PGx data sources. In demonstrating our framework, we have highlighted the strengths and weaknesses of LLMs in handling a wide range of PGx queries, providing guidance for future improvements.

4.1. *Avenues for Improving LLMs*

The main limitations we identified in LLM-based chatbots are their especially poor accuracy for queries requesting numeric answers as well as newer or less common star alleles, their tendency to

invent false information instead of refusing to answer unknown queries, and their inability to understand the quality of the underlying sources of their claims. These are broader issues in LLM research, and many techniques have been employed to address them. Prompt-engineering involves devising specific prompts to elicit more comprehensive, more accurate, and better-worded responses from LLMs, which is inexpensive and requires minimal technical expertise, making it highly accessible.[40] However, its ability to enhance results is limited, and excessive engineering can lead to increased token usage per query, potentially raising costs and complexity in processing time.[41,42] This approach was employed in many of the structured answer questions in PGxQA and yielded more concise and readily usable information.

Fine-tuning LLMs on specific datasets of PGx questions, such as those generated in this study, presents an opportunity for models to better understand and respond to domain-specific queries. This approach has been shown to improve the relevance and accuracy of LLM responses. Although fine-tuning can be expensive, requiring significant computational resources like GPUs to train and update the model, it provides a tailored solution for domain-specific prompts.[43] However, fine-tuned models can still hallucinate, as they rely on pre-trained embeddings.[44]

Retrieval augmented generation (RAG) incorporates a retrieval mechanism into LLMs, enabling the model to directly source information from an updated knowledge base. This approach is relatively cheap and straightforward to maintain, as updating the knowledge base is less resource-intensive compared to training the LLM itself.[45] This is ideal for domains such as PGx, where knowledge bases are constantly updated. This also reduces the risk of hallucinations by providing the model with direct access to accurate data sources. However, RAG systems require large context windows for effective querying and a higher degree of human intervention is involved to teach the LLM how to access and utilize these external sources.[44,46]

To address the needs efforts are underway by the PharmGKB/CPIC group at Stanford to create AI-ready data for consumption by LLMs. In addition, collaborative efforts are underway by Dr. Roxana Daneshjou and Dr. Klein's groups at Stanford to develop both clinician-forward and patient-forward tools using generative AI to disseminate this knowledge on the current PharmGKB website and in the future, in the ClinPGx resource.

4.2. *Limitations of PGxQA*

PGxQA is intended to be a framework for initial evaluation of a chatbot in answering PGx questions, particularly in answering questions concordant with pre-existing guidelines (such as information from CPIC, PharmGKB, and others). As shown above, PGxQA provides a variety of metrics that provide insight into several dimensions of the performance of LLMs. However, it is important to recognize that PGxQA has several limitations due to the way that it was devised and developed with a focus on automated assessment. First, the questions in PGxQA are largely created automatically from public PGx data sources. Most questions are query-based—requesting information that would require looking up information from one database and not synthesizing knowledge across multiple databases or fields. This facilitates automated evaluation at the expense of being able to understand this dimension of LLMs, referred to as "multi-hop reasoning". To mitigate this, handcrafted questions and actual questions asked of PGx researchers and clinicians are included through the

“External Questions" category, though LLM responses to these questions cannot fully be assessed automatically.

Our emphasis on automated scoring approaches, while valuable for large-scale evaluation, introduces other limitations as well. We engineered the prompts to instruct the LLM to return answers in our desired format to properly score responses for our information retrieval tasks, introducing a small possibility that asking for results in this strict format alters performance. As shown in the comparison between the clinical and researcher versions of our drug to genes questions, the LLMs do seem to have variable performance when similar questions are asked in different ways. However, this represents a weakness of LLMs that must also be studied prior to clinical use due to the heterogeneous nature of real-life queries. There are also limitations to our text-similarity-based scoring, as text embeddings do not fully capture the nuances of human judgment. Despite these compromises, we believe that PGxQA will still provide useful metrics for chatbot evaluation and we anticipate that future work may address many of the limitations of PGxQA and of LLM chatbots.

4.3. *Future Directions*

Going forward, we expect PGxQA to serve as an automatic evaluation framework to continually evaluate LLMs. This initial evaluation has shown dramatic improvements in performance in more recent models, such as GPT-4o, relative to older iterations such as GPT3.5. We anticipate that further advancements in model architecture and training will strengthen the ability of these models to function as a valuable resource in PGx. Using PGxQA, we can continually monitor improvements in LLM performance and assess new technologies as they are unveiled. The automatic generation of questions from the CPIC database, which is routinely updated, will also ensure that LLMs are updated with the latest information and clinical guidelines. The metrics presented in PGxQA will be continually refined to best reflect the latest evidence. As PGx is a continually evolving area of study, it is essential to have a scalable framework for ongoing evaluation to ensure that model improvements translate into tangible benefits for the field in terms of accuracy and relevance.

The future of PGx chatbots holds significant promise as LLMs become increasingly integrated into healthcare settings to provide clinical recommendations and support. These chatbots will be able to use large quantities of PGx literature and evidence to strengthen and personalize their responses to clinician, patient, and researcher queries. The development of advanced LLMs, coupled with emerging techniques like RAG, will help ensure that PGx chatbots can reliably provide personalized and accurate evidence-based guidance regarding medication intake and dosage. However, the future of these chatbots depends on rigorous continual assessment of their performance. The resources developed in PGxQA represent a first-in-class approach to guide automated LLM evaluation, prioritizing accuracy, completeness, and safety for PGx chatbots.

5. Supplemental Materials and Data Availability

Supplemental tables and the author contributions list are available at:
https://ritchielab.org/publications/supplementary-data/psb-2025/pgxqa
All code, questions, LLM answers, and scoring results are available at:
https://github.com/KarlKeat/PGxQA/

6. References

1. Lewis, J. P. & Shuldiner, A. R. Clopidogrel pharmacogenetics: Beyond candidate genes and genome-wide association studies. *Clin. Pharmacol. Ther.* **101**, 323–325 (2017).
2. Daly, A. K. Pharmacogenetics: a general review on progress to date. *Br. Med. Bull.* **124**, 65–79 (2017).
3. Klein, M. E., Parvez, M. M. & Shin, J.-G. Clinical Implementation of Pharmacogenomics for Personalized Precision Medicine: Barriers and Solutions. *J. Pharm. Sci.* **106**, 2368–2379 (2017).
4. Relling, M. V. & Evans, W. E. Pharmacogenomics in the clinic. *Nature* **526**, 343–350 (2015).
5. Nagy, M., Eirini Tsermpini, E., Siamoglou, S. & Patrinos, G. P. Evaluating the Current Level of Pharmacists' Pharmacogenomics Knowledge and its Impact on Pharmacogenomics Implementation. *Pharmacogenomics* **21**, 1179–1189 (2020).
6. Alshabeeb, M. A., Alyabsi, M., Aziz, M. A. & Abohelaika, S. Pharmacogenes that demonstrate high association evidence according to CPIC, DPWG, and PharmGKB. *Front. Med.* **9**, (2022).
7. Thorn, C. F., Klein, T. E. & Altman, R. B. PharmGKB: The Pharmacogenomics Knowledge Base. in *Pharmacogenomics: Methods and Protocols* (eds. Innocenti, F. & van Schaik, R. H. N.) 311–320 (Humana Press, Totowa, NJ, 2013). doi:10.1007/978-1-62703-435-7_20.
8. Duarte, J. D. & Cavallari, L. H. Pharmacogenetics to guide cardiovascular drug therapy. *Nat. Rev. Cardiol.* **18**, 649–665 (2021).
9. Tuteja, S. Application of Pharmacogenetics for the Use of Antiplatelet and Anticoagulant Drugs. *Curr. Cardiovasc. Risk Rep.* **17**, 27–38 (2023).

10. Green, J. S., O'Brien, T. J., Chiappinelli, V. A. & Harralson, A. F. Pharmacogenomics Instruction in US and Canadian Medical Schools: Implications for Personalized Medicine. *Pharmacogenomics* **11**, 1331–1340 (2010).
11. Karas Kuželički, N. *et al.* Pharmacogenomics Education in Medical and Pharmacy Schools: Conclusions of a Global Survey. *Pharmacogenomics* **20**, 643–657 (2019).
12. Nutter, S. C. & Gálvez-Peralta, M. Pharmacogenomics: From classroom to practice. *Mol. Genet. Genomic Med.* **6**, 307–313 (2018).
13. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. Preprint at https://doi.org/10.48550/arXiv.2303.13375 (2023).
14. Meng, X. *et al.* The application of large language models in medicine: A scoping review. *iScience* **27**, 109713 (2024).
15. Soroush, A. *et al.* Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. *NEJM AI* **1**, AIdbp2300040 (2024).
16. Clusmann, J. *et al.* The future landscape of large language models in medicine. *Commun. Med.* **3**, 1–8 (2023).
17. Rawte, V., Sheth, A. & Das, A. A Survey of Hallucination in Large Foundation Models. Preprint at http://arxiv.org/abs/2309.05922 (2023).
18. Xu, Z., Jain, S. & Kankanhalli, M. Hallucination is Inevitable: An Innate Limitation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2401.11817 (2024).
19. Zhang, Y. *et al.* Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2309.01219 (2023).
20. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and Adoption of Large Language Models in Medicine. *JAMA* **330**, 866–869 (2023).

21. Mousavi, S. M., Alghisi, S. & Riccardi, G. DyKnow:Dynamically Verifying Time-Sensitive Factual Knowledge in LLMs. Preprint at https://doi.org/10.48550/arXiv.2404.08700 (2024).
22. Ullah, E., Parwani, A., Baig, M. M. & Singh, R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagn. Pathol.* **19**, 43 (2024).
23. Kanter, G. P. & Packel, E. A. Health Care Privacy Risks of AI Chatbots. *JAMA* **330**, 311–312 (2023).
24. Murugan, M. *et al.* Empowering personalized pharmacogenomics with generative AI solutions. *J. Am. Med. Inform. Assoc.* **31**, 1356–1366 (2024).
25. cpicpgx/cpic-data. CPIC (2024).
26. psycopg/psycopg2. The Psycopg Team (2024).
27. team, T. pandas development. pandas-dev/pandas: Pandas. Zenodo https://doi.org/10.5281/zenodo.10957263 (2024).
28. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* 56–61 (2010) doi:10.25080/Majora-92bf1922-00a.
29. Generative Language API | Google AI for Developers. *Google for Developers* https://ai.google.dev/api/rest.
30. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating Text Generation with BERT. Preprint at https://doi.org/10.48550/arXiv.1904.09675 (2020).
31. He, P., Liu, X., Gao, J. & Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. Preprint at https://doi.org/10.48550/arXiv.2006.03654 (2021).
32. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at https://doi.org/10.48550/arXiv.1907.11692 (2019).

33. Li, Z. *et al.* Towards General Text Embeddings with Multi-stage Contrastive Learning. Preprint at https://doi.org/10.48550/arXiv.2308.03281 (2023).
34. Song, K., Tan, X., Qin, T., Lu, J. & Liu, T.-Y. MPNet: Masked and Permuted Pre-training for Language Understanding. Preprint at https://doi.org/10.48550/arXiv.2004.09297 (2020).
35. vmware/data-annotator-for-machine-learning. VMware (2024).
36. Kluyver, T. *et al. Jupyter Notebooks—a Publishing Format for Reproducible Computational Workflows*. *IOS Press* 87–90 (2016). doi:10.3233/978-1-61499-649-1-87.
37. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
38. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
39. Sui, Y., Zhou, M., Zhou, M., Han, S. & Zhang, D. Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* 645–654 (Association for Computing Machinery, New York, NY, USA, 2024). doi:10.1145/3616855.3635752.
40. Chen, B., Zhang, Z., Langrené, N. & Zhu, S. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *arXiv.org* https://arxiv.org/abs/2310.14735v4 (2023).
41. Patel, D. *et al.* The Limits of Prompt Engineering in Medical Problem-Solving: A Comparative Analysis with ChatGPT on calculation based USMLE Medical Questions. 2023.08.06.23293710 Preprint at https://doi.org/10.1101/2023.08.06.23293710 (2023).
42. shanepeckham. Getting started with LLM prompt engineering. https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/prompt-engineering (2024).

43. Finetuning in large language models. https://blogs.oracle.com/ai-and-datascience/post/finetuning-in-large-language-models.
44. Alghisi, S., Rizzoli, M., Roccabruna, G., Mousavi, S. M. & Riccardi, G. Should We Fine-Tune or RAG? Evaluating Different Techniques to Adapt LLMs for Dialogue. Preprint at https://doi.org/10.48550/arXiv.2406.06399 (2024).
45. Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. in *Advances in Neural Information Processing Systems* vol. 33 9459–9474 (Curran Associates, Inc., 2020).
46. Zhao, P. *et al.* Retrieval-Augmented Generation for AI-Generated Content: A Survey. Preprint at http://arxiv.org/abs/2402.19473 (2024).

Overcoming health disparities in precision medicine: Intersectional approaches in precision medicine

Francisco M. De La Vega,[1] Kathleen C. Barnes,[2] Harris Bland,[3] Todd Edwards,[3] Keolu Fox,[4] Alexander Ioannidis,[5] Eimear Kenny,[6] Rasika A. Mathias,[7] Bogdan Pasaniuc,[8] Jada Benn Torres,[9] and Digna R Velez Edwards[3]

[1]Galatea Bio, Inc.,[2]Oxford Nanopore, Ltd., [3]Vanderbilt University Medical Center, [4]University of California San Diego, [5]University of California Santa Cruz, [6]Ichan School of Medicine at Mount Sinai, [7]National Institute of Allergy and Infectious Diseases, [8]University of California Los Angeles, and [9]Vanderbilt University.

1. Overview

The "Overcoming Health Disparities: Intersectional Approaches in Precision Medicine" session at the Pacific Symposium on BioComputing 2025 was aimed to advance computational methods and data science approaches to address racial, ethnic, and gender disparities in biomedical research and healthcare. Emphasizing the role of big data and electronic health records, the session focused on how social identities and categories like race, gender, and ethnicity interact to shape individual healthcare experiences and systemic inequalities. By tackling challenges in capturing and analyzing social determinants of health and environmental risk factors, this year session's papers highlight strategies such as multi-ancestry genetic studies, enhanced data collection from large population-based cohorts, and advanced geocoding clustering techniques. These efforts are crucial for integrating complex social and biological factors to reduce health disparities and improve precision medicine.

2. Advancing multi-ancestry genetic research

Historically, genetic studies have focused predominantly on individuals of European descent, leading to disparities in risk prediction and personalized medicine. While embracing genetic diversity aims to reduce these disparities, methodological challenges persist.

Jones and Cardone *et al.* (2025) examined how different methods of combining genetic data from diverse ancestry groups affect genome wide association study (GWAS) results, finding that multi-ancestry methods can identify shared signals but may diminish ancestry-specific associations, potentially masking important genetic insights for underrepresented populations. This highlights the need for methods that consider both shared and ancestry-specific variants to ensure equitable benefits.

Addressing this, Winters *et al.* (2025) developed a multi-ancestry polygenic risk score (PRS) for uterine fibroids using GWAS data from FinnGen and Biobank Japan, which outperformed single-ancestry PRSs across diverse cohorts, demonstrating improved model transferability. The findings

demonstrate that a multi-ancestry approach captures broader genetic variation and enhances model transferability across different racial groups.

Further exploring uterine fibroids, Akerele *et al.* (2025) investigated their genetic relationship with blood pressure traits using multi-ancestry GWAS datasets, finding that higher blood pressure increases fibroid risk and vice versa, suggesting shared genetic architecture. These findings enhance understanding of the mechanisms linking these conditions, potentially leading to better diagnosis and treatment strategies.

3. Integrating social determinants of health to enhance genetic risk models

While genetic studies have traditionally focused on the influence of genetic variants on phenotypes, they have largely overlooked the role of social determinants of health (SDoH) in disease incidence and health disparities. However, genetic associations alone cannot establish causation; additionally, SDoH may have an even stronger effect than genetic variation, and thus these factors are important to consider.

Guare *et al.* (2025) investigated how SDoH and lifestyle risk factors modulate genetic susceptibility in women's health outcomes by focusing on seven disorders within the All of Us Research Program. They computed PRSs and found that nine out of twelve PRSs were significantly associated with their respective conditions. Notably, higher environmental risk groups, such as individuals with elevated body mass index (BMI), were diagnosed earlier and exhibited increased genetic susceptibility, emphasizing the importance of integrating genetic and environmental data for more precise risk models.

Similarly, Mazzotti *et al.* (2025) analyzed electronic health record data from over 1.4 million individuals to identify social risk factor clusters and their association with obstructive sleep apnea (OSA) and cardiovascular outcomes. Using latent class analysis, they defined three social burden clusters and discovered that those with the highest social burden were less likely to be diagnosed with OSA compared to those with the lowest burden. Among OSA patients, clinical predictors of cardiovascular events varied across social risk clusters, indicating that social factors differently influence cardiovascular risk. These findings highlight significant health disparities in both the diagnosis of OSA and the prediction of cardiovascular diseases, underscoring the need for tailored interventions.

4. Methods to detect and mitigate disparities

New computational methods that consider various data modalities are essential to uncover biases and disparities in healthcare data, disease incidence, and outcomes. Liu *et al.* (2025) introduce a causal inference approach using proximal mediation analysis to detect clinician implicit biases in diagnosis decisions within large-scale medical data like the UK Biobank, quantifying how biases formed by racism, ableism, and sexism impact patient outcomes.

On the other hand, the paper of Niu *et al.* (2025) presents the Gaussian Process Spatial Clustering (GPSC) method, a novel algorithm that clusters census tracts based on spatial location and socioeconomic status. GPSC captures both geographic and other characteristic patterns simultaneously, enabling the identification of meaningful clusters of census tracts based on

socioeconomic and environmental indicators associated with health and cancer risk in the Carolina Breast Cancer Study.

5. Addressing Disparities in Adverse Drug Reactions

Finally, addressing disparities in adverse drug reactions, Muse *et al.* (2025) developed a statistical approach using multistate Cox models to detect and quantify potential adverse drug events (ADEs) resulting from polypharmacy, focusing on differences between patient subgroups such as males and females. Analyzing data from nearly 2 million patients in Denmark, they computed hazard ratios for changes in laboratory test results before and after drug exposure, linking these findings to a drug-drug interaction database. Their models have potential applications for medical safety agencies and could improve efficiency in drug approval pipelines. By revealing how ADEs differ among patient subgroups, this work contributes to enhancing patient safety through precision medicine. This study complements the other research by emphasizing the importance of considering demographic factors in healthcare data analysis to reduce disparities and improve health outcomes.

6. Conclusion

Collectively, these studies highlight the role of innovative computational methods and multi-ancestry approaches in addressing health disparities across various medical domains. By integrating genetic data with social determinants of health, researchers are developing more precise risk models that account for the complex interplay of factors influencing disease outcomes, which may ultimately lead to a better understanding of causation. Advances in detecting and mitigating biases—in clinical decision-making, spatial analysis, and machine learning—contribute to reducing systemic inequalities in healthcare. The papers in this session demonstrate how intersectional and data-driven strategies in precision medicine can potentially overcome existing limitations and promote health equity.

7. Acknowledgments

We thank the anonymous reviewers that helped in the peer review process of the submissions to this session.

References

Akerele A.T., Piekos J.A., Hellwege J.N., Khankari N.K., Edwards T.L., and Velez Edwards D.R. (2025) Uterine fibroids show evidence of shared genetic architecture with blood pressure traits. In *Pacific Symposium on Biocomputing*.

Guare L.A., Das J., Caruth L., and Setia-Verma S. (2025) Social Determinants of Health and Lifestyle Risk Factors Modulate Genetic Susceptibility for Women's Health Outcomes. In *Pacific Symposium on Biocomputing*.

Jones S.C., Cardone K.M., Bradford Y., Tishkoff S.A., and Ritchie M.D. (2025) The Impact of Ancestry on Genome-Wide Association Studies. In *Pacific Symposium on Biocomputing*.

Liu K., Altman R., and Syrgkanis V. (2025) Detecting clinician implicit biases in diagnoses using proximal causal inference. In *Pacific Symposium on Biocomputing*.

Mazzotti D.R., Urbanowicz R., and Jankowska M. (2025) Social risk factors and cardiovascular risk in obstructive sleep apnea: a systematic assessment of clinical predictors in community health centers. In *Pacific Symposium on Biocomputing*.

Muse V.P, Haue A.D, Rodríguez C.L., Orozco A.A., Biel J.H., and Brunak S.. Assessment of Drug Impact on Laboratory Test Results in Hospital Settings. In *Pacific Symposium on Biocomputing 2025*.

Niu H., Troester M., and Li D.. Spatial Clustering for Carolina Breast Cancer Study. In *Pacific Symposium on Biocomputing 2025*.

Winters J.L.G., Piekos J.A., Hellwege J.N., Dikilitas O., Kullo I.J., Schaid D.J., Edwards T.L., and Velez Edwards D.R. Constructing a multi-ancestry polygenic risk score for uterine fibroids using publicly available data highlights need for inclusive genetic research. In *Pacific Symposium on Biocomputing 2025*.

The Impact of Ancestry on Genome-Wide Association Studies

Steven Christopher Jones[1]*, Katie M. Cardone[2]*, Yuki Bradford[2], Sarah A. Tishkoff[2,4], Marylyn D. Ritchie[2,3,5]

[1]Genomics and Computational Biology Graduate Group, [2]Department of Genetics, [3]Institute for Biomedical Informatics, [4]Department of Biology, [5]Department of Biostatistics, Epidemiology, and Informatics

University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA

**Equal contributions to the manuscript*

Email: marylyn@pennmedicine.upenn.edu

Genome-wide association studies (GWAS) are an important tool for the study of complex disease genetics. Decisions regarding the quality control (QC) procedures employed as part of a GWAS can have important implications on the results and their biological interpretation. Many GWAS have been conducted predominantly in cohorts of European ancestry, but many initiatives aim to increase the representation of diverse ancestries in genetic studies. The question of how these data should be combined and the consequences that genetic variation across ancestry groups might have on GWAS results warrants further investigation. In this study, we focus on several commonly used methods for combining genetic data across diverse ancestry groups and the impact these decisions have on the outcome of GWAS summary statistics. We ran GWAS on two binary phenotypes using ancestry-specific, multi-ancestry mega-analysis, and meta-analysis approaches. We found that while multi-ancestry mega-analysis and meta-analysis approaches can aid in identifying signals shared across ancestries, they can diminish the signal of ancestry-specific associations and modify their effect sizes. These results demonstrate the potential impact on downstream post-GWAS analyses and follow-up studies. Decisions regarding how the genetic data are combined has the potential to mask important findings that might serve individuals of ancestries that have been historically underrepresented in genetic studies. New methods that consider ancestry-specific variants in conjunction with the shared variants need to be developed.

Keywords: GWAS; Ancestry; Health Disparities.

1. Introduction

1.1. *Population Structure in Genome-Wide Association Studies*

Genome-wide association studies (GWAS) are a powerful tool for discovering genetic associations with traits of interest[1]. Since its introduction in 2005, the use of GWAS has become a standard method in the field of statistical genetics, offering insight into the contribution of alleles with small effect sizes for complex traits[2]. As DNA sequencing becomes more affordable, and large healthcare systems, biobanks, and consortia continue to link electronic health record (EHR) information containing disease phenotypes to patients' genetic information, larger sample sizes for complex disease show continued promise for the application of GWAS. At the time of writing this manuscript, the GWAS catalog contained summary statistics for over 5,000 phenotypes[3].

Beyond the wide application of GWAS in the field of genetics, considerable work has been done to identify the impact of quality control (QC) procedures and best practices for GWAS[4,5]. Technical decisions such as allele frequency threshold, variant quality thresholds, data missingness, and population structure are all known to impact GWAS outcomes[5]. Despite the considerable work that has been done to offer guidance on GWAS QC and study design, many decisions are made on a case-by-case basis and the approach taken can vary based on the lab and the guidance referenced[1,4,5]. We aim to focus specifically on the impact that different strategies for combining genetic data from two genetically inferred ancestry groups have on GWAS summary statistics.

An individual's genetic ancestry can be inferred from their DNA, which contains information about the genetic signatures resulting from ancestral migrations, mutations, recombination, genetic drift, and natural selection[4,6,7]. Ancestry-specific evolutionary and demographic histories can lead to linkage disequilibrium (LD) and allele frequencies that differ across populations and result in spurious associations due to the confounding effects of ancestry in GWAS[8,9]. Some standard methods to control for population structure within a GWAS study cohort are the use of a mixed model combined with a genetic relationship matrix (GRM), principal component analysis (PCA), and the subsequent inclusion of a small number of principal components (PCs) as covariates in the GWAS model[10,11]. However, even with the inclusion of PCs, population structure may not be entirely accounted for, leading to persistent spurious associations[12]. Additional methods of inferring genetic ancestry such as K-means clustering and quadratic discriminant analysis (QDA) of PCA data or the application of tools such as ADMIXTURE can provide greater resolution for decisions regarding the inference of genetic ancestry of individuals and prove useful for QC decisions for GWAS in admixed and multi-ancestry cohorts[13,14].

As the volume of genetic data combined with rich EHR phenotype data from diverse populations continues to increase, GWAS will continue to be an important tool. Subsequently, the choice between a study focused on ancestry-specific and/or multi-ancestry GWAS approaches will have important implications on the results and their interpretations, especially when GWAS summary statistics are used for downstream analyses such as transcriptome-wide association studies (TWAS), proteome-wide association studies (PWAS), or polygenic scores (PGS)[4,15,16]. Ancestry-specific GWAS may provide insight into genetic associations within specific ancestral groups, allowing for the detection of associations that may be unique or have varying effect sizes across different populations. However, these approaches can be limited due to smaller sample sizes in underrepresented global populations. Multi-ancestry mega-analysis GWAS or meta-analysis approaches can leverage larger sample sizes and provide insight into genetic associations shared across ancestrally diverse populations[4,15,17]. However, both approaches present unique challenges and opportunities that must be carefully considered in the experimental design and interpretation of results.

1.2. *Inclusion of Diverse Ancestries in Genetic Studies*

Genetic studies are predominantly focused on European ancestry, with most GWAS conducted in these populations, leading to insights that are not always generalizable to non-European groups and exacerbate health disparities[3,17–22]. The lack of diversity in genetic research limits our understanding

of genetic variation in underrepresented ancestries and its relationship with complex traits[19,21]. Initiatives like the All of Us Research Program, the Human Heredity and Health in Africa (H3Africa) Initiative, the Million Veteran Program (MVP), and the NHLBI Trans-Omics for Precision Medicine program (TOPMed) aim to address this by recruiting diverse populations and creating more representative datasets for genetic research[22–25]. However, integrating these diverse datasets into GWAS is complicated by unequal sample sizes and differences in allele frequency and LD patterns between populations, which highlight the need for robust and specialized methodologies to ensure accurate and equitable interpretation of genetic associations.

Incorporating diverse ancestries in GWAS offers opportunities to discover associations absent in European-focused studies, providing valuable insight for underrepresented populations[16,26]. It can also enhance fine mapping by leveraging genomic diversity across ancestries[17]. Multi-ancestry mega-analysis and ancestry-specific GWAS with meta-analysis offer solutions but are limited by differences in study design, sample sizes, and the model specified for the meta-analysis. Decisions between fixed effect or random effect meta-analysis will have an impact on the combined results and require assumptions regarding the heterogeneity of associations between populations[4,27,28].

1.3. *Shared and Ancestry-Specific Associations*

Most human genetic variation can be observed within all ancestry groups and many genetic associations with disease are shared across human populations[29]. However, for a small portion of the genome, associations can vary across different ancestral populations, with distinct loci contributing to the same trait in populations with distinct genetic ancestry. This is evident in Solomon Islanders, where a mutation in the *TYRP1* gene is associated with blond hair[30]. This mutation is absent outside of Oceania, and thus cannot explain blond hair in individuals of European ancestry[30]. Similarly, variants such as the G1 and G2 variants in *APOL1* have been shown to account for a substantial degree of risk for chronic kidney disease (CKD) in individuals of African ancestry while being very rare or absent in other ancestry groups[31–33]. These examples underscore the importance of conducting ancestry-specific GWAS to uncover genetic associations that may be masked, diluted, or even missing in multi-ancestry analyses.

Many GWAS of complex traits have identified associations that are shared across ancestries in which a shared variant demonstrates a similar effect size for a trait across multiple populations[26]. For example, variants in the *FTO* gene have been consistently associated with increased body mass index across diverse populations[34]. Similarly, variants in the *TCF7L2* gene are strongly associated with increased risk of type 2 diabetes (T2D) across multiple populations[35–44].

The basis of phenotypic variation and the influence of genetic ancestry is complex. Some diseases exhibit ancestry-specific genetic associations, while others share common genetic associations across populations. This complexity is further compounded by the continuous nature of admixture in natural populations. Understanding the genetic factors that influence complex traits across different populations is crucial for developing personalized medicine approaches tailored to the unique genetic makeup of diverse individuals. The present study aims to contribute to this understanding by investigating the genetic associations with chronic kidney disease (CKD) and type

2 diabetes mellitus (T2D) across European (EUR) and African (AFR) ancestries, utilizing both ancestry-specific and multi-ancestry GWAS approaches to comprehensively assess the impact of genetic variation on these traits (**Figure 1**).

2. Methods

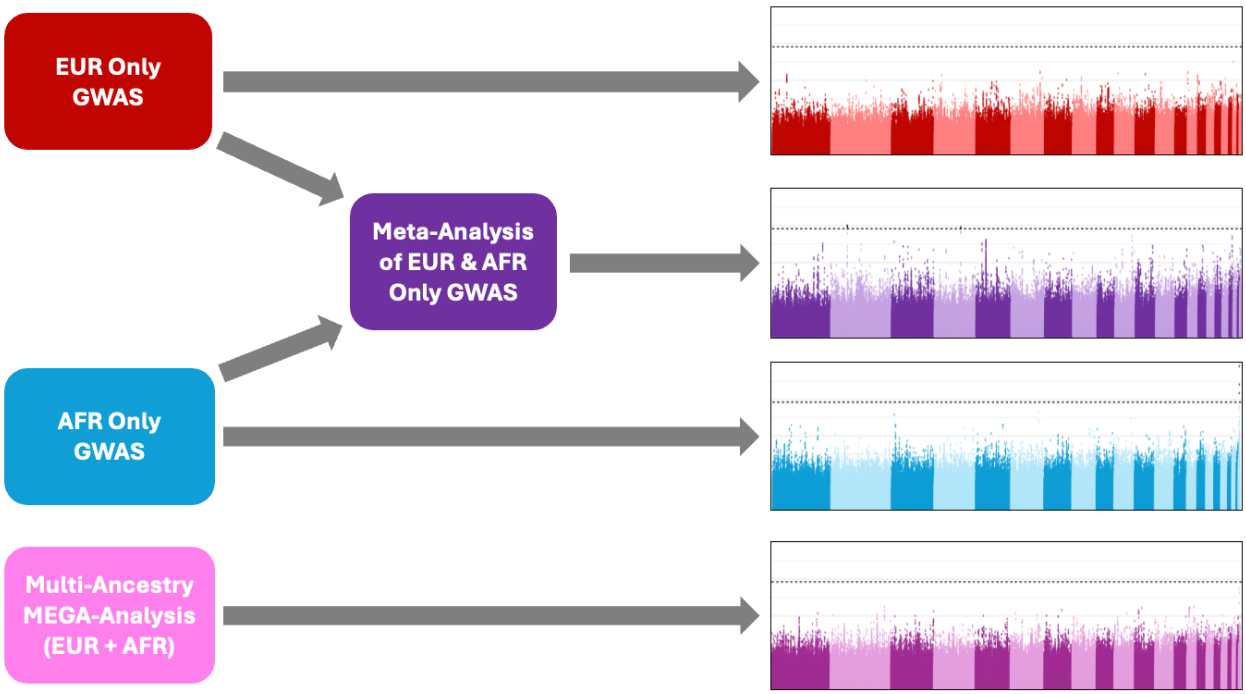


Figure 1: Study Overview: For each binary phenotype, four GWAS were run: EUR-specific, AFR-specific, EUR and AFR combined (multi-ancestry mega-analysis), and meta-analysis of EUR- and AFR-specific GWAS.

2.1. *Data and Study Participants*

The Penn Medicine BioBank (PMBB) is an electronic health record (EHR)-linked research program at the University of Pennsylvania, Perelman School of Medicine[45]. PMBB participants provided consent for research, including blood sample collection, generation of genetic data, and EHR access[45]. Individuals with imputed genotype, demographic, and EHR data were included in this study. PMBB v2.0 imputed data and v2.3 phenotype data were utilized[45].

2.2. *PMBB Centralized Genotyping, Imputation, & Quality Control*

DNA was extracted from blood samples, which were genotyped by the Regeneron Genomics Center with an Illumina Global Sequencing Array v2.0 (GSAv2) containing 654,027 fixed markers[45]. Variant and sample-level quality control was conducted prior to genotype imputation using PLINK v1.9[45,46]. Variants with genotype call rates < 95%, individuals with discordance between reported sex and genetic sex, and individuals with sample call rates < 90% were dropped[45]. Subsequently, autosomes were imputed using TOPMed version R2 genome build 38 reference panel[25,45,47]. After

imputation, PLINK v2.0 was used for additional variant and sample-level quality control[45,46]. Variants with genotype call rates < 99%, minor allele frequency (MAF) < 1%, Hardy-Weinberg Equilibrium (HWE) exact test p-value < 1e-8 or imputation R^2 scores < 0.3 were excluded[45]. Palindromic SNPs, insertions and deletions, and multiallelic variants were also excluded. In addition, individuals with sample call rates < 99% were dropped[45].

2.3. *Principal Component Analysis, Genetically Inferred Ancestry, and Ancestry-Specific Quality Control*

2.3.1 *Quality Control Prior to Principal Components Analysis*

Prior to PCA, quality control was conducted in all eligible samples using PLINK v1.9 and v2.0[46]. Individuals with sample call rates < 95% were dropped[46]. In addition, variants with genotype call rates < 95%, imputation R^2 scores < 0.80, MAF < 5%, or HWE exact test p-value < 1e-10 were excluded[46]. Subsequently, only variants in the intersection between the PMBB and 1,000 Genomes genetic datasets were included[6].

2.3.2 *Principal Component Analysis and Genetically Inferred Ancestry*

Principal component analysis (PCA) was conducted with eigensoft smartPCA on the LD pruned autosomal data[48]. PCs in PMBB were projected onto 1,000 Genomes[6,48]. Using the top two PCs, genetically inferred ancestry was computed using QDA with 1,000 Genomes super-populations as a reference[6,14]. Individuals that had >80% probability of similarity to clusters representing the 1,000 Genomes super-population of EUR of AFR were retained for inclusion in GWAS.

2.3.3 *Analysis-Specific Quality Control and Principal Components Analysis*

After computing genetically inferred ancestry, analysis-specific quality control was completed in EUR, AFR and MEGA (union of EUR and AFR) cohorts with PLINK v1.9 and v2.0[46]. Individuals with sample call rates < 95% and variants with genotype call rates < 95%, MAF < 95%, or imputation R^2 score < 0.3 were excluded. Only biallelic and non-palindromic SNPs were retained. PCA was conducted within each cohort independently following QC using eigensoft smartPCA[48]. Principal components from the cohort-specific PCA were used as covariates in the GWAS.

2.4. *Genome Wide Association Study*

GWAS were conducted using SAIGE[11]. We conducted GWAS utilizing three stratification methods: GWAS stratified to EUR individuals only (EUR-specific), GWAS stratified to AFR individuals only (AFR-specific), and GWAS with both EUR and AFR individuals (MEGA). We tested associations with two phenotypes: CKD and T2D. To phenotype individuals, ICD-9 and ICD-10 codes were mapped to PhecodeX if they had at least two separate instances of an ICD code[49]. The Phecodes used were as follows: CKD = GU_582.2, T2D = EM_202.2[49]. Eligible controls had zero instances of an ICD code used in case definition. To mitigate the effects of sample size, we randomly down sampled while matching case control ratio to ensure the same number of cases and controls across

EUR and AFR individuals for each phenotype. The multi-ancestry mega-analysis GWAS contained a balanced number of EUR and AFR individuals, and the same total sample size as ancestry-specific GWAS. Age at data release, sex assigned at birth, and PC1-7 were used as covariates. We selected the top seven PCs because this explained 79-98% of variance between individuals in the three cohorts (**Supplementary Figure 1, Supplementary Figure 2, Supplementary Figure 3**).

2.5. *Meta-Analysis*

Summary statistics from the AFR and EUR ancestry-specific GWAS analyses were meta-analyzed using METASOFT[27,28]. To compare the impact of model specification on the outcome, the meta-analyses were conducted using a fixed-effect (FE), random-effect (RE), modified random-effect (RE2_INITIAL), and modified random-effect with adjustment for mean effect and heterozygosity (RE2_CORRECTED)[27,28]. Meta-analyses were conducted on the intersection of variants included in the AFR and EUR-specific GWAS. All summary statistics from independent GWAS were adjusted using genomic control following the instructions in the METASOFT publication[27,28]. To ensure consistent sample sizes between analyses, the EUR and AFR groups were randomly down sampled prior to GWAS while maintaining balanced case control ratio such that the meta-analyses contained the same total sample size as the other GWAS. GWAS and meta-analysis results were visualized using qqman and SynthesisView[50,51]. Variants that had a p-value < 5e-8 were considered significant.

2.6. *Analysis of Effect Size Variability*

To assess changes in effect size for variants included in all analyses, we identified whether a variant's effect size changed direction in at least one analysis. We compared effect sizes in the following analyses: all analyses, ancestry-specific compared to multi-ancestry approaches, MEGA analysis compared to meta-analysis approaches, and fixed effect meta-analysis compared to random effect meta-analysis. We identified the percentage of variants that changed direction of effect in each comparison group, both genome-wide and among the variants with the most significant associations, which were visualized in SynthesisView plots[51].

3. Results

The PMBB had 43,589 individuals with genetic data that passed initial QC and were analyzed using QDA to infer genetic ancestry. Using our approach, we identified 10,631 individuals that clustered with the AFR super population and 17,495 individuals that clustered with the EUR super population from the 1,000 Genomes reference panel. **Figure 2** shows the individuals from PMBB in the PCA projection of the 1,000 Genomes. Following analysis-specific QC of these individuals, there were 10,631 individuals and 6,792,866 variants in the AFR analyses, 17,495 individuals and 4,910,840 variants in the EUR analyses, and 28,126 individuals and 5,652,287 variants in the MEGA analyses. Of these variants, 4,184,455 were shared between AFR and EUR cohorts and could be included in

meta-analyses and 3,334,796 were only found in a single ancestry after QC. **Table 1** shows the final sample sizes.

The AFR-specific GWAS of CKD replicated a known signal in the *APOL1* gene (rs73885319) on chromosome 22 (p-value = 7.92e-11) (**Figure 3, Figure 4**)[31]. This signal was not detected in the EUR-specific analysis as the MAF of this variant was 0.00869% and therefore did not pass QC. This signal was detected in the MEGA analysis with a p-value of 1.43e-7, which is below the genome-wide significance threshold. Due to the monomorphic nature of this allele in the EUR population, the variant was not included in any of the meta-analyses. The meta-analyses identified additional associations in the *ANXA5* gene on chromosome 4 and downstream of *LOC124900539* on chromosome 2.

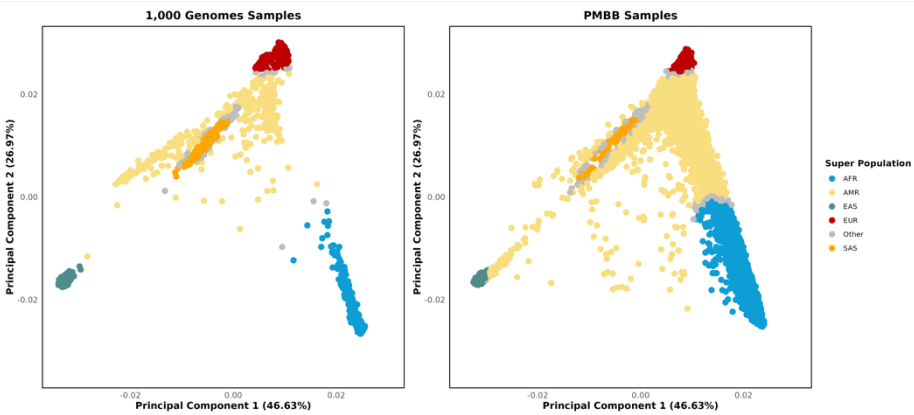


Figure 2: PCA of PMBB samples (right) **projected onto the 1,000 Genomes reference panel** (left). Colors indicate clustering with 1,000 Genomes super-population (AFR, AMR, EAS, EUR, SAS).

Phenotype	Case	Control	Total Sample Size
T2D	3,184	6,448	9,632
CKD	2,659	7,543	10,202

Table 1: Final Sample Sizes for both Ancestry groups.

The T2D GWAS replicated four known signals in the *TCF7L2* gene on chromosome 10 (rs35011184, rs7901695, rs7903146, rs34872471), and one upstream of the *CRYBA2* gene/downstream of the *MIR375* gene on chromosome 2 (rs113414093) (**Figure 3, Figure 5**)[35,36,38–44,44,52]. rs7903146 reached genome-wide significance in the AFR-specific GWAS (p-value = 6.59e-10) and the EUR-specific GWAS (p-value = 5.23e-9). This signal was detected in the MEGA and meta-analyses but was below genome-wide significance. rs34872471 was genome-wide significant in the EUR-specific GWAS (p-value = 5.00e-9) but not in the other analyses. rs35011184 and rs7901695 were detected in all GWAS iterations but were not genome-wide significant, with the EUR-specific GWAS having the lowest p-values (rs35011184 p-value = 4.05e-8, rs7901695 p-value = 1.19e-6). rs113414093 was only detected in the EUR-specific GWAS and was not genome-wide significant (p-value = 9.97e-7). This variant was not present in the other analyses as the MAF was 0.909% in the AFR-specific cohort and 3.70% in the MEGA cohort. The meta-analysis identified additional associations in the *PTPRG* gene on chromosome 3, and upstream of *LOC105374348*/downstream of *FAM53A* on chromosome 4.

In the GWAS of CKD, majority of the variants with the most significant p-values changed direction of effect in at least one analysis (**Table 2**). There was variability in the T2D analyses, but the trend was not as extreme (**Table 2**).

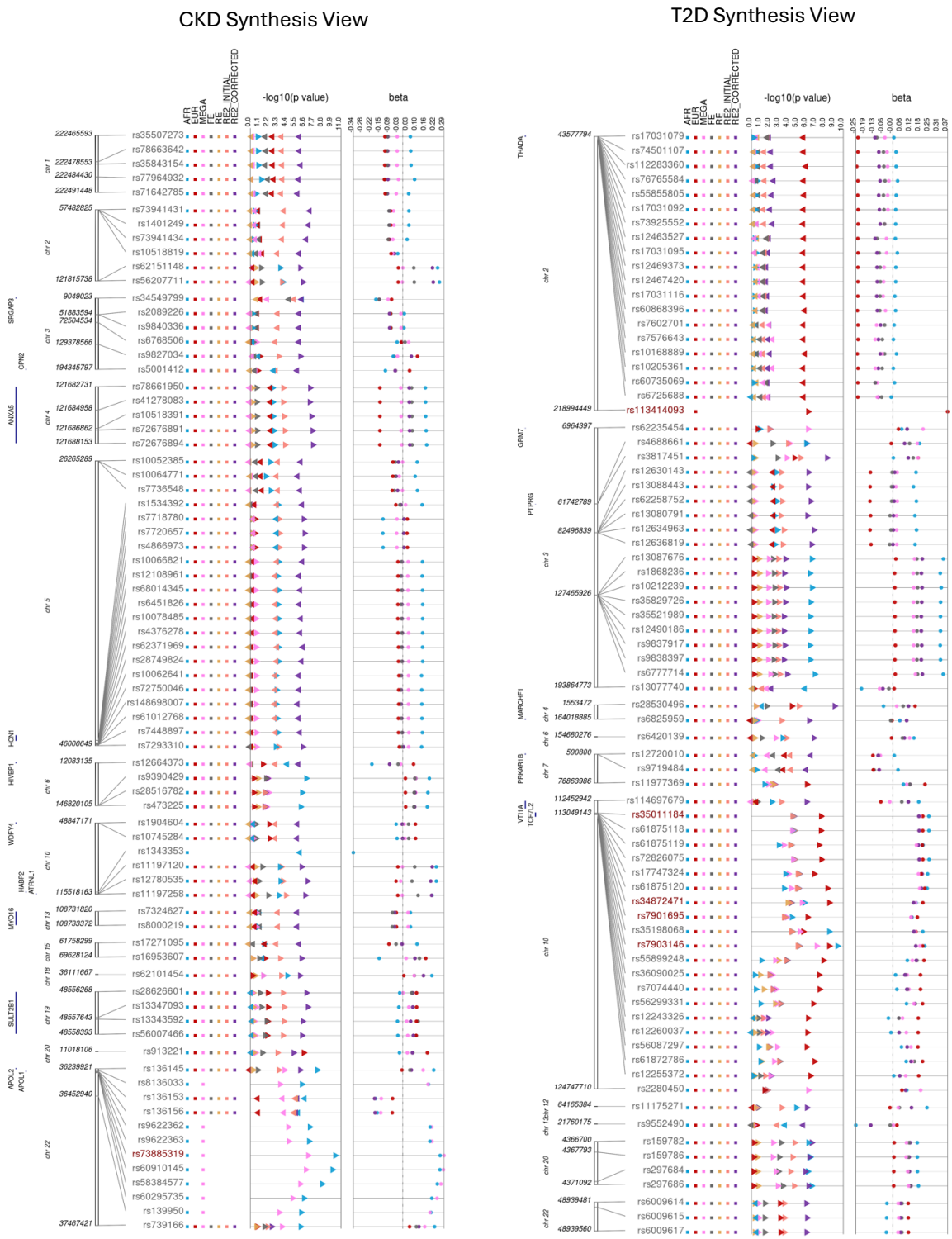


Figure 3: Significance Levels and Effect Sizes for Chronic Kidney Disease (left) and Type 2 Diabetes (right). The most significant variants for each phenotype are displayed. Variants highlighted in red are known signals. Variants are annotated with gene names (left axis).

Phenotype	All Analyses	AFR vs. Multi-Ancestry Analyses	EUR vs. Multi-Ancestry Analyses	MEGA vs. Meta Analyses	Fixed Effect vs. Random Effect Meta Analyses
Percentage of CKD Variants	86.36%	80.30%	56.06%	50%	7.58%
Percentage of T2D Variants	54.05%	47.30%	18.92%	12.16%	9.46%

Table 2: Proportion of Top Variants that Changed Direction of Effect. 75 variants were included in the T2D comparison, and 66 variants were included in the CKD comparison.

Figure 4: Chronic Kidney Disease Stacked Manhattan Plot. Top plot is AFR-specific GWAS, followed by MEGA GWAS, EUR-specific GWAS, and meta-analysis using modified random effect framework (RE2_corrected).

Additionally, direction of effect flipped less when comparing multi-ancestry methods (**Table 2**). When investigating variants genome-wide, there is a decrease in variability in CKD, but an increase in variability in T2D (**Supplementary Table 1**). Additionally, 84-98% of the most significant variants' effect sizes in multi-ancestry analyses had a value within the range of ancestry-specific effect sizes (**Supplementary Table 2**). This trend was less extreme in variants genome-wide, as

nearly 50% of effect sizes in multi-ancestry analyses were within the range of ancestry-specific effect sizes.

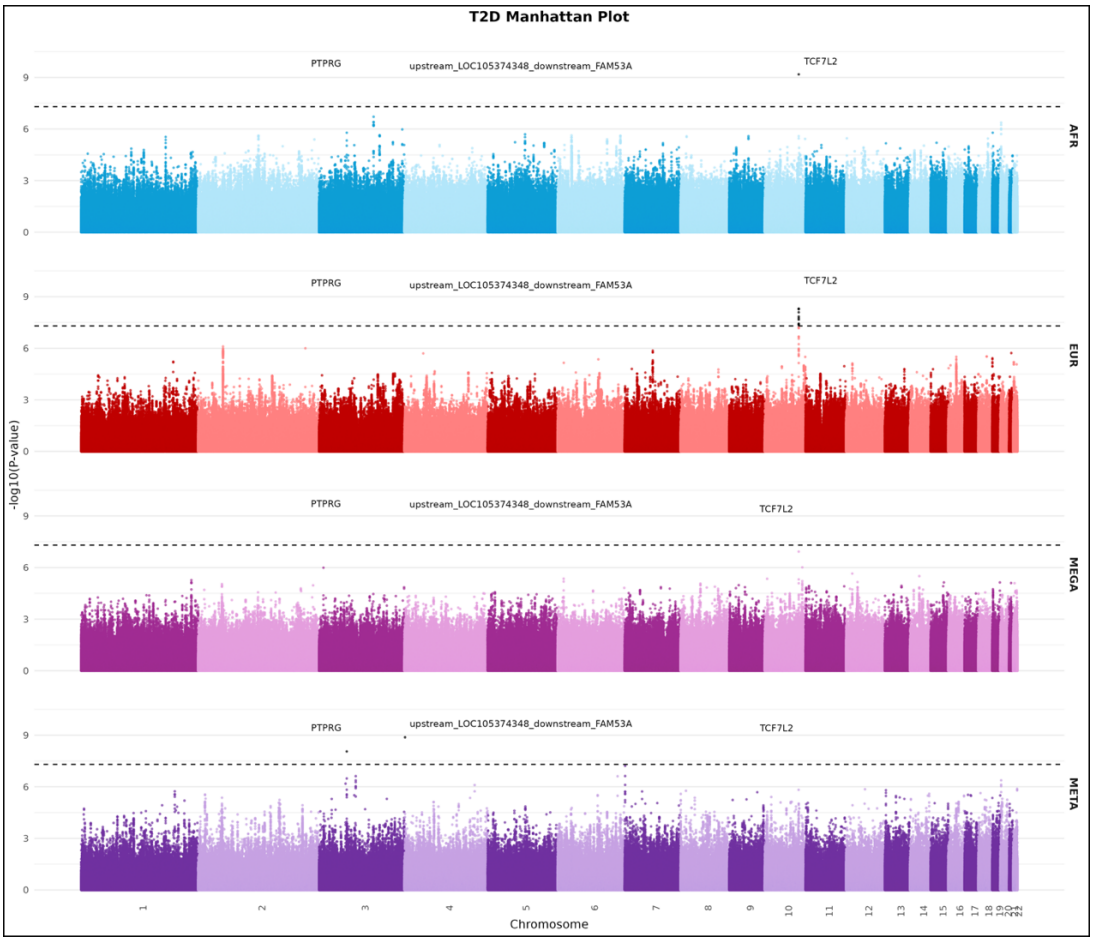


Figure 5: Type 2 Diabetes Stacked Manhattan Plot. Top plot is AFR-specific GWAS, followed by MEGA GWAS, EUR-specific GWAS, and ancestry-balanced meta-analysis using modified random effect framework (RE2_corrected).

4. Discussion

Our aim was to assess how different approaches of combining genetic data from individuals of diverse ancestries change the outcome of a GWAS. To test this, we conducted GWAS of CKD and T2D in individuals of African and European ancestry in the PMBB. We compared the differences in GWAS results through changes to the p-value and effect sizes for ancestry-specific analyses (AFR or EUR only), multi-ancestry mega-analysis (MEGA), and meta-analysis using fixed-effect (FE), random-effect (RE), and modified random-effect (RE2_INITIAL and RE2_CORRECTED). We hypothesized that while most genetic associations are shared across human populations, we would observe specific genetic associations that were statistically significant in only one ancestry and that the different multi-ancestry approaches would have inconsistent results for these variants. The results support our hypothesis as shown in **Figures 3-5** and **Table 2**.

In the GWAS of CKD, variants within the *APOL1* gene were found to be significantly associated with CKD in the AFR-specific GWAS[31]. In the mega-analysis GWAS, these variants dropped below genome-wide significance, providing evidence that multi-ancestry mega-analysis can diminish ancestry-specific signals. We also note that the use of a meta-analysis tool such as METASOFT will exclude the association observed in the AFR-specific GWAS due to this variant not passing QC in the EUR cohort. Additional variants in the *ANXA5* gene and downstream of *LOC124900539* were significantly associated in the meta-analysis (RE2_CORRECTED) but may be spurious due to genomic inflation in this approach (**Supplementary Figure 4**).

In the AFR and EUR-specific GWAS of T2D, a well-known variant (rs7903146) within the *TCF7L2* gene was significantly associated with T2D[35,43,44,52], while it dropped below genome-wide significance in all multi-ancestry analyses. The GWAS of T2D illustrates how the composition of a multi-ancestry approach can diminish the significance of ancestry-specific signals. However, we acknowledge the limitation that smaller number of cases per ancestry might have had in the multi-ancestry approaches. Additional variants in the *PTPRG* gene and upstream of *LOC105374348*/ downstream of *FAM53A* were significantly associated in the meta-analysis (RE2_CORRECTED) but may be spurious due to genomic inflation in this approach (**Supplementary Figure 5**).

Across both phenotypes, effect sizes flipped direction on many occasions, especially among variants with the lowest p-values (**Table 2, Supplementary Table 1**). This occurred more often when comparing ancestry-specific approaches to multi-ancestry approaches, rather than within multi-ancestry approaches, suggesting that observed ancestry-specific effect sizes can be altered when using multi-ancestry GWAS approaches. Additionally, effect size values in multi-ancestry results were commonly within the range of ancestry-specific effect size value for variants with the lowest p-values (**Supplementary Table 2**).

Meta-analyses can be performed using different approaches, with fixed-effect (FE) and random-effect (RE) models being most common. Fixed-effect meta-analysis assumes a homogenous effect size between studies, meaning any variation in the observed effects is attributed solely to sampling error[27]. In contrast, random-effect meta-analysis assumed that the effect size varies between studies due to differences in population or study designs, allowing for more flexibility in capturing heterogeneity across datasets[27]. We employed the RE2 method developed by Han and Eskin (2011) because it improves statistical power by relaxing the conservative assumptions of the traditional random-effect model, enabling better detection of associations in the presence of heterogeneity[27].

Our study had several limitations. Our sample sizes were limited due to down sampling to match case and control numbers across ancestry groups, so many variants did not reach genome-wide significance. This is of particular importance when considering changes to the signal in the *TCF7L2* gene in T2D between approaches. Although a higher sample size would be ideal, down sampling was a crucial step to isolate the impact of ancestry on GWAS approaches rather than sample size and statistical power. Additionally, down sampled groups were not matched by age, sex or other clinical characteristics. In addition, the modified meta-analysis in the RE2_CORRECTED analyses produced slightly inflated results which often had the most significant associations and identified several signals for CKD and T2D that had not been reported in ClinVar or the GWAS

catalog[3,53]. Due to the low sample sizes in our study compared to previously reported GWAS of T2D and CKD that had not detected these associations, it is plausible these associations may be spurious. Our meta-analyses also only included variants that intersected between the ancestry-specific GWAS, which led to the exclusion of several important ancestry-specific signals in the meta-analysis results. This can be overcome through the inclusion of more cohorts in a meta-analysis but highlights an important limitation of the meta-analysis approach under our framework for directly comparing two studies. Additionally, our method to assess variability in effect sizes was unable to fully quantify observed variability. The pattern of sample overlap between the GWAS approaches in our study violated assumptions of independence or matched dependence between studies. Quantification of this variability using a well calibrated statistical methodology is a logical next step to investigate the differences observed between approaches.

In a typical GWAS, multi-ancestry mega-analysis, or meta-analysis approaches benefit from increased sample size. Our study, however, maintained consistent sample size across approaches to isolate ancestry's impact. We found that multi-ancestry methods can diminish ancestry-specific signals, which can significantly impact downstream analyses like TWAS, PWAS, or PGS. This raises questions about the optimal approach for generating summary statistics, as results differ in meaningful ways based on initial GWAS method. Notably, many variants show striking changes in effect direction, both among those with significant p-values and genome-wide. These effect size flips are crucial, as they influence downstream analyses and biological/clinical interpretations. While many variants show consistent results across approaches, a notable subset are impacted by the choice of analysis method. As we see with variants in *APOL1*, some of these variants showing variable results or which could not be fully assessed in all approaches are essential for understanding differences in disease risk between populations. Thus, new methods that consider the ancestry-specific variants in conjunction with the multi-ancestry shared variants need to be developed.

5. Acknowledgements

We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the patient-participants of Penn Medicine who consented to participate in this research program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. Additional funding support was provided to MDR by AI077505, EY023557, AG066833, and HL169458. Additional funding was provided to SAT by ADA 1-19-VSN-02, and NIH grants 1R35GM134957, R01AR076241, and 1X01HL139409-01.

6. Supplementary Material

All supplemental data can be found at:
https://ritchielab.org/publications/supplementary-data/psb-2025/jonescardone

References

1. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).
2. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).
3. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
4. Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
5. Truong, V. Q. *et al.* Quality Control Procedures for Genome-Wide Association Studies. *Curr. Protoc.* **2**, e603 (2022).
6. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
7. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
8. Li, C. C. Population subdivision with respect to multiple alleles. *Ann. Hum. Genet.* **33**, 23–29 (1969).
9. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
10. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
11. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

12. Hellwege, J. *et al.* Population Stratification in Genetic Association Studies. *Curr. Protoc. Hum. Genet.* **95**, 1.22.1-1.22.23 (2017).
13. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
14. Qin, X., Lock, T. R. & Kallenbach, R. L. DA: Population structure inference using discriminant analysis. *Methods Ecol. Evol.* **13**, 485–499 (2022).
15. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).
16. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
17. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* **6**, 91 (2014).
18. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
19. Ju, D., Hui, D., Hammond, D. A., Wonkam, A. & Tishkoff, S. A. Importance of Including Non-European Populations in Large Human Genetic Studies to Enhance Precision Medicine. *Annu. Rev. Biomed. Data Sci.* **5**, 321–339 (2022).
20. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
21. Fatumo, S. *et al.* Diversity in Genomic Studies: A Roadmap to Address the Imbalance. *Nat. Med.* **28**, 243–250 (2022).
22. Bick, A. G. *et al.* Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).

23. The H3Africa Consortium *et al.* Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).

24. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

25. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

26. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

27. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).

28. Han, B. & Eskin, E. Interpreting Meta-Analyses of Genome-Wide Association Studies. *PLOS Genet.* **8**, e1002555 (2012).

29. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).

30. Kenny, E. E. *et al.* Melanesians blond hair is caused by an amino acid change in TYRP1. *Science* **336**, 554 (2012).

31. Genovese, G. *et al.* Association of Trypanolytic ApoL1 Variants with Kidney Disease in African-Americans. *Science* **329**, 841–845 (2010).

32. Parsa, A. *et al.* APOL1 risk variants, race, and progression of chronic kidney disease. *N. Engl. J. Med.* **369**, 2183–2196 (2013).

33. Pollak, M. R. & Friedman, D. J. APOL1 and APOL1-Associated Kidney Disease: A Common Disease, an Unusual Disease Gene − Proceedings of the Henry Shavelle Professorship. *Glomerular Dis.* **3**, 75–87 (2023).

34. Fawcett, K. A. & Barroso, I. The genetics of obesity: FTO leads the way. *Trends Genet.* **26**, 266–274 (2010).

35. Grant, S. F. A. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).

36. Haddad, S. A. *et al.* A novel TCF7L2 type 2 diabetes SNP identified from fine mapping in African American women. *PloS One* **12**, e0172577 (2017).

37. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).

38. Uribe-Salazar, J. M., Palmer, J. R., Haddad, S. A., Rosenberg, L. & Ruiz-Narváez, E. A. Admixture mapping and fine-mapping of type 2 diabetes susceptibility loci in African American women. *J. Hum. Genet.* **63**, 1109–1117 (2018).

39. Chang, Y.-C. *et al.* Association study of the genetic polymorphisms of the transcription factor 7-like 2 (TCF7L2) gene and type 2 diabetes in the Chinese population. *Diabetes* **56**, 2631–2637 (2007).

40. Ng, M. C. Y. *et al.* Replication and identification of novel variants at TCF7L2 associated with type 2 diabetes in Hong Kong Chinese. *J. Clin. Endocrinol. Metab.* **92**, 3733–3737 (2007).

41. Lehman, D. M. *et al.* Haplotypes of transcription factor 7-like 2 (TCF7L2) gene and its upstream region are associated with type 2 diabetes and age of onset in Mexican Americans. *Diabetes* **56**, 389–393 (2007).

42. Spracklen, C. N. *et al.* Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**, 240–245 (2020).

43. Zeggini, E. *et al.* Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. *Science* **316**, 1336–1341 (2007).

44. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691 (2020).
45. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *J. Pers. Med.* **12**, 1974 (2022).
46. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
47. Hinrichs, A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
48. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
49. Shuey, M. M. *et al.* Next-generation phenotyping: introducing phecodeX for enhanced discovery research in medical phenomics. *Bioinformatics* **39**, btad655 (2023).
50. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* **3**, 731 (2018).
51. Pendergrass, S. A., Dudek, S. M., Crawford, D. C. & Ritchie, M. D. Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis. *BioData Min.* **3**, 10 (2010).
52. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
53. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980-985 (2014).

Constructing a multi-ancestry polygenic risk score for uterine fibroids using publicly available data highlights need for inclusive genetic research

Jessica L.G. Winters, [1-3*†] Jacqueline A. Piekos,[1-3*] Jacklyn N. Hellwege,[1,4] Ozan Dikilitas,[6] Iftikhar J. Kullo,[6] Daniel J. Schaid,[7] Todd L. Edwards,[5] and Digna R. Velez Edwards[2-3‡]

[1]Vanderbilt Genetics Institute; [2]Department of Biomedical Informatics; [3]Division of Quantitative and Clinical Sciences, Department of Obstetrics and Gynecology; [4]Division of Genetic Medicine, [5]Division of Epidemiology, Department of Medicine; Vanderbilt University Medical Center, Nashville, TN 37203, USA

[6]Department of Cardiovascular Medicine; [7]Department of Health Sciences Research; Mayo Clinic, Rochester, MN 55905, USA

Email: todd.l.edwards@vumc.org, digna.r.velez.edwards@vumc.org

Uterine leiomyomata, or fibroids, are common gynecological tumors causing pelvic and menstrual symptoms that can negatively affect quality of life and child-bearing desires. As fibroids grow, symptoms can intensify and lead to invasive treatments that are less likely to preserve fertility. Identifying individuals at highest risk for fibroids can aid in access to earlier diagnoses. Polygenic risk scores (PRS) quantify genetic risk to identify those at highest risk for disease. Utilizing the PRS software PRS-CSx and publicly available genome-wide association study (GWAS) summary statistics from FinnGen and Biobank Japan, we constructed a multi-ancestry (META) PRS for fibroids. We validated the META PRS in two cross-ancestry cohorts. In the cross-ancestry Electronic Medical Record and Genomics (eMERGE) Network cohort, the META PRS was significantly associated with fibroid status and exhibited 1.11 greater odds for fibroids per standard deviation increase in PRS (95% confidence interval [CI]: 1.05 – 1.17, $p = 5.21 \times 10^{-5}$). The META PRS was validated in two BioVU cohorts: one using ICD9/ICD10 codes and one requiring imaging confirmation of fibroid status. In the ICD cohort, a standard deviation increase in the META PRS increased the odds of fibroids by 1.23 (95% CI: 1.15 – 1.32, $p = 9.68 \times 10^{-9}$), while in the imaging cohort, the odds increased by 1.26 (95% CI: 1.18 – 1.35, $p = 2.40 \times 10^{-11}$). We subsequently constructed single ancestry PRS for FinnGen (European ancestry [EUR]) and Biobank Japan (East Asian ancestry [EAS]) using PRS-CS and discovered a nominally significant association in the eMERGE cohort within fibroids and EAS PRS but not EUR PRS (95% CI: 1.09 – 1.20, $p = 1.64 \times 10^{-7}$). These findings highlight the strong predictive power of multi-ancestry PRS over single ancestry PRS. This study underscores the necessity of diverse population inclusion in genetic research to ensure precision medicine benefits all individuals equitably.

Keywords: Complex Traits; Health Disparities; Risk Assessment; Women's Health

---

[*] These authors contributed equally to this work.

[†] Presenting author.

[‡] Work supported by grants R01HD074711 and R03HD078567 to DRVE and R01HD093671 to DRVE/TLE.

1. Introduction

Uterine fibroids, or uterine leiomyomata, are benign tumors of the uterine smooth muscle that affect a substantial proportion of people with uteruses. While nearly all of these individuals will develop at least one fibroid in their lifetime, only about 50% will experience symptoms, leading to a condition with considerable variability in presentation.[1,2] Fibroids are recognized as a health disparity, with a higher prevalence reported among individuals identifying as Black compared to those identifying as White.[1,3] Additionally, fibroids impose a significant financial burden to the healthcare system, being the leading cause of hysterectomy and gynecological hospitalizations in the United States.[4]

Despite their common occurrence, the genetic factors contributing to fibroid development remain complex and multifactorial. Genome-wide association studies (GWAS) have enhanced our understanding of the genetic underpinnings of uterine fibroids, revealing that the condition is influenced by multiple genetic variants, each contributing a small amount to the overall risk.[5,6] This polygenic nature of fibroids means that identifying individual genes of interest through single-gene studies is insufficient. To better estimate genetic risk for polygenic diseases like fibroids, polygenic risk scores (PRS) have been developed. A PRS aggregates an individual's genetic risk across various loci, providing an overall estimate of their risk for the disease or other clinically relevant outcome.[7] In the context of uterine fibroids, PRS can refine diagnostic accuracy, help identify individuals at high genetic risk for fibroids, and predict the likelihood of treatment resistance or recurrence.[8] This personalized approach allows for more targeted interventions and pre-clinical monitoring, potentially leading to earlier and more effective management.

PRS development has traditionally relied on GWAS data from populations of European ancestry, which limits the applicability of these scores to populations of other ancestries.[9] The use of single ancestry GWAS also exacerbates issues with generalizability. There are several programs for PRS construction, and a review of the different programs and methodologies has been published elsewhere.[10] However, PRS-CSx is an approach which uses linkage disequilibrium (LD) reference panels matched to the ancestry of the GWAS population to perform continuous shrinkage across summary statistics.[11] This approach integrates multiple multi-ancestry GWAS summary statistics from different ancestry groups allowing for more genetic variability to be captured in the score. In 2022, our group published a PRS for fibroids using a European ancestry GWAS and validated it in a population of European ancestry.[12] Here, we aim to extend previous work by developing a multi-ancestry PRS for fibroids applicable to a diverse cohort. By using this method to construct a portable PRS, we hope to address and mitigate racial disparities in precision medicine by overcoming existing limitations in capturing polygenic traits.

2. Materials and Methods

2.1. *Study populations*

The Electronic Medical Records and Genomics (eMERGE) Network (2007 – present) is a national network of DNA repositories that are linked to electronic health records (EHRs). A detailed description of the organization of the eMERGE Network has been previously published.[13] Data contained in the EHR include International Classification of Disease (ICD) diagnostic and procedure codes, basic

demographics, discharge summaries, progress notes, health history, laboratory values, imaging reports, medication orders, and pathology details. Participants in the eMERGE network were genotyped separately, then imputed and merged. A detailed description of the genotyping, imputation, and quality control of the eMERGE phase III array dataset has been previously reported.[14]

The BioVU DNA Repository is a deidentified database of EHRs that are linked to patient DNA samples at Vanderbilt University Medical Center (VUMC). A detailed description about the database and its maintenance has been published elsewhere.[15] The EHR for BioVU contains the same information as stated above for eMERGE. This study also obtained Institutional Review Board (IRB) approval and was conducted in accordance with ethical standards.

While BioVU is a member of eMERGE, samples included in this study are unique to BioVU. BioVU participants were genotyped on a custom MEGA array with genotypes aligned to the forward strand. Initial quality control of both study populations excluded samples or variant sites with missingness above a 2% threshold. Samples were also excluded if consent had been withdrawn, if the sample was duplicated, if there was a failure in sex concordance, or if there was a discrepancy between reported race and genetically determined race. Genetic males were censored from analysis. Imputation was performed on the Michigan Imputation Server using Minimac4 and the 1000 Genomes Phase 3 combined reference panel.[16,17]

Phecodes within the EHR were based from ICD9 and ICD10 codes. Fibroid status in eMERGE was extracted based on phecodes recorded in EHR data.[18] Two cohorts were created in BioVU using different case and control definitions: BioVU-ICD and BioVU-imaged. The BioVU-ICD cohort classified fibroid status similarly to eMERGE, derived from phecodes, while the BioVU-imaged cohort used a previously published algorithm to identify cases or controls based on imaging records indicating the presence or absence of fibroids.[19] In the eMERGE and BioVU-ICD cohorts, cases had at least one code for fibroid diagnosis or a history of fibroid treatment, while controls had no such records. In the BioVU-imaged cohort, cases were identified by a history of fibroids or treatment procedures and at least one imaging procedure confirming fibroid presence. Controls in the BioVU-imaged cohort required two or more imaging events on separate dates without fibroid findings and no history of diagnosis or treatment. Race and ethnicity were determined via reporting through categorical options. The multi-ancestry group was comprised of all individuals that reported as White, Black, or Asian race and Hispanic or non-Hispanic ethnicity. The other two groups were based on either White or Black reported race and Hispanic or non-Hispanic ethnicity. The counts of each strata are given in Table 1.

2.2. *Polygenic risk score development*

Genetic effect weights for PRS construction were derived from uterine fibroid GWAS summary statistics from FinnGen r8 and BioBank Japan.[20,21] Both biobanks determined case and control status based on the presence or absence of ICD9/ICD10 codes or equivalent codes in their healthcare systems. For the multi-ancestry (META) PRS, posterior genetic effect weights were calculated using PRS-CSx, while weights for the single-ancestry scores, European (EUR) and East Asian (EAS) PRS, were calculated using PRS-CS.[11,22] We used linkage disequilibrium (LD) reference panels from the 1000 Genomes Project, with the EUR panel for the FinnGen cohort and the EAS panel for the BioBank Japan cohort. Both PRS-CS and PRS-CSx use a high-dimensional Bayesian framework that calculates a continuous shrinkage prior tailored to a target population, based on the selected LD reference panel.

This shrinkage prior is applied to the raw genetic weights from the source GWAS to derive posterior genetic effect weights, which are then summed to create the PRS. PRS-CS is designed for a single GWAS from a single population, whereas PRS-CSx integrates results from multiple GWAS summary statistics. The programs were applied to three target populations: eMERGE, BioVU-ICD, and BioVU-imaged. Posterior effect weights calculated for each population were summed to create a PRS using PLINK 2.0.[23,24]

2.3. *Statistical analysis*

All statistical analyses were performed using R Statistical Software (v4.2.2).[25] Samples remaining after exclusion in eMERGE and BioVU were used for ten-fold cross validation. Analysis of variance (ANOVA) test was used to determine if age and BMI differed within racial groups between all cohorts. These covariates were chosen because prior literature has revealed associations between uterine fibroid risk with both age and BMI.[1] Student's T-test was used to determine if mean META, EUR, and EAS PRS significantly differed between cases and controls for each racial group within the cohorts. Densities of each PRS stratified on case/control status, were visualized using 'ggplot2.'[26]

Table 1. Racial breakdown of cohorts and population characteristics. Listed below are total counts, mean and standard deviation (SD) of body mass index (BMI) and age, and numbers of cases and controls for each of the three groupings within all cohorts. Race consists of White reported race and non-Hispanic ethnicity (White), Black reported race and non-Hispanic ethnicity (Black), and all the above (All).

Cohort Reported Race	**N**	**BMI (SD)**	**Age (SD)**	**Controls (%)**	**Cases (%)**
eMERGE					
All	23,183	29.07 (7.49)	65.30 (18.69)	21,212 (91)	2,290 (9)
White	20,408	28.68 (7.19)	66.52 (18.52)	18,398 (91)	1,784 (9)
Black	2,775	32.44 (8.66)	56.94 (18.73)	2,306 (84)	439 (16)
BioVU - ICD					
All	33,391	29.27 (7.84)	52.53 (18.61)	32,764 (97)	1,076 (3)
White	27,141	28.69 (7.53)	54.64 (18.20)	25,812 (98)	596 (2)
Black	6,250	32.03 (8.72)	45.19 (18.19)	5,700 (93)	420 (7)
BioVU - imaged					
All	9,182	29.21 (8.08)	44.86 (17.33)	7,910 (84)	1,463 (16)
White	7,294	28.55 (7.69)	46.96 (17.45)	6,082 (86)	975 (14)
Black	1,888	31.90 (9.17)	38.02 (15.41)	1,464 (78)	410 (22)

2.4. *Ten-fold cross validation*

Ten-fold cross validation was performed using the R package 'caret.'[27] Each PRS (META, EUR, EAS) was tested for validation in each of the racial groups for every cohort, resulting in nine different

validation groups in total. Each of the nine groups was split into 80/20 training and testing sets. For each PRS, three models were applied to each of the nine validation groups. The adjusted model constructed the PRS as the main predictor with adjustments for age, BMI, and ten principal components (PCs). The unadjusted model estimated the PRS singularly, while the covariate model analyzed the model created by the covariates—age, BMI, and ten PCs—without the PRS. Odds ratios (OR) and 95% confidence intervals (CI) and pseudo-R^2 were calculated for each model. Area under receiver operator curve (AUROC) for the testing set was calculated using the 'pROC' R package.[28]

3. Results

3.1. *Population characteristics*

Out of 52,548 females in the eMERGE cohort, 23,502 samples passed quality control measures and exhibited fibroid status determinable by ICD codes (eMERGE). The average BMI of the overall group was 29.07 (standard deviation [SD] = 7.49), with 28.68 (SD = 7.19) for the White-reported race strata and 32.44 (SD = 8.66) for the Black-reported race strata. The overall average age was 65.30 (SD = 18.69), with 66.52 (SD = 18.52) for the White-reported race strata and 56.94 (SD = 18.73) for the Black-reported race strata. There were 2,290 fibroid cases in the multi-ancestry group. There were 1,784 cases in the White-reported race strata and 439 cases in the Black-reported race strata to make the prevalence of fibroids 9% and 16%, respectively (Table 1).

BioVU had 51,715 female samples of which 33,840 samples passed quality control and exhibited fibroid status determinable by ICD codes (BioVU-ICD). The average BMI of the multi-ancestry group was 29.27 (SD = 7.84). For the White-reported race strata, the average BMI was 28.69 (SD = 7.53), and for the Black-reported race strata, it was 32.03 (SD = 8.72). The average age of the overall group was 52.53 (SD = 18.61), while it was 54.64 (SD = 18.20) for the White-reported race strata and 45.19 (SD = 18.19) for the Black-reported race strata. There were 1,076 cases in the multi-ancestry group. In the White-reported race strata, there were 596 cases, and in the Black-reported race strata, there were 420 cases, for a fibroid prevalence of 2% and 7%, respectively (Table 1).

Of the 51,715 female individuals in BioVU, 9,373 samples passed quality control and had fibroid status as determined by the imaging algorithm (BioVU-imaged). The average BMI of the overall group was 29.21 (SD = 8.08). In the White-reported race strata, it was 28.55 (SD = 7.69), and in the Black-reported race strata it was 31.90 (SD = 9.17). The average age of the overall group was 44.86 (SD = 17.33). The White-reported race strata had an average age of 46.96 (SD = 17.45), and the Black-reported race strata had an average age of 38.02 (SD = 15.41). There was a fibroid prevalence of 16% out of 1,463 cases in the multi-ancestry group, whereas it was 14% of 975 cases in the White-reported race strata and 22% of 410 cases in the Black-reported race strata (Table 1).

3.2. *Polygenic risk score validation*

3.2.1. *Multi-ancestry (META) PRS*

The META PRS was validated in the multi-ancestry group of the eMERGE, BioVU-ICD, and BioVU-imaged cohorts. Student's T-tests for difference in means found mean META PRS to be significantly different between cases and controls in all multi-ancestry cohorts: $p = 9.85 \times 10^{-9}$ for eMERGE, p =

$2.50x10^{-10}$ for BioVU-ICD, and $p = 3.07x10^{-12}$ for BioVU-imaged (Table 2). For a one standard deviation increase in PRS, the OR for fibroid diagnosis was 1.11 (95% CI: 1.06 – 1.17, $p = 2.43x10^{-5}$) in eMERGE, 1.23 (95% CI 1.15 – 1.32, $p = 9.68x10^{-9}$) in BioVU-ICD, and 1.26 (95% CI: 1.18 – 1.35, $p = 2.4x10^{-12}$) in BioVU-imaged (Figure 1A). The META PRS performed best in the BioVU-imaged cohort with an AUROC of 0.74 (95% CI: 0.71 – 0.77), while the AUROC was 0.67 (95% CI: 0.64 – 0.69) in the eMERGE cohort and 0.66 (95% CI: 0.63 – 0.69) in the BioVU-ICD cohort (Figure 2A). The AUROCs for the covariate models were 0.73 (95% CI: 0.71 - 0.76), 0.66 (95% CI: 0.63 - 0.68), and 0.65 (95% CI: 0.62 - 0.69), respectively.

When the META PRS was applied to each reported race strata separately, it was validated in the White-reported race strata of each cohort but not in the Black-reported race strata (Figures 1B and 1C). The ORs for the White-reported race strata were 1.15 (95% CI 1.09 - 1.22, $p = 6.83x10^{-7}$) in eMERGE, 1.25 (95% CI: 1.15 – 1.39, $p = 5.63=10^{-7}$) in BioVU-ICD, and 1.34 (95% CI: 1.23 – 1.44, $p = 1.34x10^{-12}$) in BioVU-imaged. The META PRS performed best in the White-reported race strata of the BioVU-imaged cohort with an AUROC of 0.70 (95% CI: 0.66 - 0.73), while the AUROC was 0.63 (95% CI: 0.60 – 0.66) in eMERGE and 0.63 (95% CI: 0.58 – 0.68) in BioVU-ICD (Figure 2B). The AUROCs of the covariate model were 0.68 (95% CI: 0.65 – 0.72), 0.63 (95% CI: 0.60 – 0.65), and 0.58 (95% CI: 0.53 – 0.64), respectively. When the META PRS was modeled with covariates in the Black-reported race strata, the model itself had predictability, but the META PRS did not contribute any of the predictability (Figure 2C).

Table 2. Polygenic risk score (PRS) T-test results. Student's T-tests were used to determine if mean PRS was significantly different between cases and controls. Significance level is 0.002 (0.05/27 tests). Cases and controls in the multi-ancestry (META), European ancestry (EUR), and East Asian ancestry (EAS) PRS were stratified according to race: White reported race and non-Hispanic ethnicity (White), Black reported race and non-Hispanic ethnicity (Black), and all the above (All).

Cohort Reported Race	**META PRS**	**EUR PRS**	**EAS PRS**
eMERGE			
All	$9.85x10^{-9}$	$1.89x10^{-7}$	$7.67x10^{-17}$
White	$2.49x^{10-9}$	0.002	$9.90x10^{-12}$
Black	0.57	0.14	0.06
BioVU - ICD			
All	$2.50x10^{-10}$	$4.10x10^{-13}$	$4.64x10^{-6}$
White	$7.06x10^{-10}$	$1.06x10^{-7}$	0.00063
Black	0.07	0.21	0.06
BioVU - imaged			
All	$3.07x10^{-12}$	$6.91x10^{-11}$	$2.75x10^{-8}$
White	$1.77x10^{-13}$	$2.49x10^{-10}$	$2.21x10^{-6}$
Black	0.65	0.85	0.12

3.2.2. *European ancestry (EUR) PRS*

The EUR PRS was validated in the multi-ancestry and White-reported race strata but not in the Black-reported race strata for both BioVU cohorts. The EUR PRS was only validated in the multi-ancestry strata of the eMERGE cohort. Mean EUR PRS was significantly different between cases and controls for all multi-ancestry cohorts: $p = 1.89 \times 10^{-7}$ for eMERGE, $p = 4.10 \times 10^{-13}$ for BioVU-ICD, and $p = 6.91 \times 10^{-11}$ for BioVU-imaged (Table 2). In the multi-ancestry cohorts, the ORs were 1.18 in both BioVU-ICD (95% CI 1.09 – 1.26, $p = 8.94 \times 10^{-6}$) and BioVU-imaged (95% CI: 1.10 – 1.26, $p = 1.89 \times 10^{-6}$) (Figure 1D). The EUR PRS was not associated with the risk of fibroid diagnosis in the eMERGE cohort ($p = 0.30$). The EUR PRS performed best in the BioVU-imaged cohort with an AUROC of 0.74 (95% CI: 0.71 – 0.77), while the AUROC was 0.63 (95% CI: 0.60 – 0.66) in eMERGE and 0.67 (95%

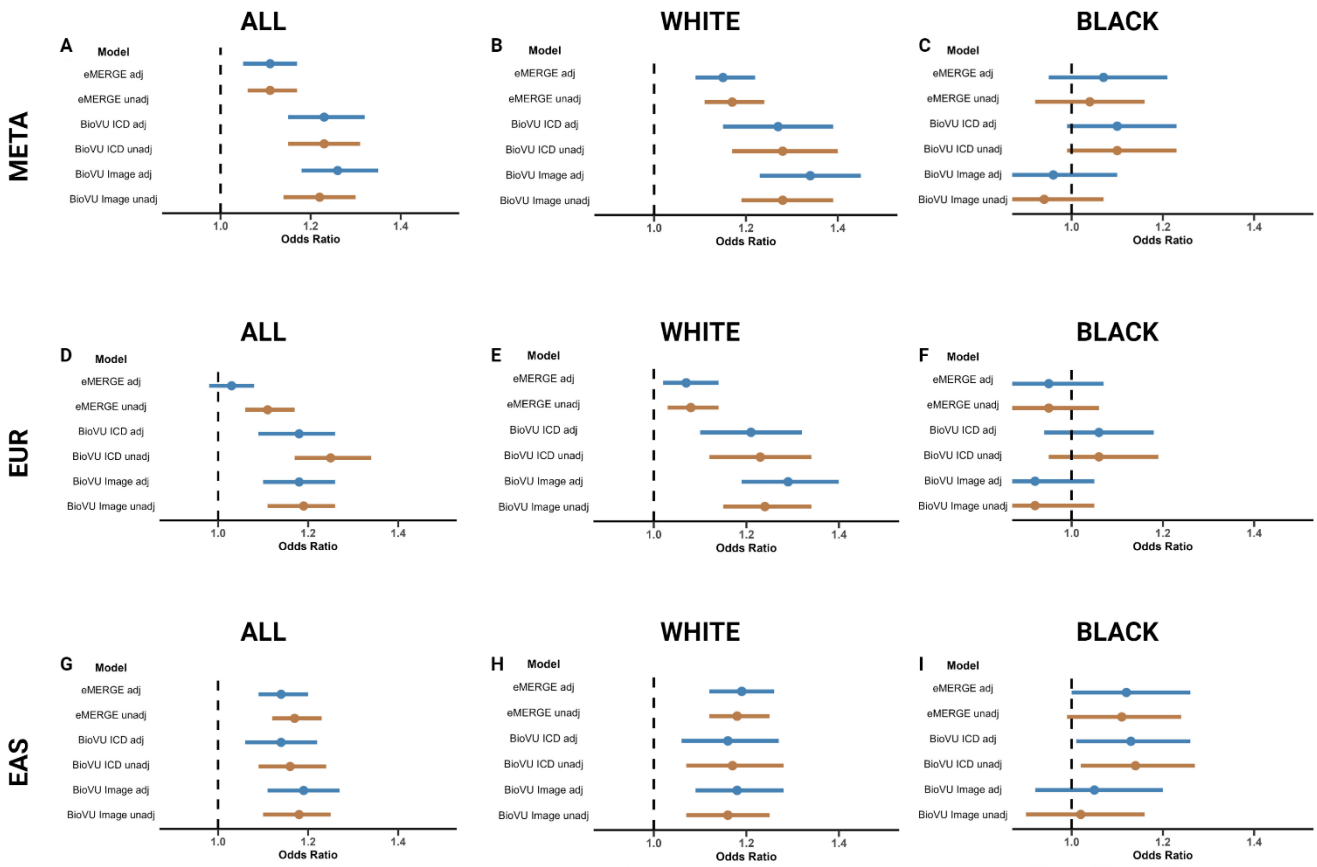


Fig. 1. Polygenic risk score (PRS) ten-fold cross validation results stratified by race for each cohort. Race refers to White reported race and non-Hispanic ethnicity (WHITE), Black reported race and non-Hispanic ethnicity (BLACK), and all the above (ALL). Odds ratios (ORs) are calculated for one standard deviation increase in PRS for adjusted and unadjusted models. **A/B/C)** ORs for all multi-ancestry (META) PRS cohorts. **D/E/F)** ORs for all European ancestry (EUR) PRS cohorts. **G/H/I)** ORs for all East Asian ancestry (EAS) PRS cohorts. Created with Biorender.com.

CI: 0.63 – 0.70) in BioVU-ICD (Figure 2D). The AUROCs for the covariate model were 0.73 (95% CI: 0.71 – 0.76), 0.66 (95% CI: 0.63 – 0.68), and 0.65 (95% CI: 0.62 – 0.69), respectively.

The EUR PRS was applied to the White-reported race strata of the cohorts, but it did not show an association with the risk of fibroid diagnosis in the eMERGE cohort ($p = 0.01$) because it did not reach the significance level of our ten-fold cross-validation for the EUR PRS ($p < 6.17 \times 10^{-4}$). The ORs for the EUR PRS in the BioVU cohorts were 1.21 (95% CI: 1.10 – 1.32, $p = 5.59 \times 10^{-5}$) in BioVU-ICD and 1.29 (95% CI: 1.19 – 1.40, $p = 4.69 \times 10^{-10}$) in BioVU-imaged (Figure 1E). The EUR PRS performed best in the BioVU-imaged cohort with an AUROC of 0.69 (95% CI: 0.66 – 0.72), while the AUROC was 0.63 (95% CI: 0.63 – 0.60 – 0.68) in eMERGE and 0.62 (95% CI: 057 – 0.67) in BioVU-ICD (Figure 2E). The AUROCs of the covariate model were 0.68 (95% CI: 0.65 – 0.72), 0.63 (95% CI: 0.60 – 0.65), and 0.58 (95% CI: 0.53 – 0.64), respectively. The EUR PRS did not associate with risk of

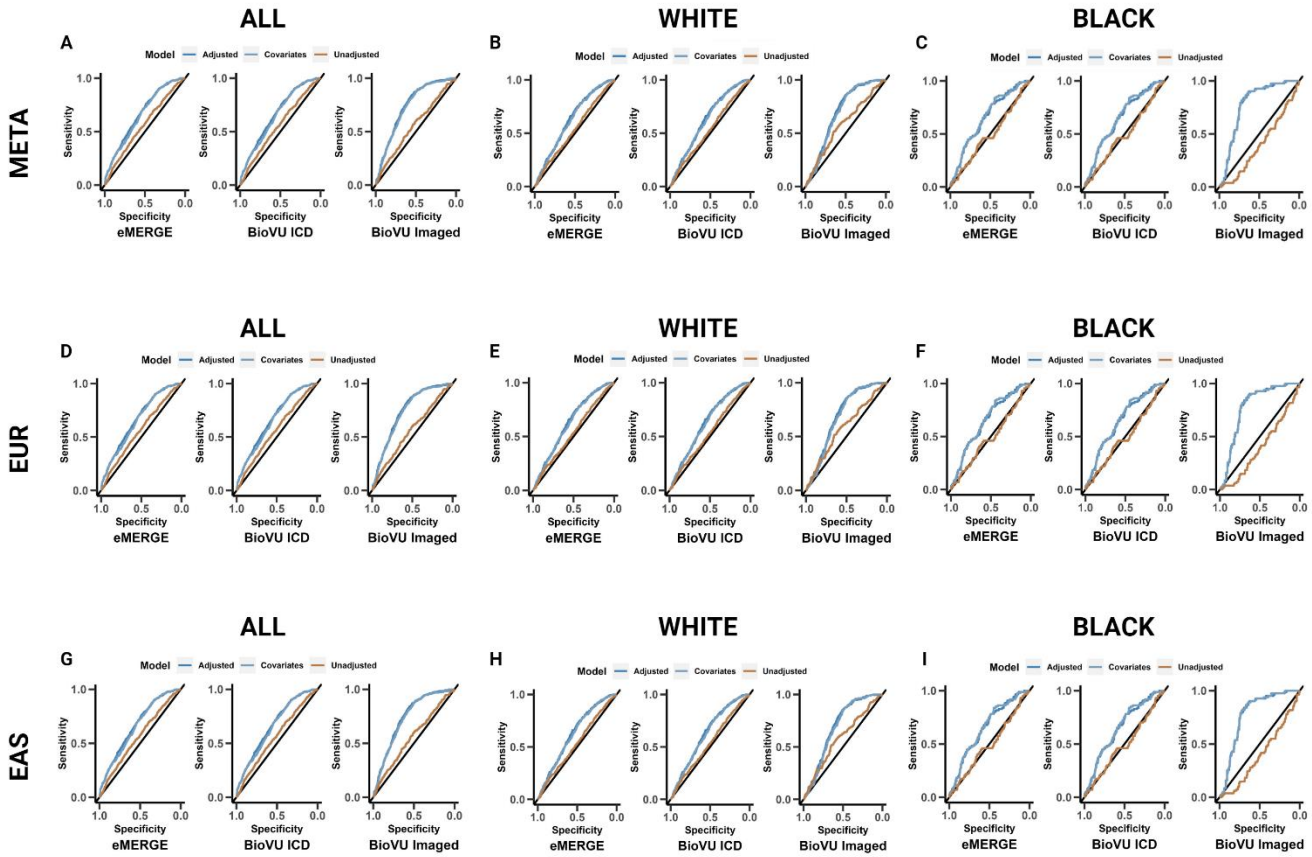


Fig. 2. Polygenic risk score (PRS) ten-fold cross validation results stratified by race for each cohort. Race refers to White reported race and non-Hispanic ethnicity (WHITE), Black reported race and non-Hispanic ethnicity (BLACK), and all the above (ALL). **A/B/C)** Area under receiver operator curve **(**AUROC) plots for each multi-ancestry (META) PRS cohort. **D/E/F)** AUROC plots for each European ancestry (EUR) PRS cohort. **G/H/I)** AUROC plots for each East Asian ancestry (EAS) PRS cohort. Created with Biorender.com.

fibroid diagnosis in the Black-reported race strata of any cohort nor did the models have predictability for fibroid status (Figures 1F and 2F).

3.2.3. *East Asian ancestry (EAS) PRS*

The EAS PRS was validated in the multi-ancestry and White-reported race strata but not the Black-reported race strata for all cohorts. There was a significant difference in mean EAS PRS between cases and controls for all multi-ancestry cohorts: $p = 7.67x10^{-17}$ for eMERGE, $p = 4.64x10^{-6}$ for BioVU-ICD, and $p = 2.75x10^{-8}$ for BioVU-imaged (Table 2). In the multi-ancestry cohorts, the ORs were 1.14 for both eMERGE (95% CI: 1.09 – 1.20, $p = 1.64x10^{-7}$) and BioVU-ICD (95% CI: 1.06 – 1.22, $p = 3.00x10^{-4}$) cohorts, while the BioVU-imaged cohort had a slightly larger OR of 1.19 (95% CI: 1.11 – 1.27, $p = 3.31x10^{-7}$) (Figure 1G). The EAS PRS performed best in the BioVU-imaged cohort with an AUROC of 0.73 (95% CI: 0.71 – 0.76), while the AUROC was 0.68 (95% CI: 0.65 – 0.70) in eMERGE and 0.66 (95% CI: 0.62 – 0.69) in BioVU-ICD (Figure 2G). The AUROCs for the covariate model were 0.73 (95% CI: 0.71 – 0.76), 0.66 (95% CI: 0.63 – 0.68), and 0.65 (95% CI: 0.62 – 0.69), respectively.

When the EAS PRS was applied to the White-reported race strata of each cohort, the ORs were similar: 1.19 (95% CI: 1.12 – 1.26, $p = 1.26x10^{-9}$) in eMERGE, 1.17 (95% CI: 1.07 – 1.28, $p = 1.00x10^{-4}$) in BioVU-ICD, and 1.18 (95% CI: 1.09 – 1.28, $p = 4.20x10^{-5}$) in BioVU-imaged. While the effect size of the EAS PRS was consistent across cohorts, the PRS had the most predictability in the BioVU-imaged cohort with an AUROC of 0.69 (95% CI: 0.66 – 0.72). Next was the eMERGE cohort with an AUROC of 0.64 (95% CI: 0.61 – 0.67) followed by the BioVU-ICD cohort with an AUROC of 0.60 95% CI: 0.54 – 0.65) (Figure 2H). The AUROCs of the covariate model were 0.68 (95% CI: 0.65 – 0.72), 0.63 (95% CI: 0.60 – 0.65), and 0.58 (95% CI: 0.53 – 0.64), respectively. The EAS PRS was not associated with risk of fibroids in the Black-reported race strata of any cohort, nor did it exhibit meaningful predictability (Figures 1I and 2I).

4. Discussion

Using current approaches to estimate PRSs and publicly available resources, we constructed and validated a multi-ancestry (META) PRS in two separate biobanks. META PRS performed better than the single ancestry PRSs, European ancestry (EUR) PRS and East Asian ancestry (EAS) PRS, in all cohorts. These findings show the utility of using a multi-ancestry approach over a single ancestry analysis for PRS. A PRS constructed from the same summary statistics may work in one target population but not others due to a variety of factors including differences in data structures, genotyped variants, and ancestry.[5] By enabling the use of two ancestries over one to construct a PRS, more genetic variation is included in the model, which is precisely what PRS-CSx was created to accomplish.[11] Including multiple different genetic ancestries in a PRS should enable the model to be transferrable to other racial groups, further attempting to answer a problem that has led to portability failures of past PRS models.

PRSs have suffered from an inability to transfer across racial and ethnic groups, resulting in concerns that use of PRS in precision medicine may further contribute to disparities observed in disease trends.[8] When our PRS was evaluated by Black-reporting and White-reporting racial strata, there were differences in validating the findings. The META PRS strongly associated with fibroid status in the

White-reported race strata among all cohorts but failed to validate in any Black-reported race strata. Yet, the AUROC of the modeled covariates in the Black-reported race strata was close to, and in some cases better than, the AUROC for the adjusted META PRS applied to the White-reported race strata. While the META PRS showed no association or predictability with fibroid diagnosis, adding the covariates of age, BMI, and ten PCs were sufficient for a prediction model. Additionally, the pseudo-R^2 was higher in the multi-ancestry group than in the racial strata, demonstrating how adding Black-reporting individuals to the overall model enhances the explained variation. We acknowledge that the smaller sample size of Black-reporting individuals may have limited statistical power, potentially affecting the precision of effect size estimates and the detection of significant associations. However, this limitation is common when studying underrepresented populations, underscoring the need for future efforts to increase sample sizes and improve cohort diversity to enhance the generalizability and accuracy of PRS in Black-reporting individuals. Excluding these populations from prediction modeling only serves to perpetuate health disparities among traditionally underrepresented populations. Thus, while META PRS does not hold any predictive power for Black-reporting individuals alone, their inclusion in the model remains essential for accurate risk assessment based upon clinical factors for all populations.

A major strength of this study is the use of publicly available resources to construct a multi-ancestry fibroid PRS, making it accessible for a broad audience. Utilizing large-scale biobank GWAS summary statistics from the FinnGen research project and the Biobank Japan, which have performed GWAS on thousands of traits, we demonstrated that these projects are sufficient for future PRS studies, sparing researchers from conducting their own on smaller populations. Despite this, we acknowledge the 'messiness' of clinical data used in these studies is often due to case-control definitions based on the presence or absence of a phenotype in an individual's EHR. In particular, case-control definitions based on EHRs are often reliant on the presence or absence of a clinical phenotype, which introduces potential inaccuracies. For example, fibroid cases may be underdiagnosed in individuals who are asymptomatic, resulting in the inclusion of false negatives among controls and subsequently impacting the accuracy and robustness of GWAS associations. A more stringent, precise set of case-control criteria, such as those incorporating diagnostic imaging, would likely improve both GWAS outcomes and PRS performance. This is demonstrated in the study, where the BioVU-imaged cohort, which confirmed fibroid diagnoses through imaging, showed improved AUROC and pseudo-R^2 compared to the ICD-defined cohort, demonstrating enhanced predictability and stability from more precise phenotyping.

Additionally, we observed significant heterogeneity across the populations studied. For instance, the Finnish population's unique genetic background, stemming from a founding bottleneck and relative isolation, may limit transferability to other populations, thereby affecting PRS-CSx program compatibility. This study primarily utilized European and East Asian ancestry data from the FinnGen research project and the Biobank Japan, but did not include African genetic ancestry, despite its known risk factor for fibroids. This highlights a broader issue in genetic research, where populations of European ancestry are often overrepresented, limiting the generalizability of findings. There has been one successful fibroid GWAS in individuals of African ancestry, which identified a unique locus associated with fibroids. This may indicate the genetic architecture of fibroids differs significantly across ancestries.[29,30] Expanding genetic studies into these underrepresented populations should help fill in this missing variance, thus increasing the predictability of PRS. We were unable to use the African

ancestry summary statistics, as that study was performed by our group with samples from BioVU. Removing the overlapping samples from the source population in the BioVU validation cohorts resulted in further insufficient sample sizes for the Black-reported race strata in this study.

In summary, we developed and validated a multi-ancestry (META) PRS in two biobanks, demonstrating superior performance compared to single ancestry PRSs (European and East Asian) across all cohorts. This underscores the advantage of a multi-ancestry approach, which incorporates a broader genetic variation and potentially increases model transferability across different racial groups. Despite the META PRS's strong association with fibroid status in White-reported race strata, it showed limited predictive power for Black-reported race strata, highlighting a persistent challenge in PRS models' applicability across racial groups. Nonetheless, including diverse ancestries in the PRS model improved overall prediction accuracy and addressed disparities in health risk assessment. Strengths of this study include the use of large-scale biobank data and imaging validation to enhance PRS robustness. However, limitations such as inaccurate case-control definitions and a lack of African genetic ancestry in the data underscore the need for more inclusive and precise research methodologies. Ultimately, while multi-ancestry PRS models hold promise for reducing health disparities, further efforts are needed to integrate diverse genetic ancestries and improve predictive accuracy for all populations.

References

1. Baird, D.D., Dunson, D.B., Hill, M.C., Cousins, D. & Schectman, J.M. High cumulative incidence of uterine leiomyoma in black and white women: ultrasound evidence. *Am. J. Obstet. Gynecol.* **188**, 1001–1007 (2003).
2. Wegienka, G., et al. Self-reported heavy bleeding associated with uterine leiomyomata. *Obstet. Gynecol.* **101**, 431–437 (2003).
3. Cramer, S.F. & Patel, A. The frequency of uterine leiomyomas. *Am. J. Clin. Pathol.* **94**, 435–438 (1990).
4. Varol, N., Healey, M., Tang, P., Sheehan, P., Maher, P. & Hill, D. Ten-year review of hysterectomy morbidity and mortality: can we change direction? *Aust. N. Z. J. Obstet. Gynaecol.* **41**, 295–302 (2001).
5. Lewis, C.M., & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med* **12**, 1-11 (2020).
6. Visscher, P.M., Yengo, L., Cox, N.J., & Wray, N.R. Discovery and implications of polygenicity of common diseases. *Science* **373**, 1468-1473 (2021).
7. Wand, H., et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**(7849), 211-219 (2021).
8. Adeyemo, A., et al. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat Med* **27**, 1876-1884 (2021).
9. Kachuri, L., et al. Principles and methods for transferring polygenic risk scores across global populations. *Nat Rev Genet* **1**, 1-18 (2023).
10. Osterman, M.D., Kinzy, T.G., & Bailey, J.N.C. Polygenic risk scores. *Curr Protoc* **1**, e126 (2021).
11. Ruan, Y. et al. Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
12. Piekos, J. A., et al. Uterine fibroid polygenic risk score (PRS) associates and predicts risk for uterine fibroid. *Hum. Genet*. **141**, 1739–1748 (2022).
13. McCarty, C.A., et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* **4**, 1-11 (2011).
14. Stanaway, I.B., et al. The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet Epidemiol* **43**, 63-81 (2019).
15. Roden, D.M., et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* **84**, 362-369 (2008).
16. Gallagher, C.S., et al. Genome-wide association and epidemiological analyses reveal common genetic origins between uterine leiomyomata and endometriosis. *Nat Commun* **10**, 4857 (2019).
17. Das, S., et al. Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
18. Denny, J. C. et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205-1210 (2010).
19. Feingold-Link, L., et al. Enhancing uterine fibroid research through utilization of biorepositories linked to electronic medical record data. *J Womens Health* **23**, 1027-1032 (2014).

20. Kurki, M.I., et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023).
21. Sakaue, S., et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415-1424 (2021).
22. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.A., & Smoller, J.W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* **10**, 1776 (2019).
23. Purcell, S., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2021).
25. Chang, C.C., et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
26. Whickam, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).
27. Kuhn, M. Building predictive models in R using the caret package. *J Stat Softw* **28**(5), 1-26 (2008).
28. Robin, X., et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **7**, 77 (2011).
29. Hellwege, J.N., et al. A multi-stage genome-wide association study of uterine fibroids in African Americans. *Hum Genet* **136**, 1363-1373 (2017).
30. Edwards, T.L., et al. A trans-ethnic genome-wide association study of uterine fibroids. *Front Genet* **10**, 511 (2019).

Uterine fibroids show evidence of shared genetic architecture with blood pressure traits

Alexis T. Akerele

School of Graduate Studies, Department of Microbiology, Immunology and Physiology, Meharry Medical College, Nashville, TN, 37208, Division of Quantitative and Clinical Science, Department of Obstetrics and Gynecology, Data Science Institute, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, 37203, U.S.A.
Email: apigg21@mmc.edu

Jacqueline A. Piekos and Jeewoo Kim

Division of Quantitative and Clinical Science, Department of Obstetrics and Gynecology, Data Science Institute Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, 37203,U.S.A.
Email: jacqueline.piekos@vanderbilt.edu, jeewoo.kim@vanderbilt.edu

Nikhil K. Khankari and Jacklyn N. Hellwege

Division of Genetic Medicine, Department of Medicine, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, 37203, U.S.A.
Email: jacklyn.hellwege@vumc.org, nikhil.khankari@vumc.org

Todd L. Edwards

Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, 37203, U.S.A.
Email: todd.l.edwards@vumc.org

Digna R. Velez Edwards

Division of Quantitative and Clinical Science, Department of Obstetrics and Gynecology, Data Science Institute Vanderbilt University Medical Center, Nashville, TN, 37203, U.S.A.
Email: digna.r.velez.edwards@vumc.org

Uterine leiomyomata (fibroids, UFs) are common, benign tumors in females, having an estimated prevalence of up to 80%. They are fibrous masses growing within the myometrium leading to chronic symptoms like dysmenorrhea, abnormal uterine bleeding, anemia, severe pelvic pain, and infertility. Hypertension (HTN) is a common risk factor for UFs, though less prevalent in premenopausal individuals. While observational studies have indicated strong associations between UFs and HTN, the biological mechanisms linking the two conditions remain unclear. Understanding the relationship between HTN and UFs is crucial because UFs and HTN lead to substantial comorbidities adversely impacting female health. Identifying the common underlying biological mechanisms can improve treatment strategies for both conditions. To clarify the genetic and causal relationships between UFs and BP, we conducted a bidirectional, two-sample Mendelian randomization (MR) analysis and evaluated the genetic correlations across BP traits and UFs. We used data from a multi-ancestry genome-wide association study (GWAS) meta-analysis of UFs (44,205 cases and 356,552 controls), and data from a cross-ancestry GWAS meta-

analysis of BP phenotypes (diastolic BP [DBP], systolic BP [SBP], and pulse pressure [PP], N=447,758). We evaluated genetic correlation of BP phenotypes and UFs with linkage disequilibrium score regression (LDSC). LDSC results indicated a positive genetic correlation between DBP and UFs (R_g=0.132, $p<5.0x10^{-5}$), and SBP and UFs (R_g=0.063, $p<2.5x10^{-2}$). MR using UFs as the exposure and BP traits as outcomes indicated a relationship where UFs increases DBP (odds ratio [OR]=1.20, $p<2.7x10^{-3}$). Having BP traits as exposures and UFs as the outcome showed that DBP and SBP increase risk for UFs (OR =1.04, $p<2.2x10^{-3}$; OR=1.00, $p<4.0x10^{-2}$; respectively). Our results provide evidence of shared genetic architecture and pleiotropy between HTN and UFs, suggesting common biological pathways driving their etiologies. Based on these findings, DBP appears to be a stronger risk factor for UFs compared to SBP and PP.

Keywords: uterine fibroids; hypertension; Mendelian randomization; genetic correlation; women's health

1. Introduction

Uterine leiomyomata (fibroids, UFs) is a highly prevalent and genetically complex disease.[1,2] UFs are the most common benign tumors in premenopausal individuals, having an estimated cumulative prevalence of up to approximately 80%, with Black women being up to three times more likely to develop UFs than White women.[2,3] Black women also experience earlier onset, more severe symptoms, greater challenges in accessing timely and effective treatment, and higher rates of surgical interventions like hysterecomy.[3,4] Uterine fibroids are characterized by the presence of fibrous masses growing in and on the smooth muscle of the uterus. People with UFs present symptoms of dysmenorrhea, heavy or abnormal uterine bleeding, and anemia, and pelvic pain.[2] Uterine fibroid symptoms have substantial overlap with other gynecologic conditions, such as ovarian cysts, endometriosis, and menstrual disorders.[1] Inevitably, the overlap in symptomology between these conditions present challenges for UFs to be accurately and timely diagnosed and treated. Studies report that up to 41% of females with UFs visit two or more providers and experience a three to five year deferment in treatment of UFs.[4,5] Nonsurgical and fertility-preserving interventions for UFs are limited; treatment strategies commonly aim at controlling symptoms or surgeries to remove the affected reproductive tissues/organs altogether. Furthermore, myomectomy surgeries are not 100% effective as reoccurrence of UFs occurs in approximately 59% of patients.[6,7] As a consequence, UFs result in over a $10-billion annual national economic burden due to direct and indirect costs associated with doctors office visits, treatments, hospitalizations, surgeries, and wage losses.[8]

UFs are associated with many comorbidities, including hypertension (HTN), which also disproportionately affects Black women. HTN is a major risk factor associated with increasing risk of UFs. Individuals with UFs are at 1.44 fold increased risk of having HTN.[9] For complex diseases such as UFs and HTN, multiple factors influence their onset, progression, and severity, with health inequities further compounding these risks. The most recent published GWAS estimated single nucleotide polymorphism (SNP)-heritability of UFs to be 13%, which is much lower than pedigree-based heritability estimates ranging between 26 and 69%, suggesting that many factors, including social determinants of health, are involved in UFs development and are not fully captured by GWAS.[10–13]

Although there are many epidemiological studies indicating strong associations between UFs and HTN, the origins and links between the two conditions remain poorly understood. Emerging evidence suggests that females with treated HTN are at a reduced risk of UFs than those with untreated HTN.[14,15] However, these studies are not without limitations, such as collider bias, selection bias, and possible incomplete covariate assessment, reducing their reliability. The limitations of these observational studies underscore the need for more robust methodologies, such as Mendelian randomization, which can better account for confounding factors and help elucidate causal relationships.

Elucidating the shared genetic architecture across UFs and HTN can help capture the biological factors contributing to disease risk. Additionally, further research is needed to determine whether these conditions are causes or consequences of one another, or perhaps arising in part from some common causes. Our study aims to clarify the genetic correlations and pleiotropy of HTN and UFs. To help clarify the potential causal association between UFs and HTN, we conducted a Mendelian randomization analysis and evaluated the genetic correlations using data from two large, multi-ancestry UFs and blood pressure (BP) trait GWAS meta-analyses.

2. Methods

2.1 *Study populations*

We used cross-ancestry, meta-analyzed uterine fibroid GWAS summary statistics from an unpublished study as well as summary statistics from a cross-ancestry GWAS meta-analysis of BP phenotypes (including diastolic BP [DBP], systolic BP [SBP], and pulse pressure[PP], [N = 447,758]).[16–18] The UFs multi-ancestry meta-analysis included diverse cohorts of female participants (with the inclusion of individuals having African/Black, European/White, and Asian ancestry or self-reported race) who were 18 years of age or older from BioVU, eMERGE, All of Us, Coronary Artery Risk Development in Young Adults, Black Women's Health Study, FinnGen, and Biobank Japan[18–25]. The BP multi-ancestry meta-analysis included cohorts of individuals in the Million Veteran Program and UK Biobank. The BP meta-analysis also included people having African/Black, European/White, and Asian ancestry or self-reported race. To ensure independent meta-analysis summary statistics for BP and UFs, summary statistics from UK Biobank were excluded from the UFs meta-analysis to avoid overlapping and 58,832 cases and 295,991 controls samples remained.

2.2 *Linkage disequilibrium score regression analysis*

With the GWAS summary statistics for UFs and BP traits, we conducted Linkage Disequilibrium Score Regression (LDSC) analyses to assess pairwise genetic correlations and estimate heritability.[26] Summary statistics were filtered under the following default parameters: imputation quality > 0.9, minor allele frequency (MAF) between 0.01 and 0, strand ambiguous SNPs, SNPs with duplicated 'rs' numbers, multi-allelic variants, insertion/deletions were removed as determined by the LDSC program. Alleles were merged with the HapMap 3 reference panel and LD scores were precomputed from 1000 Genomes European GWAS data.[27,28]

2.3 *Bidirectional, two-sample mendelian randomization*

Bidirectional, two-sample Mendelian randomization (MR) was performed with the "TwoSampleMR" R package (version 0.5.7).[29] Using the MR approach, we evaluated the relationships between UFs and BP traits. MR uses genetic variants robustly associated with exposures of interest as genetic instrumental variables to estimate the causal and unbiased association between the exposure with the outcome. The MR approach assumes the following: (1) the genetic instrument is strongly associated with the trait; (2) the genetic instrument only affects the outcome via the trait; and (3) the genetic instrument is not associated with confounders of the exposure-outcome association.[30] Bidirectional MR allows users to investigate the direction of the relationship between two phenotypes (e.g., determining if HTN is a cause or a consequence of UFs). The genetic instruments for the analysis were selected from the meta-analyzed summary-level data of BP traits and UFs by linkage disequilibrium clumping of genome-wide significant SNPs ($p<5x10^{-8}$) with an r^2 threshold of 0.01. The inverse variance weighted (IVW) method was utilized to obtain initial estimates of the associations. In addition to the standard IVW Mendelian randomization estimate, we used MR-Egger to detect directional pleiotropy and F-statistics were calculated to assess genetic instrument strength. All F-statistics of genetic instruments used in the analysis were >29.

2.4 *Functional annotation and gene set analysis*

The Functional Mapping and Annotation (FUMA) is a web-based tool that analyzes GWAS summary statistics in various post-GWAS analysis[31]. We used FUMA to conduct pathway analysis of the UFs genetic instruments used for MR. Gene mapping of SNPs present in the genetic instrument was completed using the SNP2GENE process. The UFs GWAS summary statistics were used as the GWAS input file and the UFs genetic instrument file was input as pre-defined independent lead SNPs. We utilized default parameters for SNP2GENE and opted out of the identification of additional independent lead SNPs. After the SNPs were annotated and mapped to their respective genes, we used the GENE2FUNC function, under default parameters, to obtain insight of the biological mechanisms of our prioritized gene set. GENE2FUNC uses biological information from multiple databases for each gene annotated in SNP2GENE to identify biological pathways associated with the gene set.

3. Results

3.1 *Genetic correlation across uterine fibroids and blood pressure traits*

To determine whether UFs and BP associations are due to shared genetic architecture or arise from independent genetic contributions to risk, we first used genetic correlation analysis. LDSC results (Table 1) indicated a positive genetic correlation between DBP and UFs ($R_g = 0.132$, $p < 5.0x10^{-5}$), and SBP and UFs ($R_g = 0.063$, $p < 2.5x10^{-2}$ 0.025). Genetic correlation between pulse pressure and UFs was nonsignificant ($R_g = 0.006$, $p > 0.050$). Ancestry-stratified analysis of European/White cohort displayed a similar trend where significant positive genetic correlations estimated between DBP and UFs (R_g= 0.114, $p < 5.2 \times 10^{-6}$) and SBP and UFs (R_g= 0.08, $p < 1.1 \times 10^{-3}$). Genetic correlations of UFs and BP traits in the African/Black cohort were greatly inflated, had large standard errors, and were therefore inconclusive and omitted.

Table 1: LDSC Results of Uterine Fibroids and Blood Pressure Traits. This table depicts results of genetic correlations of uterine fibroids and blood pressure traits. There were positive genetic correlations between DBP and UFs and SBP and UFs. Multi-Ancestry: includes results from GWAS summary statistics of individuals of African/Black, European/White, and Asian ancestry or self-reported race. European/White: includes results from GWAS summary statistics of individuals of European/White ancestry or self-reported race only. DBP: diastolic blood pressure; SBP: systolic blood pressure; PP: pulse pressure.

	Multi-Ancestry		**European/White**	
	Genetic Correlation (R_g)	p-value	Genetic Correlation (R_g)	p-value
DBP	0.132	$5.0x10^{-5}$	0.114	5.2×10^{-6}
SBP	0.063	$2.5x10^{-2}$	0.080	1.1×10^{-3}
PP	0.006	0.830	0.011	0.669

3.2 *Assessment of potential causal associations between uterine fibroids and blood pressure traits using bidirectional, two-sample mendelian randomization*

Having UFs as the exposure variable and BP traits as the outcomes indicated a moderate, positive relationship between UFs and DBP (odds ratio [OR] = 1.20, 95% confidence interval [CI]: 1.08-1.32, $p < 2.7x10^{-3}$) (Table 2). SBP and PP as outcomes and did not provide significant results. Using BP traits as exposure variables and UFs as the outcome showed that DBP increases risk for UFs (OR = 1.04, 95% CI: 1.01-1.06, $p < 2.2x10^{-3}$) (Table 2). With DBP as the exposure, a significant MR Egger

regression p-value ($p < 3.8x10^{-2}$) indicated horizontal pleiotropy in the analysis. Leave-one-out sensitivity analysis depicted that the DBP IVW estimate was largely influenced by a single SNP, rs78378222 with DBP as the exposure. Excluding SNP rs78378222 reduced the IVW OR estimate by 0.08 (OR = 1.12, 95% CI: 1.01-1.22, $p < 4.4x10^{-2}$) with DBP as the outcome (Table 3). A similar change in IVW estimate was observed with excluding SNP rs78378222 (OR=1.02, 95% CI: 1.01-1.04, $p < 2.8x10^{-3}$) with DBP as exposure and UFs as outcome and there was no horizontal pleiotropy (MR egger p=0.52).

Table 2. Bidirectional, Two-Sample Mendelian Randomization Results of Uterine Fibroids and Blood Pressure Traits. There were significant relationships between uterine fibroids and systolic blood pressure with systolic blood pressure as the exposure variable. There was a bidirectional relationship between uterine fibroids and diastolic blood pressure. OR: odds ratio; SE: standard error; Egger p-value: test for horizontal pleiotropy.

Exposure	Outcome	No. of Genetic Instruments	Mendelian Randomization			
			IVW OR	SE	p-value	Egger p-value
Uterine Fibroids	Diastolic Blood Pressure	108	1.20	0.061	$2.7x10^{-3}$	0.180
	Systolic Blood Pressure	110	1.15	0.111	0.220	0.892
	Pulse Pressure	110	0.91	0.089	0.310	0.289
Diastolic Blood Pressure	Uterine Fibroids	99	1.04	0.012	$2.2x10^{-3}$	$3.8x10^{-2}$
Systolic Blood Pressure		407	1.00	0.002	$4.0x10^{-2}$	0.400
Pulse Pressure		411	0.99	0.004	$9.8x10^{-3}$	0.090

Table 3. Changes in Inverse Variance Weighted Estimate After Removing SNP rs78378222 in Genetic Instruments. The IVW estimates decreased for diastolic blood pressure after removing SNP rs78378222 from the genetic instruments. Pleiotropy originally detected by the MR Egger test was eliminated. OR: odds ratio; SE: standard error; Egger p-value: test for horizontal pleiotropy.

Exposure	Outcome	No. of Genetic Instruments	Mendelian Randomization			
			IVW OR	SE	p-value	Egger p-value
Uterine Fibroids	Diastolic Blood Pressure	107	1.12	0.055	$4.4x10^{-2}$	0.433
	Systolic Blood Pressure	109	1.20	0.113	0.098	0.487
	Pulse Pressure	109	1.01	0.082	0.524	0.347
Diastolic Blood Pressure	Uterine Fibroids	98	1.02	0.008	$2.8x10^{-3}$	0.524
Systolic Blood Pressure		407	1.00	0.002	$3.5x10^{-2}$	0.400
Pulse Pressure		410	0.99	0.003	0.061	0.773

3.3 *FUMA pathway analysis*

We used the UFs genetic instruments from the MR analysis as the pre-defined lead SNPs supplementary input file for FUMA. FUMA identified 22 biological pathways, derived from Canonical Pathways, that were associated with our prioritized gene set (Table 4)[32]. There were multiple overlapping genes associated with *TP53* mediation, *P53* regulation and signaling, and cellular senescence dysfunction. Other overlapping genes are present in pathways related to androgen biosynthesis, cell cycle, DNA damage responses, and breast cancer. FUMA also identified genes previously reported from studies present in the GWAS catalog that are associated with DBP, SBP, and PP (Table 5)[33]. There were 19, 35, and 22 genes in our gene list that are significantly associated with DBP, SBP, and PP; respectively.

Table 4. FUMA Results of Biological Pathways Significantly Associated with Our Prioritized Gene List. The gene list was created from the uterine fibroid genetic instruments used for Mendelian randomization. N: number of genes; Adjusted p-value: p-value after correcting for multiple comparisons.

Gene Set	N	Adjusted p-value	Genes
Reactome Regulation of TP53 Activity Through Methylation	5	1.08×10^{-2}	*MDM4, ATM, TP53, CHEK2, EP300*
WP Glioblastoma Signaling Pathways	8	1.55×10^{-2}	*PIK3C2B, MDM4, PDGFRA, CDKN1A, ATM, FOXO1, TP53, EP300*
Biocarta G2 Pathway	5	1.55×10^{-2}	*CDKN1A, ATM, TP53, CHEK2, EP300*
PID HDAC Class III Pathway	5	1.55×10^{-2}	*CDKN1A, SIRT3, FOXO1, TP53, EP300*
WP miRNAs Involved in DNA Damage Response	4	3.46×10^{-2}	*CDKN1A, ATM, RAD52, TP53*
WP Male Infertility	9	3.67×10^{-2}	*PARP1, CLOCK, ESR1, CYP17A1, ATM, YBX2, HORMAD2, TCN2, EP300*
Reactome Sumoylation	10	3.67×10^{-2}	*PARP1, DNMT3A, THRB, ESR1, NCOA2, RAD52, TP53, EP300, L3MBTL2, RANGAP1*
Biocarta BLK3 Pathway	3	3.67×10^{-2}	*ATM, TP53, CHEK2*
PID P53 Regulation Pathway	6	3.67×10^{-2}	*MDM4, ATM, NEDD8, TP53, CHEK2, EP300*
Reactome RHO GTPases Activate PAKs	4	3.67×10^{-2}	*CDC42, MYH11, MYH10, NF2*
Biocarta ATM Pathway	4	3.67×10^{-2}	*CDKN1A, ATM, TP53, CHEK2*
Biocarta P53 Hypoxia Pathway	4	3.67×10^{-2}	*CDKN1A, ATM, TP53, EP300*
WP ATM Signaling Pathway	5	4.43×10^{-2}	*MDM4, CDKN1A, ATM, TP53, CHEK2*
PID ERA Genomic Pathway	6	4.68×10^{-2}	*GREB1, ESR1, NCOA2, NEDD8, XBP1, EP300*

WP Breast Cancer Pathway	9	4.68×10^{-2}	*WNT4, PARP1, KIT, CDKN1A, ESR1, WNT2, FGF8, ATM, TP53*
WP NAD Metabolism in Oncogene Induces Senescence and Mitochondrial Dysfunction Associated Senescence	4	4.68×10^{-2}	*PARP1, SIRT3, SLC2A4, TP53*
Reactome G1 S DNA Damage Checkpoints	6	4.68×10^{-2}	*MDM4, CDKN1A, PSMD13, ATM, TP53, CHEK2*
Reactome Androgen Biosynthesis	3	4.68×10^{-2}	*POMC, SRD5A3, CYP17A1*
Reactome Regulation of FOXO Transcriptional Activity By Acetylation	3	4.68×10^{-2}	*SIRT3, FOXO1, EP300*
KEGG P53 Signaling Pathway	6	4.82×10^{-2}	*MDM4, CDKN1A, SESN1, ATM, TP53, CHEK2*
WP DNA Damage Response	6	4.82×10^{-2}	*CDKN1A, SESN1, ATM, RAD52, TP53, CHEK2*
Reactome Circadian Clock	6	4.82×10^{-2}	*USP46, CLOCK, NCOA2, BTRC, ELOVL3, EP300*

Table 5. FUMA Results of Genes Reported in the GWAS Catalog Significantly Associated with Blood Pressure Traits. The gene list was created from the uterine fibroid genetic instruments used for Mendelian randomization. N: number of genes; Adjusted p-value: p-value after correcting for multiple comparisons.

Gene Set	N	Adjusted p-value	Genes
Diastolic Blood Pressure	19	1.69×10^{-2}	*DNM3, MDM4, OCIAD2, PDLIM5, HSPA4, ESR1, RGS17, PAX2, BTRC, ARL3, CYP17A1, CNNM2, NT5C2, SLK, SORCS3, SLC2A4, TP53, ZNRF3, TNRC6B*
Systolic Blood Pressure	35	2.17×10^{-8}	*RNF207, WNT4, DNM3, MDM4, ITPR1, TEC, SLAIN2, OCIAD1, OCIAD2, PDGFRA, TERT, ESR1, RBPMS, KANK1, PAX2, FGF8, SUFU, ARL3, WBP1L, CYP17A1, AS3MT, CNNM2, NT5C2, RPEL1, SLK, SORCS3, ARL14EP, RAD52, ITGA11, SLC2A4, ZNF208, TTC28, ZNRF3, C22orf31, TNRC6B*
Pulse Pressure	22	2.06×10^{-3}	*RNF207, TEC, SLAIN2, CHIC2, PDGFRA, PDLIM5, CDKN1A, ESR1, TRIM8, ARL3, WBP1L, CYP17A1, CNNM2, NT5C2, RPEL1, SLK, SORCS3, WT1, MYH11, SLC2A4, TP53, TNRC6B*

4. Discussion

Our study utilized summary statistics from the largest multi-ancestry UFs GWAS meta-analysis to date and a BP trait multi-ancestry GWAS meta-analysis with 447,758 individuals.[16,17] Here we report that UFs and BP have significant, positive genetic correlations and there is a bidirectional, causal relationship between UFs and BP traits. We provide evidence that genetic predisposition to UFs increases BP. We also report novel associations between DBP and UFs.

Blood pressure is a modifiable risk factor for cardiovascular disease. Approximately 20% of reproductive-aged female individuals have high BP worldwide.[34,35] HTN is often underdiagnosed in female populations, and less than 25% have this condition under control.[36] Approximately one in five deaths in females is attributed to high BP in the United States.[37] There are notable parallels between the disparities of HTN and UFs. For both conditions, Black women have the highest prevalence, the most severe symptoms, and have the poorest health outcomes.[4,5,34] So far, the epidemiological relationship between UFs and BP has been mainly evaluated by retrospective and observational studies. Some of these studies have identified that individuals being treated for HTN (specifically with angiotensin-converting enzyme [ACE] inhibitors and beta adrenoceptor antagonist [beta blockers]) are at reduced risk of UFs, implicating the renin angiotensin pathway and angiogenic processes in UFs pathology.[14,38] Yet, ACE inhibitors are prescribed less to Black patients due to poor BP responses to treatments.[39] Combination therapies that include ACE inhibitors and a calcium channel agonist or diuretic are highly effective at reducing BP in Black patients.[40] Currently, there is one U.S. Food and Drug Administration (FDA) approved combination therapy for UFs.[41,42] The causal associations identified in our study may aid in discovering new targets for developing more combination therapies for UFs.

In this study, the LDSC indicated a positive genetic correlation between UFs and SBP, UFs and DBP, but not UFs and PP, which is consistent with findings from prior research[9,43]. UFs and DBP were more correlated than SBP and UFs, implying that DBP and UFs have a stronger, common underlying genetic background. A similar trend was also observed in the MR data. In both directions, DBP exhibited a stronger relationship with UFs than SBP. SBP and DBP have differing biological mechanisms driving their pathologies. SBP is largely influenced by arterial stiffening. Elevated and high SBP are indicative of decreased compliance of the arteries, which can be attributed to the aging process or arteriosclerosis[44,45]. Conversely, DBP is a measure of arteriolar peripheral resistance. High DBP is attributed to the thickening of the vascular wall and increased tone of the smooth muscle in the arterioles.[44,46] This causes an increase in the pressure blood exhorts within the arteries the during the diastole. The relationship demonstrated in our results suggests that the shared genetics between UFs, SBP, and DBP are also through different mechanisms.

FUMA provided biological context to this study. Pathways identified by FUMA involve *p53* regulation and signaling, *TP53* mediation, and cellular response to hypoxic stress, all of which are associated with vascular alterations and endothelial dysfunction in hypertension.[47] Furthermore, *TP53* and *p53* are regulators of important cellular processes including DNA repair, cell cycle arrest, cellular senescence, and apoptosis.[48,49] Tumor suppressor genes have been implicated in mediating the induction

of sex hormone-binding globulin and other steroid binding factors and their mutations are associated with increased cancer pathogenesis.[50,51] This suggests that the vascular changes observed in HTN may intersect with the hormonal pathways influencing UFs formation, providing a plausible explanation for the shared genetic mechanisms observed between DBP, SBP, and UFs. Further investigation into the TP53 pathway and its role in both vascular and fibroid pathology could elucidate additional therapeutic targets for managing these conditions.

UFs have been associated with increasing BP in numerous studies.[43,43,52] However, the direction of the relationship remained unclear. Based on the IVW estimate in our study, the causal relationship between BP and UFs is more distinct having UFs as the exposure, suggesting that genetic predisposition to UFs increases risk for HTN. Prior studies showed that individuals with severe and symptomatic fibroids requiring surgical intervention are at higher risk of hypertension.[53] Thus, it is possible that UFs may increase BP over time. UFs depend on uterine arteries for blood supply and uterine angiogenic dysregulation plays an important role in fibroid pathophysiology.[54,55] Genetic alterations, hormonal factors, and hypoxic conditions caused by tumor growth promote angiogenesis within fibroids[56–58]. This vascular remodeling tends to be to be zone specific where more vasculature is present on the superficial layers of fibroids than in deep layers[58]. Similarly, HTN is associated with changes in artery structure and function. The arterial wall is sensitive to changes in tension and stress caused by BP elevation and it is hypothesized that arterial thickening is involved in a positive feedback loop.[59,60] This maladaptation to increased BP may drive subsequent arterial fibrosis and diminishes vascular function. Furthermore, our study demonstrates a positive genetic correlation and causal relationship between DBP and UFs, whereas previous research had only indicated SBP.[61] Interestingly here, the genetic correlation and relationship are stronger between DBP and UFs than SBP. The stronger correlation between DBP and UFs found here suggests that UFs pathophysiology shares more underlying biological mechanisms with DBP than SBP.

This study's major strength is the utilization of highly diverse and large study populations. However, there are some limitations. GWAS of UFs in African and Asian ancestry cohorts are greatly limited. Despite the large overall sample size of our study, the proportion of non-European populations were smaller, limiting the power of ancestry-stratified analyses. Therefore, the ancestry-stratified analysis could not produce reliable results. Future studies should aim to expand GWAS for these groups, particularly given the disparities in HTN and UFs burdens. Also, MR can indicate causal relationships, but results should be validated in larger-scale longitudinal studies and clinical trials. Moreover, future research should also confirm findings by exploring underlying molecular and biological mechanisms. Last, this study relies on GWAS summary statistics from published studies. Other confounding and mediating factors, such as fibroid size and number or oral contraceptive use, could not be taken into consideration.

5. Conclusions

The findings of our study provide evidence for shared genetic architecture across BP traits and UFs risk. Our analysis identified the direction of the relationship between BP and UFs where genetic risk of UFs significantly increases risk for high BP. In addition, risk for UFs influences DBP more than SBP. Clinically, UFs contribute significantly to morbidity and healthcare costs for reproductive-aged females. Accumulating information about the genetic and biological processes driving UFs formation will enhance our understanding of the disease and pave the way for improved therapeutic decisions and personalized treatments.

6. Acknowledgements

Support for this research was provided by the National Institutes of Health through the following grants: a Eunice Kennedy Shriver National Institute of Child and Human Development (NICHD) R01HD093671, R01HD074711, R03HD078567, NICHD award for the Building Interdisciplinary Research Careers in Women's Health career development program K12AR084232, and National Heart, Lung, and Blood Institute 5T32HL007737.

References

1. Sex- and Gender-Based Women's Health: A Practical Guide for Primary Care | SpringerLink. https://link.springer.com/book/10.1007/978-3-030-50695-7.
2. Farkas, A., Tilstra, S. & Gonzaga, A. M. Fibroids, Endometriosis, and Ovarian Cysts. in *Sex- and Gender-Based Women's Health* 141–155 (Springer Cham, 2021).
3. Eltoukhi, H. M., Modi, M. N., Weston, M., Armstrong, A. Y. & Stewart, E. A. The health disparities of uterine fibroid tumors for African American women: a public health issue. *American Journal of Obstetrics and Gynecology* **210**, 194–199 (2014).
4. Stewart, E. A., Nicholson, W. K., Bradley, L. & Borah, B. J. The Burden of Uterine Fibroids for African-American Women: Results of a National Survey. *Journal of Women's Health* **22**, 807–816 (2013).
5. Ghant, M. S., Sengoba, K. S., Vogelzang, R., Lawson, A. K. & Marsh, E. E. An Altered Perception of Normal: Understanding Causes for Treatment Delay in Women with Symptomatic Uterine Fibroids. *Journal of Women's Health* **25**, 846–852 (2016).
6. Donnez, J. & Dolmans, M.-M. Uterine fibroid management: from the present to the future. *Human Reproduction Update* **22**, 665–686 (2016).
7. Recurrence of Endometriosis: Risk Factors, Mechanisms and Biomarkers - Gurkan Bozdag, 2015. https://journals.sagepub.com/doi/full/10.2217/whe.15.56.
8. Al-Hendy, A., Myers, E. R. & Stewart, E. Uterine Fibroids: Burden and Unmet Medical Need. *Semin Reprod Med* **35**, 473–480 (2017).
9. Chen, Y. *et al.* Association of uterine fibroids with increased blood pressure: a cross-sectional study and meta-analysis. *Hypertens Res* **45**, 715–721 (2022).

10. Gallagher, C. S. *et al.* Genome-wide association and epidemiological analyses reveal common genetic origins between uterine leiomyomata and endometriosis. *Nat Commun* **10**, 4857 (2019).
11. Snieder, H., MacGregor, A. J. & Spector, T. D. Genes Control the Cessation of a Woman's Reproductive Life: A Twin Study of Hysterectomy and Age at Menopause1. *The Journal of Clinical Endocrinology & Metabolism* **83**, 1875–1880 (1998).
12. Luoto, R. *et al.* Heritability and risk factors of uterine fibroids — The Finnish Twin Cohort Study. *Maturitas* **37**, 15–26 (2000).
13. Kurbanova, M. K., Koroleva, A. G. & Sergeev, A. S. [Genetic-epidemiologic analysis of uterine myoma: assessment of repeated risk]. *Genetika* **25**, 1896–1898 (1989).
14. Fischer, N. M., Nieuwenhuis, T. O., Singh, B., Yenokyan, G. & Segars, J. H. Angiotensin-Converting Enzyme Inhibitors Reduce Uterine Fibroid Incidence in Hypertensive Women. *The Journal of Clinical Endocrinology & Metabolism* **106**, e650–e659 (2021).
15. Mitro, S. D. *et al.* Hypertension, Cardiovascular Risk Factors, and Uterine Fibroid Diagnosis in Midlife. *JAMA Network Open* **7**, e246832 (2024).
16. Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet* **51**, 51–62 (2019).
17. Kim, J. *et al.* 430 Genome-wide meta-analysis identifies novel risk loci for uterine fibroids across multiple ancestry groups. *Journal of Clinical and Translational Science* **8**, 128–129 (2024).
18. Edwards, T. L. *et al.* A Trans-Ethnic Genome-Wide Association Study of Uterine Fibroids. *Frontiers in Genetics* **10**, (2019).
19. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* **15**, 761–771 (2013).
20. Roden, D. *et al.* Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clinical Pharmacology & Therapeutics* **84**, 362–369 (2008).
21. Pulley, J., Clayton, E., Bernard, G. R., Roden, D. M. & Masys, D. R. Principles of Human Subjects Protections Applied in an Opt-Out, De-identified Biobank. *Clinical and Translational Science* **3**, 42–48 (2010).
22. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
23. FinnGen provides genetic insights from a well-phenotyped isolated population | Nature. https://www.nature.com/articles/s41586-022-05473-8.
24. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415–1424 (2021).
25. Hellwege, J. N. *et al.* A multi-stage genome-wide association study of uterine fibroids in African Americans. *Hum Genet* **136**, 1363–1373 (2017).
26. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
27. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).

28. A global reference for human genetic variation | Nature. https://www.nature.com/articles/nature15393.
29. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
30. Mendelian Randomization | Genetics and Genomics | JAMA | JAMA Network. https://jamanetwork.com/journals/jama/article-abstract/2664027?casa_token=iqpp1cjno3oAAAAA:WAkrKHD6_G1Z3a7A27XeOeUIg6IP7qp8vdi2Dcey-yg8xZXI3-jH25QSKzeBv241mXnrxo8pCl8.
31. Functional mapping and annotation of genetic associations with FUMA | Nature Communications. https://www.nature.com/articles/s41467-017-01261-5.
32. Kutmon, M. *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res* **44**, D488-494 (2016).
33. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977–D985 (2023).
34. Zhou, B., Perel, P., Mensah, G. A. & Ezzati, M. Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension. *Nat Rev Cardiol* **18**, 785–802 (2021).
35. Ford, N. D., Robbins, C. L., Hayes, D. K., Ko, J. Y. & Loustalot, F. Prevalence, Treatment, and Control of Hypertension Among US Women of Reproductive Age by Race/Hispanic Origin. *Am J Hypertens* **35**, 723–730 (2022).
36. CDC. Hypertension Prevalence in the U.S. | Million Hearts®. *Centers for Disease Control and Prevention* https://millionhearts.hhs.gov/data-reports/hypertension-prevalence.html (2023).
37. Wenger, N. K. *et al.* Hypertension Across a Woman's Life Cycle. *J Am Coll Cardiol* **71**, 1797–1813 (2018).
38. Fischer, N. M. *et al.* Beta blockers reduce uterine fibroid incidence in hypertensive women. *European Journal of Obstetrics & Gynecology and Reproductive Biology* **287**, 119–125 (2023).
39. Helmer, A., Slater, N. & Smithgall, S. A Review of ACE Inhibitors and ARBs in Black Patients With Hypertension. *Ann Pharmacother* **52**, 1143–1151 (2018).
40. Flack, J. M., Nasser, S. A. & Levy, P. D. Therapy of Hypertension in African Americans. *Am J Cardiovasc Drugs* **11**, 83–92 (2011).
41. Stewart, E. A. *et al.* Relugolix combination therapy in Black/African American women with symptomatic uterine fibroids: LIBERTY Long-Term Extension study. *American Journal of Obstetrics & Gynecology* **230**, 237.e1-237.e11 (2024).
42. Al-Hendy, A. *et al.* Treatment of Uterine Fibroid Symptoms with Relugolix Combination Therapy. *New England Journal of Medicine* **384**, 630–642 (2021).
43. Haan, Y. C. *et al.* The risk of hypertension and cardiovascular disease in women with uterine fibroids. *J Clin Hypertens (Greenwich)* **20**, 718–726 (2018).

44. Williams, D. L. & Burgess, E. J. 34 - Hypertension. in *Scientific Foundations of Biochemistry in Clinical Practice (Second Edition)* (eds. Williams, D. L. & Marks, V.) 585–600 (Butterworth-Heinemann, 1994). doi:10.1016/B978-0-7506-0167-2.50039-X.
45. Magder, S. The meaning of blood pressure. *Crit Care* **22**, 257 (2018).
46. Smulyan, H. & Safar, M. E. The Diastolic Blood Pressure in Systolic Hypertension. *Ann Intern Med* **132**, 233–237 (2000).
47. Chan, G. H.-H. *et al.* The role of p53 in the alternation of vascular functions. *Front Pharmacol* **13**, 981152 (2022).
48. Aubrey, B. J., Strasser, A. & Kelly, G. L. Tumor-Suppressor Functions of the TP53 Pathway. *Cold Spring Harb Perspect Med* **6**, a026062 (2016).
49. Lieberman, H. B., Panigrahi, S. K., Hopkins, K. M., Wang, L. & Broustas, C. G. p53 and RAD9, the DNA Damage Response, and Regulation of Transcription Networks. *Radiat Res* **187**, 424–432 (2017).
50. Huang, Y. *et al.* TP53-specific mutations serve as a potential biomarker for homologous recombination deficiency in breast cancer: a clinical next-generation sequencing study. *Precis Clin Med* **7**, pbae009 (2024).
51. Charni, M. *et al.* Novel p53 target genes secreted by the liver are involved in non-cell-autonomous regulation. *Cell Death Differ* **23**, 509–520 (2016).
52. Radin, R. G. *et al.* Hypertension and risk of uterine leiomyomata in US black women. *Hum Reprod* **27**, 1504–1509 (2012).
53. Haan, Y. C. *et al.* Hypertension risk in Dutch women with symptomatic uterine fibroids. *Am J Hypertens* **28**, 487–492 (2015).
54. Tal, R. & Segars, J. H. The role of angiogenic factors in fibroid pathogenesis: potential implications for future therapy. *Hum Reprod Update* **20**, 194–216 (2014).
55. Kirschen, G. W. *et al.* Vascular biology of uterine fibroids: connecting fibroids and vascular disorders. *Reproduction* **162**, R1–R18 (2021).
56. Mayer, A. *et al.* HIF-Mediated Hypoxic Response is Missing in Severely Hypoxic Uterine Leiomyomas. in *Oxygen Transport to Tissue XXXI* (eds. Takahashi, E. & Bruley, D. F.) 399–405 (Springer US, Boston, MA, 2010). doi:10.1007/978-1-4419-1241-1_58.
57. Olson, S. L., Akbar, R. J., Gorniak, A., Fuhr, L. I. & Borahay, M. A. Hypoxia in Uterine Fibroids: Role in Pathobiology and Therapeutic Opportunities. *Oxygen* **4**, 236–252 (2024).
58. Konarska, M. *et al.* Angiogenesis and pro-angiogenic factors in uterine fibroids - facts and myths. *Folia Med Cracov* **56**, 37–43 (2016).
59. Humphrey, J. D. Mechanisms of Vascular Remodeling in Hypertension. *American Journal of Hypertension* **34**, 432–441 (2021).
60. Humphrey, J. D., Harrison, D. G., Figueroa, C. A., Lacolley, P. & Laurent, S. Central Artery Stiffness in Hypertension and Aging. *Circulation Research* **118**, 379–381 (2016).

61. Xu, H. *et al.* Uterine leiomyoma causes an increase in systolic blood pressure: a two-sample Mendelian randomization study. *Front. Endocrinol.* **15**, (2024).

Social Determinants of Health and Lifestyle Risk Factors Modulate Genetic Susceptibility for Women's Health Outcomes

Lindsay A Guare, Jagyashila Das, PhD, Lannawill Caruth, Shefali Setia-Verma, PhD

Department of Pathology and Laboratory Medicine, University of Pennsylvania
Philadelphia, PA 19104

Emails: lindsay.guare@pennmedicine.upenn.edu, jagyashila.das@pennmedicine.upenn.edu, lanna.caruth@pennmedicine.upenn.edu, shefali.setiaverma@pennmedicine.upenn.edu

Abstract

Women's health conditions are influenced by both genetic and environmental factors. Understanding these factors individually and their interactions is crucial for implementing preventative, personalized medicine. However, since genetics and environmental exposures, particularly social determinants of health (SDoH), are correlated with race and ancestry, risk models without careful consideration of these measures can exacerbate health disparities. We focused on seven women's health disorders in the All of Us Research Program: breast cancer, cervical cancer, endometriosis, ovarian cancer, preeclampsia, uterine cancer, and uterine fibroids. We computed polygenic risk scores (PRSs) from publicly available weights and tested the effect of the PRSs on their respective phenotypes as well as any effects of genetic risk on age at diagnosis. We next tested the effects of environmental risk factors (BMI, lifestyle measures, and SDoH) on age at diagnosis. Finally, we examined the impact of environmental exposures in modulating genetic risk by stratified logistic regressions for different tertiles of the environment variables, comparing the effect size of the PRS. Of the twelve sets of weights for the seven conditions, nine were significantly and positively associated with their respective phenotypes. None of the PRSs was associated with different ages at diagnoses in the time-to-event analyses. The highest environmental risk group tended to be diagnosed earlier than the low and medium-risk groups. For example, the cases of breast cancer, ovarian cancer, uterine cancer, and uterine fibroids in highest BMI tertile were diagnosed significantly earlier than the low and medium BMI groups, respectively). PRS regression coefficients were often the largest in the highest environment risk groups, showing increased susceptibility to genetic risk. This study's strengths include the diversity of the All of Us study cohort, the consideration of SDoH themes, and the examination of key risk factors and their interrelationships. These elements collectively underscore the importance of integrating genetic and environmental data to develop more precise risk models, enhance personalized medicine, and ultimately reduce health disparities.

Keywords: Polygenic Risk Scores, Social Determinants of Health, Health Disparities, Genetic Risk, Disease Prediction, Women's Health, Breast Cancer, Endometriosis, Ovarian Cancer, Preeclampsia, Uterine Cancer, Uterine Fibroids

1 Introduction

Since the completion of the Human Genome Project in 2003, countless studies have been conducted to associate genetic variants with diseases[1–3]. However, genetic factors accompanied by environmental factors collectively contribute to pathogenesis and progression of diseases. Therefore, quantifying the effects of multimodal risk factors separately and together will help to improve disease risk models. Accurate stratification of individual disease risk is an essential step in the way to reduce the burden of health disparities and implement personalized preventative care.

For many highly heritable diseases, such as coronary artery disease and type 2 diabetes, PRSs are useful for stratifying patients into risk groups based on their genetics. However, in the context of women's health diseases, which have historically been underfunded[4] and understudied[5], the predictive accuracy of PRSs has been inconsistent, especially across diverse populations[6]. Globally, large efforts have been undertaken to build diverse resources to support such studies, including the UK Biobank[7], Finngen[8], BioVU[9], BioBank Japan[10], the Penn Medicine Biobank[11], and a newer resource funded by the NIH, the All of Us (AOU) Research Program[12]. The growth of large genomic datasets has enabled not only the detection of disease-associated genetic variations but also the possibility of using genetic and non-genetic risk factors to predict disease risk before the onset. Numerous studies, like the WISDOM trial[13] focusing on breast cancer and the eMERGE network examining PRS results for 10 disease outcomes[14], are underway to investigate how PRSs can be incorporated into clinical practice.

Environmental risk factors are multi-faceted, including lifestyle measurements as well as social determinants of health (SDoH). Most of these variables are measured through survey participation. Lifestyle aspects, like alcohol use, smoking, and physical activity, have been linked to disease risk for endometriosis[15], breast cancer[16], and uterine fibroids[17], respectively. SDoH are define measurements for social inequities which can impact a person's health. These include neighborhood disorder, stress, and loneliness. Chronic stress and loneliness have been shown to increase lifetime risk of many serious diseases, like Alzheimer's[18], cardiovascular disease[19], etc. Additionally, SDoH impact diseases affecting women specifically[20–22]. Interactions between genetic and environmental effects have been studied previously, with respect to both individual genetic variants[23] and PRSs[24]. It has been shown that incorporating PRS with environment measurements such as stress improves model performance for other complex disorders[25]. Therefore, understanding the influence of lifestyle and environmental factors alongside genetic factors is crucial for predicting women's health outcomes.

One important aspect of predictive modeling in personalized medicine is to examine the disease progression, including the onset of diseases. Both categories of risk factors (genetic and environmental) are most often studied in the context of lifetime disease risk. Time-to-event analyses are growing in popularity to evaluate longitudinal risk, utilizing survival analysis methodologies to evaluate the impact of risk factors on disease progression, including the onset of the disease.

The aim of this study is to identify and quantify interactions between genetic risk of women's health conditions and external variables in a diverse cohort of women within the AOU. We hypothesize that an individuals' susceptibility to disease risks is not solely dictated by their genetic composition but is greatly influenced by these environmental and social determinants. Understanding how environmental contexts impact the efficacy and clinical utility of PRSs will help to ensure that they are implemented in equitable ways.

2 Methods

2.1 Study Dataset – All of Us Research Program

The All of Us Research Program (AOU) is a dataset supported by the NIH comprised of 409,420 participants with electronic health record (EHR) data, 245,400 of whom have short-read whole genome sequencing (WGS) data. In our study, we included 145,563 of the WGS individuals who were assigned female at birth[26] For study individuals, genetic ancestry was assigned by the AOU data team, who computed genetic similarity with the 1000 genomes reference populations based on genetic principal components.

The EHR data for AOU are stored as billing codes in tables that follow the Observational Medical Outcomes Partnership (OMOP) structure[27]. For our focus on women's health conditions, we selected breast cancer (BC), cervical cancer (CC), endometriosis (Endo), ovarian cancer (OC), preeclampsia (PE), uterine cancer (UC), and uterine fibroids (UF). Each of these diseases has ICD-9 and ICD-10 diagnosis codes (Results, Table 1). Case/control status was determined by the presence of one or more ICD codes for each phenotype.

2.2 Calculating PRSs for women's health outcomes

The PGS Catalog[28] is a public repository of PRS weights that have been published and validated. We browsed the PGS catalog for PRSs for each condition. In cases when more than one PRS was available, we prioritized sets of weights that had been tested on large, multi-ancestry validation cohorts and that have shown promising results based on metrics such as AUROC. The accession numbers for the weights we selected are shown in Figure 1. We computed all 12 scores from the downloaded files in genome build 38 with Plink 2.0's --score function[29]. The scores for each phenotype were then standardized by genetic ancestry group.

2.3 Environmental variables (SDoH and lifestyle measures)

AOU issued several surveys to its participants, including SDoH and Lifestyle questionnaires, combining instruments from other well-studied surveys. To compute continuous scales for neighborhood physical disorder, neighborhood social disorder, stress, and loneliness, we followed procedures as described in Tesfaye et al 2024[30]. The other two survey-derived lifestyle variables were smoking and alcohol use. For smoking, there were seven questions. For the three quantitative questions (ranging from 0-99), we assigned these values: responses of zero (1), then the remaining quartiles (2-5). For the other four smoking questions, we assigned numeric values to the responses: Not At All (1), Some Days (3), Every Day (5). There were three questions pertaining to alcohol use, and we assigned responses numerical values of one to five, with five corresponding to heavier drinking.

We aimed to capture other health measurements using biometrics and wearables data. Per individual, we used median Body Mass Index (BMI) measurement over time. We quantified activity levels using two Fitbit-derived measurements: daily steps (ST) and daily sedentary minutes (SM), as both have been linked to health risks[31,32]. Similarly to BMI, we took the median across each day that had measurements to obtain one value per individual. Once we computed each of the nine continuous environmental factors, we visualized the Pearson correlation between them to examine how they relate to each other and potentially eliminate any that were highly correlated.

2.4 Statistical analyses

2.4.1 Stratified time-to-event analyses for age at diagnosis

For each case of the six phenotypes, we assigned the age of first diagnosis code of a condition as "age at diagnosis". This age variable was used as outcome for time-to-event analyses. Time-to-event analyses were performed in two different contexts: stratified by genetic risk and stratified by environmental variable level. For each phenotype, we looked at three curves defined by the tertiles of the stratifying variable (low/medium/high). Those curves (1 = low, 2 = medium, 3 = high) were fit to survival functions[33] using KaplanMeierFitter from the lifelines Python package[34]. The three survival functions were compared in a pairwise scheme using the log rank test, which results in a chi-squared test statistic.

2.4.2 Quantifying effects of PRSs in environmental contexts

Association testing was performed for each of the twelve PRSs with their corresponding phenotype. The odds ratio (OR) coefficient was estimated using a logistic regression (with an intercept) in which the outcome was the phenotype, the risk score was the independent variable, and age at the time of the EHR data extraction was included as a covariate (Equation 1).

$$Logit(Phenotype) \sim Intercept + PRS + Age \quad (1)$$

For the phenotypes with more than one set of PRS weights (breast cancer, endometriosis, ovarian cancer, and uterine fibroids), we selected the PRS with the largest regression coefficient, resulting in six phenotypes with significant PRS effects (Results, Figure 1).

Next, for each phenotype and environmental risk factor, we divided our study population into nine groups based on environmental variable tertiles (low, medium, high) and genetic risk tertile (low, medium, high). To illustrate the differences in risk levels among various environmental and genetic risk groups, we used the medium/medium subgroup as a reference and computed the odds ratio (and 95% confidence interval) for the phenotype in each of the other eight subgroups, displayed in 3x3 grids for comparison.

Finally, to examine whether the impact of the polygenic risk score (PRS) on disease risk varied across different levels of environmental risk, we conducted stratified regression analyses. By dividing the study population into subgroups based on environmental factors, we assessed how the association between PRS and disease outcomes changed within each subgroup, allowing us to determine if the PRS effect size was influenced by the level of environmental risk. Each environmental variable was divided into tertiles, and then the logistic regression was performed as described previously (Equation 1) for each of the three sub-groups. In a similar manner, we tested the effect of each environmental risk factor on the phenotypes, stratified by genetic risk tertile (Equation 2).

$$Logit(Phenotype) \sim Intercept + Environment + Age \quad (2)$$

3 Results

3.1 PRSs for women's health phenotypes

Our study cohort consisted of female AOU participants with short-read WGS (N = 145,563). We assigned case/control phenotypes in AOU using hierarchical diagnosis billing codes, Table 1considering both ICD-9 and ICD-10 codes, as shown in Table 1.

Table 1: The seven women's health phenotypes tested. The root ICD codes used for case definitions, the number of cases in the female AOU WGS dataset, and the mean age at diagnosis (Dx) for those cases.

Phenotype	ICD-9 Code	ICD-10 Code	AOU Cases	Dx Age Mean (std)
Breast Cancer (BC)	174	C50	6,444	58.4 (11.7)
Cervical Cancer (CC)	180	C53	546	51.1 (13.3)
Endometriosis (Endo)	617	N80	4,306	43.5 (11.6)
Ovarian Cancer (OC)	183	C56	815	55.1 (13.2)
Preeclampsia (PE)	642	O14	1,966	30.3 (7.0)
Uterine Cancer (UC)	182	C55	715	59.1 (11.1)
Uterine Fibroids (UF)	218	D25	10,829	48.2 (11.1)

12 sets of weights selected from PGS catalog with reported associations to our phenotypes of interest were selected (Table 2).

Table 2 : The PRSs evaluated along with their reported traits, number of variants, and the percentage of the population reported as European in development/training (dev) and testing set. Those reported as "Unspecified" did not provide ancestry specific population reporting

Score	Reported Trait	Year	Number of Variants	% EUR in Dev	% EUR in Validation
PGS000004	Breast Cancer	2018	313	100	76.4
PGS004611	Breast Cancer	2023	76	58.6	Unspecified
PGS001299	Cervical cancer	2022	24	100	40
PGS003447	Endometriosis	2021	14	98	54.5
PGS002077	Endometriosis	2022	14	100	37.5
PGS001866	Endometriosis	2022	399	100	37.5
PGS002250	Epithelial ovarian cancer	2022	27,240	100	60
PGS003394	Epithelial ovarian cancer	2022	36	100	50
PGS004593	Preeclampsia	2022	1,102,059	Unspecified	100
PGS001795	Uterine cancer	2023	911,692	83.9	100
PGS001032	Uterine fibroids	2022	161	100	40
PGS002263	Uterine fibroids	2022	4,457	100	100

We tested logistic regressions for each of the 12 sets of weights selected from the PGS catalog. The PRS for each phenotype with the most significant positive effect was chosen for downstream analysis (Figure 1).

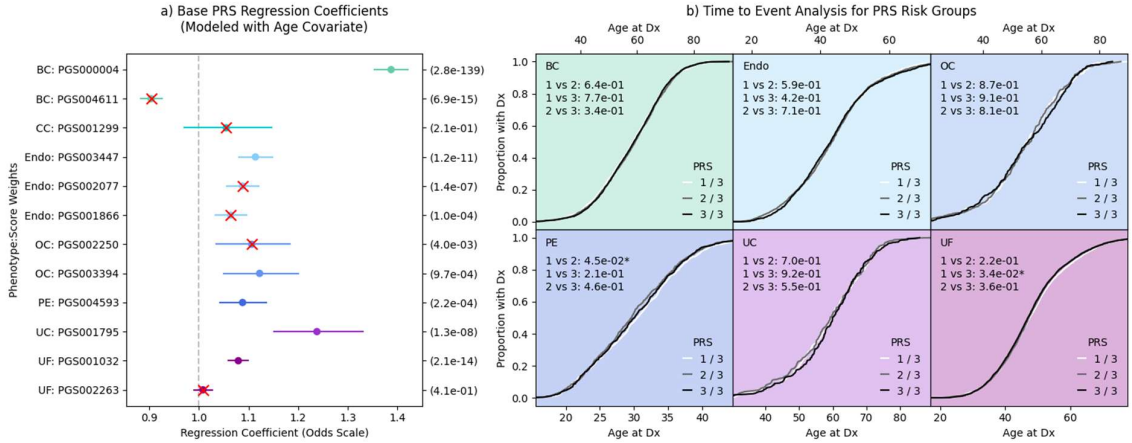


Figure 1: Testing the effects of the PRSs on the women's health outcomes. (a) Coefficients (in odds ratio scale) for logistic regressions based on each PRS. The left axis labels indicate phenotype and PGS Catalog Weights. The right axis labels show the p-value. Scores that were not considered in downstream analyses have a red "X". (b) Time-to event analyses with one curve per PRS risk tertile. Pairwise log rank comparison p values are indicated as text. X-axes above and below each panel are age at diagnosis (Dx). BC: Breast Cancer; UF: Uterine Fibroids; CC: Cervical Cancer; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia.

Based on the logistic regression coefficients for each of the 12 PRSs, we dropped any PRS with odds coefficient <1 (PGS004611 for breast cancer[35]) and any PRS whose p-value for the coefficient was >0.05 (PGS001299 for cervical cancer[36], PGS003394 for ovarian cancer[37], and PGS002263 for uterine fibroids[38]). Since Cervical cancer PRS could not meet these filtering criteria, the phenotype was removed from downstream analysis. In addition, although both PGS002077[39] and PGS001866[39] were significantly associated with endometriosis, only the score that had the strongest effect (PGS003447[40]) was retained.

3.2 Environmental risk factor measurements

The influence of environmental factors, namely, stress level (SL), loneliness level (LL), neighborhood physical disorder (NPD), and neighborhood social disorder (NSD), one biometric measurement (median BMI), two lifestyle scores — alcohol use (AU) and smoking (SK), and two Fitbit measurements — daily steps (ST) and daily sedentary minutes (SM) were tested on susceptibility to genetic risk. We tested these variables for correlation (Figure 2a). Since some measurements were unavailable on all participants, we report the smaller case numbers for each phenotype-measurement combination in Figure 2b.

The most highly correlated variables were NSD and NPD (0.73). Since a higher/greater number of daily steps (ST) is beneficial to health, it was found to be negatively correlated with all other variables except AU. LL was moderately correlated with three other measures, NSD (0.28), NPD (0.21), and SL (0.29). Since some measurements were unavailable for some participants, we report the smaller case numbers for each phenotype-measurement combination. The Fitbit measurements had the fewest participants, so the numbers of cases were small, especially for the rarer phenotypes such as cervical cancer, uterine cancer, ovarian cancer, and preeclampsia. Nearly every participant had BMI measurements, so tests with BMI had the largest sample sizes.

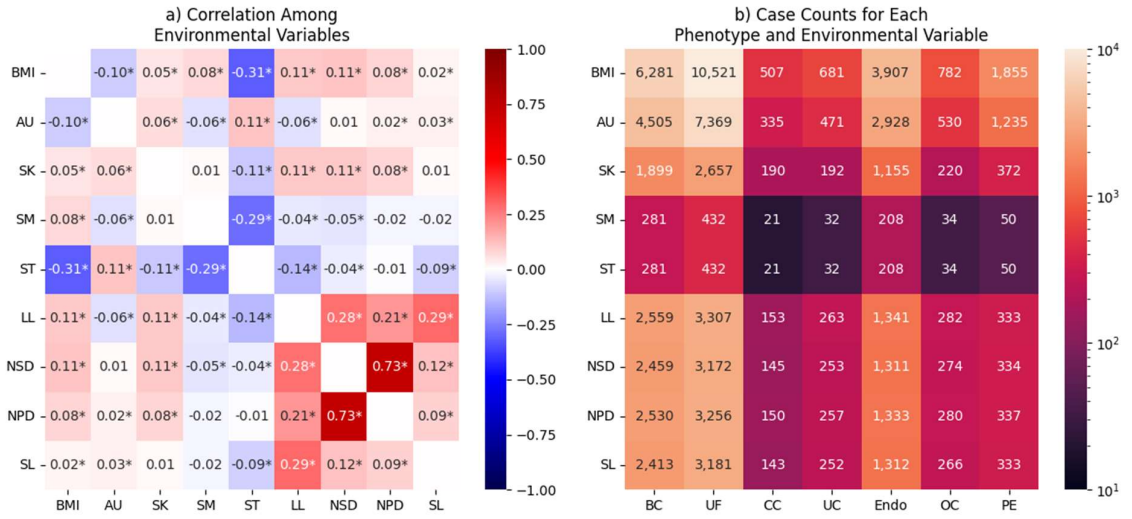
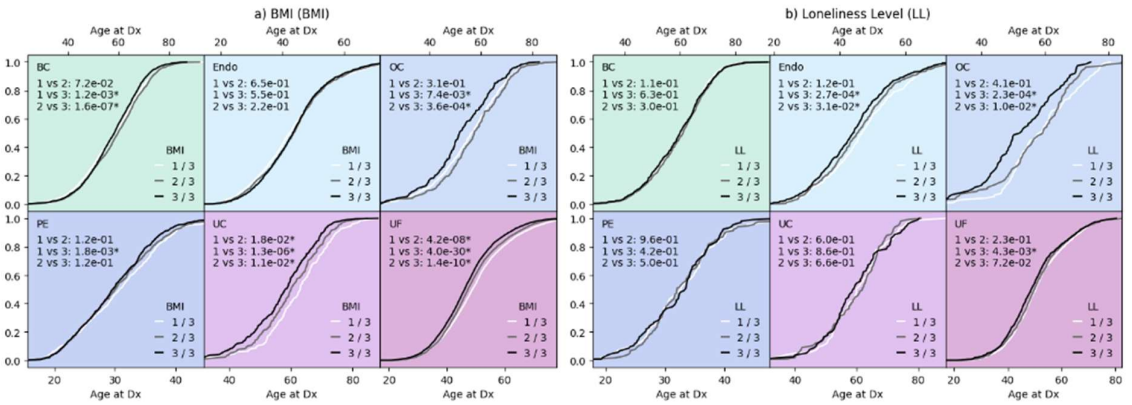


Figure 2: (a) heatmap showing correlation between all nine measurements considered. Correlation values significantly different from zero ($p < 0.05$) are marked with an asterisk. (b) heatmap showing the number of cases for a given phenotype (column) and measurement (row) combination. BC: Breast Cancer; UF: Uterine Fibroids; CC: Cervical Cancer; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia. BMI: Body Mass Index; AU: Alcohol Use; SK: Smoking ; SM: Sedentary Minutes; ST: Steps; LL: Loneliness; NSD: Neighborhood Social Deprivation; NPD: Neighborhood Physical Deprivation and SL: Stress Level.

3.3 Environmental effects on age at diagnosis with time-to-event curves

We estimated the effect of different levels of environmental exposures, categorized into low/medium/high tertiles, on the age at diagnosis for each phenotype. Among the four social determinants of health (SDoH) factors, Neighborhood Social Deprivation (NSD) was removed from the analysis due to its high correlation with Neighborhood Physical Deprivation (NPD), as illustrated previously in Figure 2a. The survival functions, which depict the probability of remaining disease-free over time for each tertile of environmental exposure, are presented in Figure 3. Additionally, the pairwise p-values indicate the statistical significance of the differences between the survival curves for each tertile, highlighting the impact of varying levels of environmental exposures on disease onset.



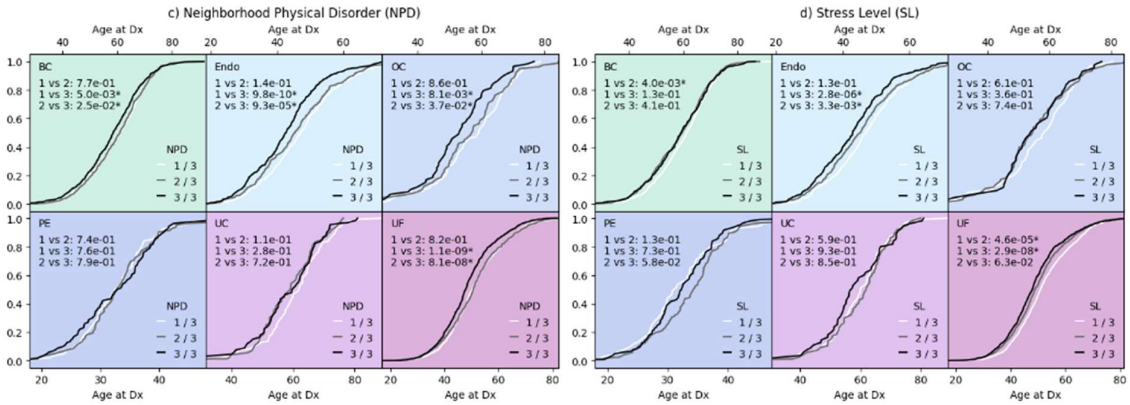


Figure 3: Time-to-event analyses for BMI and the SDoH themes (a - BMI, b - loneliness, c - neighborhood physical disorder, and d - stress). Each panel shows three "survival" curves per phenotype, stratified by the value of the environmental measure where 1 is the lowest tertile and 3 is the highest tertile. The x-axes represent age at diagnosis (Dx). Also indicated in each grid cell are the p-values of pairwise log rank comparisons between those three curves. Any p-values less than 0.05 are annotated with an asterisk. BC: Breast Cancer; UF: Uterine Fibroids; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia.

Of all the environmental risk factors, BMI had the most significant effect on the age at diagnosis. High BMI corresponded to earlier diagnoses of uterine cancer and uterine fibroids (three out of three pairwise comparisons significant), breast cancer and ovarian cancer (two out of three significant), and preeclampsia ($P = 1.8 \times 10^{-3}$ comparing first and third tertiles). Those with high LL scores tended to have earlier diagnoses of endometriosis, ovarian cancer, and uterine fibroids. The high NPD tertile (3) resulted in a significantly earlier diagnosis than the other tertiles for breast cancer, endometriosis, ovarian cancer, and uterine fibroids. No phenotypes had three out of three significant comparisons between the SL tertiles, but the highest SL tertile was associated with earlier diagnosis of endometriosis, while the lowest SL tertile was associated with a later diagnosis of uterine fibroids.

Next, we performed the same time-to-event analyses for the lifestyle variables: AU, SK, ST, and SM (Figure 4). The different AU tertile groups didn't show significant differences for age at diagnosis, except for between the first and second tertiles in breast cancer ($P = 2.2 \times 10^{-3}$); those who drink lightly get diagnosed with breast cancer earlier than those that drink moderately. Similarly, different levels of sedentary minutes also didn't significantly impact diagnosis except for between the first and third tertiles in breast cancer ($P = 4.4 \times 10^{-2}$), with those in the high SM curve, get diagnosed later than the low SM group. Smokers in the third tertile get diagnosed with uterine fibroids earliest (P vs Low = 2.3×10^{-3}, P vs Medium = 1.8×10^{-11}). Breast cancer cases in the lowest tertile of steps get diagnosed latest (P vs Medium = 8.6×10^{-5}, P vs High = 1.4×10^{-2}), this could be confounded by age as older women likely take fewer daily steps. For preeclampsia and uterine cancer cases, those in the third tertile of steps get diagnosed latest.

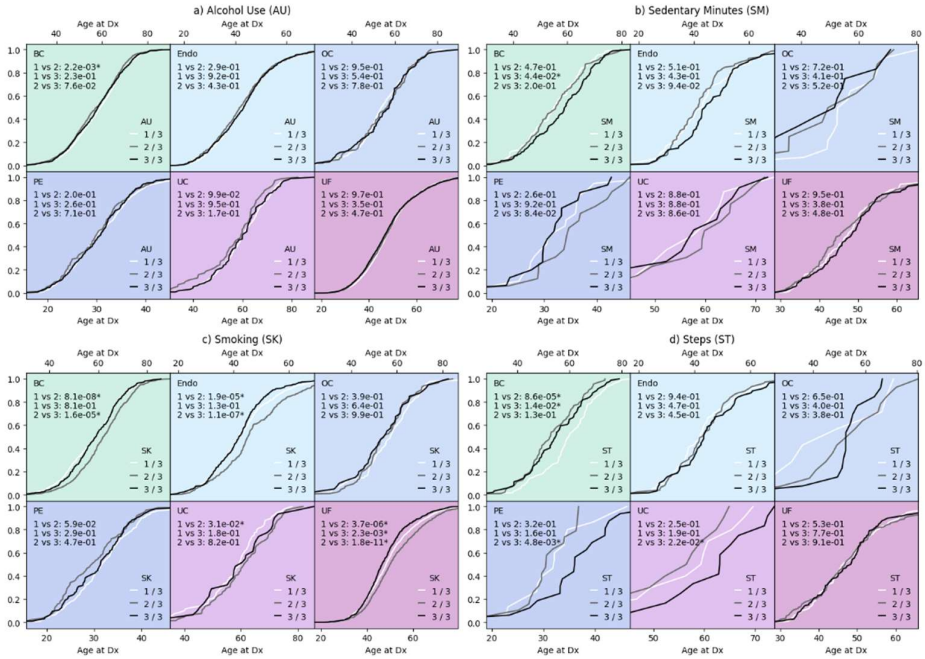
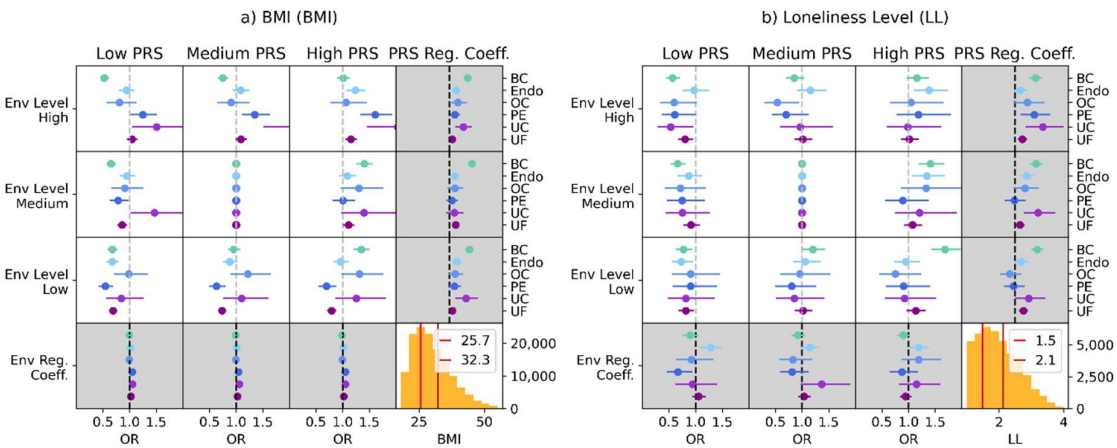


Figure 4: time-to-event analyses for lifestyle measurements (a - alcohol use, b - sedentary minutes, c - smoking, and d - steps). Each panel shows three "survival" curves per phenotype, stratified by the value of the environmental measure where 1 is the lowest tertile and 3 is the highest tertile. The x-axes represent age at diagnosis (Dx). Also indicated in each grid cell are the p-values of pairwise log rank comparisons between those three curves. Any p-values less than 0.05 are annotated with an asterisk. BC: Breast Cancer; UF: Uterine Fibroids; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia

3.4 Genetic risk effects vary by environmental context

We assigned every individual to a genetic risk tertile (low, medium, high) and an environmental exposure level (low, medium, high), the combinations of which resulted in nine sub-groups. Within each of the sub-groups, we computed the odds ratio of the phenotype relative to the medium-medium group. We also performed stratified logistic regressions to estimate the PRS and environmental measurement effects. Because NPD and NSD scores were highly correlated, we only tested NPD. First, we focused on the three remaining SDoH and BMI (Figure 5).



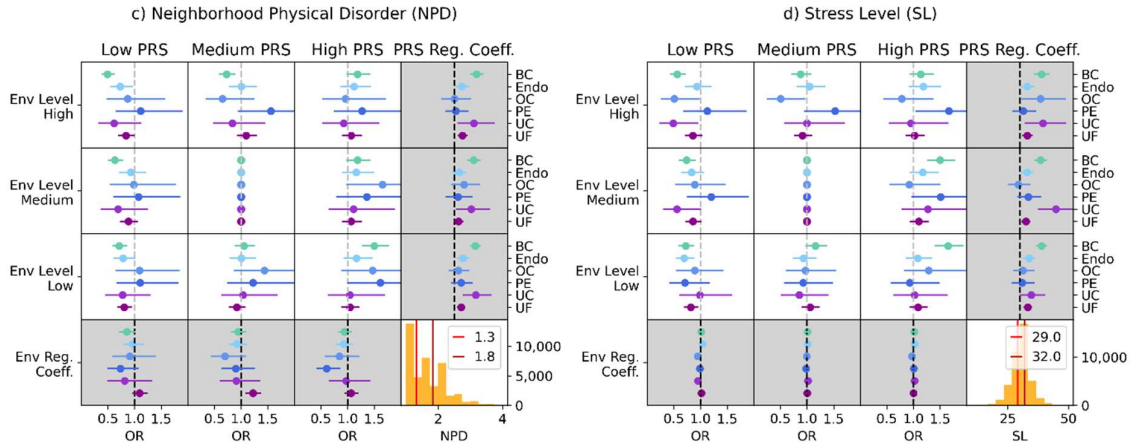
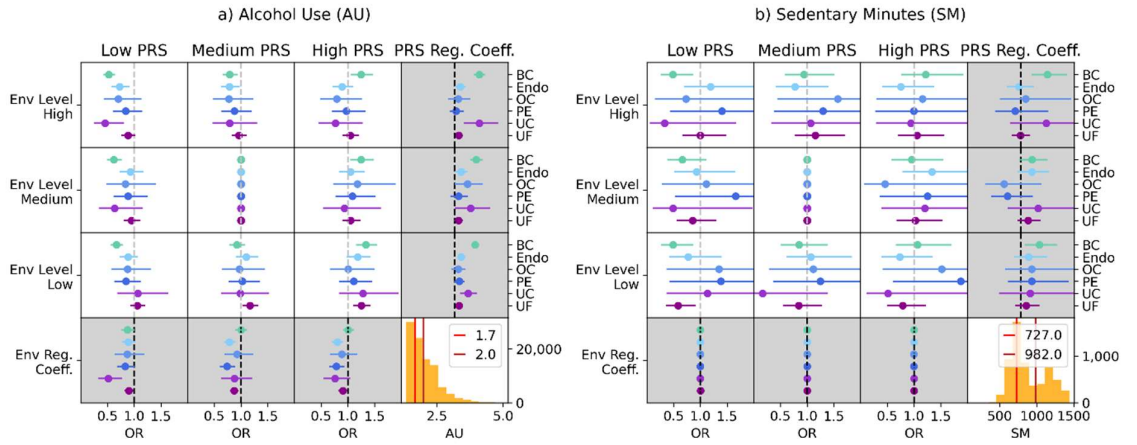


Figure 5: All odds ratio and logistic regression tests performed for BMI and SDoH. The environmental factors are (a) BMI, (b) loneliness, (c) neighborhood physical disorder, and (d) stress. The upper left 3x3 grid in each pane shows the odds ratios of the phenotypes in each cell. The rightmost column shows regression coefficients stratified by environmental tertile. The bottom row shows regression coefficients stratified by genetic risk. The bottom right cell shows a histogram of the environmental variable, with the cutoffs between the tertiles marked. BC: Breast Cancer; UF: Uterine Fibroids; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia

The BMI tertiles were split at 25.7 and 32.3, which are near the conventional cutoffs for overweight (25) and obese (30). At all levels of genetic risk (low, medium, and high), BMI was positively associated with preeclampsia, uterine cancer, and uterine fibroids. BMI was negatively associated with breast cancer. Chronic loneliness and stress are known to be detrimental to long-term health. In the lowest genetic risk group, loneliness was positively associated with endometriosis. Those in the medium and high loneliness groups were more susceptible to genetic risk of ovarian cancer, preeclampsia, and uterine cancer.

Next, we focused on modulating effects of lifestyle factors, including the two Fitbit variables, smoking, and alcohol use (Figure 6).



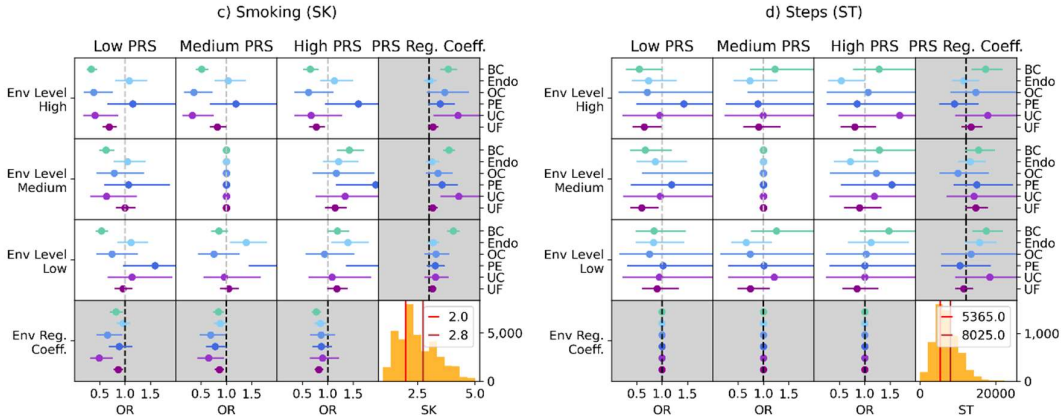


Figure 6: All odds ratio and logistic regression tests performed for the lifestyle variables. The environmental factors are (a) alcohol use, (b) sedentary minutes, (c) smoking, and (d) steps. The upper left 3x3 grid in each pane shows the odds ratios of the phenotypes in each cell. The rightmost column shows regression coefficients stratified by environmental tertile. The bottom row shows regression coefficients stratified by genetic risk. The bottom right cell shows a histogram of the environmental variable, with the cutoffs between the tertiles marked. BC: Breast Cancer; UF: Uterine Fibroids; UC: Uterine Cancer; Endo: Endometriosis; OC: Ovarian Cancer; PE: Preeclampsia

AU had a highly skewed distribution, so the cutoffs between the three tertiles were close together (1.7 vs 2.0). The effect sizes of the PRSs for breast cancer, endometriosis, and uterine cancer were strongest in the tertile with the highest drinking scores. Notably, SK had an inverse effect on breast cancer and uterine fibroids at all levels of genetic risk. Since the models were adjusted for age, it is unlikely that age is confounding these results. Additionally, within the lowest smoking group, the PRS coefficient was not significant, but it was significant for the medium and high smokers. SM had a bimodal distribution. Due to the smaller sample size of the Fitbit data, most of the test statistics were not significant. However, the breast cancer PRS was significantly associated with breast cancer for those who were the most sedentary. Similarly, most of the effect sizes for the steps tests were not significant, but the effect of the breast cancer PRS was significant in the group that took the fewest daily steps on average.

4 Discussion

In this study, we evaluated the effects of environmental variables on women's health outcomes. Specifically, we looked at effects on age at diagnosis and modulation of genetic risk. In 145,563 women in AOU, we analyzed six risk models for women's health diseases. From there, we calculated stratified effect sizes for each PRS for tertiles of each environmental measurement. Overall, we showed that genetic risk models are significantly impacted by different environmental contexts. In general, the most severely affected group of the environment had the strongest effect of the PRS and often resulted in the earliest. These findings underscore the necessity of integrating diverse environmental and social factors into disease risk models to capture the full spectrum of influences on health.

Of the 12 PRSs tested based on their performance in the PGS catalog, nine showed significant positive associations with their respective phenotypes, with breast cancer demonstrating the strongest association. The disparity between the sample population used to create these risk scores and the AOU biobank likely influenced these results, as PRS performance is highly sensitive to population mismatch[41]. There were differences between the

derivation datasets and AOU's unique composition, with about half of the genomic dataset comprising participants of non-European ancestry[42]. This highlights a key drawback of existing PRSs, which are often based on European populations, limiting their relevance for non-European individuals. Notably, genetic risk did not significantly affect age at diagnosis for the six best risk scores, aligning with expectations, as these scores were derived from studies evaluating lifetime disease risk rather than onset. Factors such as SDoH and environmental influences, often correlated with race and ancestry, also play a role in disease susceptibility.

BMI has been significantly associated with a multitude of gynecological conditions[43]. In the current study, we have demonstrated that high BMI can serve as a risk factor for earlier diagnosis of breast, ovarian, and uterine cancer as well as uterine fibroids. Furthermore, BMI was found to be associated with preeclampsia, uterine cancer and uterine fibroids, across all genetic risk groups. Preeclampsia is a pregnancy-related condition, so it is possible that several of the environmental risk factor measurements (BMI, activity levels) may not be representative of the woman's environment at the time of onset as these variables are affected by pregnancy. However, we aimed to evaluate average lifestyle trends, including time leading up to pregnancy. These findings, in conjunction with previous reports on metabolism-related genes on various female cancer types[44,45], emphasize the importance of incorporating environmental factors, especially BMI, for a holistic understanding of disease risk and health outcomes.

The lowest genetic risk groups for endometriosis, preeclampsia, ovarian cancer, and uterine cancer showed positive associations at multiple levels of loneliness. This highlights the profound impact that social and psychological factors can have on physical health. By considering and stratifying risk factors based on both genetic and environmental factors, we can potentially facilitate earlier detection of health burden across diverse population groups. It allows us to identify individuals who, despite having a low genetic risk, may still be at high overall risk due to adverse environmental or social conditions, and ultimately enhance health outcomes for a broader spectrum of the population.

Our study has several limitations. One limitation is that EHR-based phenotyping can be challenging for complex disorders, especially in women's health diseases which are often under-diagnosed, such as uterine fibroids[46] and endometriosis[47]. Phenotyping algorithms have been previously designed to compute phenotypes more accurately than ICD codes alone. Their use in our study is restricted by reliance on clinical notes[48], which are not available in AOU. Other large genomic biobank studies, have leveraged ICD- or PheCode-based case-control phenotyping[1,49,50]. While the accuracy of ICD codes alone varies across the phenotypes, a key advantage of large biobank data is that the substantial sample size can help mitigate the impact of noise introduced by imprecise phenotyping, leading to more robust statistical associations[51].

Another limitation of our study was that we used age at the first diagnosis code of a condition as a proxy for disease onset. Depending on how patients move between healthcare systems, a common occurrence in the EHR is that a condition may have been diagnosed earlier at a different facility, but the corresponding diagnosis code is entered into the EHR only after the patient joins a new healthcare system. This introduces potential noise into the age variable, as the true onset might have been recorded elsewhere or at a different time. However, since many of our sample sizes were large enough to yield significant effects, which should have counteracted the noise. We found that higher-risk environmental groups typically had earlier diagnoses. Given the EHR data, it can be hard to disentangle earlier diagnosis due to earlier onset versus earlier diagnosis due to increased vigilance based on existing risk factors.

Survey data are notoriously challenging to work with, so we were also limited by potential noise introduced by the self-reporting process. To mitigate error, we divided the participants into subgroups by environmental variable tertiles rather than relying on the exact quantitative measures. However, stratifying the individuals into subgroups reduced the sample size and statistical power for each regression. The observations that smoking levels seemed to have non-monotonic effects (medium smokers get diagnosed later with breast cancer, endometriosis, and uterine fibroids) may stem from confounders in the survey measurements. Our overall approach, though it has a few limitations, has provided a practical and scalable way to examine multi-modal predictive and progression models of women's health diseases.

Due to systemic challenges faced by marginalized communities, such populations find themselves exposed to environmental stressors at greater rates[52]. Differing odds ratios for those with similar levels of genetic risk but different levels of environmental risk suggest that not including environmental risk factors in predictive models utilizing PRS could lead to inaccurate risk assessments and potentially overlook significant contributors to disease susceptibility. The current study identifies the dangers in reductionist approach to disease stratification and risk prediction, based solely on either genetics or environmental factors. This suggests that integrating both the genetic and environmental components into a specific disease model would help better classify individual risk.

In the future, using nonlinear approaches for risk modeling which capture variable interactions such as multilayer perceptron could aid in more accurately representing complex relationships between genetics, environmental risk factors, and the phenotypes. While those types of models are harder to train, we can now take advantage of growing data repositories, including AOU, to develop generalizable models that capture important modalities of risk variables. We included eight environmental risk factors, four SDoH and four lifestyle measurements, which capture some, but not all, external influences. Future methodologies may include more risk factors but also should account for potential missing data, as it can be challenging to administer surveys and/or collect wearables data on a large scale. In the future, we also hope to replicate these results in additional biobanks.

Complex systems approaches to incorporate multi-directional interactions between patients and their environment, such as those modeled here, are better suited to leverage the power of genomic data in making widely applicable, clinically relevant tools. Further attempts to strengthen the predictive ability of PRS models need not focus solely on improving the identification of relevant loci, but also relevant environmental risk factors including SDoH. By improving our understanding and application of PRSs, especially in underrepresented areas like women's health, we can enhance disease prediction, prevention, and personalized treatment strategies.

5 Acknowledgments

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study.

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number R01HD110567.

6 References

1. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022).
2. Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024).
3. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
4. Mirin, A. A. Gender Disparity in the Funding of Diseases by the U.S. National Institutes of Health. *Journal of Women's Health* **30**, 956–963 (2021).
5. Schubert, K. G., Bird, C. E., Kozhimmanil, K. & Wood, S. F. To Address Women's Health Inequity, It Must First Be Measured. *Health Equity* **6**, 881–886 (2022).
6. Shah, P. D. Polygenic Risk Scores for Breast Cancer—Can They Deliver on the Promise of Precision Medicine? *JAMA Network Open* **4**, e2119333 (2021).
7. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
8. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
9. Pulley, J., Clayton, E., Bernard, G. R., Roden, D. M. & Masys, D. R. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin Transl Sci* **3**, 42–48 (2010).
10. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology* **27**, S2–S8 (2017).
11. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *Journal of Personalized Medicine* **12**, 1974 (2022).
12. The "All of Us" Research Program. *New England Journal of Medicine* **381**, 668–676 (2019).

13. Eklund, M. *et al.* The WISDOM Personalized Breast Cancer Screening Trial: Simulation Study to Assess Potential Bias and Analytic Approaches. *JNCI Cancer Spectr* **2**, pky067 (2019).

14. Lennon, N. J. *et al.* Selection, optimization, and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse populations. *medRxiv* 2023.05.25.23290535 (2023) doi:10.1101/2023.05.25.23290535.

15. Zhang, Y. & Ma, N.-Y. Environmental Risk Factors for Endometriosis: An Umbrella Review of a Meta-Analysis of 354 Observational Studies With Over 5 Million Populations. *Front. Med.* **8**, (2021).

16. Daly, A. A., Rolph, R., Cutress, R. I. & Copson, E. R. A Review of Modifiable Risk Factors in Young Women for the Prevention of Breast Cancer. *Breast Cancer: Targets and Therapy* **13**, 241–257 (2021).

17. Vafaei, S., Alkhrait, S., Yang, Q., Ali, M. & Al-Hendy, A. Empowering Strategies for Lifestyle Interventions, Diet Modifications, and Environmental Practices for Uterine Fibroid Prevention; Unveiling the LIFE UP Awareness. *Nutrients* **16**, 807 (2024).

18. Sundström, A., Adolfsson, A. N., Nordin, M. & Adolfsson, R. Loneliness Increases the Risk of All-Cause Dementia and Alzheimer's Disease. *The Journals of Gerontology: Series B* **75**, 919–926 (2020).

19. Ajibewa, T. A. *et al.* Chronic Stress and Cardiovascular Events: Findings From the CARDIA Study. *American Journal of Preventive Medicine* **67**, 24–31 (2024).

20. Crear-Perry, J. *et al.* Social and Structural Determinants of Health Inequities in Maternal Health. *Journal of Women's Health* **30**, 230–235 (2021).

21. Katon, J. G., Plowden, T. C. & Marsh, E. E. Racial disparities in uterine fibroids and endometriosis: a systematic review and application of social, structural, and political context. *Fertility and Sterility* **119**, 355–363 (2023).

22. Kurani, S. S. *et al.* Association of Neighborhood Measures of Social Determinants of Health With Breast, Cervical, and Colorectal Cancer Screening Rates in the US Midwest. *JAMA Network Open* **3**, e200618 (2020).

23. Kim, S. *et al.* A comprehensive gene–environment interaction analysis in Ovarian Cancer using genome-wide significant common variants. *International Journal of Cancer* **144**, 2192–2205 (2019).

24. Domingue, B. W., Trejo, S., Armstrong-Carter, E. & Tucker-Drob, E. M. Interactions between Polygenic Scores and Environments: Methodological and Conceptual Challenges. *Sociol Sci* **7**, 465–486 (2020).

25. Musliner, K. L. *et al.* Polygenic liability, stressful life events and risk for secondary-treated depression in early life: a nationwide register-based case-cohort study. *Psychological Medicine* **53**, 217–226 (2023).

26. Data Browser | All of Us Public Data Browser. https://databrowser.researchallofus.org/.

27. Hallinan, C. M. *et al.* Seamless EMR data access: Integrated governance, digital health and the OMOP-CDM. *BMJ Health Care Inform* **31**, e100953 (2024).

28. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* **53**, 420–425 (2021).

29. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).

30. Tesfaye, S. *et al.* Measuring social determinants of health in the All of Us Research Program. *Sci Rep* **14**, 8815 (2024).

31. Park, J. H., Moon, J. H., Kim, H. J., Kong, M. H. & Oh, Y. H. Sedentary Lifestyle: Overview of Updated Evidence of Potential Health Risks. *Korean J Fam Med* **41**, 365–373 (2020).

32. Inoue, K., Tsugawa, Y., Mayeda, E. R. & Ritz, B. Association of Daily Step Patterns With Mortality in US Adults. *JAMA Network Open* **6**, e235174 (2023).

33. Rich, J. T. *et al.* A practical guide to understanding Kaplan-Meier curves. *Otolaryngol Head Neck Surg* **143**, 331–336 (2010).

34. Davidson-Pilon, C. lifelines: survival analysis in Python. *Journal of Open Source Software* **4**, 1317 (2019).

35. Shieh, Y. *et al.* Development and testing of a polygenic risk score for breast cancer aggressiveness. *npj Precis. Onc.* **7**, 1–11 (2023).

36. Tanigawa, Y. *et al.* Significant sparse polygenic risk scores across 813 traits in UK Biobank. *PLOS Genetics* **18**, e1010105 (2022).

37. Dareng, E. O. *et al.* Polygenic risk modeling for prediction of epithelial ovarian cancer risk. *Eur J Hum Genet* **30**, 349–362 (2022).

38. Piekos, J. A. *et al.* Uterine fibroid polygenic risk score (PRS) associates and predicts risk for uterine fibroid. *Hum Genet* **141**, 1739–1748 (2022).

39. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics* **109**, 12–23 (2022).

40. Kloeve-Mogensen, K. *et al.* Polygenic Risk Score Prediction for Endometriosis. *Frontiers in Reproductive Health* **3**, (2021).

41. Wang, Y. *et al.* Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. *Cell Genomics* **3**, (2023).

42. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).

43. Venkatesh, S. S. *et al.* Obesity and risk of female reproductive conditions: A Mendelian randomisation study. *PLOS Medicine* **19**, e1003679 (2022).

44. Hua, Y., Gao, L. & Li, X. Comprehensive Analysis of Metabolic Genes in Breast Cancer Based on Multi-Omics Data. *Pathol Oncol Res* **27**, 1609789 (2021).

45. M, M., Tj, R.-F., A, K. & Rj, S. Genetics of enzymatic dysfunctions in metabolic disorders and cancer. *Frontiers in oncology* **13**, (2023).

46. Ahmad, A. *et al.* Diagnosis and management of uterine fibroids: current trends and future strategies. *Journal of Basic and Clinical Physiology and Pharmacology* **34**, 291–310 (2023).

47. Soliman, A. M., Fuldeore, M. & Snabes, M. C. Factors Associated with Time to Endometriosis Diagnosis in the United States. *Journal of Women's Health* **26**, 788–797 (2017).

48. Hoffman, S. R. *et al.* Optimizing research in symptomatic uterine fibroids with development of a computable phenotype for use with electronic health records. *American Journal of Obstetrics and Gynecology* **218**, 610.e1-610.e7 (2018).

49. GWAS and enrichment analyses of non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE Network | SpringerLink. https://link.springer.com/article/10.1186/s12916-019-1364-z.

50. Rahmioglu, N. *et al.* The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. *Nat Genet* **55**, 423–436 (2023).

51. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102–1111 (2013).

52. Evans, G. W. & Kantrowitz, E. Socioeconomic Status and Health: The Potential Role of Environmental Risk Exposure. *Annual Review of Public Health* **23**, 303–331 (2002).

Social risk factors and cardiovascular risk in obstructive sleep apnea: a systematic assessment of clinical predictors in community health centers[a]

Diego R. Mazzotti[1]; Ryan Urbanowicz[2]; Marta Jankowska[3]

[1]Division of Medical Informatics, Division of Pulmonary Critical Care and Sleep Medicine, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, United States Email: droblesmazzotti@kumc.edu; [2]Department of Computational Biomedicine, Cedars Sinai Medical Center, Los Angeles, CA; Department of Biostatistics Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, United States, Email: Ryan.Urbanowicz@cshs.org; [3]Population Sciences, Beckman Research Institute, City of Hope, Duarte, CA, United States Email: mjankowska@coh.org

We leveraged electronic health record (EHR) data from the Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network (CRN) to identify social risk factor clusters, assess their association with obstructive sleep apnea (OSA), and determine relevant clinical predictors of cardiovascular (CV) outcomes among those experiencing OSA. Geographically informed social indicators were used to define social risk factor clusters via latent class analysis. EHR-wide diagnoses were used as predictors of 5-year incidence of major adverse CV events (MACE) using STREAMLINE, an end-to-end rigorous and interpretable automated machine learning pipeline. Analyses among over 1.4 million individuals revealed three major social risk factor clusters: lowest (35.7%), average (43.6%) and highest (22.7%) social burden. In adjusted analyses, those experiencing highest social burden were less likely to have received a diagnosis of OSA when compared to those experiencing lowest social burden (OR [95%CI]=0.85[0.82-0.88]). Among those with OSA and free of prior CV diseases (N=4,405), performance of predicting incident MACE reached a ROC-AUC of 0.70 [0.03] overall but varied when assessed within each social risk factor cluster. Feature importance also revealed that different clinical factors might explain predictions among each cluster. Results suggest relevant health disparities in the diagnosis of OSA and across clinical predictors of CV diseases among those with OSA, across social risk factor clusters, indicating that tailored interventions geared toward minimizing these disparities are warranted.

Keywords: Health disparities; Social risk factors; sleep disorders; cardiovascular risk; electronic health records.

[a] This research was, in part, funded by the National Institutes of Health (NIH) Agreement NO. 1OT2OD032581-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH. The research reported in this work was powered by PCORnet®. PCORnet has been developed with funding from the Patient-Centered Outcomes Research Institute® (PCORI®) and conducted with the Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network (CRN). ADVANCE is a Clinical Research Network in PCORnet® led by OCHIN in partnership with Health Choice Network, Fenway Health, University of Washington, and Oregon Health & Science University. ADVANCE's participation in PCORnet® is funded through the PCORI Award RI-OCHIN-01-MC.

1. Introduction

Sleep problems disproportionally affect populations experiencing health disparities[1]. Racial, ethnic and socioeconomically disadvantaged minorities are more likely to experience insufficient sleep[2-8], sleep disorders[9], and negative cardiovascular (CV) outcomes[10,11]. Yet, many of these conditions go unnoticed in these populations, largely due to lack of healthcare access focused on diagnosing and treating sleep disorders. Consequently, pathways linking health disparities to sleep disturbances and CV outcomes are largely underexplored, particularly among underrepresented populations.

Obstructive sleep apnea (OSA) is a heterogeneous sleep disordered breathing condition and one of the most prevalent sleep disorders, affecting approximately 1 billion adults worldwide[12]. Epidemiological and experimental evidence supports a major role of OSA towards increasing CV risk[13-16]. However, prior studies were mostly focused on population or community-based cohorts that generally underrepresented important groups known to be at greater risk of experiencing health disparities. The identification of clinical predictors of major adverse CV events (MACE) in these populations is a necessary step towards design tailored and equitable sleep-promoting interventions towards improved CV health.

Efforts supporting the integration and availability of electronic health record (EHR) data linked with relevant social risk information is essential to better characterize the effects of health disparities. Towards that goal, initiatives such as the Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network (CRN) led by the OCHIN network of community health organizations enable such studies[17], with a great potential to inform public health. As such, the current study leveraged data from the ADVANCE CRN and demonstrated an approach to dissect the heterogeneity of geographically informed social risk factors by applying clustering techniques and identifying social risk factor clusters. This data-driven approach supports the identification of population subgroups experiencing similar levels of social exposures and can offer an exploratory perspective on the impact of socio-environmental burden on health. We further assessed the association between social risk factors clusters and evidence of OSA diagnosis. Next, by employing a robust, end-to-end, and interpretable automated machine learning (ML) pipeline, we assessed clinical predictors of 5-year incidence of new onset MACE among individuals with OSA belonging to different clusters. We hypothesized that 1) individuals experiencing higher social burden were less likely to have received a diagnosis of OSA; and 2) clinical predictors of incident MACE varied across social risk clusters, likely reflecting different pathways towards CV risk depending on socio-environmental exposure.

2. Methods

2.1. *Study Design and Population*

This is a retrospective clinical cohort study of patients at risk for sleep disorders that were part of the ADVANCE CRN with available geographically informed social risk factor data ascertained between 2012 and 2021. Data was sourced from the OCHIN Epic EHR system. Data is representative of outpatient community-based health care organizations delivering high-quality primary care services for communities impacted by health disparities in the U.S. Clinical institutions

include Federally Qualified Health Centers or other federally supported community health centers. The ADVANCE CRN is part of PCORnet®, the National Patient-Centered Clinical Research Network, thus data is organized according to the PCORnet® Common Data Model. Access was requested and facilitated by the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) program. The study has been approved by Institutional Review Boards from the University of Kansas Medical Center and Harvard Medical School with non-human subjects determination, as only de-identified data was made available.

Out of a dataset of over 3.2 million adults (age ≥18 years), we identified a cohort with at least one year of interactions with community health centers, a minimum of 3 encounters, and non-missing geographically informed social risk factors. Among those patients, we further created a subset of those with evidence of OSA and at least 5 years of interaction with the community health centers and without prior evidence of CV diseases to determine clinical predictors of incident MACE.

2.2. *Geographically informed social risk factors*

Geographically linked neighborhood-level indicators at census tract and/or ZCTA levels[18] were made available through OCHIN as part of the ADVANCE CRN data warehouse. Linkage was performed by matching participant's address ZIP code with publicly available data sources from the U.S. Census Bureau and American Community Survey, and used to impute the following area-level social indicators: income inequality coefficient, or Gini coefficient, a measure ranging from 0 (perfectly equal geographical region where all income is equally shared) and 1 (perfectly unequal society where all income is earned by 1 individual)[19,20]; median household income (in U.S. dollars); percent of adults age >25 years who graduated from college; percent of total population in poverty (<100% federal poverty level [FPL]); and rate of unemployment among population age ≥16 years. These indicators were categorized into quartiles prior to downstream analyses.

2.3. *Computable phenotypes for OSA*

A validated EHR algorithm was used to identify individuals with evidence of OSA, as described by Keenan et al. 2020[21]. Individuals with 2 or more International Classification of Diseases (ICD)-9 or 10 codes for OSA at different dates were classified as having OSA (ICD-9: 327.20, 327.23, 327.29, 780.51, 780.53, 780.57; ICD-10: G47.30, G47.33, G47.39). This algorithm presented optimal predictive performance across six health systems in the U.S., with overall positive predictive value (95% CI) of 97.1% (95.6, 98.2) and negative predictive value of 95.5% (93.5, 97.0)[21]. Individuals not meeting these criteria were defined as not having evidence of OSA diagnosis.

2.4. *Phecode mapping*

The phecode framework[22] is a high-throughput EHR phenotyping method with the goal of representing a wide range of clinical phenotypes. Structured as an ontology-based classification system, phecodes combine groups of ICD codes into clinically relevant groups, thus minimizing the dimensionality of clinical diagnosis. In this study, we focused on phecodes observed in at least 1,000

participants in our final cohort, resulting in a total of 932 phecodes included as predictors in our ML analyses. Phecode maps can be queried elsewhere (https://www.phewascatalog.org/phecodes).

2.5. *Study outcomes*

We report the results of two analyses. Our primary analysis consisted of investigating the association between social risk factors clusters and evidence of OSA. Thus, our outcome was prevalence of OSA. Secondly, we assessed clinical predictors of 5-year incidence of MACE, defined as a composite of myocardial infarction, coronary artery disease, cerebrovascular disease, heart failure or stroke, using validated computable phenotypes as previously described[23-29]. A list of ICD and Current Procedural Terminology codes used to define these conditions are available elsewhere (https://raw.githubusercontent.com/RWD2E/phecdm/main/res/valueset_curated/vs-osa-comorb.json).

2.6. *Statistical analyses*

All analyses were conducted within the AIM-AHEAD Service Workbench cloud infrastructure. Initial cohort characterization was performed through a data request with OCHIN. A database schema was created in Microsoft SQL Server and access was provided to the author. A series of tables resulting from this database schema were generated to capture the following data domains: patient demographics, social risk factors, diagnosis, and procedures. Queries used to create analysis-ready can be found elsewhere (https://github.com/mazzottidr/AIMAHEAD_Fellowship_Mazzotti).

First, we determined univariate associations between OSA and sociodemographic characteristics (sex, race, ethnicity, gender identity, current FPL, marital status, homeless status, and sexual orientation), as well as between OSA and quartiles of geographically informed social risk factors (Gini coefficient, median household income, percent of college graduates; percent of total population in poverty; and rate of unemployment) using chi-squared tests or t tests. Next, we used latent class analysis (LCA) to identify clusters of social risk factors using quartiles of the geographically informed social risk factors listed above. Due to the large computational requirements of performing LCA on large datasets, we assessed the optimal number of clusters by sub-setting the data into 10 random subsamples of N=5,000 participants and performing LCA using 1 through 5 clusters. We used the Bayesian Information Criterion and the elbow method to determine the optimal number of clusters. Based on these analyses, we determined that a 3-cluster solution was as the optimal in all 10 iterations. We further re-ran LCA in the complete dataset using only this solution, setting the maximum number of iterations through each estimation algorithm (maxiter) as 1,000 and the number of times to estimate the model with different class-conditional response probabilities (nrep) as 25, with default parameters otherwise. We used the poLCA package in R[30]. Cross-sectional associations between social risk factor clusters and OSA were assessed using chi-squared test and unadjusted and adjusted logistic regression. Covariates included age, sex, language, race, marital status, ethnicity, and urban/rural status.

We proceeded to determine whether different social risk factor clusters would prioritize different clinical risk factors towards predicting MACE risk among a cohort of individuals with evidence of OSA. For this analysis, we included only participants with evidence of OSA, at least 5 years of

follow-up data, to allow for ascertainment of MACE incidence. Phecode feature sets were used as predictors of 5-year incidence of MACE (binary outcome) using STREAMLINE, an end-to-end rigorous and interpretable auto-ML pipeline (https://github.com/UrbsLab/STREAMLINE)[31,32], which has been implemented in a SageMaker instance of the AIM-AHEAD Service Workbench. Data were split into training/testing (90%) and validation (10%), maintaining proportions for both the outcome and social risk factor clusters. For each cluster, we optimized four different ML methods (logistic regression [LR], random forest [RF], Light Gradient Boost Machine [LightGBM], and Extreme Gradient Boosting [XGB]), as well as evaluated models with area under the receiver operating characteristics curve (ROC-AUC) and area under the precision-recall curve (PRC-AUC) using a 3-fold cross-validation design. Feature importance scores were determined, along with social risk factor cluster-specific final models for independent validation. The top performing features in each subgroup were then selected and compared across clusters. Analyses were conducted using R (v 4.1.3) and Python (v 3.10.8).

3. Results

3.1. *Sample characterization*

In our initial analysis focused on assessing the association between social risk factors and prevalence of sleep disorders, our primary cohort consisted of 1,476,358 adults with encounters in community health centers across the U.S. **Figure 1** represent the study flowchart.

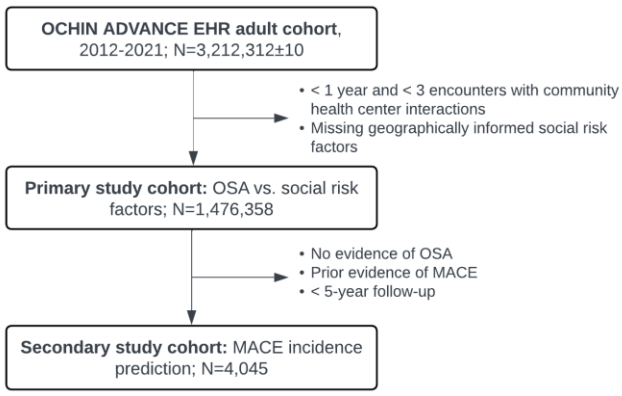


Figure 1. Study flowchart representing sample sizes for each included study cohort.

Among those, 63.2% were female, 69.9% spoke English as the primary language, 67.4% were White, 20.9% were Black, 5.1% were Asian, 67.8% had a current FPL <100%, 3.3% reported being homeless and 16.5% lived in rural areas. These characteristics highlight the sociodemographic diversity of the included cohort. **Table 1** provides descriptive statistics of the overall sample, as well as by evidence of OSA status. Individuals with evidence of an OSA diagnosis represented 2.3% of the included cohort (N=33,064), and univariate analyses suggest they were older, more likely to be males and with male gender identity, more likely to speak English as primary language, more likely to be White, less likely among those who were single, less likely among those with current FPL

<100%, less likely to be Hispanic or Latino, less likely among those reporting homelessness, more likely among those reporting heterosexual orientation, and more likely among those living in rural areas. Geographically informed social risk factors mostly differ between those with and without evidence of OSA, suggesting that those with a diagnosis are more likely to live in areas with lower social risk (**Table 1**).

Table 1. Sample characteristics, overall and stratified by evidence of obstructive sleep apnea (OSA).

Variable	Category	Overall (N=1,476,358)	Evidence of OSA		p[a]
			No (N=1,443,294)	Yes (N=33,064)	
Age, years		44.3 (15.9)	44.1 (15.9)	52.1 (12.8)	<0.001
Sex	Female	932,903 (63.2)	917,179 (63.6)	15,724 (47.6)	<0.001
	Male	543,036 (36.8)	525,703 (36.4)	17,333 (52.4)	
Primary language	English	1,031,528 (69.9)	1,002,947 (69.5)	28,581 (86.4)	<0.001
	Spanish	354,107 (24.0)	350,674 (24.3)	3433 (10.4)	
	Other	90,723 (6.1)	89,673 (6.2)	1,050 (3.2)	
Race	White	993,964 (67.4)	969,574 (67.3)	24,390 (73.9)	<0.001
	American Indian or Alaska Native	12,696 (0.9)	12,335 (0.9)	361 (1.1)	
	Asian	75,552 (5.1)	74,646 (5.2)	906 (2.7)	
	Black or African American	307,628 (20.9)	301,653 (20.9)	5,975 (18.1)	
	Multiple Race	20,259 (1.4)	19,774 (1.4)	485 (1.5)	
	Native Hawaiian or Other Pacific Islander	9,899 (0.7)	9,622 (0.7)	277 (0.8)	
	Refuse to answer	54,396 (3.7)	53,771 (3.7)	625 (1.9)	
Marital status	Current Partnership	363,381 (24.6)	355,745 (24.6)	7,636 (23.1)	<0.001
	Divorced/Separated	86,176 (5.8)	83,422 (5.8)	2,754 (8.3)	
	Single	496,312 (33.6)	488,703 (33.9)	7,609 (23.0)	
	Unknown	497,609 (33.7)	483,544 (33.5)	14,065 (42.5)	
	Widowed	32,880 (2.2)	31,880 (2.2)	1,000 (3.0)	
Current FPL	101-150 %	217,701 (14.7)	212,929 (14.8)	4,772 (14.4)	<0.001
	≤100 %	1,000,448 (67.8)	979,666 (67.9)	20,782 (62.9)	
	151-200 %	94,042 (6.4)	91,871 (6.4)	2,171 (6.6)	
	>200 %	164,167 (11.1)	158,828 (11.0)	5,339 (16.1)	
Ethnicity	Not Hispanic or Latino	912,928 (62.8)	886,753 (62.4)	26,175 (80.5)	<0.001
	Hispanic or Latino	540,075 (37.2)	533,727 (37.6)	6,348 (19.5)	
Gender identity	Female	535,485 (36.3)	523,697 (36.3)	11,788 (35.7)	<0.001
	Male	316,468 (21.4)	304,107 (21.1)	12,361 (37.4)	
	Transgender, Gender Queer, Other	19,768 (1.3)	19,591 (1.4)	177 (0.5)	
	Unknown	604,637 (41.0)	595,899 (41.3)	8,738 (26.4)	
Homelessness status	No/Unknown	1,427,117 (96.7)	1,394,342 (96.6)	32,775 (99.1)	<0.001
	Yes	49,241 (3.3)	48,952 (3.4)	289 (0.9)	
Sexual orientation	Heterosexual	731,141 (49.5)	710,476 (49.2)	20,665 (62.5)	<0.001
	Homosexual	24,422 (1.7)	23,584 (1.6)	838 (2.5)	
	Bisexual	15,222 (1.0)	14,837 (1.0)	385 (1.2)	
	Other	6,006 (0.4)	5,856 (0.4)	150 (0.5)	
	Unknown	699,567 (47.4)	688,541 (47.7)	11,026 (33.3)	
Rural/urban status	Urban	1,233,127 (83.5)	1,209,391 (83.8)	23,736 (71.8)	<0.001
	Rural	243,231 (16.5)	233,903 (16.2)	9,328 (28.2)	
Geographically informed indicators					
Unemployment rate, %		7.33% (3.41)	7.33% (3.42)	7.14% (3.29)	<0.001
Median household income, U.S. dollars		$53,994 (20,242)	$53,968.92 (20,269)	$55,086 (19,020)	<0.001
% of college graduates		26.91% (15.06)	26.91% (15.07)	26.86% (14.63)	0.518
Gini coefficient		0.45 (0.05)	0.45 (0.05)	0.45 (0.05)	<0.001
% of population below FPL		18.89 (9.37)	18.92 (9.38)	17.68 (8.68)	<0.001

[a] Chi-squared tests or t-tests. Categorical variables are represented as N (%) and continuous variable as mean (SD).
Abbreviations: OSA, obstructive sleep apnea; FPL, Federal Poverty Level; SD: standard deviation.

3.2. *Clusters of social risk factors*

Results of LCA revealed three major social risk factor clusters: lowest (N=489,191; 35.7%), average (N= 642,973; 43.6%) and highest (N=335,194; 22.7%) social burden. **Table 2** describes the differences between each geographically informed social risk factor and the 3-cluster solution used to inform the names of each cluster. The highest social burden cluster had the greatest proportion of the highest quartiles of unemployment rates, Gini coefficient, and proportion of individuals living below poverty level, and the lowest quartiles of median household income and proportion of individuals that are college graduates.

Table 2. Association between social risk factor quartiles and identified social risk clusters.

Social risk factor quartiles	Category	Lowest Social Burden (35.7%)	Average Social Burden (43.6%)	Highest Social Burden (22.7%)	p[a]
Unemployment rate	Q1 [<5.1%]	299,777 (60.2)	61,207 (9.5)	8,503 (2.5)	<0.001
	Q2 [5.1-6.6%]	139,264 (28.0)	202,799 (31.5)	23,253 (6.9)	
	Q3 [6.6-8.9%]	51,053 (10.2)	267,651 (41.6)	49,361 (14.7)	
	Q4 [≥8.9%]	8,097 (1.6)	111,316 (17.3)	254,077 (75.8)	
Median household income	Q1 [<40.7k]	424 (0.1)	56,959 (8.9)	313,600 (93.6)	<0.001
	Q2 [40.7-50.0k]	1,856 (0.4)	343,123 (53.4)	21,594 (6.4)	
	Q3 [50.0-63.8k]	135,449 (27.2)	230,266 (35.8)	<11	
	Q4 [≥63.8k]	360,462 (72.4)	12,625 (2.0)	<11	
% of college graduates	Q1 [<16.7%]	17,317 (3.5)	144,164 (22.4)	205,989 (61.5)	<0.001
	Q2 [16.7-23.1%]	46,037 (9.2)	247,449 (38.5)	74,078 (22.1)	
	Q3 [23.1-33.5%]	146,377 (29.4)	179,683 (27.9)	39,874 (11.9)	
	Q4 [≥33.5%]	288,460 (57.9)	71,677 (11.1)	15,253 (4.6)	
Gini coefficient	Q1 [<0.42]	204,591 (41.1)	140,407 (21.8)	20,896 (6.2)	<0.001
	Q2 [0.42-0.45]	104,607 (21.0)	199,895 (31.1)	63,231 (18.9)	
	Q3 [0.45-0.48]	93,785 (18.8)	169,999 (26.4)	104,347 (31.1)	
	Q4 [≥0.48]	95,208 (19.1)	132,672 (20.6)	146,720 (43.8)	
% below poverty level	Q1 [<12.1%]	355,133 (71.3)	1,4870 (2.3)	1,067 (0.3)	<0.001
	Q2 [12.1-17.6%]	130,118 (26.1)	236,982 (36.9)	<11	
	Q3 [17.6-23.7%]	4,322 (0.9)	349,902 (54.4)	14,214 (4.2)	
	Q4 [≥23.7%]	8,618 (1.7)	41,219 (6.4)	319,913 (95.4)	

[a] Chi-squared tests. Categorical variables are represented as N (%).

3.3. *Associations between social risk factors clusters and OSA*

We proceeded to determine the association between social risk factors clusters and evidence of OSA. Univariate analysis indicated that individuals with evidence of OSA were less likely to belong to the highest social burden cluster (16.8%) when compared to those without evidence of OSA (22.8%, p<0.001). On the other hand, those with evidence of OSA were more likely to belong to both the lowest and average social burden clusters when compared to those without evidence of OSA (34.9% vs. 33.7% and 48.3% vs. 43.4%, respectively, both p<0.001). Logistic regression adjusted for relevant confounders, including individual level social risk factors, indicated that individuals belonging to the lowest social burden cluster were less likely to have received a diagnosis of OSA (OR [95%CI] = 0.85 [0.82-0.88]) when compared to those belonging to the highest social burden cluster. On the other hand, individuals belonging to the average social risk burden were slightly more likely to have received a diagnosis of OSA (1.03 [1.01-1.06]) compared to those in the highest

social burden cluster. Results suggest important socio-environmental contributions to potential disparities in the diagnosis of OSA in community health centers.

3.4. *MACE prediction among individuals with OSA*

Next, we proceeded to understand the clinical factors contributing to increased CV risk among individuals with OSA in the included sample, without taking into consideration their social risk cluster. A cohort of 4,045 individuals with OSA, without prior evidence of MACE and with at least 5 years of follow-up since their first OSA diagnosis was included in this analysis. Among those, 327 (8.1%) individuals had evidence of a MACE within the 5-year follow-up.

Using a robust ML pipeline, we proceeded to create our training (90%) and testing (10%) sets, maintaining the proportions of incident MACE cases and social risk clusters. Our training dataset consisted of 3,641 individuals (294 [8.1%] cases) and our testing dataset consisted of 404 individuals (33 [8.2%] cases). We determined these training/testing splits to allow greater representation of the dataset during training, due to the limited sample size of the cohort.

First, we assessed the performance of clinical risk factors (represented as phecodes) to predict incident MACE in the training and testing datasets, regardless of social risk cluster membership, using four different ML methods (LR, RF, LightGBM, and XGB). **Figure 2** summarizes the prediction performances in terms of ROC-AUC and PRC-AUC across the four methods. While XGB demonstrated the best performance in the training dataset for both performance metrics (mean [SD across cross-validation] ROC-AUC = 0.67 [0.03]; PRC-AUC = 0.14 [0.02]), LR was the best performing method in the testing dataset (ROC-AUC = 0.70 [0.03]; PRC-AUC = 0.19 [0.02]).

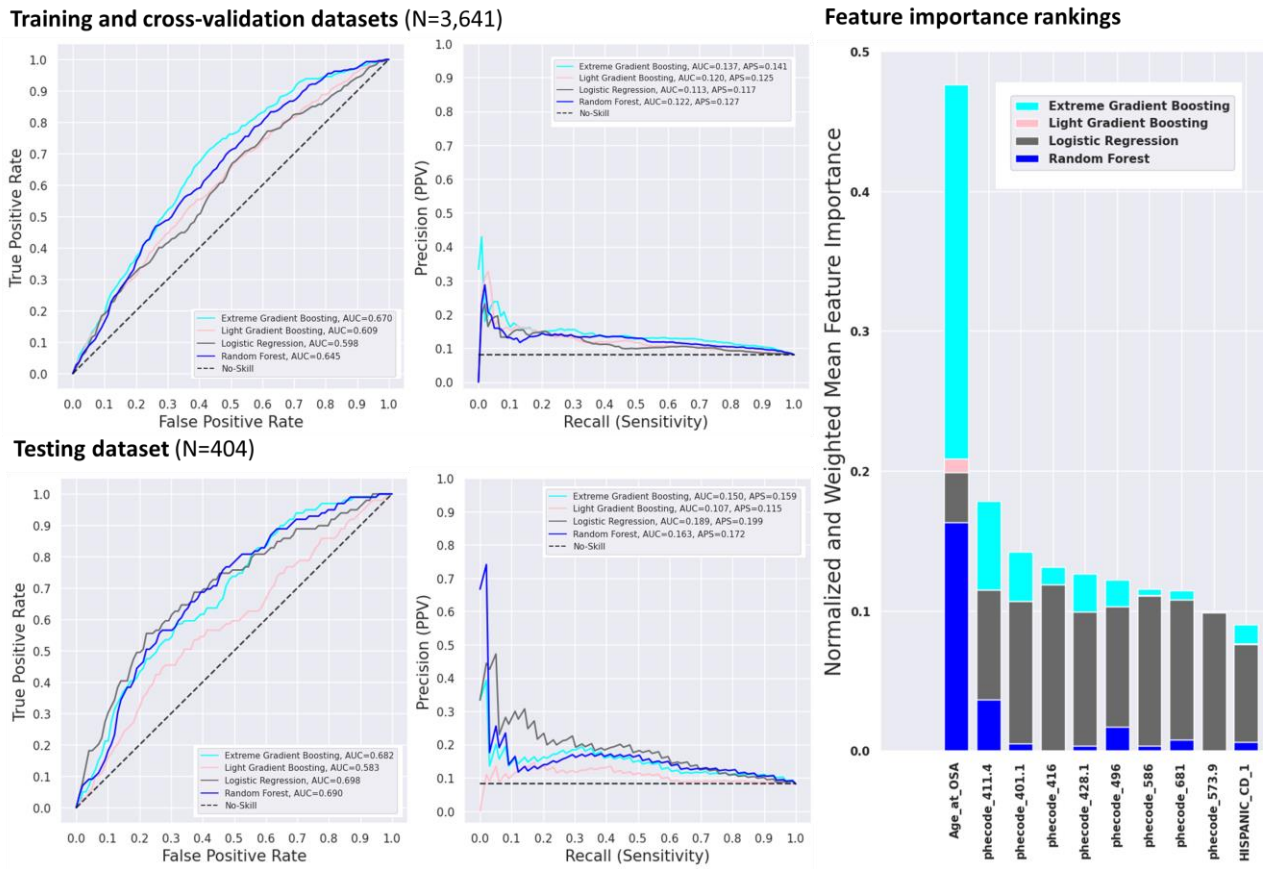


Figure 2. Summary of incident MACE prediction performance and feature importance (top 10 features).

Inspection of normalized and balanced accuracy-weighted feature importance plots (**Figure 2**) indicates that age was the most important predictor across all methods, except for LR. For this method, the most important feature was phecode 416 (cardiomegaly). Other relevant features listed among the top 10 included phecodes 411.4 (coronary atherosclerosis), 401.1 (essential hypertension), 428.1 (congestive heart failure), 496 (chronic airway obstruction), 586 (other disorders of the kidney and ureters), 681 (superficial cellulitis and abscess), 573.9 (abnormal serum enzyme levels) and ethnicity.

3.5. *MACE prediction after social risk factor cluster stratification*

Finally, we proceeded to explore how these models would perform within specific subgroups according to the assigned social risk factor clusters, and whether top clinical predictors would be similar or different across clusters. For this analysis we trained and evaluated ML models using the same methods described above, but within each social risk factor cluster. Training and testing dataset sample sizes for each cluster were as follows: lowest social burden cluster (N_{train}=1,136; N_{test} = 126), average social burden cluster (N_{train} =1,791; N_{test} = 199), and highest social burden cluster (N_{train} = 467; N_{test} = 52).

Table 3 summarizes the results of the predictive performance in the testing dataset from models trained and evaluated within each social risk factor cluster separately. According to the ROC-AUC, within participants assigned to the lowest social burden clusters, LR was the best performing method, while RF performed the best in both the average and highest social burden clusters. According to the PRC-AUC, within participants assigned to the lowest social burden clusters, LR was also the best performing method, while XGB performed the best in both the average and highest social burden clusters.

Table 3. Summary of prediction performance metrics in the testing datasets using models trained within social risk factor clusters.

Method	Metric	Cluster		
		Lowest social burden	Average social burden	Highest social burden
XGB	ROC-AUC	0.500	0.617	0.564
	PRC-AUC	0.117	0.133	0.189
LightGBM	ROC-AUC	0.616	0.606	0.504
	PRC-AUC	0.126	0.098	0.177
LR	ROC-AUC	0.689	0.631	0.522
	PRC-AUC	0.213	0.114	0.103
RF	ROC-AUC	0.634	0.634	0.628
	PRC-AUC	0.163	0.127	0.141

Abbreviations: XGB, Extreme Gradient Boosting; LightGBM, Light Gradient Boost Machine; LR, logistic regression; RF, random forest; ROC-AUC, area under the receiver operating characteristics curve; ROC-PRC, area under the precision-recall curve.

We then inspected differences in the normalized and balanced accuracy-weighted feature importance plots (**Figure 3**) across the models and social risk factor clusters to investigate whether clinical risk factors that predict incident MACE would be different depending on individuals' socio-economic exposures. Results suggest that while age at diagnosis of OSA was an important predictor across all social risk factor clusters, being listed among the top 10 features in all groups, there were important differences in the comorbidity profile linked to incident MACE within each group. For example, among those with lowest social burden, some more conventional CV comorbidities or risk factors were observed, such as essential hypertension (401.1), nonspecific chest pain (418) and both type 1 and 2 diabetes (250.1 and 250.2). However, among those with highest social burden, top predictors included symptoms such as malaise and fatigue (798), pain in joint (745), and dizziness and giddiness (light-headedness and vertigo, 386.9), in addition to a more metabolic comorbidity profile (244.4, hypothyroidism and 250.2, type 2 diabetes). Among those with average social burden, features included both conventional ones (401.1, essential hypertension and 416, cardiomegaly) as well as other infectious and parasitic diseases (136) and Lyme disease (130.1). Anxiety disorder (300.1) was also observed as an important predictor among those with lowest and average social burden.

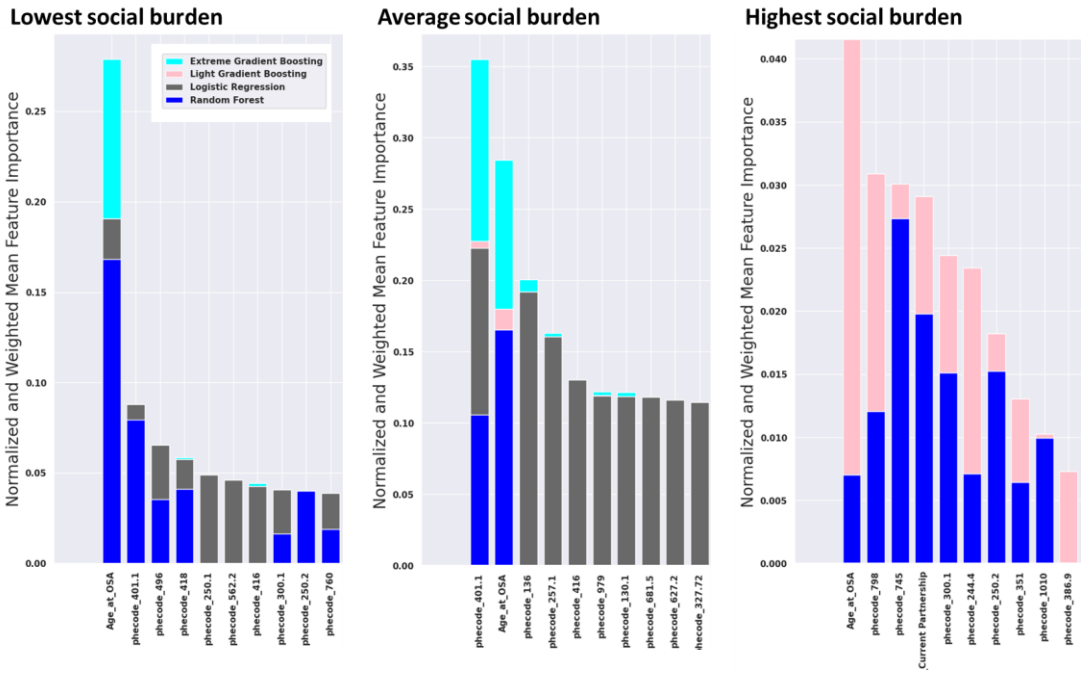


Figure 3. Top feature important comparison across models evaluated within different social risk factor clusters.

4. Discussion

Our main findings highlight important social disparities related to the identification and diagnosis of OSA in community health centers, as well as important differences in clinical factors that contributed to the prediction of incident CV diseases among participants with a diagnosis of OSA. We applied an innovative approach to identify social risk factor clusters derived from relevant geographically informed social indicators estimated from national surveys. We identified three clusters (lowest, average, and highest social burden), consistent with observed individual-level sociodemographic characteristics. Individuals belonging to the highest social burden cluster were less likely to have received a diagnosis of OSA, even after adjusting for relevant confounders such as sex, race, and ethnicity – factors that have been consistently demonstrated to affect health disparities within sleep disorders[1-8]. Our study also demonstrated that a LR-based incident MACE prediction model trained on hundreds of clinical features (i.e., phecodes) had a reasonable, yet not optimal performance in testing sets. Nevertheless, performance varied across subgroups defined by social risk factor clusters, as well as the top features contributing to those predictions, suggesting different pathways towards CV risk depending on socio-environmental exposure.

The study provides novel insights about the clinical prevalence and recognition of OSA within community health centers in a diverse population at greater social burden. Our dataset was composed of a large proportion of underrepresented minorities according to sociodemographic characteristics, including race, ethnicity, gender and sexual identity. More importantly, 67.8% of the cohort were below the Federal poverty level. In this context, the observed clinical prevalence of OSA (2.3%) is lower than other clinical cohorts defined using EHR-based methods, such as within the National COVID Cohort Collaborative (3.9%), comprised of a sample of individuals that have

been tested positively for SARS-CoV-2 through encounters within academic health systems[33]. The prevalence is even lower than the expected population prevalence of OSA, estimated to affect nearly 1 billion people worldwide[12]. It is well-established that OSA is underdiagnosed[34,35], and our analysis in community health centers identified even further differences. Clustering of geographically informed social indicators revealed that individuals at greater social burden (i.e., highest social burden cluster) were significantly less likely to have received a diagnosis of OSA, even after adjusting for other established individual-level social risk factors. When assessing individual-level sociodemographic characteristics, those receiving a diagnosis of OSA were more likely to speak English as primary language, more likely to be White, less likely to be among those with current FPL <100%, less likely to be Hispanic or Latino, less likely to report homelessness, although more likely to live in rural areas. These findings suggest important socio-environmental contributions to potential disparities in the diagnosis of OSA in community health centers and that underrepresented minorities may not be receiving adequate sleep care. Thus, screening of sleep disorders particularly within this subgroup at greater risk is necessary. While it might seem impractical to offer screening and treatment of chronic sleep disorders such as OSA in community health centers, preventing high risk individuals from obtaining access to quality sleep health care might exacerbate disparities related to metabolic, neurological, and psychiatric conditions, all of which have been associated with OSA[36].

In this context, CV diseases are particularly relevant due to the worsening of CV disparities over several decades, despite efforts of addressing health needs of vulnerable populations[37]. Due to the major epidemiological and experimental evidence supporting the role of OSA towards increasing CV risk[13-16], understanding and addressing these needs are of high importance. Towards this goal, we assessed whether an incident MACE prediction model trained on a broad range of clinical features within individuals with OSA had adequate performance and could be used to prioritize clinical profiles based on most relevant features. Despite our best model, a LR with a ROC-AUC of 0.70 and a PRC-AUC of 0.19, not being necessarily optimal for deployment, it helped us identify important features contributing to the prediction, many of them with established associations with OSA. For example, in our overall analysis, top features included cardiomegaly, atherosclerosis, essential hypertension, congestive heart failure, chronic airway obstruction, disorders of the kidney and ureters, and abnormal serum enzyme levels. Many of these features are established CV risk factors, supporting internal validity of our approach. More importantly, therapies focused on mitigating the effects of OSA have been demonstrated to improve some of these risk factors[38].

When assessing the prediction performance of models across strata of social risk factor clusters, we continued to identify similar, although slightly lower performance across groups with testing ROC-AUC ranging from 0.63 to 0.69 and PRC-AUC ranging from 0.13 to 0.21 for the best models. This is likely explained by the smaller sample size used for training in the stratified analyses, preventing models from learning relationships between clinical features and outcome. Some key clinical factors contributing to these predictions are observed across all social risk factor clusters, such as age at OSA diagnosis, cardiometabolic conditions (e.g., type 2 diabetes, hypertension), and anxiety disorders. However, among those with highest social burden, top predictors included symptom-related factors such as malaise and fatigue, pain in joint, and light-headedness and vertigo, while among those with average social burden, features included infectious and parasitic diseases. These presentations might reflect primary reasons or exposure to different healthcare specialists. In this context, the study provided a systematic data driven approach to identify these factors, where

future studies could further explore, under a more hypothesis-driven methodology whether these conditions could be suggestive of higher CV risk within vulnerable populations.

Our study has several strengths, such as providing an analysis in a large, racial, ethnic, and socioeconomically diverse clinical cohort of individuals observed in community health centers, a target population often neglected from epidemiological and experimental studies. In addition, we use a robust ML pipeline comparing, in a systematic way, different sets of ML methods and features towards understanding clinical factors of incident CV diseases. However, our study also present important limitations that should be considered when interpreting the findings. Access to OSA therapies, such as continuous positive airway pressure or mandibular advancement devices are likely not offered by this care modality and therefore not necessarily recorded in the ADVANCE EHR data warehouse, thus they could not be considered as confounders. More granular information about severity of OSA based on the apnea-hypopnea index or other metrics was not available, as it required parsing of clinical sleep study reports. Similarly, phecodes are not necessarily always precise, granular measures of diagnoses and may lack sensitivity and specificity of validated computable phenotypes. However, as part of a data-driven EHR-wide analysis, they may offer an initial set of hypotheses that could be assessed with more robust phenotypes in future investigations. Despite our observed signals, incidence rates of MACE were relatively low, possibly due to the relative short, 5-year follow-up time, resulting in a very imbalanced classification problem. However, longer follow-up windows would substantially reduce sample size and was not a feasible alternative.

In conclusion, this study leveraged heterogeneous EHR data from community health centers in the United States and described sociodemographic and geographically informed social disparities as they relate to diagnosis of OSA. Prediction models of incident MACE among individuals experiencing OSA also disparities in across clinical predictors of CV diseases. Thus, tailored interventions geared toward minimizing these disparities are warranted.

5. Acknowledgments

The research reported in this work was powered by PCORnet®. PCORnet has been developed with funding from the Patient-Centered Outcomes Research Institute® (PCORI®) and conducted with the Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network (CRN). ADVANCE is a Clinical Research Network in PCORnet® led by OCHIN in partnership with Health Choice Network, Fenway Health, University of Washington, and Oregon Health & Science University. ADVANCE's participation in PCORnet® is funded through the PCORI Award RI-OCHIN-01-MC.

References

1.Jackson CL, Walker JR, Brown MK, Das R, Jones NL. A workshop report on the causes and consequences of sleep health disparities. *Sleep*. Aug 12 2020;43(8)doi:10.1093/sleep/zsaa037
2.Carnethon MR, De Chavez PJ, Zee PC, et al. Disparities in sleep characteristics by race/ethnicity in a population-based sample: Chicago Area Sleep Study. *Sleep Med*. Feb 2016;18:50-5. doi:10.1016/j.sleep.2015.07.005
3.Jackson CL, Patel SR, Jackson WB, 2nd, Lutsey PL, Redline S. Agreement between self-reported and objectively measured sleep duration among white, black, Hispanic, and Chinese adults in the United States: Multi-Ethnic Study of Atherosclerosis. *Sleep*. Jun 1 2018;41(6)doi:10.1093/sleep/zsy057
4.Johnson DA, Lisabeth L, Hickson D, et al. The Social Patterning of Sleep in African Americans: Associations of Socioeconomic Position and Neighborhood Characteristics with Sleep in the Jackson Heart Study. *Sleep*. Sep 1 2016;39(9):1749-59. doi:10.5665/sleep.6106
5.Kaufmann CN, Mojtabai R, Hock RS, et al. Racial/Ethnic Differences in Insomnia Trajectories Among U.S. Older Adults. *Am J Geriatr Psychiatry*. Jul 2016;24(7):575-84. doi:10.1016/j.jagp.2016.02.049
6.Liu Y, Wheaton AG, Chapman DP, Cunningham TJ, Lu H, Croft JB. Prevalence of Healthy Sleep Duration among Adults--United States, 2014. *MMWR Morb Mortal Wkly Rep*. Feb 19 2016;65(6):137-41. doi:10.15585/mmwr.mm6506a1
7.Petrov ME, Lichstein KL. Differences in sleep between black and white adults: an update and future directions. *Sleep Med*. Feb 2016;18:74-81. doi:10.1016/j.sleep.2015.01.011
8.Roane BM, Johnson L, Edwards M, Hall J, Al-Farra S, O'Bryant SE. The link between sleep disturbance and depression among Mexican Americans: a Project FRONTIER study. *J Clin Sleep Med*. Apr 15 2014;10(4):427-31. doi:10.5664/jcsm.3622
9.Dudley KA, Patel SR. Disparities and genetic risk factors in obstructive sleep apnea. *Sleep Med*. Feb 2016;18:96-102. doi:10.1016/j.sleep.2015.01.015
10.Seixas AA, Trinh-Shevrin C, Ravenell J, Ogedegbe G, Zizi F, Jean-Louis G. Culturally tailored, peer-based sleep health education and social support to increase obstructive sleep apnea assessment and treatment adherence among a community sample of blacks: study protocol for a randomized controlled trial. *Trials*. Sep 24 2018;19(1):519. doi:10.1186/s13063-018-2835-9
11.Jean-Louis G, Newsome V, Williams NJ, Zizi F, Ravenell J, Ogedegbe G. Tailored Behavioral Intervention Among Blacks With Metabolic Syndrome and Sleep Apnea: Results of the MetSO Trial. *Sleep*. Jan 1 2017;40(1)doi:10.1093/sleep/zsw008
12.Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. Aug 2019;7(8):687-698. doi:10.1016/S2213-2600(19)30198-5
13.Drager LF, Togeiro SM, Polotsky VY, Lorenzi-Filho G. Obstructive Sleep Apnea. *Journal of the American College of Cardiology*. 2013;62(7):569-576. doi:10.1016/j.jacc.2013.05.045
14.Mazzotti DR, Keenan BT, Lim DC, Gottlieb DJ, Kim J, Pack AI. Symptom Subtypes of Obstructive Sleep Apnea Predict Incidence of Cardiovascular Outcomes. *Am J Respir Crit Care Med*. Aug 15 2019;200(4):493-506. doi:10.1164/rccm.201808-1509OC
15.Azarbarzin A, Sands SA, Stone KL, et al. The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the Osteoporotic Fractures in Men Study and the Sleep Heart Health Study. *European Heart Journal*. 2019-04-07 2019;40(14):1149-1157. doi:10.1093/eurheartj/ehy624

16.Azarbarzin A, Sands SA, Younes M, et al. The Sleep Apnea–Specific Pulse-Rate Response Predicts Cardiovascular Morbidity and Mortality. *American Journal of Respiratory and Critical Care Medicine*. 2021;203(12):1546-1555. doi:10.1164/rccm.202010-3900OC
17.DeVoe JE, Gold R, Cottrell E, et al. The ADVANCE network: accelerating data value across a national community health center network. *J Am Med Inform Assoc*. Jul-Aug 2014;21(4):591-5. doi:10.1136/amiajnl-2014-002744
18.Hughes LS, Phillips RL, DeVoe JE, Bazemore AW. Community Vital Signs: Taking the Pulse of the Community While Caring for Patients. *The Journal of the American Board of Family Medicine*. 2016;29(3):419-422. doi:10.3122/jabfm.2016.03.150172
19.Tan AX, Hinman JA, Abdel Magid HS, Nelson LM, Odden MC. Association Between Income Inequality and County-Level COVID-19 Cases and Deaths in the US. *JAMA Network Open*. 2021;4(5)doi:10.1001/jamanetworkopen.2021.8799
20.De Maio FG. Income inequality measures. *Journal of Epidemiology & Community Health*. 2007;61(10):849-852. doi:10.1136/jech.2006.052969
21.Keenan BT, Kirchner HL, Veatch OJ, et al. Multisite validation of a simple electronic health record algorithm for identifying diagnosed obstructive sleep apnea. *J Clin Sleep Med*. Feb 15 2020;16(2):175-183. doi:10.5664/jcsm.8160
22.Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annual Review of Biomedical Data Science*. 2021;4(1):1-19. doi:10.1146/annurev-biodatasci-122320-112352
23.Singer DE, Chang Y, Borowsky LH, et al. A New Risk Scheme to Predict Ischemic Stroke and Other Thromboembolism in Atrial Fibrillation: The ATRIA Study Stroke Risk Score. *Journal of the American Heart Association*. 2013;2(3)doi:10.1161/jaha.113.000250
24.Go AS, Hylek EM, Chang Y, et al. Anticoagulation Therapy for Stroke Prevention in Atrial Fibrillation. *Jama*. 2003;290(20)doi:10.1001/jama.290.20.2685
25.Sidney S, Sorel M, Quesenberry CP, DeLuise C, Lanes S, Eisner MD. COPD and Incident Cardiovascular Disease Hospitalizations and Mortality: Kaiser Permanente Medical Care Program. *Chest*. 2005;128(4):2068-2075. doi:10.1378/chest.128.4.2068
26.Go AS, Yang J, Ackerson LM, et al. Hemoglobin Level, Chronic Kidney Disease, and the Risks of Death and Hospitalization in Adults With Chronic Heart Failure. *Circulation*. 2006;113(23):2713-2723. doi:10.1161/circulationaha.105.577577
27.Gurwitz JH, Magid DJ, Smith DH, et al. Contemporary Prevalence and Correlates of Incident Heart Failure with Preserved Ejection Fraction. *The American Journal of Medicine*. 2013;126(5):393-400. doi:10.1016/j.amjmed.2012.10.022
28.McKee PA, Castelli WP, McNamara PM, Kannel WB. The Natural History of Congestive Heart Failure: The Framingham Study. *New England Journal of Medicine*. 1971;285(26):1441-1446. doi:10.1056/nejm197112232852601
29.Chen W, Yao J, Liang Z, et al. Temporal Trends in Mortality Rates among Kaiser Permanente Southern California Health Plan Enrollees, 2001-2016. *Perm J*. 2019;23doi:10.7812/TPP/18-213
30.Linzer DA, Lewis JB. **poLCA** : An *R* Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software*. 2011 2011;42(10)doi:10.18637/jss.v042.i10
31.Urbanowicz R, Zhang R, Cui Y, Suri P. STREAMLINE: A Simple, Transparent, End-To-End Automated Machine Learning Pipeline Facilitating Data Analysis and Algorithm Comparison. *Genetic Programming Theory and Practice XIX*. 2023:201-231:chap Chapter 9. *Genetic and Evolutionary Computation*.

32.Urbanowicz RJ, Bandhey H, Keenan BT, et al. STREAMLINE: An Automated Machine Learning Pipeline for Biomedicine Applied to Examine the Utility of Photography-Based Phenotypes for OSA Prediction Across International Sleep Centers. 2023:arXiv:2312.05461. doi:10.48550/arXiv.2312.05461 Accessed December 01, 2023. https://ui.adsabs.harvard.edu/abs/2023arXiv231205461U
33.L Mandel H, Colleen G, Abedian S, et al. Risk of post-acute sequelae of SARS-CoV-2 infection associated with pre-coronavirus disease obstructive sleep apnea diagnoses: an electronic health record-based analysis from the RECOVER initiative. *Sleep*. 2023;46(9)doi:10.1093/sleep/zsad126
34.Kapur V, Strohl KP, Redline S, Iber C, O?Connor G, Nieto J. Underdiagnosis of Sleep Apnea Syndrome in U.S. Communities. *Sleep and Breathing*. 2002;6(2):49-54. doi:10.1055/s-2002-32318
35.Finkel KJ, Searleman AC, Tymkew H, et al. Prevalence of undiagnosed obstructive sleep apnea among adult surgical patients in an academic medical center. *Sleep Medicine*. 2009;10(7):753-758. doi:10.1016/j.sleep.2008.08.007
36.Gleeson M, McNicholas WT. Bidirectional relationships of comorbidity with obstructive sleep apnoea. *European Respiratory Review*. 2022;31(164)doi:10.1183/16000617.0256-2021
37.Walton-Moss B, Samuel L, Nguyen TH, Commodore-Mensah Y, Hayat MJ, Szanton SL. Community-Based Cardiovascular Health Interventions in Vulnerable Populations. *Journal of Cardiovascular Nursing*. 2014;29(4):293-307. doi:10.1097/JCN.0b013e31828e2995
38.Sircu V, Colesnic S-I, Covantsev S, et al. The Burden of Comorbidities in Obstructive Sleep Apnea and the Pathophysiologic Mechanisms and Effects of CPAP. *Clocks & Sleep*. 2023;5(2):333-349. doi:10.3390/clockssleep5020025

Detecting clinician implicit biases in diagnoses using proximal causal inference

Kara Liu†, Russ Altman, Vasilis Syrgkanis

Computer Science Department, Stanford University,
Stanford, CA 94305, USA
†*E-mail: karaliu@stanford.edu*

Clinical decisions to treat and diagnose patients are affected by implicit biases formed by racism, ableism, sexism, and other stereotypes. These biases reflect broader systemic discrimination in healthcare and risk marginalizing already disadvantaged groups. Existing methods for measuring implicit biases require controlled randomized testing and only capture individual attitudes rather than outcomes. However, the "big-data" revolution has led to the availability of large observational medical datasets, like EHRs and biobanks, that provide the opportunity to investigate discrepancies in patient health outcomes. In this work, we propose a causal inference approach to detect the effect of clinician implicit biases on patient outcomes in large-scale medical data. Specifically, our method uses proximal mediation to disentangle pathway-specific effects of a patient's sociodemographic attribute on a clinician's diagnosis decision. We test our method on real-world data from the UK Biobank. Our work can serve as a tool that initiates conversation and brings awareness to unequal health outcomes caused by implicit biases.*

Keywords: Implicit bias, proximal causal inference, fairness, healthcare

1. Introduction

Implicit bias refers to unconscious and automatic associations that affect how we perceive, evaluate, and interact with people from different social groups.[1] Outside of mere cognitive distortions, these biases held by healthcare professionals influence clinical decisions and alter a patient's quality of care. Implicit biases have been shown to be both harmful and pervasive in modern-day medicine, exacerbating existing inequality in the treatment and health outcomes of marginalized groups.[2,3] For instance, unconscious attitudes held by clinicians result in disparate outcomes where women are less likely then men to be diagnosed with myocardial infarction,[3] Black women in the UK and US experience higher maternal mortality than White women,[4] and low socioeconomic (SES) and non-White patients receive sub-optimal pain management treatment compared to high SES and White patients.[5,6]

The recent integration of machine learning (ML) models into clinical decision-making has highlighted the prevalence of biases in medicine. By replicating the patterns from real-world medical data, ML models perpetuate and risk amplifying existing disparities in the medical treatment of marginalized groups.[7,8] While much attention has been given to the statistical

*Our method is available at `https://github.com/syrgkanislab/hidden_mediators`

objectives of fairness and the development of fair models, there has been comparatively less focus on investigating the biases present in the underlying data. A method capable of detecting implicit clinician bias in observational datasets would prevent ML models from unintentionally perpetuating biased decisions.

However, measuring implicit bias is challenging. Existing methods for quantifying implicit bias rely on the Implicit Association Test (IAT)[9] and randomized psychological experiments like affective priming.[10] While these tests are useful for initiating dialogue, they only provide a snapshot of individual clinician attitudes and do not guarantee a causal link to behavior or larger systemic discrepancies of care.[10]

In this work, we propose a computational tool to detect clinician implicit bias in observational datasets by measuring the causal effect of patient attributes, like race, SES, and other social determinants of health (SDoH), on medical diagnoses. By decomposing the causal effect into two pathways, we can separate the *biological effect* (the influence of a demographic attribute on diagnosis as mediated by valid biological traits) from the *implicit bias effect* (how the patient's attribute affects a clinician's judgement independent of their actual health state). As it is unlikely to observe a patient's true health state, we use observed medical data as proxies using proximal causal inference.[11] To estimate the effect of implicit bias, we propose a novel proximal mediation method that guarantees identifiability under several assumptions. Using real patient data from the UK Biobank, we validate our method can robustly detect several clinician implicit biases identified from prior works. We aim for the proposed method to serve as a bias-detection tool in dataset audits and initiate discussion on reducing systemic discrimination in medicine.

Disclaimer: While we use the UK Biobank data for method validation, we emphasize that this work is not a commentary on specific examples of discrimination within the UK healthcare system. Additionally, it is crucial to clarify that our method of estimating implicit bias is not intended to target clinicians but rather reflect on clinician behaviors within the context of discriminatory healthcare systems.

2. Method

2.1. *Background*

2.1.1. *Overview*

According to the Hippocratic Oath, clinicians should base their diagnostic decisions on each patient's history and current health status, unaffected by biases or stereotypes of the perceived patient identity. However, even in the ideal scenario of unbiased treatment, patient sociodemographic attributes will still influence diagnosis. Attributes including race, sex, or SES have been shown to influence a patient's true health status via mechanisms like genetics, lifestyle, and weathering from systemic oppression.[12–14] These biologically-mediated effects increase the risk of certain medical conditions. For instance, patients from lower SES backgrounds experience higher levels of stress and reduced access to healthcare, increasing their risk of cardiovascular disease.[15] In light of these known biological influences, the causal effect of a patient's sociodemographic attribute on their diagnosis by a clinician is therefore comprised of two pathway effects: the *biological effect* and the *implicit bias effect*, the latter referring to

the clinician's subjective biases of the sociodemographic attribute not mediated through the patient's actual health state.

We present the assumed causal relationships between variables as the directed acyclic graph (DAG) in Figure 1. Dashed arrows denote optional edges, and bi-directional arrows denote indirect confounding paths through latent variables. Let D be the binary sociodemographic attribute and Y the diagnosis decision we wish to measure implicit bias with respect to. M represents the latent variables encoding a patient's true underlying health state. However, as M is typically unknown, we instead observe Z and X as multivariate proxies of M. We differentiate these proxies into the variables Z that do not affect the diagnostic decision Y but could be affected by the attribute D; and the variables X which are not directly affected by the attribute D but can influence diagnosis Y. For example, X could be recent lab reports a clinician uses to make their diagnosis, and Z could be a patient's survey responses to a sleep questionnaire (assuming the survey does not influence the clinician's diagnosis). Finally, let W be sociodemographic confounders to control for.

We can now reframe *biological* and *implicit bias effects* using pathway causal effects. The *biological effect* of attribute D on diagnosis Y is the indirect effect as mediated through the true underlying health state M: $D \to M \to Y$. The *implicit bias effect* we wish to measure is the direct effect of $D \to Y$ that flows through the edge θ and is defined as the residual of the biological effect. We formally define bias in terms of controlled direct effects in Equation (D.1).

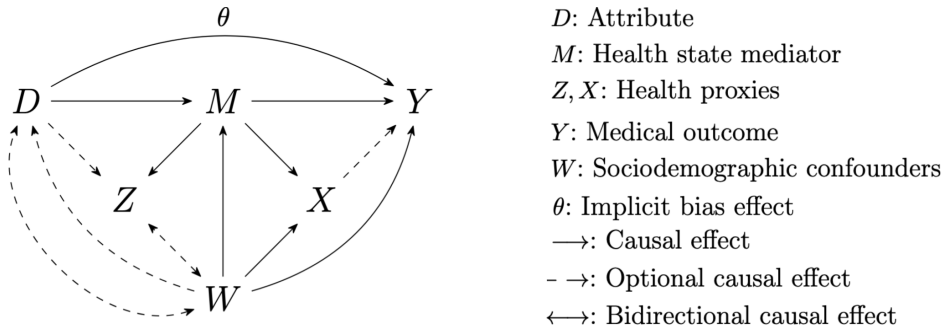


Fig. 1: Assumed causal graph.

2.1.2. *Related works*

Measuring implicit biases requires detecting the unconscious and automatic attitudes that shape behavior. The predominant method for implicit bias measurement thus far has been the Implicit Association Test (IAT),[9] a questionnaire developed in 1998 intended to measure group association through word categorization. To capture clinician biases, several works have linked clinician attitudes via their IAT score to behavioral manifestation.[10,16] Other methods for detecting implicit clinician bias include affective priming, which measures biased associations after stimulus priming; and the assumption method, which surveys clinicians' decisions after reading patient vignettes.[17] While association tests like the IAT have been integral in bringing awareness to medical biases, they are criticized for their arbitrary scoring system, inability to predict real-world patient outcomes, and context-dependency.[3,10,16,17] Furthermore,

administering these controlled tests in every clinical encounter is impractical and unscalable.

Computational methods present a promising and scalable alternative for detecting implicit bias in real-world medical data. While the field of ML fairness has explored bias detection, the focus has been on identifying and mitigating bias in models rather than the data.[7] In causal inference, disentangling a causal effect into natural indirect and direct pathway effects has led to methods that control for "fair" and "unfair" causal pathway effects. [18–21] propose metrics for measuring fair pathway influence on outcomes and develop methods that mitigate the effect of unfair pathways on the predicted outcome. [22] leveraged the Fairness-Aware Causal paThs (FACTS)[23] algorithm to quantify disparate pathway influence of SDoH attributes on mortality using real-world health data. While these methods recognize an attribute's influence on an outcome contains both fair and unfair effects, prior works are limited to simple scenarios where all variables are known and observed. Our work is the first to extend pathway inference to large-scale observational data with potentially unobserved variables.

Finally, a few recent methods have explored proximal mediation analysis, where pathway effects can be measured despite unobserved mediators by using proxy variables.[24,25] However, by relying on natural direct and indirect pathway effects, these works rely on more stringent assumptions, require learning complicated bridge functions, and limit their analysis to simple datasets. In comparison, our method makes several relaxations that enable application to observational data. First, we identify controlled instead of natural effects, which presents an equally good measurement of a biased decision yet lends to a much simpler statistical problem. Additionally, we assume partially linear equations instead of requiring the identification of a complex bridge function. Finally, we do not require uniqueness of the parameters unrelated to implicit bias (i.e., the nuisance parameters for the outcome bridge function). These relaxations enable our approach to be effective at analyzing large-scale real-world medical data.

2.2. *Our method*

Our goal is to identify and estimate the following controlled direct effect:

$$\theta = \int_{m,w} \mathbb{E}[Y(1,m) - Y(0,m) \mid W = w]\ p(m,w)\ dm\ dw \tag{1}$$

where $Y(d,m)$ is the potential (or counterfactual) outcome when we intervene on the attribute D and the mediator M and set them to values (d,m); and $p(m,w)$ is the natural probability distribution in the data. If the controlled direct effect is nonzero, then there exists a direct influence of the attribute D on the outcome Y, which is evidence of implicit bias.

If we observe M, the above controlled direct effect can be identified by a simple g-formula that "controls" for M and W: $\theta = \mathbb{E}[\mathbb{E}[Y \mid D = 1, M, W] - \mathbb{E}[Y \mid D = 0, M, W]]$. Unfortunately, this equation is intractable if M is unobserved. However, we show that under a few reasonable assumptions the controlled direct effect is still identifiable.

Theorem 1 (Identification). [a] *Consider a non-parametric structural causal model (SCM) that respects the causal relationships encoded in Figure 1 (see Appendix D.1) and assume there exists a "bridge function" q that solves $\mathbb{E}[Y \mid D, M, W] = \mathbb{E}[q(D, X, W) \mid D, M, W]$. Then*

[a]We present more intuitive interpretations of each theorem and lemma in the Appendix.

q also solves the Non-Parametric Instrumental Variable (NPIV) problem defined by the set of conditional moment restrictions

$$\mathbb{E}[Y - q(D, X, W) \mid D, Z, W] = 0 \tag{2}$$

and the controlled direct effect can be identified as $\theta = \mathbb{E}[q(1, X, W) - q(0, X, W)]$.

Identifying parameters θ using a bridge function q (where q also solves an NPIV problem) has been extensively studied in proximal causal inference literature.[26–38] However, these approaches rely on solving saddle-point problems with adversarial training or require learning conditional density functions, both of which are statistically daunting.

We can avoid these difficult statistical tasks if we assume that the bridge function is partially linear in D and X. The following lemma shows that partial linearity of q is implied by a more primitive assumption of partial linearity of two other functions (proof in Appendix D.4).

Lemma 1 (Identification under partial linearity). *Consider a non-parametric SCM that respects the constraints encoded in Figure 1 and assume that X has dimension p_X at least as large as the dimension p_M of M. Moreover, assume that the following functions are partially linear:*

$$\mathbb{E}[Y \mid D, M, X, W] = D\,c + M^T b + X^T g + f_Y(W) \tag{3}$$

$$\mathbb{E}[X \mid M, W] = F\,M + f_X(W) \tag{4}$$

where F is a $p_X \times p_M$ matrix, b is a p_M-dimensional vector, g is a p_X-dimensional vector and f_Y, f_X are arbitrary non-parametric functions. If we assume the matrix F has full column rank, then there exists a partially linear outcome bridge function

$$q(D, X, W) = D\,\theta + X^T h + f(W) \tag{5}$$

that satisfies Equation (2), where parameter $h = F^{+}b + g$[b] and $\theta = c$.

Under the assumption of partial linearity, we can simplify the estimation problem by first removing the effect of W from all the remaining variables (see Appendix D.5), where for any variable V we define the residual $\tilde{V} = V - \mathbb{E}[V \mid W]$. Partial linearity of q from Equation (5), when combined with the NPIV Equation (2), implies that θ can be identified using linear instrumental variable (IV) regression where $(\tilde{Z}; \tilde{D})$[c] are the instruments and $(\tilde{X}; \tilde{D})$ are the treatments:

$$\mathbb{E}\left[(\tilde{Y} - \tilde{X}^T h - \tilde{D}\,\theta)\begin{pmatrix}\tilde{Z}\\ \tilde{D}\end{pmatrix}\right] = 0 \qquad \text{(Primal Equation)}$$

Unique identification of θ seemingly requires unique identification of the other "nuisance" parameters like h, which might be difficult to achieve as the covariance matrix $\mathbb{E}[(\tilde{X}; \tilde{D})\,(\tilde{Z}; \tilde{D})^T]$ is usually not full rank[d]. We invoke and simplify ideas from the recent proximal inference

[b] F^{+} is the Moore-Penrose pseudoinverse of F.
[c] We denote $(A; B)$ to be concatenation of vectors A and B.
[d] This could be the case if the number of proxies is much larger than the dimensionality of the latent mediator M.

literature[35,39] to show that θ can be point-identified even if h is not. To achieve this, we construct a moment restriction equation that is Neyman orthogonal to the nuisance parameters h but still point-identifies θ, given sufficient quality of the proxy Z. Intuitively, we learn a new instrument $V = (\tilde{D} - \gamma^\top \tilde{Z})$ such that V is uncorrelated with $\tilde{X}$ and thus estimation of θ is not sensitive to h. Existence of such a γ is sufficient for point-identification of θ. We provide the proof for the point identification of θ in Appendix D.7 and for Neyman orthogonality in D.8.

Theorem 2. *Let h_* be the minimum norm solution to the* (Primal Equation) *and assume that the following dual equation also admits a solution γ_*:*

$$\mathbb{E}[\tilde{X}\,(\tilde{D} - \gamma^T \tilde{Z})] = 0 \qquad \text{(Dual Equation)}$$

Furthermore, assume $\mathbb{E}[\tilde{D}\,(\tilde{D} - \gamma_*^T Z)] \neq 0$. Then the solution θ_* to the equation:

$$\mathbb{E}[(\tilde{Y} - \tilde{X}^T h_* - \tilde{D}\,\theta)\,(\tilde{D} - \gamma_*^T \tilde{Z})] = 0 \tag{6}$$

uniquely identifies the controlled direct effect θ. Furthermore, this moment restriction is Neyman orthogonal with respect to nuisance parameters γ_*, h_*.

Theorem 2.2 allows us to invoke the general framework of [40] to construct an estimate and confidence interval for the controlled direct effect θ. The full estimation algorithm is presented in Appendix D.9.

2.3. *Testing and Removing Weak Instruments*

Our method for uniquely identifying the controlled direct effect θ relies on several assumptions, e.g., $(\tilde{Z}; \tilde{D})$ are good instruments for $(\tilde{X}; \tilde{D})$. To assess the validity of these assumptions, we developed a suite of tests that must pass for the estimate θ to be valid and can be used as validity checks by practitioners. These tests are further described in Appendix C:

(1) *Primal equation violation* - We develop a χ^2-test to check if the primal equation admits a solution, i.e., $\mathbb{E}[(\tilde{Y} - \tilde{X}^T h_* - \tilde{D}\,\theta_*)(\tilde{Z}; \tilde{D})] \approx 0$. Intuitively, violation of the primal test implies either the variables X are insufficient proxies of the health state M or the residual proxy $\tilde{Z}$ has a direct path to $\tilde{Y}$.

(2) *Dual equation violation* - We develop a χ^2-test to check if the dual equation admits a solution, i.e., $\mathbb{E}[\tilde{X}(\tilde{D} - \gamma_*^T \tilde{Z})] \approx 0$. Violation of the dual implies the variables Z are insufficient proxies of the health state M or that the residual proxy $\tilde{X}$ has a direct path from $\tilde{D}$.

(3) *Strength of identification* - We perform two tests to check if $V = (\tilde{D} - \gamma^T \tilde{Z})$ is a good instrument for (i.e., retains enough information about) $\tilde{D}$. (a) We develop an effective F-test[41,42] to check the correlation strength of V with $\tilde{D}$. (b) We develop a z-test to check if the quantity $\mathbb{E}[\tilde{D}(\tilde{D} - \gamma_*^T \tilde{Z})]$ is substantially bounded away from zero (see assumption in Theorem 2.2). Intuitively, these tests will fail if the hidden mediator is a very deterministic function of the attribute D.

(4) *Proxy covariance rank test* - To ensure the health proxies are sufficiently related, we check the rank of the covariance matrix of $\tilde{X}$ and $\tilde{Z}$ by identifying the number of statistically significant singular values. This rank can be viewed as an upper bound on the dimensionality of the hidden mediator M that we can control for.

2.3.1. *Proxy selection algorithm*

In practice, the initial selection of proxies X, Z may violate key assumptions, which can be detected by the failure of one or more of the aforementioned tests. In Appendix B, we provide an algorithm for identifying subsets of X and Z that satisfy the necessary assumptions and thus produce valid estimates. This proxy selection algorithm should be performed on a separate dataset from the one used to estimate θ.

3. Experiments

3.1. *Data*

To validate our approach, we use the UK Biobank, a rich and accessible repository containing genomic, imaging, and tabular health data from over 500,000 patients. Our work uses its tabular data, which includes survey questions and biometrics collected upon an individual's enrollment into the biobank. In addition, several health outcomes, including medical diagnoses via ICD10 codes, have been linked to most patients. We note and discuss the caveats of applying our method to biobank data in Section 5.2.

		Prevalence in UK Biobank (n=502411)	Prior works on implicit bias
Sociodemographic attribute D	*Race* - Asian	2.4%	43–45
	Race - Black	1.8%	3–5,46,47
	Gender - Female	54.4%	3,48
	Disability status - On disability allowance	6.2%	49,50
	Income - Household income <18,000£	20.3%	5,14,15,51
	Education - No post-secondary education	67.3%	5,51
	Weight - BMI >30	24.3%	52,53
	Insurance - Not on private insurance	31.4%	54
Medical diagnosis Y	Osteoarthritis	18.0%	47,49
	Rheumatoid arthritis	1.9%	55
	Chronic kidney disease	5.0%	56,57
	Complications during labor	2.4%	3,4
	Heart disease	10.7%	3,15,48
	Depression	6.0%	46,58
	Melanoma	1.2%	59,60

Table 1: Selected sociodemographic attributes D and diagnoses Y

Prior works have proposed sociodemographic attributes that might bias clinical decisions. For example, [48] showed that clinicians exhibited greater uncertainty when diagnosing coronary heart disease in women compared to men. We list in Table 1 most of the attributes D and diagnoses Y we test for implicit bias, and present the full list of the 102 (D, Y) pairs in Appendix E.2. To highlight the influence of clinician subjectivity, we concentrate on diagnoses that require clinician interpretation of patient-reported symptoms, e.g., chronic pain.

Selecting health proxies for Z and X relies user intuition and medical expertise to determine which variables have a direct relationship with attribute D and outcome Y, respectively. In general, proxies X could be observed by the clinician during their diagnostic decision, and proxies Z are not accessible during diagnosis but might have a direct causal relationship with attribute D. In the UK Biobank, we select X to be the biometric variables collected by the biobank at patient enrollment, which includes lab results and blood pressure readings. For Z, we use survey responses of self-reported pain levels, mental health, and sleep. We list all variables, including the sociodemographic confounders W, in Appendix E.1. Note our data

contains a mix of binary, integer, and continuous variable types.

3.2. *Evaluation metrics*

3.2.1. *Semi-synthetic data validation*

We test if our method can retrieve a known implicit bias effect using semi-synthetic data. We use real data from the UK Biobank for attribute D, confounders W, and health proxies X,Z. We develop a model that computes M and a synthetic diagnosis Y with a known implicit bias effect $\theta = 0.5$ using linear structural equations. We test against fully continuous (Experiment 1) and both binary and continuous (Experiment 2) semi-synthetic data, the latter being more realistic in real-world medical data. Our semi-synthetic data generation method is fully described in Appendix A. As a baseline, we compare two variants of ordinary least squares (OLS): (a) given we know M, we fit an OLS model over W, D, X, and M to predict Y; (b) in the more realistic scenario where M isn't known, we learn over W, D, X, and Z. We compute the average effect estimate and confidence interval based on $\pm 1.96\ \sigma$ where the average and standard deviation σ is taken over K=100 iterations.

3.2.2. *Calculating the implicit bias effect in the UK Biobank*

We next run our method on the full UK Biobank data. We compute the residuals of Z, X, Y, D fitted on W using Lasso regression. For all models, the regularization term is chosen via semi-cross fitting[61,62] over 3 splits. We fit all models using the `scikit-learn` Python package. For nuisance parameters h_* and γ_* we used regularized adversarial IV estimation[35,63] with linear functions and a theoretically driven penalty choice that decays faster than the root of the number of samples.

In cases where the data may not meet the method's assumptions, we developed a proxy selection algorithm (see Section 2.3.1) that identifies an optimal subset of X, Z proxies for each (D, Y) pair using the assumption tests from Section 2.3. Although we recommend separate data splits for proxy selection and effect estimation, we use the same dataset as our intent is method demonstration rather than robust effect estimates. Details of the hyperparameters used for the selection algorithm are provided in Appendix B.

For each of the 102 pairs of attribute D and diagnosis Y, we report seven metrics: the implicit bias effect θ, the 95% confidence interval, as well as our five proposed tests from Section 2.3: (1) the primal and (2) dual violation, (3-4) the strength of identification, and (5) the $\tilde{Z}, \tilde{X}$ covariance rank test. In addition, we also run the following five analyses:

Weak identification confidence interval - If the instrument identification tests from 2.3 are violated, then effect estimation can be unstable and normality-based confidence intervals inaccurate. We thus compute an alternative confidence interval[62] developed under the assumption of weak instruments (see Appendix C.5 for the description).

Bootstrapping analyses - We perform several bootstrapping analyses to test the sensitivity of the estimate. In the first analysis, given the computational complexity of recomputing the full estimate, we compare K=10 bootstrapped iterations re-estimating the full pipeline (stage 1); K=100 iterations using the pre-computed residuals but re-estimating all other parameters

(stage 2); and K=1000 iterations re-computing only the final Equation (6) (stage 3). Each iteration samples 50% of the data without replacement. In the second analysis, we compare sampling $10\%, 25\%, 50\%$ or 75% of the original data for K=10 bootstrapped iterations, re-estimating over full pipeline (stage 1). Finally, we compare different sample sizes for $K = 1000$ iterations re-estimating from stage 3 of the pipeline.

Influence points - Inspired by [64], we analyze influence scores, which measure how influential each data point is in the effect estimate. A significant change to the estimate after removing a small set of highly-influential points indicates the implicit bias calculation is highly sensitive to a few (potentially) outlier patients. We also include a preliminary interpretability analysis that explores the distinguishing phenotypes of highly influential patients, which could aid in determining if these subsets of patients correspond to some interpretable outlier group. We describe how we calculate the influence score and identify highly-influential patient sets in Appendix C.6.

Income stratification - To investigate intersectionality in implicit biases, we perform a stratified effect estimate over different income groups where $D \neq$Income.

Partial non-linearity of W - Our identification theorem allows for partial non-linearity in the effect of W. We thus re-compute the point estimate allowing for non-linear interactions with W using XGBoost[65] models instead of Lasso.

4. Results

4.1. *Synthetic data validation*

The results in Table 2 demonstrate that our method is able to retrieve the true implicit bias effect $\theta = 0.5$ with high certainty for both fully continuous and mixed-type data, with comparable performance to the best-case OLS where M is known. We report our method's coverage, RMSE, bias, standard deviation, mean confidence interval, and performance on our five tests (from Section 2.3), as well as testing other values of θ, in Appendix F.1.

	θ	Our method	OLS(D, W, M, X)	OLS(D, W, Z, X)
Experiment 1: Continuous	0.5	0.54 ± 0.003	0.5 ± 0.01	1.10 ± 0.01
Experiment 2: Continuous and binary	0.5	0.53 ± 0.003	0.5 ± 0.01	1.385 ± 0.01

Table 2: Semi-synthetic data estimates θ and confidence interval over K=100 iterations.

4.2. *Calculating the implicit bias effect in the UK Biobank*

In Appendix F.2, we show the effect estimates for the (D, Y) pairs using all proxies Z, X, adjusting the confounders W by excluding the column corresponding to the attribute D. However, as evidenced by the failure of the dual and primal tests, we found the initial sets of proxies Z, X did not meet our method's necessary assumptions. As discussed further in Appendix F.3, we believe these test failures indicate there might exist some features in X with a causal path from D that does not go through M or features within Z with a causal path to Y that doesn't flow through M. Such paths invalidate the resulting effect estimates.

We thus found applying our proxy selection algorithm (see 2.3.1) necessary for producing valid effect estimates. After running the algorithm to select subsets of admissible X, Z proxies (the description and interpretation of the selected proxies can be found in Appendix F.3), we found 34 (D, Y) pairs that pass all tests with narrow confidence intervals. We report six in Table 3 and include the remaining estimates in Appendix F.4. Note that $\theta > 0$ implies a patient with D is more likely to be diagnosed with Y due to clinician bias, and conversely $\theta < 0$ implies a patient is less likely to be diagnosed. In Section 5.2, we offer a framework for interpreting the implications of these results.

4.2.1. *Weak instrument confidence interval*

As shown in Figure 2A, the confidence interval predicted under the weak instrument regime consistently aligns with the interval under our method, thus indicating our estimate's robustness to weak instruments.

(D, Y)	$\theta \pm 95\%$ **CI**	**(1) Primal statistic < critical**	**(2) Dual statistic < critical**	**(3)** $\mathbb{E}[\tilde{D}V] \neq 0$ **statistic > critical**	**(4)** V **strength F-test statistic > critical**	**(5)** $\mathrm{Cov}(\tilde{X}, \tilde{Z})$ **rank**
Low income, Depression	0.03 ± 0.02	59.9<60.5	31.9<40.1	84.1>0.4	3332.1>23.1	3
Disability insurance, Rh. Arthritis	0.06 ± 0.0	67.3<75.6	3.4<11.1	29.2>0.4	801.1>23.1	3
Female, Heart disease	-0.19 ± 0.06	115.8<118.8	23.3<23.7	18.8>1.3	92.5>23.1	4
Black, Chronic kidney disease	0.14 ± 0.03	56.9<58.1	10.6<21.0	9.8>0.3	23.3>23.1	4
Obese, Osteoarthritis	0.09 ± 0.02	90.5<100.7	24.8<28.9	76.5>1.7	254.9>23.1	3
Asian, Osteoarthritis	-0.06 ± 0.03	94.7<101.9	33.1<33.9	13.9>0.3	74.6>23.1	5

Table 3: Six of the 34 valid UK Biobank implicit bias effect estimates after applying our X, Z proxy selection algorithm. Tests (1-5) are detailed in 2.3, where *statistic* is the given data's statistic and *critical* is the necessary critical value to be greater or less than to pass. $V = \tilde{D} - \gamma^T \tilde{Z}$.

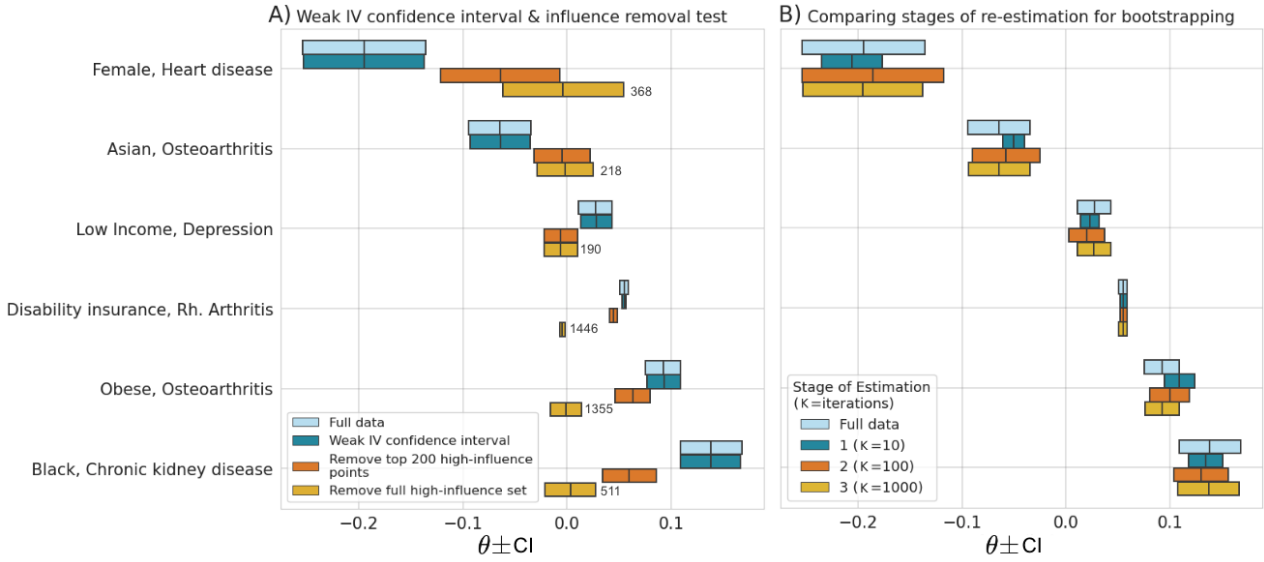


Fig. 2: Comparing effect estimates for six (D, Y) pairs using all data with: A) weak instrument and influence set removal (where the numbers next to the yellow bar reflect the set size of high-influence points); B) bootstrapped subsampling 50% of the data at different stages of re-estimation.

4.2.2. *Bootstrapping analyses*

In Figure 2B, we show the results of the first bootstrap analysis comparing different stages of re-estimation. We observe that, regardless of the estimation stage, bootstrapped estimates are consistent with the estimate from the full dataset. The consistency of the bootstrapped estimates over different sample sizes, as shown in Appendix F.6, further support the robustness of our method.

4.2.3. *Influence points*

In Figure 2A, we see that removing only a few highly-influential points leads to a significant decrease in the magnitude of the estimated effect. To investigate, we run a preliminary interpretabilty analysis where we analyze the univariate differences between patients with high influence and those with low influence. In Figure 3A patients that strongly influence the negative implicit bias estimate for (D=Female, Y=Heart disease) are more likely to be low income, unemployed due to disability, and suffer from depression. It is plausible such patients are the "outliers" driving the strong negative bias estimate.

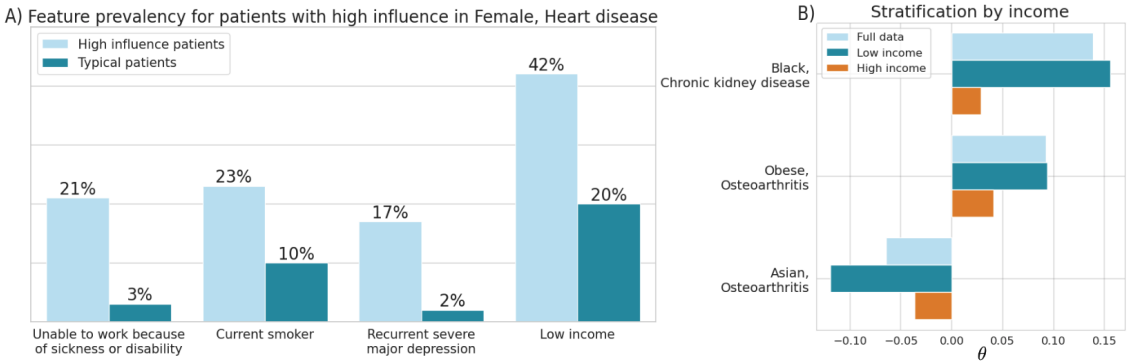


Fig. 3: A) Interpretability into high influence points. B) Income stratification

4.2.4. *Income stratification*

In Figure 3B we analyze the effect of stratification based on income. We see a general increase in bias effect estimate for the low income strata and a corresponding decrease in effect for high income strata, demonstrating potential evidence of intersectional discrimination.[5,66]

4.2.5. *Partial non-linearity of W*

In Appendix F.9, we show our implicit bias estimate with non-linear W interactions leads to a similar effect estimates of θ.

5. Discussion

5.1. *Limitations*

In this work, we propose a robust causal inference method designed to detect clinician implicit bias by estimating pathway-specific causal effects. We demonstrate the applicability of our approach to large-scale medical data by validating on both semi-synthetic and real-world datasets.

However, our work contains several limitations. First, while the UK Biobank is a rich and accessible source of medical data, most patient information is collected once upon signing up for the biobank. Although UK Biobank has synced their records to a handful of outcomes provided by EHR data (like ICD10 codes), it is unclear to what extent the available proxies for X (which were collected at patient enrollment) are used by clinicians for diagnoses. Additionally, the synced ICD10 codes are from hospital records, thus excluding primary care visits. We plan to validate our method with time-series EHR data in follow-up work.

Second, while the assumption of partially linear structural equations is crucial for enabling better identifiability of the outcome bridge function under minimal conditions, it is possible the ground truth equations are non-linear.

Finally, it is well known that intersectional identities shape complex patterns of discrimination in healthcare.[5,66] A more comprehensive analysis on the effect of implicit bias from intersectional attributes on patient treatment would be valuable for improving equity in healthcare outcomes.

5.2. *Interpretation and application of results*

While we re-iterate the intent of this work is not to diagnose specific cases of implicit bias in the UK Biobank, our method did flag several areas of clinical inequity that have been reported in literature. For instance, many works have reported gender-based inequality in cardiovascular health,[67] and we similarly detected an estimate of θ = -0.19 indicating clinicians are less likely, due to implicit biases, to diagnose D=Females with Y=heart disease. In another example, our estimate θ = -0.06 suggested clinicians are less likely to diagnose D=Asian patients with Y=osteoarthritis, and many works have highlighted both patient- and clinician-stigmas regarding pain-associated disorders, like osteoarthritis, in Asians.[68–70]

However, we did find several estimates contrary to what we expected. For example, our estimate $\theta = 0.14$ indicated clinicians are *positively* biased towards diagnosing Black patients with chronic kidney disease. However, at the time of UK Biobank data collection, many doctors relied on a race-based equation for kidney function now known to have under-detected kidney disease in Black patients.[71]

To understand a discrepancy between a produced estimate and literature (or user intuition), we recommend (1) ensuring the data used contains sufficient health proxies and satisfy all assumptions (e.g., see biobank data limitations in 5.1); (2) investigating all mechanisms creating the medical outcome Y (e.g., hospital-specific diagnosis protocol); and (3) exploring how the discovered bias estimate fits in context, rather than opposed, to those found in literature. While our method does not offer a solution on *how* to tackle implicit biases, by bringing awareness to potential areas of discrimination within a given healthcare system, detecting biases is the first step towards creating systemic-level change through interdisciplinary collaboration and targeted anti-bias training programs.

6. Appendix

The appendix can be found at `https://github.com/syrgkanislab/hidden_mediators`.

References

1. J. Holroyd, J. Sweetman, M. Brownstein and J. Saul, The heterogeneity of implicit bias, *Implicit bias and philosophy* **1**, 80 (2016).
2. M. B. Vela, A. I. Erondu, N. A. Smith, M. E. Peek, J. N. Woodruff and M. H. Chin, Eliminating explicit and implicit biases in health care: evidence and research needs, *Annual review of public health* **43**, 477 (2022).
3. D. P. Gopal, U. Chetty, P. O'Donnell, C. Gajria and J. Blackadder-Weinstein, Implicit bias in healthcare: clinical practice, research and decision making, *Future healthcare journal* **8**, 40 (2021).
4. B. Saluja and Z. Bryant, How implicit bias contributes to racial disparities in maternal morbidity and mortality in the united states, *Journal of women's health* **30**, 270 (2021).
5. T. M. Anastas, M. M. Miller, N. A. Hollingshead, J. C. Stewart, K. L. Rand and A. T. Hirsh, The unique and interactive effects of patient race, patient socioeconomic status, and provider attitudes on chronic pain care decisions, *Annals of Behavioral Medicine* **54**, 771 (2020).
6. J. A. Sabin and A. G. Greenwald, The influence of implicit bias on treatment recommendations for 4 common pediatric conditions: pain, urinary tract infection, attention deficit hyperactivity disorder, and asthma, *American journal of public health* **102**, 988 (2012).
7. D. Ueda, T. Kakinuma, S. Fujita, K. Kamagata, Y. Fushimi, R. Ito, Y. Matsui, T. Nozaki, T. Nakaura, N. Fujima *et al.*, Fairness of artificial intelligence in healthcare: review and recommendations, *Japanese Journal of Radiology* **42**, 3 (2024).
8. S. R. Pfohl, A. Foryciarz and N. H. Shah, An empirical characterization of fair machine learning for clinical risk prediction, *Journal of biomedical informatics* **113**, p. 103621 (2021).
9. A. G. Greenwald, D. E. McGhee and J. L. Schwartz, Measuring individual differences in implicit cognition: the implicit association test., *Journal of personality and social psychology* **74**, p. 1464 (1998).
10. S. A. Arif and J. Schlotfeldt, Gaps in measuring and mitigating implicit bias in healthcare, *Frontiers in Pharmacology* **12**, p. 633565 (2021).
11. E. J. T. Tchetgen, A. Ying, Y. Cui, X. Shi and W. Miao, An introduction to proximal causal learning, *arXiv preprint arXiv:2009.10982* (2020).
12. A. T. Forde, D. M. Crookes, S. F. Suglia and R. T. Demmer, The weathering hypothesis as an explanation for racial disparities in health: a systematic review, *Annals of epidemiology* **33**, 1 (2019).
13. D. R. Williams, Stress and the mental health of populations of color: Advancing our understanding of race-related stressors, *Journal of health and social behavior* **59**, 466 (2018).
14. X. Cui and C.-T. Chang, How income influences health: decomposition based on absolute income and relative income effects, *International Journal of Environmental Research and Public Health* **18**, p. 10738 (2021).
15. A. M. K. Minhas, V. Jain, M. Li, R. W. Ariss, M. Fudim, E. D. Michos, S. S. Virani, L. Sperling and A. Mehta, Family income and cardiovascular disease risk in american adults, *Scientific reports* **13**, p. 279 (2023).
16. A. G. Greenwald, N. Dasgupta, J. F. Dovidio, J. Kang, C. A. Moss-Racusin and B. A. Teachman, Implicit-bias remedies: Treating discriminatory bias as a public-health problem, *Psychological Science in the Public Interest* **23**, 7 (2022).
17. C. FitzGerald and S. Hurst, Implicit bias in healthcare professionals: a systematic review, *BMC medical ethics* **18**, 1 (2017).
18. S. Chiappa, Path-specific counterfactual fairness, **33**, 7801 (2019).
19. R. Nabi and I. Shpitser, Fair inference on outcomes, **32** (2018).
20. L. Zhang, Y. Wu and X. Wu, A causal framework for discovering and removing direct and indirect discrimination, *arXiv preprint arXiv:1611.07509* (2016).

21. A. I. Naimi, M. E. Schnitzer, E. E. Moodie and L. M. Bodnar, Mediation analysis for health disparities research, *American journal of epidemiology* **184**, 315 (2016).
22. I. Jun, S. E. Ser, S. A. Cohen, J. Xu, R. J. Lucero, J. Bian and M. Prosperi, Quantifying health outcome disparity in invasive methicillin-resistant staphylococcus aureus infection using fairness algorithms on real-world data, in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, 2023.
23. W. Pan, S. Cui, J. Bian, C. Zhang and F. Wang, Explaining algorithmic fairness through fairness-aware causal path decomposition, in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.
24. A. Ghassami, A. Yang, I. Shpitser and E. T. Tchetgen, Causal inference with hidden mediators, *arXiv preprint arXiv:2111.02927* (2021).
25. A. Ghassami, I. Shpitser and E. T. Tchetgen, Partial identification of causal effects using proxy variables, *arXiv preprint arXiv:2304.04374* (2023).
26. X. Chen and D. Pouzo, Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals, *Econometrica* **80**, 277 (2012).
27. C. Ai and X. Chen, Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica* **71**, 1795 (2003).
28. C. Ai and X. Chen, The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions, *Journal of Econometrics* **170**, 442 (2012).
29. G. Lewis and V. Syrgkanis, Adversarial generalized method of moments, *arXiv preprint arXiv:1803.07164* (2018).
30. N. Dikkala, G. Lewis, L. Mackey and V. Syrgkanis, Minimax estimation of conditional moment models, *Advances in Neural Information Processing Systems* **33**, 12248 (2020).
31. A. Bennett, N. Kallus and T. Schnabel, Deep generalized method of moments for instrumental variable analysis, *Advances in neural information processing systems* **32** (2019).
32. A. Bennett and N. Kallus, The variational method of moments, *arXiv preprint arXiv:2012.09422* (2020).
33. W. Miao and E. T. Tchetgen, A confounding bridge approach for double negative control inference on causal effects (supplement and sample codes are included), *arXiv preprint arXiv:1808.04945* (2018).
34. Y. Cui, H. Pu, X. Shi, W. Miao and E. T. Tchetgen, Semiparametric proximal causal inference, *arXiv preprint arXiv:2011.08411* (2020).
35. A. Bennett, N. Kallus, X. Mao, W. Newey, V. Syrgkanis and M. Uehara, Inference on strongly identified functionals of weakly identified functions, *arXiv preprint arXiv:2208.08291* (2022).
36. A. Bennett, N. Kallus, X. Mao, W. Newey, V. Syrgkanis and M. Uehara, Source condition double robust inference on functionals of inverse problems, *arXiv preprint arXiv:2307.13793* (2023).
37. A. Bennett, N. Kallus, X. Mao, W. Newey, V. Syrgkanis and M. Uehara, Minimax instrumental variable regression and l_2 convergence guarantees without identification or closedness, *arXiv preprint arXiv:2302.05404* (2023).
38. J. Zhang, W. Li, W. Miao and E. T. Tchetgen, Proximal causal inference without uniqueness assumptions, *Statistics & Probability Letters* **198**, p. 109836 (2023).
39. Q. Chen, Robust and optimal estimation for partially linear instrumental variables models with partial identification, *Journal of Econometrics* **221**, 368 (2021).
40. V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen and W. Newey, Double/debiased/neyman machine learning of treatment effects, *American Economic Review* **107**, 261 (2017).
41. J. L. M. Olea and C. Pflueger, A robust test for weak instruments, *Journal of Business & Economic Statistics* **31**, 358 (2013).
42. D. W. Andrews and J. H. Stock, *Inference with Weak Instruments*, Working Paper 313, National Bureau of Economic Research (August 2005).

43. C. L. McMurtry, M. G. Findling, L. S. Casey, R. J. Blendon, J. M. Benson, J. M. Sayde and C. Miller, Discrimination in the united states: Experiences of asian americans, *Health services research* **54**, 1419 (2019).
44. O. Bougie, M. I. Yap, L. Sikora, T. Flaxman and S. Singh, Influence of race/ethnicity on prevalence and presentation of endometriosis: a systematic review and meta-analysis, *BJOG: An International Journal of Obstetrics & Gynaecology* **126**, 1104 (2019).
45. C. Wu, Y. Qian and R. Wilkes, Anti-asian discrimination and the asian-white mental health gap during covid-19, in *Race and Ethnicity in Pandemic Times*, (Routledge, 2021) pp. 101–117.
46. H. N. Garb, Race bias and gender bias in the diagnosis of psychological disorders, *Clinical Psychology Review* **90**, p. 102087 (2021).
47. J. L. W. Taylor, C. M. Campbell, R. J. Thorpe Jr, K. E. Whitfield, M. Nkimbeng and S. L. Szanton, Pain, racial discrimination, and depressive symptoms among african american women, *Pain Management Nursing* **19**, 79 (2018).
48. N. N. Maserejian, C. L. Link, K. L. Lutfey, L. D. Marceau and J. B. McKinlay, Disparities in physicians' interpretations of heart disease symptoms by patient gender: results of a video vignette factorial experiment, *Journal of women's health* **18**, 1661 (2009).
49. J. McClendon, U. R. Essien, A. Youk, S. A. Ibrahim, E. Vina, C. K. Kwoh and L. R. Hausmann, Cumulative disadvantage and disparities in depression and pain among veterans with osteoarthritis: the role of perceived discrimination, *Arthritis Care & Research* **73**, 11 (2021).
50. L. VanPuymbrouck, C. Friedman and H. Feldner, Explicit and implicit disability attitudes of healthcare providers., *Rehabilitation psychology* **65**, p. 101 (2020).
51. I. Stepanikova and G. R. Oates, Perceived discrimination and privilege in health care: the role of socioeconomic status and race, *American journal of preventive medicine* **52**, S86 (2017).
52. S. M. Phelan, D. J. Burgess, M. W. Yeazel, W. L. Hellerstedt, J. M. Griffin and M. van Ryn, Impact of weight bias and stigma on quality of care and outcomes for patients with obesity, *Obesity Reviews* **16**, 319 (2015), Open Access, Citations: 737.
53. M. Fulton, S. Dadana and V. N. Srinivasan, Obesity, stigma, and discrimination, in *StatPearls [Internet]*, (StatPearls Publishing, 2023)
54. X. Han, K. T. Call, J. K. Pintor, G. Alarcon-Espinoza and A. B. Simon, Reports of insurance-based discrimination in health care and its association with access to care, *American journal of public health* **105**, S517 (2015).
55. C. A. McBurney and E. R. Vina, Racial and ethnic disparities in rheumatoid arthritis, *Current rheumatology reports* **14**, 463 (2012).
56. K. Evans, J. Coresh, L. D. Bash, T. Gary-Webb, A. Köttgen, K. Carson and L. E. Boulware, Race differences in access to health care and disparities in incident chronic kidney disease in the us, *Nephrology Dialysis Transplantation* **26**, 899 (2011).
57. K. A. Jenkins, S. Keddem, S. B. Bekele, K. E. Augustine and J. A. Long, Perspectives on racism in health care among black veterans with chronic kidney disease, *JAMA Network Open* **5**, e2211900 (2022).
58. M. Kim, Racial/ethnic disparities in depression and its theoretical perspectives, *Psychiatric Quarterly* **85**, 1 (2014).
59. Z. Rizvi, V. Kunder, H. Stewart, P. Torres, S. Moon, N. Lingappa, M. Kazaleh, V. Mallireddigari, J. Perez, N. John *et al.*, The bias of physicians and lack of education in patients of color with melanoma as causes of increased mortality: a scoping review, *Cureus* **14** (2022).
60. L. Krueger, E. Hijab, J.-A. Latkowski and N. Elbuluk, Clinical decision-making bias in darker skin types: a prospective survey study identifying diagnostic bias in decision to biopsy., *International Journal of Dermatology* **62** (2023).
61. A. Ahrens, C. B. Hansen, M. E. Schaffer and T. Wiemann, ddml: Double/debiased machine learning in stata, *The Stata Journal* **24**, 3 (2024).

62. V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler and V. Syrgkanis, Applied causal inference powered by ml and ai, *arXiv preprint arXiv:2403.02467* (2024).
63. N. Dikkala, G. Lewis, L. Mackey and V. Syrgkanis, Minimax estimation of conditional moment models, *Advances in Neural Information Processing Systems* **33**, 12248 (2020).
64. T. Broderick, R. Giordano and R. Meager, An automatic finite-sample robustness metric: when can dropping a little data make a big difference?, *arXiv preprint arXiv:2011.14999* (2020).
65. T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (ACM, New York, NY, USA, 2016).
66. O. Ogungbe, A. K. Mitra and J. K. Roberts, A systematic review of implicit bias in health care: A call for intersectionality, *IMC Journal of Medical Science* **13**, 5 (2019).
67. J. Bosomworth and Z. Khan, Analysis of gender-based inequality in cardiovascular health: An umbrella review, *Cureus* **15** (2023).
68. S.-Y. Yang, E. Y. S. Woon, K. Griva and B. Y. Tan, A qualitative study of psychosocial factors in patients with knee osteoarthritis: insights learned from an asian population, *Clinical Orthopaedics and Related Research®* **481**, 874 (2023).
69. G. C. Gee, M. S. Spencer, J. Chen and D. Takeuchi, A nationwide study of discrimination and chronic health conditions among asian americans, *American journal of public health* **97**, 1275 (2007).
70. K. Kumar, R. J. Stack, A. Adebajo and J. Adams, Health-care professionals' perceptions of interacting with patients of south asian origin attending early inflammatory arthritis clinics, *Rheumatology Advances in Practice* **3**, p. rkz042 (2019).
71. M. A. Marzinke, D. N. Greene, P. M. Bossuyt, A. B. Chambliss, L. R. Cirrincione, C. R. McCudden, S. E. Melanson, J. H. Noguez, K. Patel, A. E. Radix *et al.*, Limited evidence for use of a black race modifier in egfr calculations: a systematic review, *Clinical chemistry* **68**, 521 (2022).
72. J. H. Stock, Weak instruments, weak identification, and many instruments: Part 1 and part 2 (2018), NBER Summer Institute Methods Lectures.
73. H. Weyl, Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung), *Mathematische Annalen* **71**, 441 (1912).
74. I. Montagni, T. Cariou, C. Tzourio and J.-L. González-Caballero, "i don't know","i'm not sure","i don't want to answer": a latent class analysis explaining the informative value of non-response options in an online survey on youth health, *International Journal of Social Research Methodology* **22**, 651 (2019).

Spatial Clustering for Carolina Breast Cancer Study

Hongqian Niu[1], Melissa Troester[2], and Didong Li[1,†]
[1]*Department of Biostatistics,* [2]*Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA.* [†]*E-mail: didongli@unc.edu*

In the Carolina Breast Cancer Study (CBCS), clustering census tracts based on spatial location, demographic variables, and socioeconomic status is crucial for understanding how these factors influence health outcomes and cancer risk. This task, known as spatial clustering, involves identifying clusters of similar locations by considering both geographic and characteristic patterns. While standard clustering methods such as K-means, spectral clustering, and hierarchical clustering are well-studied, spatial clustering is less explored due to the inherent differences between spatial domains and their corresponding covariates. In this paper, we introduce a spatial clustering algorithm called Gaussian Process Spatial Clustering (GPSC). GPSC leverages the flexibility of Gaussian Processes to cluster unobserved functions between different domains, extending traditional clustering techniques to effectively handle geospatial data. We provide theoretical guarantees for GPSC's performance and demonstrate its capability to recover true clusters through several empirical studies. Specifically, we identify clusters of census tracts in North Carolina based on socioeconomic and environmental indicators associated with health and cancer risk.

Keywords: Census tracts; Gaussian process; Socioeconomic status.

1. Introduction

There is growing research suggesting that socioenvironmental factors can play a key role in affecting health outcomes, potentially contributing to health disparities in marginalized groups, and may even predictably impact outcomes at the molecular level with diseases such as cancer.[1,2] However, identifying areas of such risk can be a difficult task. In the community-wide socioeconomic and environmental indicators dataset, the spatial locations of North Carolina census tracts were paired with socioeconomic data from the American Community Survey[3] from 2014 chosen to reflect socioeconomic advantage and disadvantage,[4] as well as environmental pollution data from the U.S. Environmental Protection Agency (EPA) National Air Toxics Assessment (NATA[2,5]). This then poses the problem: how can geographically spread NC census tracts be clustered together based on risk factors including socioeconomic indicators and environmental pollution? North Carolina is known to be an ethnically diverse state,[6] with a wide range of spatially dependent differences in socioeconomic status such as access to healthcare, poverty rates, and education, while meaningful clusterings must take into consideration all these differences.[6] A standard clustering algorithm applied to the data collected from the patients in each tract or to the environmental variables alone fails to necessarily capture the

significant spatial dependence inherent in the data collected in the studies. This problem is known as spatial clustering or geospatial clustering.[7]

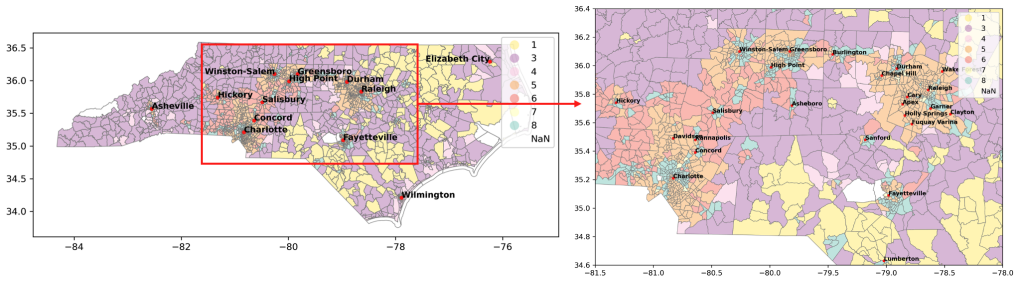


Fig. 1: Socioeconomic and environmental advantage-disadvantage latent class map of NC.

In spatial clustering, the goal is to identify clusters of similar locations based on regionalization, as well as patterns in characteristics over those locations. Clustering of geospatial data is a common unsupervised learning problem with many applications to areas, e.g., public health,[8] urban planning,[9] or transportation,[10] where geography plays an essential role.

Furthermore, spatial data, also known as geospatial data, is commonly characterized by having a distinct geographic component.[11] Unlike traditional data that only include observations as a single set of features x, spatial data may be considered as a vector $[s, x]$, where $s \in \mathbb{R}^2$ represents the spatial location of the observation and $x \in \mathbb{R}^p$ is the set of features or covariates. The analysis of such spatial datasets poses challenges, such as accurately capturing the relative effects between the spatial and covariate domains.[11] Importantly, geographically close areas may still have very different patterns of characteristics, while separated areas may share similarities and constitute a single functional cluster. Together, this can pose challenges to traditional clustering methods that equally treat the separate domains inherent to geospatial data such as K-means, as the geographic locations of distinct clusters may be well mixed, or the measurements themselves of different variables at those locations may be well mixed.

Without the spatial component, clustering itself is a well-studied problem with many established techniques such as K-means clustering,[12] spectral clustering,[13] hierarchical clustering,[14] and density-based spatial clustering of applications with noise (DBSCAN[15]), to name a few popular algorithms. Each of these algorithms offers distinct advantages based on their modeling assumptions when performed on different types of data. Additionally, common extensions of these algorithms include supervised fuzzy C-means,[16] spatial hierarchical clustering,[17] and the generalized DBSCAN (GDBSCAN[18]) algorithm. These algorithms are able to better incorporate either response labels or spatial data directly through customized distance metrics or connectivity constraints.

However, in this paper, we consider the case of supervised spatial data, with observations consisting of three components (s, x, y), where $s \in \mathbb{R}^2$ is the spatial component, $x \in \mathbb{R}^p$ is the feature component, while $y \in \mathbb{R}$ is the response variable of particular interests. Assuming that in the data there is a relationship between features x, or between features and geography (s, x), and the response y, we propose a new spatial clustering algorithm based on Gaussian Processes (GPs), called Gaussian Process Spatial Clustering (GPSC), which groups together clusters

based on each group's ability to predict the response variable y. We focus on single-output cases in this paper for simplicity, but the extension to multi-output cases where $y \in \mathbb{R}^d$ with $d > 1$ is straightforward.

For the motivating example from NC census tracts data, s is the longitude/latitude pairs defining each state census tract, x is the set of environmental pollution variables such as levels of hexane, lead, mercury, etc, as well as average socioeconomic indicators such as unemployment rates, poverty rates, or education, and the y response to be predicted is a previously defined latent class[2] measuring socioeconomic and environmental advantage-disadvantage .

In order to do so, GPSC leverages the flexibility of GPs, well-studied near-universal function approximators,[19,20] to fit the true functional relationships within each clustering and to cluster tract locations and features pertaining to socioeconomic status. Simulation studies show that the GPSC algorithm is capable of accurately recovering and clustering these functional relationships even in cases of limited spatial dependencies such as in the case of irregular cluster shapes or sizes, and regardless of any dependencies in the covariate domain. This is important because, as in Figure 1, clusters may not always be completely separated, so it is essential to control the relative influence of each domain in the clustering done in GPSC by choosing the kernel. Furthermore, GPSC is less sensitive to dependencies in the covariate domain compared to traditional clustering methods such as K-means clustering. We prove that GPSC is able to find the true clusters as long as the functional relationships between the clusters are distinct. When applied to community-wide study, GPSC successfully clusters tracts in NC with finer detail than traditional methods and can be interpreted by domain experts.

In summary, our contributions in this paper are 1) a novel spatial clustering GPSC algorithm, 2) theoretical support to GPSC and 3) application to NC tract level data with new interpretable discoveries. Full proofs of theorems, implementation details, as well as extended simulations are presented in the Supplementary Material at `https://github.com/hong-niu/gpsc-psb25`.

2. Model

2.1. *Gaussian Process Regression*

In this section, we review the GP model and its application towards regression and classification. By definition, a GP is a random function for which any finite realization follows a multivariate Gaussian distribution:[21]

Definition 2.1. *f follows GP in domain Ω with mean function μ and covariance function K, denoted by $f \sim GP(\mu, K)$, where $\mu : \Omega \to \mathbb{R}$, $K : \Omega \times \Omega \to \mathbb{R}$, if for any $x_1, \cdots, x_n \in \Omega$,*

$$[y_1, \cdots, y_n]^\top := [f(x_1), \cdots, f(x_n)]^\top \sim N(v, \Sigma),$$

where $v = [\mu(x_1), \cdots, \mu(x_n)]^\top$ and $\Sigma_{ij} = K(x_i, x_j)$.

A GP is completely determined by the mean function μ and the covariance function K, also known as the kernel. In this paper, we assume $\mu = 0$ for simplicity and use the radial basis function (RBF), also known as the squared exponential kernel, defined as: $K(x, x') = \sigma^2 e^{-\frac{d^2(x,x')}{2b}}$, but our model can be extended to other kernels. The two parameters, i.e., spatial variance σ^2 and length scale b are estimated by maximizing the likelihood (MLE). Given

training data $(x_i, y_i)_{i=1}^n$ with MLE $\theta_n = (\sigma_n^2, b_n)$ and a new observation x_*, the best unbiased linear predictor (BLUP[22]) of $y_* = f(x_*)$ is given by $\widehat{y}_* = K_{\theta_n}(x_*, X)K_{\theta_n}(X, X)^{-1}Y$, where $K_{\theta_n}(x_*, X)_i = K_{\theta_n}(x_*, x_i)$, $K_{\theta_n}(X, X)_{ij} = K_{\theta_n}(x_i, x_j)$ and $Y = [y_1, \cdots, y_n]^\top \in \mathbb{R}^n$. As a flexible regression algorithm, GP can be modified into a classifier using a link function[21] for a discrete response variable y, so we will not distinguish between Gaussian process regression (GPR) and Gaussian process classification (GPC) in this paper.

2.2. *GP Spatial Clustering*

Now we will consider observations $\{(s_i, x_i, y_i)\}_{i=1}^n$, where $s_i \in \mathcal{S} \subset \mathbb{R}^2$ is the spatial location, $x_i \in \Omega \in \mathbb{R}^p$ is the covariate, and y_i is the response variable. Let $l_i \in \{1, \cdots, L\}$ be the unobserved cluster label such that $l_i = j \iff s_i \in \mathcal{S}_j \subset \mathcal{S}$, where $\mathcal{S}_1, \cdots, \mathcal{S}_L$ is a partition of Ω. We focus on the following model. $y_i = \sum_{j=1}^L \mathbf{1}_{\{s_i \in \mathcal{S}_i\}} f_j(x_i) = \sum_{j=1}^L \mathbf{1}_{\{l_i = j\}} f_j(x_i)$, where f_j is unknown function on Ω in certain function class that will be discussed in Section 3. That is, the functional relation between y_i and x_i varies across spatial clusters supported by $\mathcal{S}_i$. The goal is to recover the cluster label l_i, called spatial clustering since the clusters are rooted in the spatial domain $\mathcal{S}$.

For example, in the NC tracts data, each $\mathcal{S}_i$ consists of tracts in NC, while the relationship between the latent class and the socioeconomic and environmental covariates varies across the tracts spatially. The goal is to partition NC into several clusters so that each cluster admits a unique functional relationship.

For a given observation x_i in cluster j with response y_i, we expect the prediction error of f_j to be the lowest among all f_j's, and hence we can assign x_i to the cluster with the lowest prediction error. However, neither the cluster label l_i or domain partition $\mathcal{S}_i$, nor the functions f_j is observed. Motivated by the flexibility of GP models, we use GP to approximate the unobserved functions f_j, denoted by $\widehat{f}_j$, and assign x_i to the cluster labeled by $\widehat{l}_i$ with the lowest prediction error: $\widehat{l}_i = \text{argmin}_j \ (\widehat{f}_j(s_i, x_i) - y_i)^2$. Then we update the cluster and $\widehat{f}_j$ iteratively. The GPSC algorithm is summarized in algorithm 1.

Algorithm 1 Gaussian Process Spatial Clustering

Input: data $(s_i, x_i, y_i)_{i=1}^n$, number of clusters L, maximum number of iterations T
Initialize $\hat{l}_i = \text{randomInt}(1, 2, \cdots, L)$
for $t = 1$ **to** T **do**
  for $j = 1$ **to** L **do**
    $(S_j, X_j, Y_j) = \{(s_i, x_i, y_i) : \widehat{l}_i = j\}$, $\widehat{f}_j = \text{GPR}(([S_j, X_j], Y_j))$
  end for
  for $i = 1$ **to** n **do**
    $\hat{l}_i = \text{argmin}_j \ (\widehat{f}_j((s_i, x_i)) - y_i)^2$
  end for
end for

In this flexible construction, it is also possible to extend the reassignment function for different applications, such as reinforcing spatial contiguity constraints as is common in

geographical clustering:

$$\widehat{l}_i = \underset{j=1,\cdots,L}{\operatorname{argmin}} \{(\widehat{f}_j(s_i, x_i) - y_i)^2 + \lambda \|s_i - C_j\|\}.$$

Here, C_j is the center in the spatial domain of the current cluster $\mathcal{S}_j$, while λ is a tuning parameter that controls the penalization of assigning points to clusters that are spatially distant. For the rest of the paper, we will focus on the case $\lambda = 0$, but will demonstrate the effects of adding such penalties in the simulation studies.

In summary, the inputs to the algorithm are observations $\{(s_i, x_i, y_i)\}_{i=1}^n$, along with tuning parameters including the number of iterations T and the number of clusters L. In practice the number of iterations T need not necessarily be large, and can be replaced with the stopping criterion when the cluster assignments stabilize. The proper choice of the number of clusters L is a typical challenge in the field of clustering,[23] which is beyond the scope of this paper. The choice of L often requires domain expertise specific to the application at hand, see Section 5 for more detailed discussion. In practice, we also typically bound the parameters of the covariance function during optimization to prevent overfitting.

3. Theory

In this section, we provide theoretical support to the GPSC algorithm. We start with the necessary definitions to state the assumptions and theorems.

Definition 3.1. *Let K be a positive definite kernel on $\Omega \subset \mathbb{R}^p$, then $\mathcal{F}_K(\Omega) \coloneqq \operatorname{span}\{K(\cdot, x) : x \in \Omega\}$ with inner product form $\left(\sum_{i=1}^n a_i K(\cdot, x_i), \sum_{j=1}^m b_j K(\cdot, \widetilde{x}_j)\right)_K \coloneqq \sum_{i,j} a_i b_j K(x_i, \widetilde{x}_j)$, so that $\mathcal{F}_K(\Omega)$ is a pre-Hilbert space with a reproducing kernel K. The linear mapping $\Phi : \mathcal{F}_K(\Omega) \to C(\Omega) : \Phi(f)(x) \coloneqq (f, K(\cdot, x))_K$, is injective. Then the image of Φ, $\mathcal{N}_K(\Omega) \coloneqq \Phi(\mathcal{F}_K(\Omega))$ is a Hilbert space with a reproducing kernel K equipped with the inner product $(f, g)_K \coloneqq (\Phi^{-1} f, \Phi^{-1} g)_K$.*

For simplicity, we fix K_θ to be the RBF kernel with $\theta = (\sigma^2, b)$ from now on.

Definition 3.2. *Given observations X and x_0 with unobserved y_0 to be predicted. Let $\psi_{X,x_0} : Y \mapsto K_{\theta(Y)}(x_0, X)^\top K_{\theta(Y)}(X, X)^{-1} Y$, where $\theta(Y) = \operatorname{argmax}_\theta N(Y|0, K(X, X))$ is the maximum likelihood estimator of θ based on potential observations Y. That is, ψ is the BLUP of y_0 based on observations (X, Y). By the definition of ψ, the smoothness of the Gaussian density function and the linearity of BLUP, ψ is differentiable.[22] We also introduce the following assumptions:*

(A1) $\Omega \subset \mathbb{R}^p$ is compact and $p(x) > 0,\ \forall x \in \Omega$, where $p(x)$ is the density function of x.
(A2) $f_j \in \mathcal{N}_K(\Omega),\ j = 1, \cdots, L$.

Theorem 3.3. *Under assumptions (A1)-(A2), at any iteration in Algorithm 1, let $n_{jk} \coloneqq \left|\{i : l_i = j, \widehat{l}_i = k\}\right|$, $n_j \coloneqq \left|\{i : \widehat{l}_i = j\}\right|$ then the current x_i is a assigned to the correct cluster if for any $k \neq j$,*

$$\frac{\sum_{m \neq j} n_{mj}}{\sum_{m \neq j} n_{mk}} < \frac{D_l E_l}{D_u E_u} - \frac{\|f\|_K e^{-c_1 n_j^{\frac{1}{p}}} + \|f\|_K e^{-c_2 n_k^{\frac{1}{p}}}}{D_u E_u n_{22}}, \tag{1}$$

where c_1 and c_2 are constants, and

$$D_l := \inf \|\nabla\psi(Y)\| \le D_u := \|\nabla\psi(Y)\|_\infty,$$
$$E_l := \inf_{x\in\Omega, j,k=1,\cdots,L} |f_j(x) - f_k(x)| \le E_u := \sup_{x\in\Omega, j,k=1,\cdots,L} |f_j(x) - f_k(x)| < \infty.$$

In particular, let $L = 2$, $j = 1$, $k = 2$ and let $n_1, n_2 \to \infty$, Equation (1) becomes: $\frac{n_{21}}{n_{22}} < \frac{D_l E_l}{D_u E_u}$. That is, the mis-clustered proportion is small enough.

The right-hand side of inequality (1) is highly interpretable. The ratio $\frac{D_l}{D_u}$ measures the robustness of the BLUP, that is, how the BLUP changes with training data Y. The less robust the BLUP, the smaller the ratio, and the harder it is to find the correct clusters. The ratio $\frac{E_l}{E_u}$ measures the separation between functions $f_1, \cdots, f_L$. The smaller the separation, the smaller the ratio, and the harder it is to find the correct clusters. Theorem 3.3 also implies that the state of correct clustering is an absorbing state, that is, if the current clusters are close enough to the true clusters, then perfect clustering results will be achieved in the next iteration. Note that even if the inequality does not hold, the algorithm may still converge to a better state with more correctly clustered data, although not within one single step. This is because even when the right-hand side of Equation (1) is small, there might be some region $\Omega_0 \subset \Omega$ where the f_j's are relatively well separated so that the right-hand side is relatively large on Ω_0, so that samples within Ω_0 will be assigned to true clusters. Meanwhile, for the region where f_j's are well mixed, it is challenging for all clustering algorithms.

In practice, the response variable y is often subject to measurement error, leading to a more realistic model: $y = f(x) + \epsilon$, where $\epsilon \sim N(0, \tau^2)$ represents noise. The following theorem serves as the counterpart to Theorem 3.3 in the presence of Gaussian noise:

Theorem 3.4. *Under the same assumption and notation as of Theorem 3.3, with the addition of Gaussian noise, the current x_i is assigned to the correct cluster if for any $k \neq j$,*

$$\frac{\sum_{m\neq j} n_{mj}}{\sum_{m\neq j} n_{mk}} < \frac{D_l E_l}{D_u E_u} - \frac{\|f\|_K e^{-c_1 n_j^{\frac{1}{p}}} + \|f\|_K e^{-c_2 n_k^{\frac{1}{p}}} + \xi}{D_u E_u n_{22}}, \tag{2}$$

where ξ is the sum of independent χ-distributions with degrees of freedom $1, n_1$ and n_2 rescaled by 2τ, τ and τ respectively.

In particular, when $L = 2$, $j = 1$, $k = 2$, and $n_1, n_2 \to \infty$, the right-hand side simplifies to $\frac{D_l E_l}{D_u E_u}$ with probability one. When $\tau = 0$, that is, the noise vanishes, then $\xi = 0$ so Theorem 3.4 coincides with Theorem 3.3.

4. Simulation Studies

To evaluate the performance of GPSC, we present three simulation studies in this section, with detailed implementation details in the Supplementary Materials. The first simulation will demonstrate an application of Algorithm 1 in the case of responses generated by linear functions with two clusters, while the second simulation shows the performance of GPSC in the case of responses generated by nonlinear functions. The third simulation shows the robustness of GPSC to noisy data and overspecified number of clusters. In all simulations, we

compare the performance of GPSC with traditional clustering algorithms: K-means, spectral clustering, hierarchical clustering, and DBSCAN, as well as spatial or supervised analogs: supervised fuzzy C-means, spatial hierarchical clustering, generalized GDBSCAN, and also the Gaussian mixture model (GMM[24]). We evaluate the performance using the adjusted Rand index (ARI[25]) and adjusted mutual information (AMI[26]) against the true labels. The data used in these simulations take the form $\{(s_i, x_i, y_i)\}_{i=1}^n$, where $s_i \in \mathbb{R}^2$ is the spatial domain, $x_i \in \mathbb{R}^2$ is the covariate domain, and $y_i \in \mathbb{R}$ is the response domain, taken for visualization purposes. Note that for all algorithms, including GPSC and the aforementioned traditional, nonspatial clustering algorithms, the input is taken to be the full vector (s, x, y) with the spatial domain included, so that all competitors always use the full information. The results can be directly extended to higher p and multivariate responses.

4.1. *Simulation 1 - Linear Functions*

In this simulation, y is a linear function of x for visualization purposes, where both s_i and x_i are generated from independent uniform distributions. After generating the data $\{(s_i, x_i)\}_{i=1}^n$, the spatial domain is subdivided into two clusters, the center ball and the background region. The $y_i \in \mathbb{R}$ are then generated as distinct linear functions of x_i for each cluster. For visualizations of the resulting clusters in the XY domain and all ARI/AMI scores, see Supplement D.1.

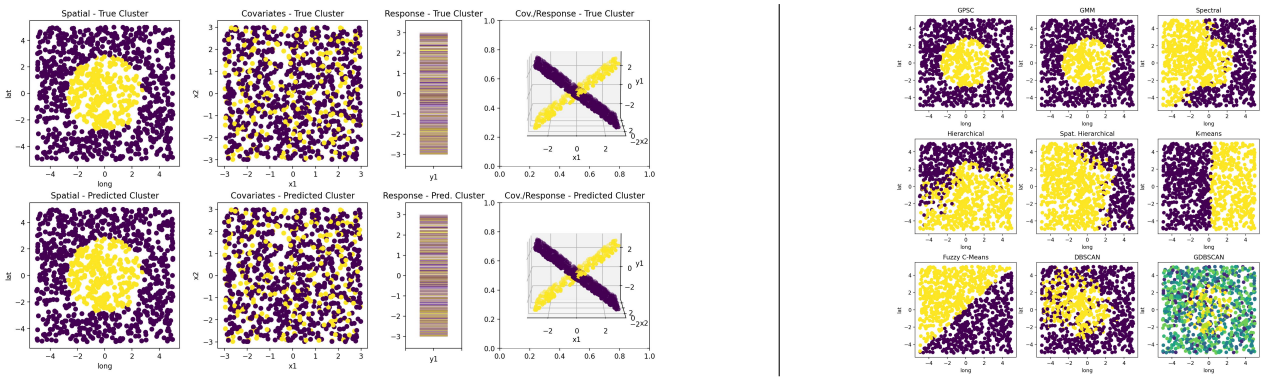


Fig. 2: [Left] GPSC results for Simulation 1, colored by cluster. The first column plots the spatial domain $s_i \in \mathbb{R}^2$, the second column plots the covariate space $x_i \in \mathbb{R}^2$, the third column plots the response space $y_i \in \mathbb{R}$, while the right-most column plots $y_i \in \mathbb{R}$ against $x_i \in \mathbb{R}^2$. The first row shows the ground truth generated data. The second row shows the predicted clusters from GPSC after randomized initialization. [Right] Clusters for Simulation 1 by nine clustering algorithms visualized in the spatial domain.

It can be seen that this simulation is challenging for several reasons. First, there is almost no separation considering any dimension s, x, or y on its own as in the first three columns in Figure 2 (left); the separation is solely in the functional domain XY. As a result, most traditional algorithms cannot capture this functional relationship, as supported by Panels 3-7 in Figure 2 (right). Although it can seen that the Gaussian mixture model is able to rediscover the clusters in this case (Panel 2), this is due to GMM's ability to estimate the pairwise

linear correlation between each domain. However, we expect GMM to fail to capture nonlinear functional relationships, as shown in the following Simulation 2. It is also noted that DBSCAN and GDBSCAN (Panels 8 and 9) also perform reasonably well, but have challenges of their own such as GDBSCAN greatly overestimating the number of clusters.

4.2. *Simulation 2 - Nonlinear Functions*

In this simulation, we will show that in an irregular spatial distribution with nonlinear relationships between the covariates and the response variable, GPSC is still able to recover the true functional relationships in contrast to the competitors. After generating the data $\{(s_i, x_i)\}_{i=1}^n$ from independent uniform distributions, the spatial domain is subdivided into two clusters, the ring and the background region. The $y_i \in \mathbb{R}$ are then generated as distinct nonlinear functions of x_i for each cluster (the first row of Figure 3).

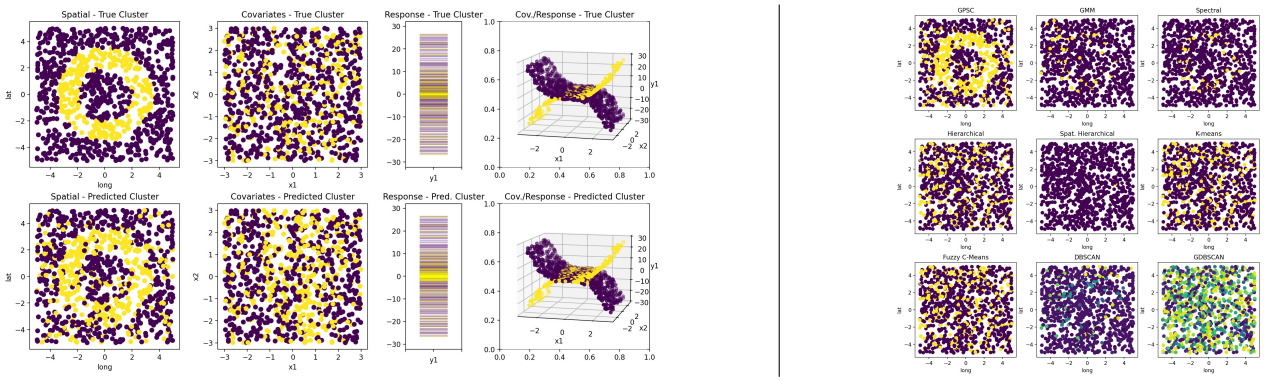


Fig. 3: [Left] Results for Simulation 2 with true generated data (top) and results of GPSC (bottom). [Right] Clusters by nine different algorithms visualized in the spatial domain.

It can be seen that in this more challenging simulation, only GPSC is able to recover the true functional clusters, with the results of each clustering algorithm plotted in the spatial domain in Figure 3 (see Supplement D.2 for more details).

4.3. *Simulation 3 - Model Robustness*

In Simulation 3, we present a more realistic scenario of three clusters that have some degree of spatial separation. Motivated by our real-world application of clustering North Carolina census tracts, the sun and moon clusters could be interpreted to represent two urban centers surrounded by a larger rural region. By applying the spatially penalized version of GPSC, we will show that the clustering results remain stable across both increasing levels of noise, as well as to overspecification of the input number of clusters. Full visualization and comparisons can be found in Supplement D.3, D.4 and D.5.

After generating the data $\{(s_i, x_i)\}_{i=1}^n$ from independent uniform distributions, the spatial domain is subdivided into the three clusters, the sun and moon shape, and the background region. The $y_i \in \mathbb{R}$ are then generated as distinct nonlinear functions of x_i for each cluster with varying degrees of zero-mean Gaussian noise. For an extension of Simulation 3 to nonlinear

functions of both s_i and x_i, see Supplement D.5.

Noisy Responses

We first show that GPSC works under noisy conditions as per Theorem 3.4. In Figure 4, we present Simulation 3 with noise variance = 100, showing that the spatially penalized version of GPSC still performs well under noisy conditions. In particular, GPSC is able to outperform competitors at all tested noise levels, where no other competitor is able to recover the true clusters (with exact ARI/AMI scores and additional details in Supplement D.3).

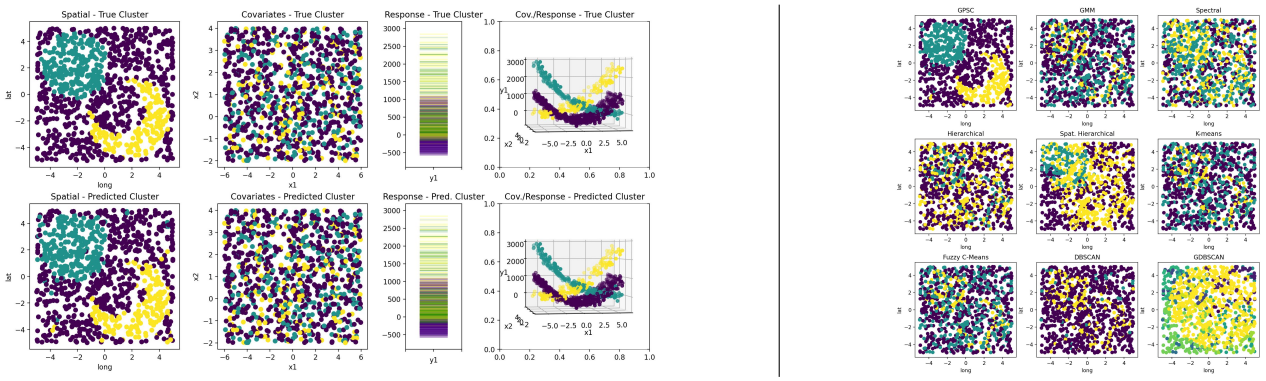


Fig. 4: [Left] Results for Simulation 3 with true generated data (top) and results of GPSC (bottom). [Right] Clusters by nine different algorithms visualized in the spatial domain.

Overspecified Number of Clusters

Finally, we show that GPSC is stable when the number of clusters is overspecified. Specifically, it can be seen in Figure 5 when the number of specified clusters is 5, the sun (teal) and moon (yellow) clusters remain stable, while the background cluster (originally purple) is split into three purple, indigo, and light green clusters. In contrast, the competitors are unable to recover the true clusters when the number of clusters are overspecified, while further visualizations and comparisons to the competitor models are presented in Supplement D.4.

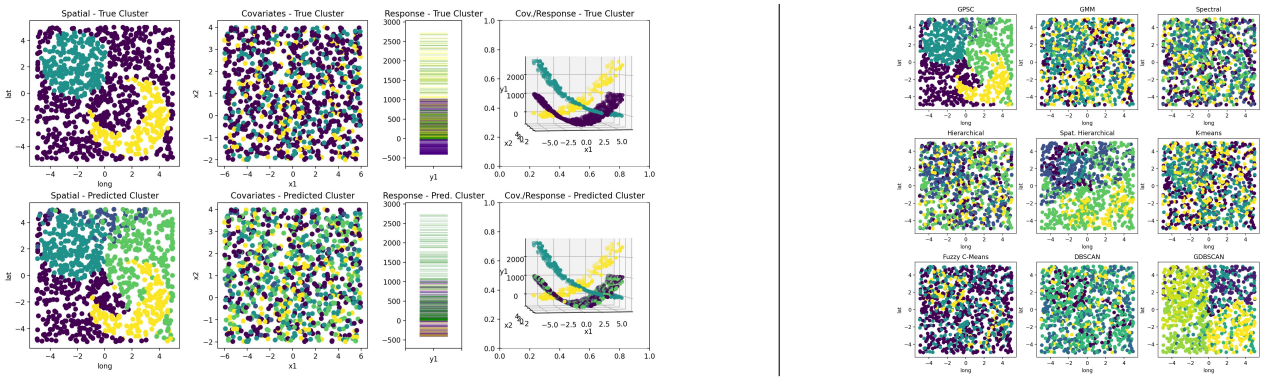


Fig. 5: [Left] GPSC results for Simulation 3 with overspecified number of clusters as 5. [Right] Results of nine algorithms with overspecified input presented in the spatial domain.

5. Applications to NC Tract Data

This dataset consists of 29 community-wide covariates aggregated by census tracts in North Carolina. Such covariates ranged from measures of environmental pollution to averages of socioeconomic indicators such as unemployment, housing environment, education, etc (see Supplement E for a full list). Each census tract is associated with a single (longitude, latitude) pair of coordinates. The overall socioeconomic indicators were previously aggregated using latent class analysis into a single advantage/disadvantage class with 8 categories.[2]

Based on the distribution of the full latent classes seen in Figure 1, we can see that there is some degree of separation in the spatial domain between certain groups. Thus, we initialized our GPSC algorithm by performing traditional K-means clustering on solely the spatial domain. We then applied our GPSC algorithm using this latent class as the response variable, taking all other features as the set of covariates. Here, we focus on K-means clustering for comparison due to its interpretable results from previous studies,[2] with results from other clustering algorithms presented in Supplement E. Based on our results, we find that $L = 3$ produced the most interpretable clusters, and thus aggregated the 8 latent classes into 3 as a baseline against GPSC seen in Figure 6. Using the language of Larson et al. (2020)[2] for our predicted 3 clusters, we will consider the overall socioeconomic and environmental advantage to be three levels: low (pink), medium (gray), and high (green).

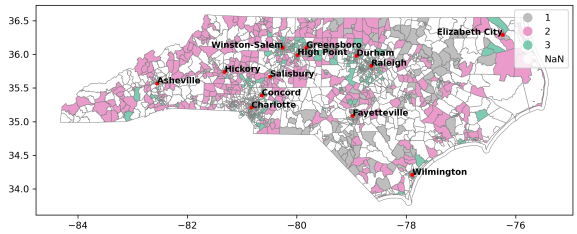


Fig. 6: Baseline aggregate groups of socioeconomic and environmental latent class indicator.

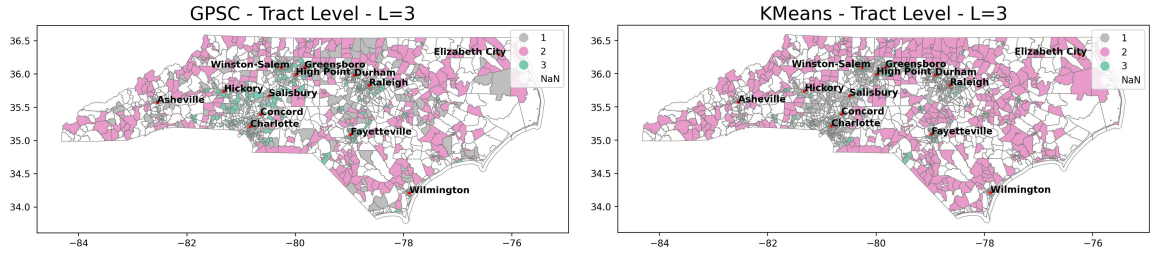


Fig. 7: Clusters by GPSC and K-means for tract data, interpreted as overall socioeconomic and environmental advantage between levels of low (pink), medium (grey), and high (green).

At first glance, the general spatial distribution of our GPSC and K-means algorithms tends to agree. However, the GPSC predicted clusters differ from K-means and baseline in several meaningful ways. First, in the central region depicted in the first row of Figure 8, GPSC identifies more areas of high advantage (green). Notably, this includes the area surrounding cities such as Chapel Hill, Cary, and the capital city Raleigh (Research Triangle Park), as well as Greensboro and High Point (the Piedmont Triad), which are known to be wealthier and more urbanized regions of the state, whereas the K-means algorithm puts tracts within this region in the medium (gray) advantage group.

Towards the edges of the state we can also see significant differences as the GPSC algorithm tends to further differentiate tracts around the extremities between low and medium advantage. Most notably, around Asheville and Wilmington, two more prominent cities in North Carolina, we are able to distinguish further differences between low and medium advantage tracts, as seen in the second and third rows in Figures 8. Considering the ARI and AMI scores between the two clusterings, we find the scores to be both 0.002, suggesting that clusterings, despite visually seeming to separate the tracts spatially in similar patterns, are actually very different. One challenge of K-means clustering when determining the original 8 latent classes[2] was

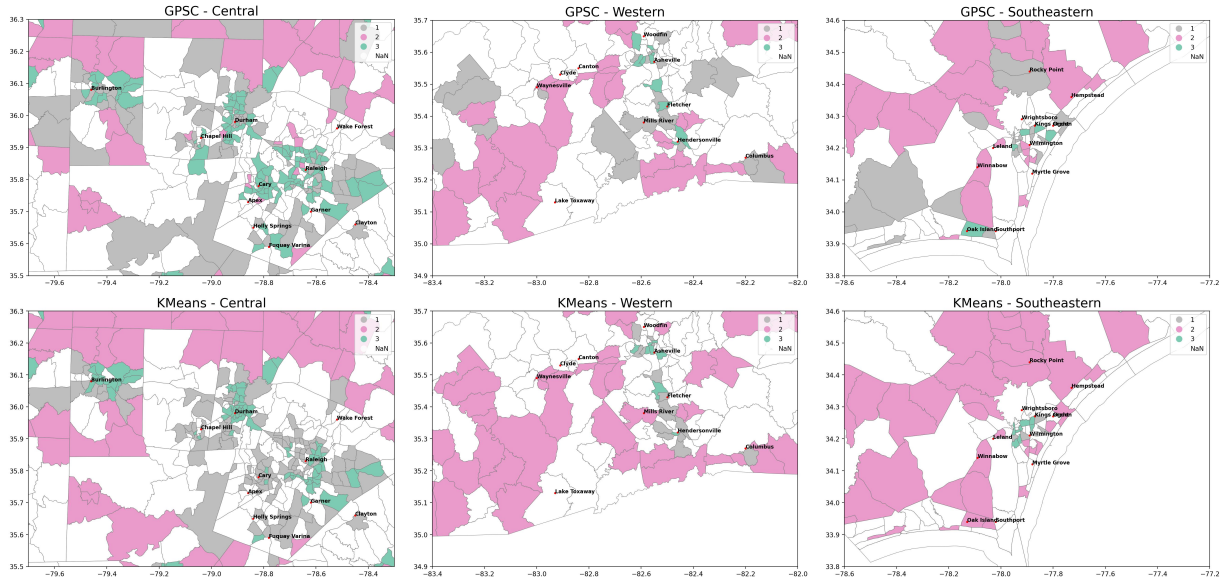


Fig. 8: GPSC and K-means cluster results for NC tracts. Column 1: Central NC; Column 2: Western NC (Asheville); Column 3: Southeastern NC (Wilmington)

a potential lack of finer detail from the K-means predicted clusters. However, here we have shown that despite using the same $L = 3$ clusters, GPSC is able to further differentiate between areas of low and medium disadvantage, in less dense areas of the state along the coast and the western region. Furthermore, there is reason to believe that not all 8 classes are necessary to describe the different advantage groups. In the original grouping, the latent class 2 is actually an empty group, as seen in Figure 1. Thus, the results from GPSC in comparison to K-means and baseline suggest that the algorithm is able to better balance nuance against a traditional clustering algorithm, while also retaining simpler interpretability by using fewer clusters.

6. Discussion

Spatial clustering offers unique challenges in comparison to traditional clustering problems due to the spatial domain inherent to geographic data. In our application, the census tract data have distinctly different properties compared to the measured covariates over the tracts. In this paper, we propose a GP-based clustering algorithm and demonstrate its performance in both simulation studies and a real data application. The advantages of GPSC include being

able to capture the relative effects between the spatial domain and the measured covariates, largely independent of intersections in the covariate domain as long as the clustered functions themselves have some degree of separation. We also provide theoretical guarantees to the convergence of GPSC and extend it to noisy settings.

GPSC can also be highly scalable; the complexity of the algorithm stems from the fitting of each GP in each iteration, where standard Gaussian processes regression is $O(n^3)$ in the size of the input. In our case, we applied a standard Gaussian process regression model from the scikit-learn package[27] since our sample size was relatively small. However, in cases of large sample size, scalable GP methods can be applied for a reduction in runtime to $O(n \log n)$.[28] The GPSC model also has few tuning parameters, notably the number of clusters, optional spatial penalty for data thought to contain spatially contiguous clusters, and and can also be highly flexible through the choice of GP kernel. Although the form of our theorem is independent of the specific choice of kernel (only the convergence rate will differ), in practice more nuanced anisotropic or nonstationary kernels may be more suitable for datasets with strong heterogeneity, for which the actual design of such kernels remains an open problem.

In the real-world application, we applied GPSC to a North Carolina socioeconomic and environmental indicator dataset and found distinct patterns of advantage-disadvantage across the state that captured finer details around the less dense outer regions of the state in comparison to K-means and other clustering methods (presented in Supplement E), while our method also offered simpler interpretability than previous analysis. When utilized by domain experts, the goal of the results of these models is to supplement the identification of marginalized communities, which could be targeted with interventions. Furthermore, in context of our long-term goal of designing interventions, ensuring the accuracy of these models is also of high ethical importance. Therefore in our case, before any application, we can perform sensitivity analyses that tile the geographic region with alternative regional classifiers (county, AHEC region, latitude and longitude tiles of uniform size) to confirm that the same areas arise in multiple boundary definitions. This will confirm that the boundary definitions are not driving artifactual associations. More broadly, it is important that in these high-stakes applications we do not over-rely on any one method. We envisage the possibility of using these clustering results (and GPSC in general) as a supplementary tool for experts to potentially better identify marginalized communities and areas that may be otherwise overlooked.

Acknowledgements

The authors would like to thank the Carolina Breast Cancer Study for the data, which is supported by the Susan G Komen Foundation (OGUNC1202, OG22873776, SAC210102, TREND21686258), National Cancer Institute (R01CA253450), the National Cancer Institute Specialized Program of Research Excellence (SPORE) in Breast Cancer (NIH/NCI P50-CA058223), the Breast Cancer Research Foundation (HEI-23-003), the US Department of Defense (HT94252310235), NIH (P30CA016086) via Rapid Case Ascertainment of NC Central Registry and UNC Lineberger, University of North Carolina BioSpecimen Processing Facility (http://bsp.web.unc.edu/), and the Breast Cancer Research Foundation (HEI-23-003). HN, MT and DL were supported by NIH grant P30ES010126.

References

1. B. D. Lord, A. R. Harris and S. Ambs, The impact of social and environmental factors on cancer biology in black americans, *Cancer Causes & Control* , 1 (2022).
2. A. Larsen, V. Kolpacoff, K. McCormack, V. Seewaldt and T. Hyslop, Using latent class modeling to jointly characterize economic stress and multipollutant exposure, *Cancer Epidemiology, Biomarkers & Prevention* **29**, 1940 (2020).
3. American community survey, *U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies* (2014).
4. A. Palumbo, Y. Michael and T. Hyslop, Latent class model characterization of neighborhood socioeconomic status, *Cancer Causes & Control* **27**, 445 (2016).
5. National air toxics assessment, *U.S. Environmental Protection Agency* (2014).
6. M. A. Emerson, Y. M. Golightly, X. Tan, A. E. Aiello, K. E. Reeder-Hayes, A. F. Olshan, H. S. Earp and M. A. Troester, Integrating access to care and tumor patterns by race and age in the Carolina Breast Cancer Study, 2008–2013, *Cancer Causes & Control* **31**, 221 (2020).
7. J. Aldstadt, Spatial clustering, in *Handbook of applied spatial analysis*, (Springer, 2010) pp. 279–300.
8. O. Fonseca-Rodríguez, P. E. Gustafsson, M. San Sebastián and A.-M. F. Connolly, Spatial clustering and contextual factors associated with hospitalisation and deaths due to covid-19 in sweden: a geospatial nationwide ecological study, *BMJ Global Health* **6**, p. e006247 (2021).
9. X. Tu, C. Fu, A. Huang, H. Chen and X. Ding, Dbscan spatial clustering analysis of urban "production–living–ecological" space based on poi data: a case study of central urban wuhan, china, *International Journal of Environmental Research and Public Health* **19**, p. 5153 (2022).
10. H. Yue, Q. Guan, Y. Pan, L. Chen, J. Lv and Y. Yao, Detecting clusters over intercity transportation networks using k-shortest paths and hierarchical clustering: a case study of mainland china, *International Journal of Geographical Information Science* **33**, 1082 (2019).
11. S. Kisilevich, F. Mansmann, M. Nanni and S. Rinzivillo, Spatio-temporal clustering, in *Data mining and knowledge discovery handbook*, (Springer, 2009) pp. 855–874.
12. J. MacQueen, Classification and analysis of multivariate observations, in *5th Berkeley Symp. Math. Statist. Probability*, 1967.
13. J. Shi and J. Malik, Normalized cuts and image segmentation, *IEEE Transactions on pattern analysis and machine intelligence* **22**, 888 (2000).
14. F. Nielsen, Hierarchical clustering, in *Introduction to HPC with MPI for Data Science*, (Springer, 2016) pp. 195–211.
15. M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, A density-based algorithm for discovering clusters in large spatial databases with noise., in *KDD*, (34)1996.
16. E. Yasunori, H. Yukihiro, Y. Makito and M. Sadaaki, On semi-supervised fuzzy c-means clustering, in *2009 IEEE International Conference on Fuzzy Systems*, 2009.
17. A. X. Y. Carvalho, P. H. M. Albuquerque, G. R. de Almeida Junior and R. D. Guimaraes, Spatial hierarchical clustering, *Revista Brasileira de Biometria* **27**, 411 (2009).
18. J. Sander, M. Ester, H.-P. Kriegel and X. Xu, Density-based clustering in spatial databases: The algorithm gdbscan and its applications, *Data mining and knowledge discovery* **2**, 169 (1998).
19. H. Wendland, *Scattered data approximation* (Cambridge university press, 2004).
20. S. Ghosal and A. Van der Vaart, *Fundamentals of nonparametric Bayesian inference* (Cambridge University Press, 2017).
21. C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning* (MIT press Cambridge, MA, 2006).
22. M. L. Stein, *Interpolation of spatial data: some theory for kriging* (Springer Science & Business Media, 1999).

23. B. Mirkin, Choosing the number of clusters, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**, 252 (2011).
24. N. E. Day, Estimating the components of a mixture of normal distributions, *Biometrika* **56**, 463 (1969).
25. D. Steinley, Properties of the hubert-arable adjusted Rand index., *Psychological methods* **9**, p. 386 (2004).
26. A. Strehl and J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, *Journal of machine learning research* **3**, 583 (2002).
27. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825 (2011).
28. H. Liu, Y.-S. Ong, X. Shen and J. Cai, When gaussian process meets big data: A review of scalable gps, *IEEE transactions on neural networks and learning systems* **31**, 4405 (2020).
29. J. Warner and J. Sexauer, scikit fuzzy, twmeggs, alexsavio, a, *Unnikrishnan, G. Castelão, FA Pontes, T. Uelwer, pd2f, laurazh, F. Batista, alexbuy, WV den Broeck, W. Song, TG Badger, RAM Pérez, JF Power, H. Mishra, GO Trullols, A. Hörteborn, and* **99991**.

Assessment of Drug Impact on Laboratory Test Results in Hospital Settings

Victorine P. Muse and Amalie D. Haue
The Disease Systems Biology Group, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B Copenhagen, 2200, Denmark
Email: victorine.muse@cpr.ku.dk; adh@cpr.ku.dk

Cristina L. Rodríguez, Alejandro A. Orozco, and Jorge H. Biel
The Disease Systems Biology Group, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B Copenhagen, 2200, Denmark

Søren Brunak
The Disease Systems Biology Group, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B Copenhagen, 2200, Denmark
Email: soren.brunak@cpr.ku.dk

Patients experiencing adverse drug events (ADE) from polypharmaceutical regimens present a huge challenge to modern healthcare. While computational efforts may reduce the incidence of these ADEs, current strategies are typically non-generalizable for standard healthcare systems. To address this, we carried out a retrospective study aimed at developing a statistical approach to detect and quantify potential ADEs. The data foundation comprised of almost 2 million patients from two health regions in Denmark and their drug and laboratory data during the years 2011 to 2016. We developed a series of multistate Cox models to compute hazard ratios for changes in laboratory test results before and after drug exposure. By linking the results to data from a drug-drug interaction database, we found that the models showed potential for applications for medical safety agencies and improved efficiency for drug approval pipelines.

Keywords: adverse drug events, polypharmacy, electronic patient records, population-wide data

1. Introduction

1.1. *Electronic health record data to overcome health disparities in precision medicine*

Population-wide electronic health record (EHR) data present an important source to overcome health disparities in precision medicine. Adverse drug events (ADEs) describe known and yet unknown effects of a drug that may be due to undiscovered drug effects in specific population subgroups, or due to an unexpected interaction with one or more additional drugs. This is a particular area of interest within pharmacovigilance since most drugs are only clinically tested as monotherapies and additionally mostly on healthy men.[1,2]

As such, population-wide EHR data present an important source to identify potential ADEs among users of healthcare irrespective of e.g. co-morbidity burden and socioeconomic status. Models for detection of ADEs based on EHR data could overcome health disparities in precision medicine by identification of potential ADEs in real-world settings. Generally, only 10% of ADEs are reported and several studies have stated that up to 30% of ADE-related hospital admissions are preventable.[3–7] As the risk of ADE increases with the burden of polypharmacy, the phenomenon translates into an additional risk in multi-morbid patients. Therefore, methods for detection of ADEs are an integral aspect of overcoming health disparities in precision medicine.

Denmark's comprehensive laboratory, pharmaceutical, and disease registries represent a unique opportunity to explore associations between polypharmaceuticals, laboratory data and potential ADEs (pADEs).[8,9] To our knowledge, only one study, conducted in South Korea, has directly investigated ADE risk within drug-laboratory test pairs, but with the goal of identifying new signals for known ADR events.[10] We present a potential strategy for large-scale monitoring of drug effects when administered in combinations, which is of increasing interest in ageing, multi-morbid populations.[11–13]

2. Materials and Methods

2.1. *Data Availability and Sources*

Population-wide laboratory healthcare and pharmaceutical data from two of the five Danish healthcare regions (approximately 50% of the entire population) were collected and processed for this study. Data covered all in-patient, out-patient, and emergency room settings at public hospitals, in total 1,987,180 patients. As only 1% of healthcare costs originate from private hospitals in Denmark, these data were considered population-wide.[14] Due to Denmark's person identification system (initiated in 1968) we were able to completely link data records across hospitals and data sets, fully integrating the laboratory and pharmaceutical data for the study cohort.[15,16]

We defined the study period from 2011-10-28 to 2016-06-30 corresponding the period where all relevant hospital system data overlapped in an ideal manner. The cohort was further reduced, removing tourists and other short-term residents with unknown study exit dates and standing. It is important to note, that as of 2014 it was possible to legally change your person identification number to reflect a change in self-identified gender (restricted to a binary system of male/female), our data reflects the gender that was legally registered at the time of hospital interaction.

2.2. *Laboratory Healthcare Data*

The processed laboratory data applicable in this study consisted of 1,924,869 patients and 310,455,299 laboratory measurements. These data were systematically cleaned and conformed to a more centralized naming and coding system of which is thoroughly described in Muse et al.[17,18] In summary, typos and symbols (=, >, < etc.) were removed or corrected and naming systems were conformed and translated to English.[19–23] Typical test coding systems made use of the Nomenclature, Properties, and Units (NPU) classifications as is common in the Nordic countries. Failed or incomplete tests were removed from the data set.[19]

Tests were categorized as within range, normal (0), or out of range, abnormal (-1 or 1), based on the test result within national health authorities reference intervals that are calculated to be the 95% confidence interval for healthy patients. Tests were labelled as "-1" if the value was below the reference interval or "1" if they were above it, or otherwise abnormal for binary testing only (positive vs. negative). This labelling system was used throughout the study and in figures to distinguish an adverse change when the laboratory test in question was decreased (-1) or increased (1) in relation to the relevant reference interval. A unique laboratory test was defined as a unique analyte taken from a unique source: for example, B-LEUKOCYTES-1 indicates leukocytes taken from blood were abnormally low in relation to the given reference interval that may be age-dependent for some tests.

2.3. *Pharmaceutical Data*

The pharmaceutical data were matched to the timeframe and patient IDs of the laboratory data. The pharmaceutical data used for this study is the first and last date of the confirmed administration of the drug and the respective Anatomical Therapeutic Chemical (ATC) group classification code. The data set only included drugs administered at the hospital (in-patient, out-patient, and emergency room data).[24,25] ATC codes are alpha-numeric codes used internationally as a tool for drug utilization monitoring and research. The codes are formula specific meaning that ATC codes may be the same for certain drugs, even though the route of administration differs. For example, the ATC code for the antibiotic drug moxifloxacin is J01MA14 irrespective of route of administration, e.g. orally or intravenously.[26]

ATC codes are seven characters long (representing the anatomical main group, therapeutic group, pharmacological subgroup, chemical subgroup, and chemical structure, respectively), but may in practice be registered with fewer characters. This study made use of the most specific codes available (i.e. preferably ATC codes containing seven characters) as to pinpoint possible relevant drug mechanisms. Dosing information was available, but not used or needed for the purpose of this study.

2.4. *Multistate Cox Model, Monotherapies*

The core model of the study was developed using a multistate approach Cox model.[27,28] In this model, the main hazard ratio (HR) calculation is defined as $\lambda_{12}/\lambda_{02}$, outlined in Figure 1.This risk can be intuitively understood as the increased risk of an event given an exposure, compared to those who never had the exposure (Figure 1). Here, we considered drug administrations and exposure and extrapolated from abnormal laboratory tests events. That is, the first position in the subscript indicates in the study whether the subject was exposed (0: "no"; 1: "yes").

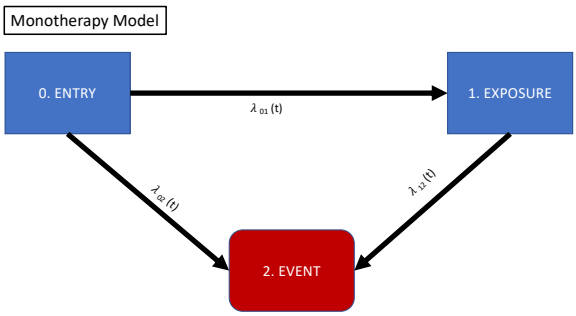


Figure 1: Schematic overview of potential pathways for each patient to take. The labeled paths exemplify the hazard at time t for a given individual. This study specifically examines the hazard ratio of $\lambda 12 / \lambda 02$ for any given time t. Different model state assumptions are clarified in Table 1.

The second position indicates if the event happened (2: "event) (Table 1). Two versions of the model were established: Model A based on abnormal tests followed by drug administration within a certain number of days, and Model B based on drug administration followed by a newly returned abnormal test (i.e., it was not documented as abnormal before the medication started) within a certain number of days, further detailed in Table 1. To capture the different types of pADEs that can develop within hours, days, weeks, or months, each model was created for different time frames: 24 hours, 48 hours, 72 hours, 7 days, 14 days, 30 days, 60 days, and 90 days.[29,30] Time was calculated from the first date of exposure; drug administration duration was therefore not included in these models.

In addition, only the earliest known administration of a drug within the time frame of the study was included. Similarly, only the first instance a patient had an abnormal test was retained. In Model A, the calculated HRs should follow typical diagnostic protocols. Model B included the inverse approach where single drugs can be investigated for their pADEs.

Table 1: Defined states for each model as outlined in figure 1

*x can be any of the defined time frames listed in Methods

State	Model A	Model B
0. Entry	Entering study with normal lab test and no drug administration	Entering study with normal lab test and no drug administration
1. Exposure	An abnormal lab test result	Being administered a drug
2. Event	Being administered a drug within x* days of exposure date	Receiving an abnormal lab test result within x* days of exposure date

2.5. *Monotherapy Model Parameters*

For the monotherapy multistate Cox model there are two required entries: time to exposure and time to event. Patients can have four paths, 1: never having the exposure or the event, 2: having the exposure but not the event, 3: having the exposure and the event, or 4: never having the exposure but having the event (Figure 1). The patients in the cohort were therefore always included in the model in question because they would always be categorized in one of these four paths. Times associated to each input was calculated as the patients age at the time, correcting for immortal time bias and accounting for age in the model. The HR outputs of these iterated models over the different time frames were corrected for multiple testing using the false discover rate (FDR) (full results for all models provided in supplemental Table 1). Models of all ATC codes and test combinations were run if at least 100 unique patients experienced both during the study window, removing drugs and laboratory tests that are least common, increasing the power of the study and relevance of the results.

2.6. *Multistate Cox Model, Polypharmaceutical Therapies*

The same core multistate Cox model approach was used to study pADE in the context of polypharmacy, here modelled by simultaneous administration of two different drugs. Figure 2 shows a schematic outline for the polypharmaceutical model where patients experiencing monotherapies and polypharmacy can be accounted for in the same design.

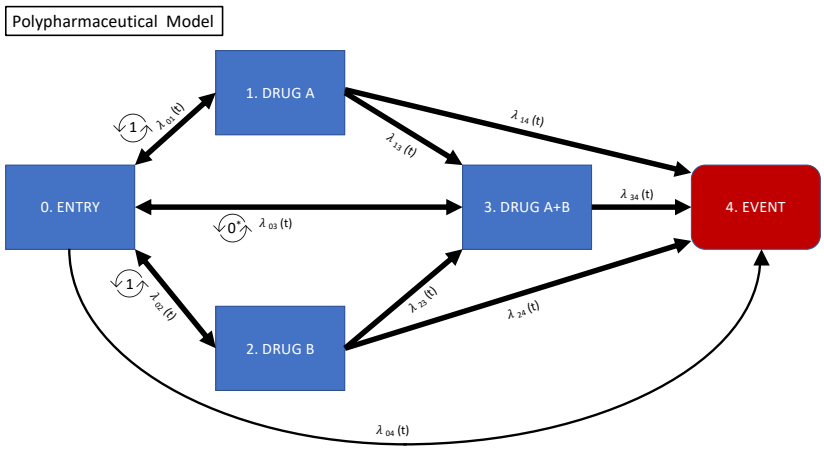


Figure 2: Schematic diagram of the polypharmaceutical model. All patients begin the study at state 0. The double-sided arrows indicate that a patient can move back to state 0 one time, the single sided arrows indicate a uni-directional path for patients. The event/exit indicates that the study window is over either with a patient incurring an adverse outcome, or the study window ending. Further details of each possible path are found in supplemental table 2. * Symbol indicates the patient can only move to state 4 after moving back to 0 (i.e. monotherapy exposures not considered after concomitant administration).

Again, this model only considered the first time an abnormal test was recorded for each patient. This means that the model only "counts" an abnormal test after concomitant therapy if the patient in question had no previous record of that abnormal test occurring. Drug pairs studied included only those listed in the Danish Drug-Drug Interaction (DDI) database (maintained by the Danish Medicines Agency) due to computational feasibility limitations (it would require ~6 months to compute all possible iterations).[25,31] Data from this database is then used to analyze results from this model. Other conditions for this model can be related to time frames. Drugs administered

sequentially were considered in the model if the two drugs were administered within six hours of each other (from last administration of the first drug). This is done under the assumption that the pharmacokinetic (PK) profile of a given drug will remain in the patients' blood stream for at least several hours and therefore has an interaction with the second drug in question.[32] Given the complexity of the system, the polypharmaceutical model only examined possible ADEs after 24 hours after co-administration time (i.e., when the second drug was administered).

2.7. *Polypharmaceutical Model Parameters*

As the case for the monotherapy model, there are several paths a patient can take in the multistate model. To adjust for potential immortal time bias, each patient record was broken into several records that account for a patient's movement between states as outlined in Figure 2. As exemplified there, every patient enters at state 0 and can move to state 1, 2, or 3. In this model, patients can also move back to state 0 from state 1 or 2 if no ADE was detected, and then move back into any state again. This event can only happen once per patient to consider if the patient in question did not experience a monotherapy induced ADE. From any of the states, the given patient can move to state 4, i.e., the exit state as this indicates an ADE occurred (or end of study) and the patient "exits" the study window. Patients' ages for each pathway were recorded for all patients and used as inputs to the Coxph model; this approach therefore integrates age into the model directly.[28] Because of the complexity of this approach, patients can have their record broken up into up to five entries in preprocessing steps. All possible pathways are summarized in supplemental table 2. Models of all possible ATC code pairs and test combinations were run if at least 5 unique patients experienced both during the study window, as per research approval guidelines (see data approval section). The HR outputs of these iterated models were corrected for multiple testing using FDR (processed results of reported models provided in supplemental table 3). All analyses were performed in R version 4.0.0 with the "The Coxph package" as the main resource.[33]

2.8. *Correlations between Laboratory Tests and Polypharmaceutical Drug Dosage Changes*

To substantiate the results of our multistate Cox models, we compared our findings with the results of a recent study published by Rodríguez et al.[24] In brief, in a data set of 77,494 potential drug pairs, Rodríguez et al. identified 694 drug pairs where drug dosage changes are more likely to happen during co-administration, compared to when they were administered as a monotherapy. Moreover, these 694 drug pairs had not previously been reported in 15 different drug-drug interaction databases. We assessed the overlap of drug pairs identified in the present study and the drug pairs identified by Rodríguez et al.

3. Results

A total of 1,634,655 patients were included in the model for all interactions of 462 medications and 323 possible biochemical outcomes at the 90-day time frame, decreasing for the other time frames due to cohort restrictions (see Methods). All possible iterations of these data were fed through multi-state Cox models A and B as outlined in Figure 1 and detailed in Table 1. Resulting HRs were

analyzed and presented in Figures 3-5; all monotherapy HRs and corresponding p values are reported in supplemental table 1.

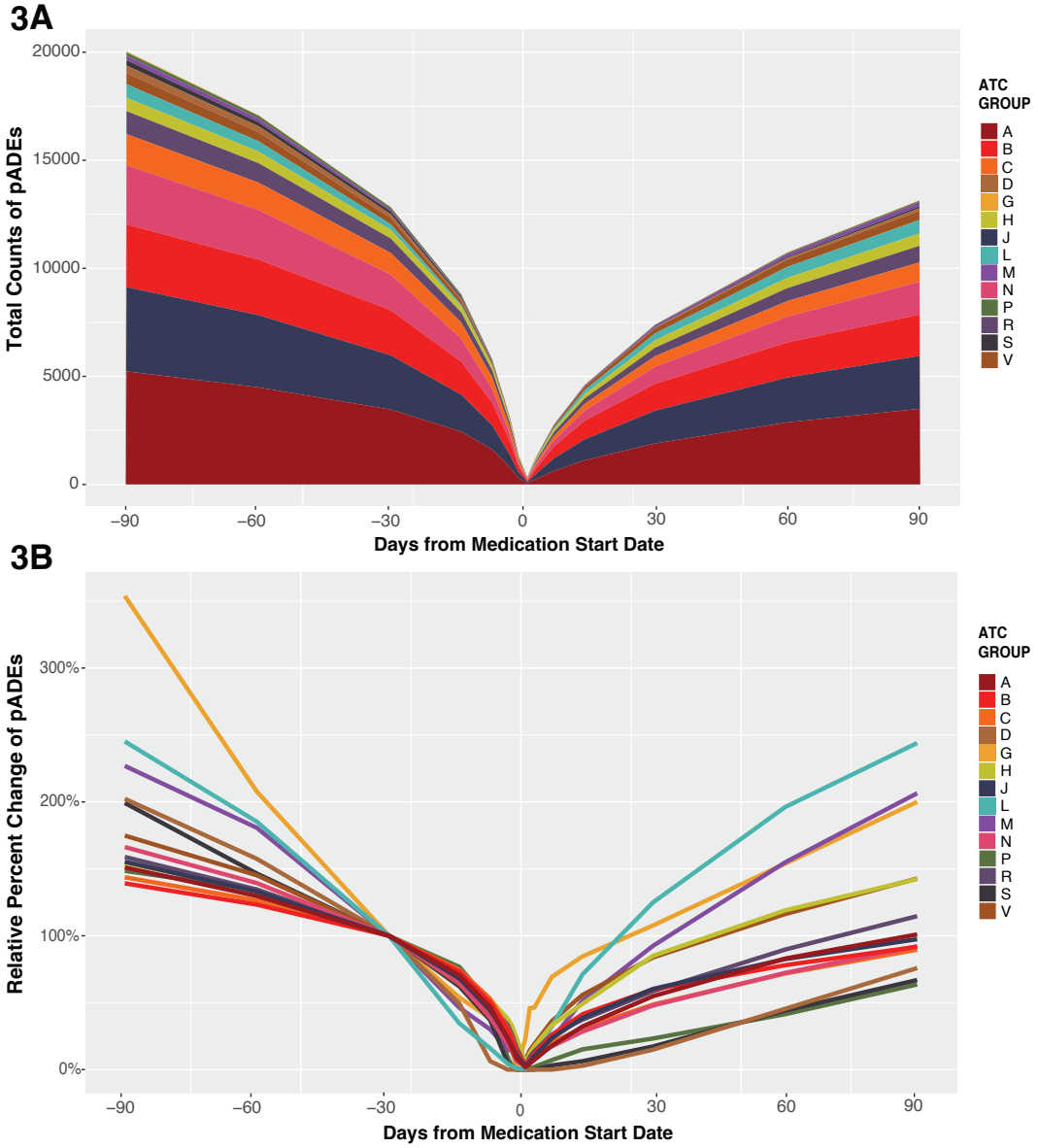


Figure 3: The total detected pADEs across the study period, in relation to medication start date (day 0). Figure 3a summarizes the total pADEs identified in the cohort for significant pADEs with HR > 2 (pval<0.05, FDR corrected). The same data is then assessed using relative percentage of pADEs as compared to day "-30" by ATC group in Figure 3b. The colors indicate the different ATC groups. Full legend provided in supplemental figure 1.

Figures 3a and 3b combine results from model A and B to visualize total HR counts and cumulative counts, respectively. In Figure 3b we can see expected trends in laboratory value changes, where more significant pADEs are identified before drug administration start dates than after which reflects the diagnostic period followed by reduced symptoms (after day 0) as the

medication takes effect; this finding serves as a proof-of-principle for the method approach. Further, the proportion of HRs after the drug start date increases in group L (Figure 3b) which mainly consists of cancer-fighting drugs; chemotherapies are widely known to cause severe side-effects, namely in white blood cell counts. Lastly there is a notable short increase in group G proportionally around days 1-7 post medication date of misoprostol (G02AD06), an abortion medication which typical involves several side effects that subside within a few days to a week.[34]

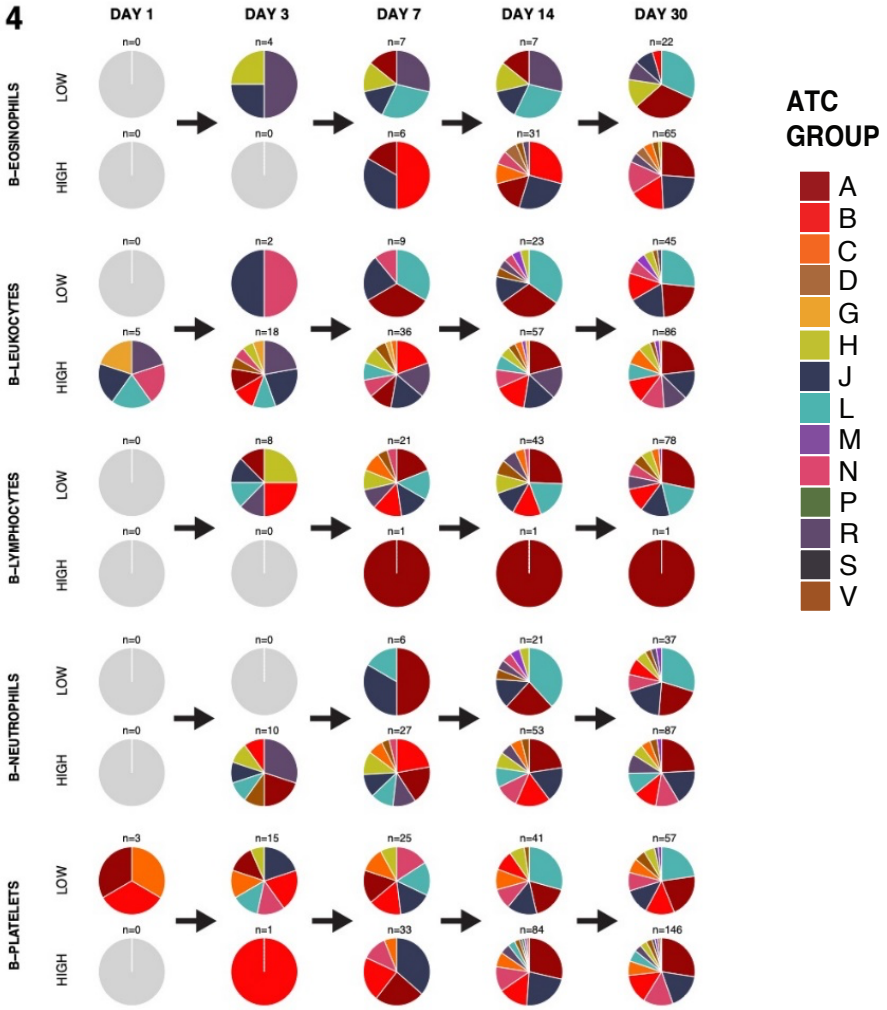


Figure 4: A time-frame overview of pADEs to each laboratory test over selected timeframes. The selection of tests highlighted here are selected white blood cell count tests and platelets, with results for both abnormally high and low counts. Grayed out pie charts indicate that no significant pADEs (HR>2, pval <0.05, FDR corrected) were detected at the given time point and lab test combination. Coefficient values for all data are reported in supplemental table 1. Full legend provided in supplemental figure 1.

Figure 4 expands on pADEs identified in Figure 3 by looking at specific laboratory tests by specified timeframes. The figure provides a trajectory over time of newly abnormal laboratory tests and their relation to different anatomical groups.

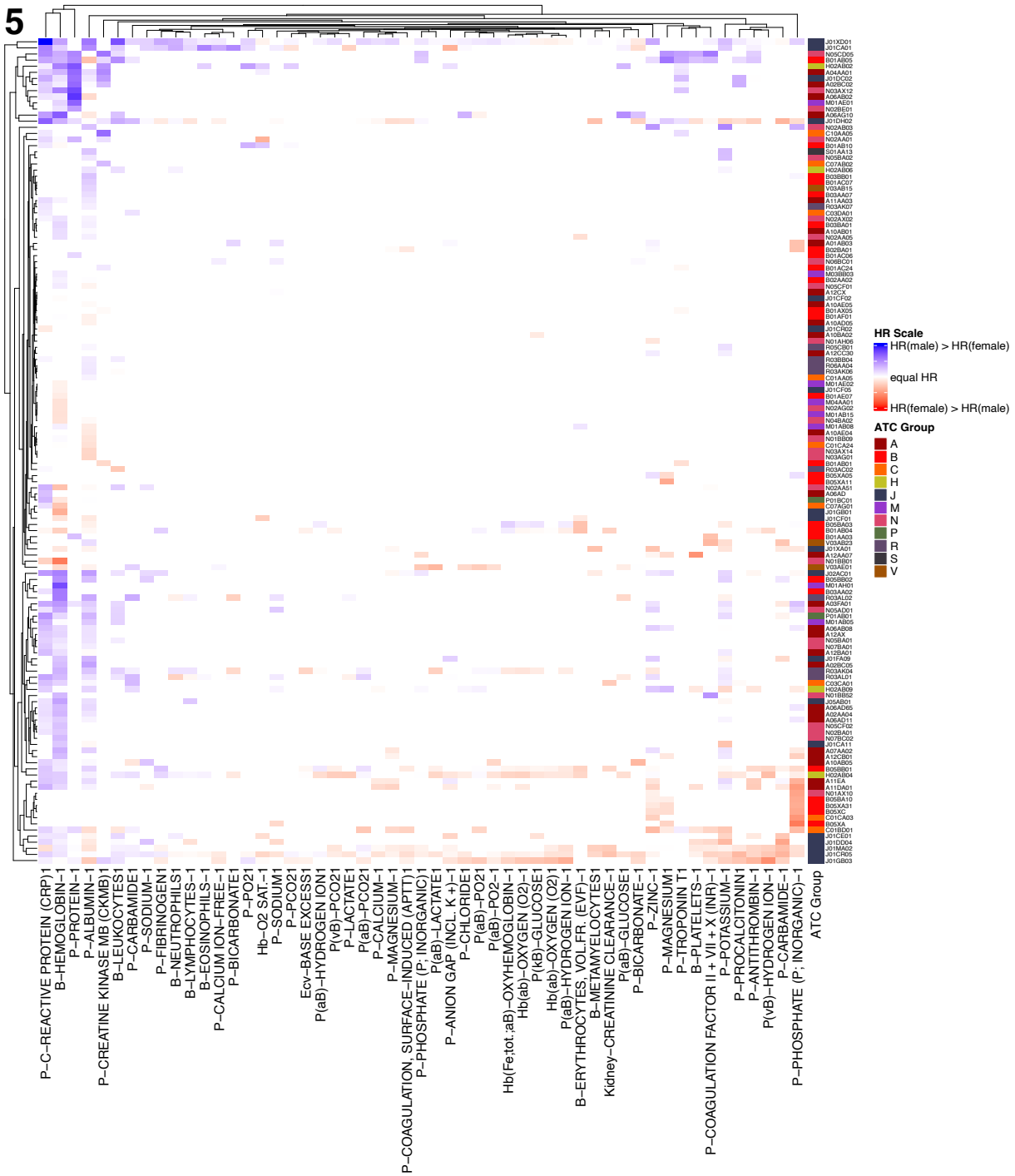


Figure 5: Heatmap providing a summary of sex-driven differences of pADEs that occur within two days of medication start. X-axis: Blood tests. Y-axis: Drugs. Bluer colors indicate a higher HR value for male vs female patients, while redder colors indicate the inverse. White: No significant difference between HRs for sexes or no pADEs were identified by the model (pval >0.05, FDR corrected). The actual HR values for all data can be found in supplemental table 1. Full legend provided in supplemental figure 1.

The timeframes of this pADE development are clearer here as the trends from hourly, daily, weekly, and monthly pADEs are elucidated, again primarily in group L around the 7-to-14-day mark in various white blood cell and platelet count tests. Except for lymphocytosis (high lymphocytes), the number of drugs that associates with an abnormal lab test increases as the time from exposure increases for all blood tests. However, the compositions of drug classes are only similar across time for some tests, e.g., thrombocytotosis (high platelets). Generally, the trajectories for leukocytes and neutrophils resemble each other, reflecting the fact that 40-60% of leukocytes are neutrophils. In both cases, drugs in anatomical groups J and L compromise at least 50% of the pairs, consistent with the characteristics of the population where these drugs are typically administered (i.e., treatment of infections and malignancies). Notably, the trajectory for lymphocytopenia remains constant over the observation time.

Figure 5 examines pADEs detected within 48 hours of drug administration by sex. Notably, HRs for hemoglobin, C-reactive protein, and albumin were generally higher for males than for females (left columns). Conversely, associations for antibiotics often administered as second or third line of treatment have many associations with abnormal lab values where HRs are higher for females (bottom rows). Yet, the trends for C-reactive protein, hemoglobin and leukocytes show higher HRs for males. In contrast, procalcitonin generally associates with higher HRs for females. In sum, these observations are consistent with existing knowledge of differences in inflammatory responses between sexes.[35–37] Importantly, the observed trends are evidence that these differences also affect treatment response.

Drug pairs are investigated in Figure 6 where a network approach is used to summarize the overlap between findings from the multi-state Cox model approach and known DDIs as reported by public health authorities. Generally, the network is dominated by drugs from ATC groups A, B, C, J, and N consistent with the trends in Figure 3. I.e., these drug classes are also the drugs classes that correlated with most abnormal lab tests when administered as a monotherapy. The figure shows that the severity of pADE is high for several drug pair therapies from the J chapter, when combined with drugs from the A and B chapter. Specifically, the J01 and J02 groups have a lot of edges (which we loosely denote “hub”) with a relatively large fraction being dark and indicating that there is a high risk of pADE when antibiotics for systemic used are combined with drugs used to treat bacterial and fungal infections. Further, Figure 6 illustrates that there likely are underreported drug pair therapies with pADE as indicated by “hubs” within the J chapter and chapter N. In addition, some of the trends displayed in the figure reflect clinical practice. For example, there is a dark edge between fentanyl (N01AH01) and triazolam (N05CD05) which are often administered jointly to patients subjected to surgery. Overall, there are a lot of thick and dark connections with ciprofloxacin (J01MA02) and similar group J drugs.

In addition, corticosteroid for systemic use (H02) is represented in four nodes making it the most prevalent drug class from chapter H (80% of drugs from chapter H). While the analysis does not present evidence that the indication for corticosteroids in these cases were allergic reactions, it is worth noting that voriconazole (J02AC03) pairs with corticosteroid for systemic use (H02). Voriconazole is a systemic antimycotic drug with a narrow therapeutic index often used to treat invasive aspergillus in immunocompromised patients.

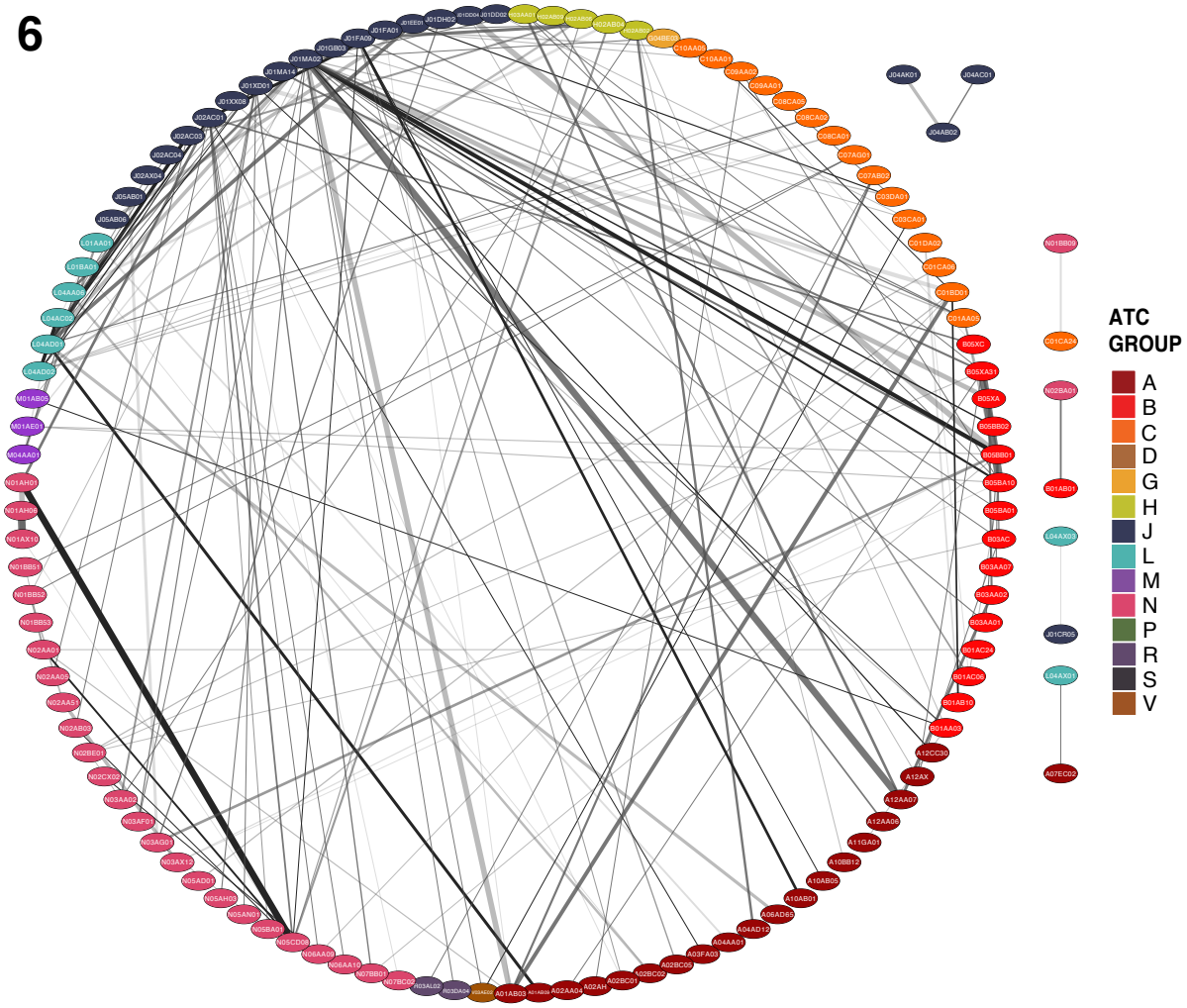


Figure 6: Circle network overview of pADEs by drug pairs. Line thickness correlate to the total number of pADEs detected using the model shown in figure 2, while the coloring corelates to the known severity of the drug interaction in accordance with the Danish Drug-Drug interaction database (lighter= less severe, darker= more severe). Data used to create this figure are provided in supplemental table 3. Only pADEs with HR>3 are included in this figure; any interactions that were known to not cause side effects (level = "ingen") were removed. A minimum of 4 pADEs were also required for each pair to be included in this visualization. Full legend provided in supplemental figure 1.

Another interesting pair is the amiodarone (C01BD01) and metronidazole (J01XD01), which has no documented DDI warning (therefore hidden from Figure 6) but has 46 pADEs listed in supplemental table 3. Owing to the risk of pro-arrythmias, amiodarone treatment is typically initiated under tight monitoring[38]. The fact that co-administration of amiodarone and metronidazole correlates with many abnormal blood tests might indicate that administration of amiodarone to infected patients presents yet another risk.

To assess if the trends identified in the present study were also reflected in drug dosage changes, we compared the 694 pairs defined as pADE in the present study with the potential undescribed

drug interaction pairs identified by Rodríguez et al.[24] Of the 694 pairs, there were 357 that were also identified as pADE in the present study. The most prevalent pADE was ampicillin and dexamethasone, which was also described by Rodríguez et al. The fact that this drug combination was identified in both studies is consistent with the fact, that infections and inflammatory responses are not necessarily trivial to distinguish in clinical practice. For example, in cancer patients you would expect dexamethasone discontinuation, if antibiotics (e.g. ampicillin) is initiated. We also noted that the co-administration of, for example, morphine and insulins was overlapping in the two datasets.

4. Discussion

In this study, we have developed and presented an approach that systematically assesses correlations of drug impact on laboratory test results for secondary care patients managed at hospitals. Outcomes shown in Figures 3 and 4 present positive control cases of the model providing initial validation for the statistical approach. These results build the foundation for Figure 5 where the model directly compares pADEs for male vs female patients in cases where one sex experiences a significant risk as compared to the other sex. Notably inflammation markers were most consistently different between sexes which has been established in the literature previously.[35–37] This is important because it highlights that the physiological response to drugs can be fundamentally different between sexes, emphasizing the need for improved representation in clinical trial approval protocols. Drugs have historically been mostly approved using men as test subjects and additionally at a singular dose, regardless of BMI or other differing features.[39] Reasonings from these findings, with similar findings in the literature, suggest that in several instances it is likely that physicians are over or under medicating females as opposed to their male counterparts for the same disease.[40] Further, since gender changes are included in this dataset starting from 2014, it is possible that there are some hidden transgender population trends that would be interesting to investigate separately once a more robust data set is developed. Transgender populations are often identified as a group that suffers greatly from health disparities and as such should be a focus in similar studies going forward as data availability increases . At this time, we are also unable to stratify by race or ethnicity in this dataset, but this would be an important future research question as well.

In Figure 6, the first attempt is made to model pADEs driven by concomitant therapies using a multi-state Cox model, whose use is validated using the monotherapy model presented here and confirmed results in Figures 3-5. Figure 6 takes a network view of drug pairs already known in the Danish DDI database and overlaid with the results generated in the model overviewed in Figure 2. These results open the door for future applications of this method where specific sub-groups of patients can be compared for increased risk of certain ADEs, better informing their physicians when determining the proper therapeutic approach to follow. Additionally, the polypharmaceutical model captures pADEs at both the monotherapy and polypharmaceutical level in the same statistical test, allowing for more direct assessments for additive, synergistic, or antagonistic drug pairs.

While still a preliminary approach, this study demonstrates the potential for identifying and alerting authorities more efficiently to possible DDIs that are yet unknown especially when new drugs enter the market. Technically, we introduced model restrictions in the multistate Cox models

(cf. Methods) to reduce the impact of potential bias from the physicians who had already seen and reacted to the respective patient's symptoms or side effects, effectively focusing the model on pADEs associated to first time drug exposures. We further identified overlapping trends in lab values and drug dosage changes, which exemplifies a novel way of assessing potential drug effect and adverse drug effects. In a population with an increasing age and prevalence of polypharmacy, we argue that it is of uttermost importance to develop methods for monitoring drug effects.

The method also provides the foundation for a tool for exploring which targets and mechanisms of action are more prone to severe ADEs and can therefore be studied more thoroughly when testing new drugs in the pre-clinical phase, as to avoid costly human trials that eventually may end in the removal of drugs from the market.

In summary, this study presents the first retrospective study investigating how a patient's laboratory data history can be used to investigate possible drug-induced biochemical changes within specific population groups, improving their safety and health in the long run. Further potential benefits include reduced hospital admittance for the treatment of these same ADEs, reducing both the cost and physical/mental toll on these patients.[42]

Supplemental material

Supplemental material is available at https://github.com/vmuse12/ADE_data

Acknowledgments

The authors acknowledge funding from the Novo Nordisk Foundation (grants NNF17OC0027594 and NNF14CC0001).

References

1. Pin C, Collins T, Gibbs M, Kimko H. Systems Modeling to Quantify Safety Risks in Early Drug Development: Using Bifurcation Analysis and Agent-Based Modeling as Examples. AAPS J. 2021 Jul;23(4):77.

2. Vergetis V, Skaltsas D, Gorgoulis VG, Tsirigos A. Assessing Drug Development Risk Using Big Data and Machine Learning. Cancer Res. 2021 Feb 15;81(4):816–9.

3. Hazell L, Shakir SAW. Under-Reporting of Adverse Drug Reactions: A Systematic Review. Drug Saf. 2006;29(5):385–96.

4. Shalviri G, Mohebbi N, Mirbaha F, Majdzadeh R, Yazdizadeh B, Gholami K. Improving adverse drug event reporting by health care professionals. Practice CE, Group O of C, editors. Cochrane Database Syst Rev [Internet]. 2017 Mar 11; Available from: https://doi.wiley.com/10.1002/14651858.CD012594

5. Bénard-Laribière A, Miremont-Salamé G, Pérault-Pochat MC, Noize P, Haramburu F, the EMIR Study Group on behalf of the French network of pharmacovigilance centres. Incidence of hospital admissions due to adverse drug reactions in France: the EMIR study. Fundam Clin Pharmacol. 2015 Feb;29(1):106–11.

6. Overhage JM, Gandhi TK, Hope C, Seger AC, Murray MD, Orav EJ, et al. Ambulatory Computerized Prescribing and Preventable Adverse Drug Events. J Patient Saf. 2016 Jun;12(2):69–74.

7. Field TS, Gurwitz JH, Harrold LR, Rothschild JM, Debellis K, Seger AC, et al. Strategies for Detecting Adverse Drug Events among Older Persons in the Ambulatory Setting. J Am Med Inform Assoc. 2004 Nov;11(6):492–8.

8. Lynge E, Sandegaard JL, Rebolj M. The Danish national patient register. Scand J Public Health. 2011;39(7):30–3.

9. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. Eur J Epidemiol. 2014 Aug 1;29(8):541–9.

10. Jeong E, Park N, Choi Y, Park RW, Yoon D. Machine learning model combining features from algorithms with different analytical methodologies to detect laboratory-event-related adverse drug reaction signals. PloS One. 2018;13(11):e0207749.

11. Chowdhury SR, Chandra Das D, Sunna TC, Beyene J, Hossain A. Global and regional prevalence of multimorbidity in the adult population in community settings: a systematic review and meta-analysis. eClinicalMedicine. 2023 Mar;57:101860.

12. Divo MJ, Martinez CH, Mannino DM. Ageing and the epidemiology of multimorbidity. Eur Respir J. 2014 Oct;44(4):1055–68.

13. Khan MohdR, Malik MA, Akhtar SN, Yadav S, Patel R. Multimorbidity and its associated risk factors among older adults in India. BMC Public Health. 2022 Dec;22(1):746.

14. Sundhed Denmark. Foreningen af danske sundhedsvirksomheder [Internet]. Available from: https://www.sundheddanmark.nu/media/1195/sundheddanmark-kort-om-privathospitaler-2018.pdf

15. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. Scand J Public Health. 2011 Jul;39(7_suppl):30–3.

16. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol. 2015 Nov;449.

17. Muse VP, Aguayo-Orozco A, Balaganeshan SB, Brunak S. Population-wide analysis of hospital laboratory tests to assess seasonal variation and temporal reference interval modification. Patterns. 2023 Jun;100778.

18. Muse VP, Brunak S. Protocol for EHR laboratory data preprocessing and seasonal adjustment using R and RStudio. STAR Protoc. 2024 Mar;5(1):102912.

19. Bietenbeck A, Boeker M, Schulz S. NPU, LOINC, and SNOMED CT: a comparison of terminologies for laboratory results reveals individual advantages and a lack of possibilities to encode interpretive comments. LaboratoriumsMedizin. 2018 Dec;42(6).

20. Properties, IFCC U (C SN of the, Nomenclature IJC on, Pontet F, Petersen UM, Fuentes-Arderiu X, et al. Clinical laboratory sciences data transmission: the NPU coding system. Stud Health Technol Inform. 2009;150.

21. Statens Serum Institut. NPU Terminologien BRUGERMANUAL [Internet]. 2014. Available from: http://www.ssi.dk/

22. Epidemiology D of C. Manual for using the LABKA database for research projects [Internet]. Available from: https://kea.au.dk/fileadmin/KEA/filer/Other_reports/119_LABKA_manual.pdf

23. Grann, Erichsen R, Nielsen, Frøslev, Thomsen R. Existing data sources for clinical epidemiology: The clinical laboratory information system (LABKA) research database at Aarhus University, Denmark. Clin Epidemiol. 2011 Apr;

24. Rodríguez CL, Haue AD, Mazzoni G, Eriksson R, Biel JH, Cantwell L, et al. Drug dosage modifications in 24 million in-patient prescriptions covering eight years: A Danish population-wide study of polypharmacy. PLOS Digit Health. 2023 Sep 7;2(9):e0000336.

25. Leal Rodríguez C, Kaas-Hansen BS, Eriksson R, Biel JH, Belling KG, Andersen SE, et al. Drug interactions in hospital prescriptions in Denmark: Prevalence and associations with adverse outcomes. Pharmacoepidemiol Drug Saf. 2022 Jun;31(6):632–42.

26. World Health Organization. Anatomical Therapeutic Chemical (ATC) Classification [Internet]. [cited 2023 Apr 17]. Available from: https://www.who.int/tools/atc-ddd-toolkit/atc-classification

27. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer; 2000. 350 p. (Statistics for biology and health).

28. coxph: Fit Proportional Hazards Regression Model [Internet]. Available from: https://www.rdocumentation.org/packages/survival/versions/3.5-5/topics/coxph

29. Leroy F, Dauxois JY, Théophile H, Haramburu F, Tubert-Bitter P. Estimating time-to-onset of adverse drug reactions from spontaneous reporting databases. BMC Med Res Methodol. 2014 Dec;14(1):17.

30. Phillips EJ, Bigliardi P, Bircher AJ, Broyles A, Chang YS, Chung WH, et al. Controversies in drug allergy: Testing for delayed reactions. J Allergy Clin Immunol. 2019 Jan;143(1):66–73.

31. Danish Medical Agency. Danish drug interaction databases [Internet]. Available from: https://www.interaktionsdatabasen.dk/

32. Holford N. Absorption and Half-Life. Transl Clin Pharmacol. 2016;24(4):157.

33. Therneau T. A package for survival analysis in R.

34. NIH National Library of Medicine. Misoprostol [Internet]. [cited 2023 Aug 3]. Available from: https://medlineplus.gov/druginfo/meds/a689009.html

35. Klein SL, Flanagan KL. Sex differences in immune responses. Nat Rev Immunol. 2016 Oct;16(10):626–38.

36. Takahashi T, Iwasaki A. Sex differences in immune responses. Science. 2021 Jan 22;371(6527):347–8.

37. Jacobsen H, Klein SL. Sex Differences in Immunity to Viral Infections. Front Immunol. 2021 Aug 31;12:720952.

38. Siddoway LA. Amiodarone: guidelines for use and monitoring. Am Fam Physician. 2003 Dec 1;68(11):2189–96.

39. Raz L, Miller VM. Considerations of Sex and Gender Differences in Preclinical and Clinical Trials. In: Regitz-Zagrosek V, editor. Sex and Gender Differences in Pharmacology [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013 [cited 2024 Feb 26]. p. 127–47. (Handbook of Experimental Pharmacology; vol. 214). Available from: https://link.springer.com/10.1007/978-3-642-30726-3_7

40. Sørup FKH, Eriksson R, Westergaard D, Hallas J, Brunak S, Ejdrup Andersen S. Sex differences in text-mined possible adverse drug events associated with drugs for psychosis. J Psychopharmacol (Oxf). 2020 May;34(5):532–9.

41. Kachen A, Pharr JR. Health Care Access and Utilization by Transgender Populations: A United States Transgender Survey Study. Transgender Health. 2020 Sep 1;5(3):141–8.

42. Greene SM, Reid RJ, Larson EB. Implementing the Learning Health System: From Concept to Action. Ann Intern Med. 2012 Aug 7;157(3):207.

Precision Medicine: Multi-modal and multi-scale methods to promote mechanistic understanding of disease

Hannah Carter
Department of Medicine, University of California San Diego, 9500 Gilman Dr. La Jolla, CA 92093, USA
Email: hkcarter@ucsd.edu

Steven Brenner
Department of Plant & Microbial Biology, University of California Berkeley, 461 Koshland Hall, Berkeley, California 94720-3102, USA
Email: brenner@compbio.berkeley.edu

Yana Bromberg
Department of Biology, Emory University, 1510 Clifton Rd, USA
Department of Computer Science, Emory University, 400 Dowman Dr. Atlanta, GA 30317, USA
Email: yana.bromberg@emory.edu

Precision medicine focuses on developing treatments and preventative strategies tailored to an individual's genomic profile, lifestyle, and environmental context. The Precision Medicine sessions at the Pacific Symposium on Biocomputing (PSB) have consistently spotlighted progress in this domain. Our 2025 manuscript collection features algorithmic innovations that integrate data across scales and diverse data modalities, presenting novel techniques to derive clinically relevant insights from molecular datasets. These studies highlight recent advances in technology and analytics and their application toward realizing the potential of precision medicine to enhance human health outcomes and extend lifespan.

Keywords: Precision medicine; variants; GWAS; genomics; proteomics; machine learning.

1. Introduction

Precision medicine leverages data-driven approaches to personalize medical decisions and treatments for individual patients. Novel technologies enabling rapid and systematic production of molecular measurements and imaging data have facilitated new mechanistic insights into complex disease processes, ultimately translating into better clinical decision-making. Breakthroughs in genomic technologies, such as spatial transcriptomics and single-cell analysis, have enabled the discovery of genetic biomarkers associated with drug responses, disease susceptibility, and other key medical outcomes. Concurrently, the vast scale of these data has spurred the development of

novel computational techniques, exemplified by the surge in deep learning-based approaches for biological and medical data analysis.

As the richness of datasets characterizing molecules, cells, and tissues grows, there are new opportunities to combine them across data modalities and scales. Methods to synthesize these data into mechanistic understanding and better biomarkers for clinically relevant outcomes are needed. In our 2025 session, we highlight ground-breaking research from a wide range of disciplines that integrate divergent data to offer novel insight into disease mechanisms, diagnosis, and treatment. These methods provide a preview of how computation can continue to advance precision medicine in the coming years.

2. Session Contributions

2.1. *Integrating imaging data with genotype to investigate mechanism*

Imaging data is increasingly available and can provide valuable information about how genetic factors are linked to disease mechanism through effects on the organization of biological systems. Two papers in this collection showcase the use of imaging to study genotypes. In one, Blennemann *et al* use live cell imaging to obtain longitudinal and spatially resolved information about T cell interaction with tumor cells across 3 genotypes[1]. In the other, Chandio and colleagues use diffusion MRI-based tractometry to obtain 3D quantitative measurements of white matter tracts across individuals with different risk genotypes and diagnoses, finding different structural features in carriers of risk versus protective genotypes[2].

2.2. *Biological understanding through linked diseases*

This collection also includes two studies where investigators use known disease associations as a starting point to identify genes and pathways underlying shared mechanisms. To gain insight into genes driving the link between Down Syndrome and obesity, Nandi *et al* derive latent variables from RNA sequencing analysis and analyze them in the context of karyotype and BMI using causal inference[3]. Ball and team use multi-disease modeling to analyze transcriptomic data from brain samples of mouse models of AD, T2D, both simultaneously and postmortem human brain to uncover a link between these diseases mediated by estrogen and inflammatory pathways[4].

2.3. *Finding drug targets and mediators of adverse drug responses*

Identification of disease-specific drug targets and understanding of the mechanisms mediating drug resistance and adverse responses are essential to inform drug development and clinical matching of patients to drugs. Three papers address these aspects of precision therapy. Orlenko *et al* implicate putative drug targets for Alzheimer’s disease by integrating information about interaction partners of known Alzheimer's genes with known gene-drug associations from a drug database[5]. By integrating methylation QTLs with GWAS loci associated with drug response phenotypes, Smith *et al* identify a number of candidate genes where genetics and epigenetics converge to generate adverse drug response[6]. Wen *et al* propose spherical PCA for single cell imaging data as a strategy to identify

cancer cells that evade cell cycle blocking drugs[7]. They find that cells that evade arrest after treatment express key genes that may represent additional drug targets.

2.4. *Using polygenic scores for molecular and intermediate phenotypes to uncover disease mechanisms*

Polygenic risk scores (PRS) have featured prominently in precision medicine research but have provided limited mechanistic understanding of disease due to their associative nature. Three papers in this session show that polygenic scores linked to protein levels or cellular activities, as opposed to disease incidence, can provide new biological insight. Phillips *et al* use genotype and RNA sequencing data to develop a polygenic score quantifying astrocyte activation, then use the PRS to study Alzheimer's-associated characteristics in a large cohort of elderly individuals[8]. They find that the score correlates with effects on memory and high-level cognition. By integrating information across genotype associations with protein versus gene expression levels, Moore *et al* reveal pathways underlying MRI-derived characteristics of heart function linked to heart failure[9]. Woerner *et al* show that PRS and polygenic predictors of protein levels can be combined to improve prediction of inflammatory bowel disease[10]. They find that polygenic protein scores are even more predictive when polygenic risk is high.

2.5. *Identifying environmental modifiers of traits and risk*

Environmental variables can modify traits and disease risk. Two papers in our session use computational strategies to better understand environmental factors. The study by Rico *et al* uses "environment by environment" associations and lipid measurements to explore interactions among environmental factors that affect lipid phenotypes[11]. They find several cases where combinations of two environmental exposures associate with significant differences in HDL levels. To study the effects of salt intake on risk of chronic kidney disease (CKD), Shivakumar *et al* use polygenic risk scores for CKD to stratify individuals in the UKBioBank, then examine the association between salt consumption and incidence of CKD in each subgroup[12].

2.6. *Methods addressing computational challenges for multimodal health data analysis*

More broadly, the availability of rich multi-modal and sensitive health data presents new computational challenges for advancing precision medicine. Two papers in this session propose solutions to specific challenges in this space. Colombo and team develop a strategy to predict cancer type from a low dimensional representation of 2 data types, SNVs and CNVs, while ensuring preservation of spatial relationships between genes in CNV regions[13]. They further demonstrate that they can operate on encrypted data to ensure patient privacy which remains a major concern for precision medicine approaches that require genomic data. Golovanevsky *et al* propose a one-versus-others attention approach to address computational bottlenecks in neural network-based integration across the rich array of data modalities available for clinical applications[14]. These advances address more general barriers to scaling and implementing computational approaches for precision medicine.

References

1. M. Blennemann, A. Verma, S. Bachl, J. Carnevale, B. E. Engelhardt, Understanding TCR T cell knockout behavior using interpretable machine learning *Pacific Symposium on Biocomputing* (2024)
2. B. Q. Chandio, J. E. Villalon-Reina, T. M. Nir, S. I. Thomopoulos, Y. Feng, S. Benavidez, N. Jahanshad, J. Harezlak, E. Garyfallidis, P. M. Thompson, Amyloid, Tau, and APOE in Alzheimer's Disease: Impact on White Matter Tracts *Pacific Symposium on Biocomputing* (2024)
3. S. Nandi, Y. Z., L. A Gillenwater, M. S. Ganes, H. Zhang, N. Janani, C. Greene, M. Pividori, M. Chikina, J. C. Costello, A Pathway-Level Information ExtractoR (PLIER) framework to gain mechanistic insights into obesity in Down syndrome *Pacific Symposium on Biocomputing* (2024)
4. B. K. Ball, E. A. Proctor, D. K. Brubaker, Cross-Species Modeling Identifies Gene Signatures in Type 2 Diabetes Mouse Models Predictive of Inflammatory and Estrogen Signaling Pathways Associated with Alzheimer's Disease Outcomes in Humans *Pacific Symposium on Biocomputing* (2024)
5. A. Orlenko, M. Venkatesan, L. Shen, M. D. Ritchie, Z. P. Wang, T. Obafemi-Ajayi, J. H. Moore, Biologically Enhanced Machine Learning Model to uncover Novel Gene-Drug Targets for Alzheimer's Disease *Pacific Symposium on Biocomputing* (2024)
6. D. A. Smith, S. A. Arteaga, M.C. Sadler, R. B. Altman, Identifying DNA methylation sites affecting drug response using electronic health record–derived GWAS summary statistics *Pacific Symposium on Biocomputing* (2024)
7. O. Wen, S. C. Wolff, W. Stallaert, D. Li, J. E. Purvis, T. M. Zikry, Spherical Manifolds Capture Drug-Induced Changes in Tumor Cell Cycle Behavior *Pacific Symposium on Biocomputing* (2024)
8. J. M Phillips, J. A. Schneider, D. A. Bennett, P. K. Crane, S. L. Risacher, A. J. Saykin, L. C. Dumitrescu, T. J. Hohman, Astrocyte Reactivity Polygenic Risk Score May Predict Cognitive Decline in Alzheimer's Disease *Pacific Symposium on Biocomputing* (2024)
9. A. Moore, R. Venkatesh, M. G. Levin, S. M. Damrauer, N. Reza, T. P. Cappola, M. D. Ritchie, Connecting intermediate phenotypes to disease using multi-omics in heart failure *Pacific Symposium on Biocomputing* (2024)
10. J. Woerner, T. Westbrook, S. Jeong, M. Shivakumar, A. R. Greenplate, S. A. Apostolidis, S. Lee, Y. Nam, D. Kim, Genomics and Computational Biology Graduate Group, University of Pennsylvania, "Plasma protein-based and polygenic risk scores serve complementary roles in predicting inflammatory bowel disease *Pacific Symposium on Biocomputing* (2024)
11. A. L. G. Rico and N. Palmiero, M. D. Ritchie, M. A. Hall, Integrated exposomic analysis of lipid phenotypes: leveraging GE.db in environment by environment association studies *Pacific Symposium on Biocomputing* (2024)
12. M. Shivakumar, Y. Kim, S. Jung, J. Woerner, D. Kim, Salt intake is a stronger predictor of chronic kidney disease in genetically susceptible population *Pacific Symposium on Biocomputing* (2024)

13. A. A. F. Colombo, L. Colombo, A. Falcetta, and M. Roveri, Enhancing Privacy-Preserving Cancer Classification with Convolutional Neural Networks *Pacific Symposium on Biocomputing* (2024)
14. M. Golovanevsky, E. Schiller, A. Nair, E. Han, R. Singh, C. Eikchoff, One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data *Pacific Symposium on Biocomputing* (2024)

Understanding TCR T cell knockout behavior using interpretable machine learning

Marcus Blennemann

Gladstone Institutes
San Francisco, CA 94158, USA
E-mail: marcus.blennemann@gladstone.ucsf.edu

Archit Verma

Gladstone Institutes
San Francisco, CA 94158, USA
E-mail: archit.verma@gladstone.ucsf.edu

Stefanie Bachl

University of California, San Francisco
San Francisco, CA 94143, USA
E-mail: stefanie.bachl@ucsf.edu

Julia Carnevale

University of California, San Francisco
San Francisco, CA 94143, USA
E-mail: julia.carnevale@ucsf.edu

Barbara E. Engelhardt

Gladstone Institutes
San Francisco, CA 94158, USA
Biomedical Data Science, Stanford University
Stanford, CA 94305, USA
E-mail: bengelhardt@stanford.edu

Genetic perturbation of T cell receptor (TCR) T cells is a promising method to unlock better TCR T cell performance to create more powerful cancer immunotherapies, but understanding the changes to T cell behavior induced by genetic perturbations remains a challenge. Prior studies have evaluated the effect of different genetic modifications with cytokine production and metabolic activity assays. Live-cell imaging is an inexpensive and robust approach to capture TCR T cell responses to cancer. Most methods to quantify T cell responses in live-cell imaging data use simple approaches to count T cells and cancer cells across time, effectively quantifying how much space in the 2D well each cell type covers, leaving actionable information unexplored. In this study, we characterize changes in TCR T cell's interactions with cancer cells from live-cell imaging data using explainable artificial intelligence (AI). We train convolutional neural networks to distinguish behaviors in TCR T cell with CRISPR knock outs of CUL5, RASA2, and a safe harbor control knockout. We use explainable AI to identify specific interaction types that define different knock-out conditions. We find that T cell and cancer cell coverage is a strong marker of TCR T cell modification when comparing similar experimental time points, but differences in cell aggre-

gation characterize CUL5KO and RASA2KO behavior across all time points. Our pipeline for discovery in live-cell imaging data can be used for characterizing complex behaviors in arbitrary live-cell imaging datasets, and we describe best practices for this goal.

Keywords: Explainable AI, Grad-CAM, machine learning, live cell imaging.

1. Introduction

Since FDA approval in 2017, chimeric antigen receptor (CAR) T cell immunotherapies have proven effective at treating advanced leukemias and lymphomas.[1,2] CAR T cell therapy uses *ex vivo* modification of native patient T cells to express a chimeric antigen receptor (CAR), capable of binding to surface markers of cancerous cells, to enhance T cell immunological response. T-cell receptor (TCR) therapy is a related method of treatment that uses naturally existing TCRs, protein complexes that bind to a cell's major histocompatibility complex (MHC), as an alternative to CAR proteins. TCR therapy targets various cancers by recognizing a specific antigen presented by a human leukocyte antigen (HLA) on cancer cell surfaces. This reduces the risk of toxicity associated with CAR T cell therapy, which currently struggles to distinguish between solid cancer cells and normal tissues.[1] TCR therapy, in contrast, has demonstrated effective responses against multiple solid cancer types such as melanoma and lung carcinoma with reduced off-target effects.[3]

Genetic editing of CAR and TCR T cells with CRISPR-based tools is an emerging approach to engineer improved T cell therapies. CRISPR knock out of the *RASA2* (RASA2KO) or *CUL5* (CUL5KO) genes, for example, has been demonstrated to improve T cell performance against cancer cells *in vitro*.[4] *RASA2* is a signalling checkpoint in human T cells and increases in response to chronic antigen exposure. TCR and CAR T cells without *RASA2* show better activation, higher cytokine production, and increased metabolic activity, en route to improved cancer cell removal. These RASA2KO T cells also have a survival advantage in mouse models of leukemia and other cancers.[4] *CUL5* is known to be a negative regulator of the signaling pathways in cytotoxic T lymphocytes. Knocking out *CUL5* has been shown to effectively inhibit tumor growth in mouse studies.[5] Although these genes have been identified as effective modifications in TCR T cells in *in vivo* mouse studies, understanding the biological mechanisms underlying these positive outcomes remains a challenge due to the complex, multi-scale nature of T cell and cancer cell interactions in humans.

Live-cell imaging is a common approach for evaluating the success of different types of modified T cells. Live-cell imaging with high-resolution 2D imaging from one or more channels, usually a bright field along with fluorescent marker channels, across days at fixed time intervals (e.g., every four minutes) captures the dynamics of co-cultures of cancer and modified T cells. Traditional analyses quantify the total amount of cancer cell-specific fluorescent markers as a proxy of tumor response to treatment.[4–6] Live-cell imaging has been used to identify dynamic behavior such as morphological changes during T cell killing, or differences in response to

liquid or solid tumors, using deep learning methods to segment sequential images.[7] Even with existing approaches, many questions about dynamic cellular behaviors are difficult to answer.

Computer vision, a subfield of AI, is advancing rapidly in biomedical imaging. Deep learning models, especially convolutional neural networks (CNNs), enable extraction of complex phenotypes from live-cell imaging data. This includes cell segmentation, single-cell tracking, spatiotemporal pattern recognition, and predictive modeling, all of which may be used to study the therapeutic behavior of these modified T cells. Efforts are underway to integrate CNN-driven platforms with patient-derived organoids (PDOs) for personalized drug research, exemplified by projects like OrganoID[8] and OrBITS.[9] While these tools are powerful, their prediction processes are black-box and challenging to understand. Interpreting a CNN's decision-making process should provide important information for researchers attempting to gain biological insights from their live-cell experiments. Explainable AI techniques have emerged that allow researchers to interrogate the features of images that most directly explain deep learning models' predictions and performance.[10,11]

In this work, we demonstrate the ability of explainable AI to characterize modified T cell behavioral changes under genetic perturbation. We identify phenotypic differences between TCR T cells with beneficial *RASA2* or *CUL5* knock-outs from live-cell imaging data versus TCR T cell negative controls. We use a suite of CNN classifiers trained to predict one of three genetic perturbations captured in live-cell imaging of TCR T cells co-cultured with cancer cells. We use Grad-CAM, an image explainable AI technique that estimates the change in prediction as a function of changes in pixel space, to identify the specific regions in held-out live-cell images that inform prediction for control Safe Harbor KO, RASA2KO, and CUL5KO TCR T cells. Grad-CAM highlights the regions of the image that contribute to classification as each output class. By highlighting regions that contribute to classification decisions, the Grad-CAM interpretation of images allows us to identify the cell-level phenotypic changes associated with each TCR T cell experiment, and we use these interpretable image markers to characterize the distinct T cell behaviors in the three experimental conditions. Our work develops an interpretable deep learning workflow for the analysis of live-cell imaging data, and we show the benefits of our approach by characterizing the differential behavior of SHKO (control), RASA2KO, and CUL5KO TCR T cells.

2. Methods

2.1. *Data Generation*

2.1.1. *Isolation of primary T cells from healthy donors*

Leukopaks from deidentified healthy donors with approved IRBs were purchased from StemCell Technologies. Primary human T cells were isolated with the EasySep Human T Cell Isolation Kit (StemCell Technologies) according to the manufacturer's protocol. T cells were seeded at a density of 1 million cells per mL maintained in X-Vivo-15 medium supplemented with 5% fetal bovine serum, 50 μM beta-mercaptoethanol, and 10 mM N-acetyl-L-cysteine plus 100 IU/mL of IL-2 and activated with Dynabeads Human T-Activator CD3/CD28 (Gibco) at a 1:1 bead-to-cell ratio.

2.1.2. *CRISPR KO in primary human T cells using Cas9–RNP electroporation*

T cell transduction was accomplished by adding concentrated lentivirus directly to the T cells 24 hours after activation with Dynabeads Human T-Activator CD3/CD28, 40 μL virus per 1×10^6 T cells in X-Vivo-15. At 48 h post-activation, Cas9–sgRNA–RNP electroporation was conducted with the Amaxa P3 Primary Cell 96-well 4D-Nucleofector Kit (Lonza). The safe harbor T cells were targeted using the *AAVS1* sequence GGGCCACTAGGGACAGGAT, the *RASA2*-ablation T cells with the sequence AGATATCACACATTACAGTG, and the *CUL5*-ablation T cells with the sequence ATTGGAGTAAGAGAATCCTA. crRNAs and tracrRNAs were then complexed 1:1 by volume and incubated for 30 minutes at 37C to form sgRNAs. The sgRNAs were then mixed with Cas9 (stock concentration of 40 μM, QB3 Macrolab) at 1:1 by volume for 15 minutes at 37C to produce ribonucleoproteins (RNPs) complexes. After counting, T cells were resuspended in P3 buffer at 1×10^6 per 20 μl, mixed with 3 μl of RNPs, and added to a 96-well electroporation plate. Electroporation was performed using using the EH115 protocol and recovered by adding 80 μl T cell medium (X-Vivo-15, Lonza) at 37C for 15 min. Cells were transferred to appropriate culture vessels containing X-Vivo-15 medium supplemented with IL-2 containing 100 IU per mL.

2.1.3. *Repetitive stimulation assay*

Tumor cells were maintained in a complete RMPI (Gibco) consisting of 1% penicillin-streptomycin (Gibco), GlutaMAX supplement (Gibco) and 10% fetal bovine serum (Corning), and then resuspended in T cell medium. T cells were seeded on top of the cancer cells at a 1:1 E:T ratio with IL-2 at 100 IU mL^{-1}. Subsequent repeated co-cultures were set up every 48 h. For each co-culture, T cells were counted using the Cellaca MX High-throughput Cell Counter (Revity), percentage of TCR+ cells was measured via flow cytometry, and T cells were replated onto fresh tumor cells every 48 hours maintaining a 1:1 E:T ratio.

2.1.4. *In vitro cancer killing assay by TCR T cells*

Antigen-specific T cells were co-cultured in X-VIVO-15 plus supplements – 100 IU IL-2 per mL and 1X Glucose (Gibco) – with mKate+ A375 cells pre-seeded in a 96-well flat-bottom plates at a 1:1 E:T ratio. Images were captured every 4 minutes over a 24-hour span using the IncuCyte S3 live-cell imaging platform (Essen Bioscience). The mKate+ object counts for each well were recorded over time.

2.2. *Model architecture, training, and evaluation*

A convolution neural network was trained to identify the TCR T cell genetic perturbation – RASA2KO, CUL5KO, or SHKO – from a single 300 by 300 pixel subsection of each image. The network was trained on images from nine of the replicates, three from each condition, and validated on the remaining three held-out wells. The model consists of a ResNet50[12] block, a fifty layer residual convolutional neural network (CNN), that feeds into a fully connected linear layer to predict the weights for each class. ResNet50 is a CNN designed for image classification tasks. The first stage consists of 64 7×7 convolutional filters, followed by four

stages of residual blocks. These stages contain filters configured as follows: the first has 64, 64, and 256 filters; the second has 128, 128, and 512 filters; the third has 256, 256, and 1,024 filters; and the fourth has 512, 512, and 2,048 filters. The network ends with a fully connected layer with one neuron per possible output class, which is three in our application. The initial convolutional layer and first layer of each stage uses a stride size of 2, while all other layers use a stride size of 1. The weights and biases of the final, fully connected output linear layer were trained to minimize the cross-entropy loss of the predicted probability of each class, a softplus of the linear output layer, to the true data label.

The untrained parameters of the ResNet50 block were initialized as the parameters of ImageNet.[13] These weights capture high-level features, such as edges and shapes, allowing us to reach accurate classification faster. The last layer of the model was fine tuned on 12,600 unique frames of our training data, evenly split among the three conditions, with two frames per batch. We use the quarter-sectioned 300×300 pixel images to minimize the effects of downsampling, as ResNet50 takes as input 224×224 pixel images and downsamples larger inputs. The brightfield phase images were converted from grayscale to RGB to match the required input parameters of Resnet50. The CNN was fine tuned with the Adam optimizer[14] for forty epochs with a learning rate of 1×10^{-3}. The same procedure and architecture was also used to train a CNN classifier on a subset of frames from between 800 and 9996 minutes (frames 200 through 249 out of 350 total) into the experiment, a total of 1800 images, to evaluate the time dependence of the predictions. The model was trained on an NVIDIA A30 GPU using CUDA, PyTorch, and PyTorch lightning.

To obtain "visual explanations" for the classification of each frame, we applied the gradient-weighted class activation mapping (Grad-CAM) technique[10] to the model for each frame of the validation set. This technique computes the gradients of the target class score with respect to the feature maps of the final convolutional layer of the network. These gradients are pooled across the convolutional filter to provide a spatial-average importance value for different regions of the input image that contribute to the target class score. Grad-CAM returns an "importance" of each pixel to the final prediction that can be superimposed onto the original images and visually inspected to identify relevant image details. Model training and analysis code is available at `https://github.com/25marcusb/Understanding-TCR-T-cell-knockout-behavior-using-interpretable-machine-learning`.

3. Results

3.1. *Convolutional neural networks can distinguish between different genetic perturbations from a single frame*

We first validate the predictive ability of the trained CNN classifier to distinguish the genetic perturbation given a section of a live-cell imaging frame. The CNN was trained to distinguish between three classes - RASA2KO T cells, CUL5KO T cells, and SHKO T cells. RASA2KO and CUL5KO are known to improve T cell anti-cancer activity after repetitive stimulation.[4,5] The SHKO T cells, with *AAVS1* knockouts as a negative control, should be "exhausted" after repetitive stimulation, leading to less anti-cancer activity.[4] On held-out validation data, the model assigns more than 50% probability to the correct class on 2,974 out of 4,200 test

images evenly balanced across classes, an accuracy of 71%. The CNN model outperforms a traditional support vector machine (SVM) classifier that predicts the perturbation from cell counts from segmentation, which has a test accuracy of only 50%. We find that the model has consistent precision around 70%, e.g., the fraction of true RASA2KO frames out of the set of all frames predicted to be RASA2KO, across the genetic perturbations (Table 1). However, we observe that the ability to recall the SHKO control condition is much worse than the ability to recall the genetic perturbations (Table 1). While the "confusion" with the safe harbor control indicates that many features of T cell and cancer cell dynamics are maintained after perturbation, the relatively low number of incorrect cross predictions between the two genetic knock-outs suggests the model can differentiate the changes from CRISPR perturbation and be used as a tool to interrogate the different behaviors.

	Predicted SHKO	Predicted CUL5KO	Predicted RASA2KO	Recall
True SHKO	387	609	373	28%
True CUL5KO	67	1318	13	94%
True RASA2KO	114	7	1268	91%
Precision	68%	68%	76%	

Table 1: **Full prediction model confusion matrix.** The rows represent the true labels for the three experiment types; the columns represent the predicted labels. The last row and column of the matrix are the precision and recall, respectively, for each experiment class label.

We observe a relationship between the collection time of the image and the ability to accurately classify its genetic perturbations. For the control SHKO images, the model tends to classify early time frames as CUL5KO and later images as RASA2KO (Figure 1). To better understand how time affects classification performance, we trained a *limited-time* model with the same architecture, but we restricted the training data to include only images from frames 200 to 249, between 800 and 996 minutes post culture, around the inflection point of RASA2KO and CUL5KO mis-classification. We find that this limited-time model has higher validation accuracy of 89% on held-out data (also in the same time window), and makes relatively few misclassifications (Table 2). CNN-based prediction again outperforms a cell count based SVM classifier, which has an overall test accuracy of 64%. This model does not generalize well to early time frames, but has above 75% accuracy in the 200 minute periods before and after its training data (Figure 2). The inability to generalize well to early time frames is expected given the lack of differentiation between all three conditions in the early parts of the experiment. More generally, this change in predictive ability over time reveals that genetic perturbations may affect the dynamics and timing of immune and cancer cell interactions.

3.2. *Explainable AI techniques reveal differences in T cell interactions with cancer cells under genetic perturbation*

To better understand the differences in behavior across genetic perturbations, we applied the Grad-CAM technique[10] to both full- and restricted-time models and testing with held-out

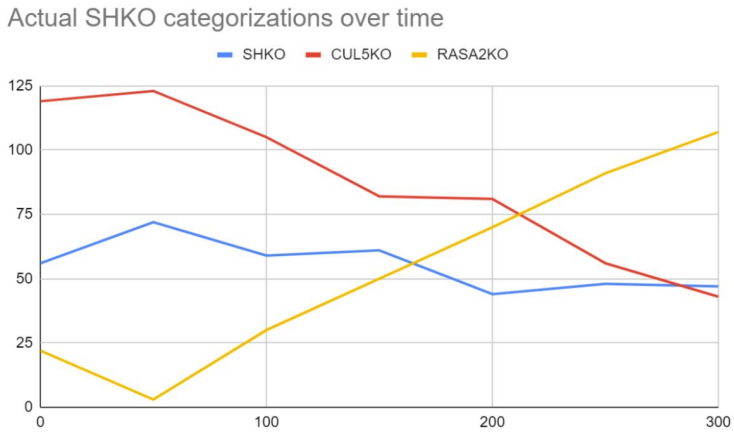


Fig. 1: **SHKO categorizations over time.** Each point corresponds to the 50 frame time bucket starting at that frame (time). A total of 200 images per time bucket are categorized.

	Predicted SHKO	Predicted CUL5KO	Predicted RASA2KO	Recall
True SHKO	184	9	6	92%
True CUL5KO	1	198	1	99%
True RAS2KO	41	1	154	77%
Precision	81%	95%	86 %	

Table 2: **Limited time (frames 200 - 250) test confusion matrix.** Rows represent true labels for the three experiment types; columns represent predicted labels. The last row and column of the matrix are the precision and recall, respectively, for each experiment class label.

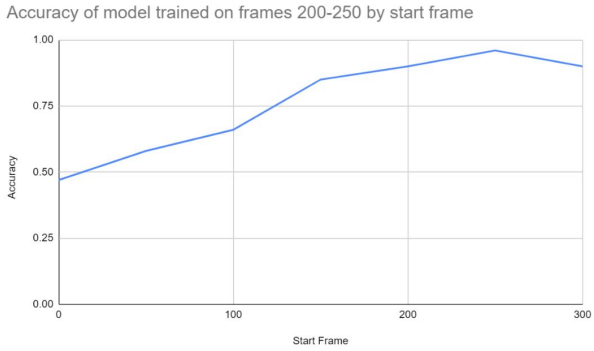


Fig. 2: **Accuracy of the limited-time model across all held-out time points data.** These held-out test accuracy results (y-axis) were aggregated by time (x-axis) into five groups.

validation data. For an individual sample's prediction, Grad-CAM combines the gradients of the model's weight to calculate the influence of each pixel feature to the prediction. These values generate a "feature importance heatmap" that identifies the most important regions of an image for classifications.

We analyzed the output of Grad-CAM across different time points and different conditions

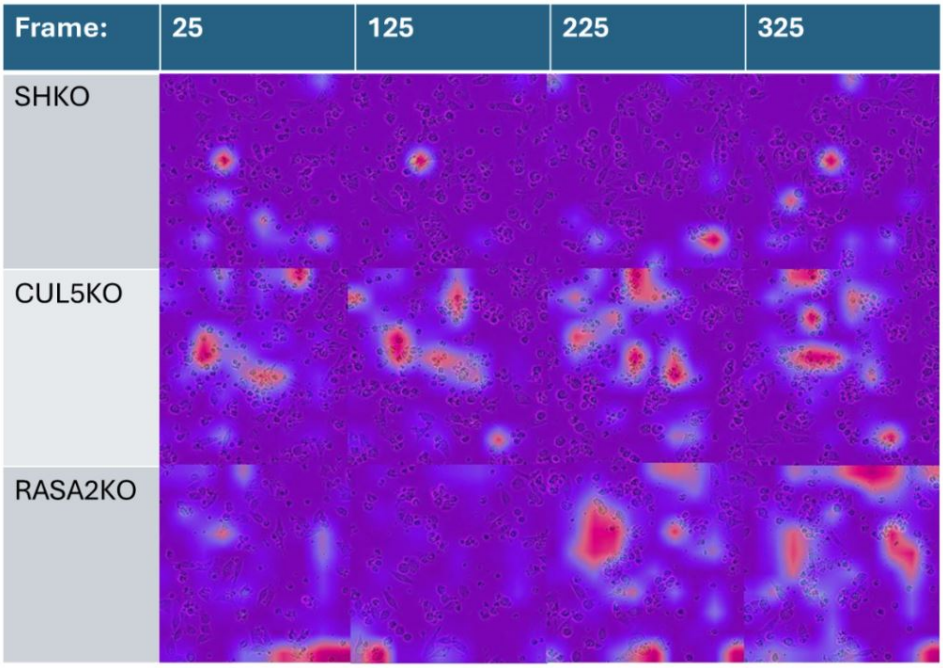


Fig. 3: **Grad-CAM importance scores for the limited-time model across condition and time on held-out images.** Frame label indicates both the true label and associated Grad-CAM class label. The purple areas represent the lowest impact areas, blue represents the medium impact, and red represents the highest impact. This color gradient is consistent for all of the Grad-CAM visualizations throughout this paper.

to identify changes recognized by CNN classifiers. For the model trained on all time points, we observe that Grad-CAM highlights interactions between cancer cells and T cells, focusing its attention on the cellular aggregates to recognize CUL5KO (Figure 3). In the RASA2KO Grad-CAM visualizations, on the other hand, the highlighted regions are focused almost exclusively on the areas between cells and cellular aggregates (Figure 3). Moreover, we observed that the highlighted regions in the SHKO group seem to be distributed randomLy, but each time focused on individual cancer cells (Figure 3). These Grad-CAM visualizations suggest specific characteristics of behavior of each of the three experiments.

We quantified the enrichment of these patterns on a small scale in the three experiments by manually annotating the number of healthy and interacting T cells in the highlighted regions of each type on the frame interval 150-160 in the second quadrant from the full time frame model in the held-out images. Across all three sets of heatmaps, the CUL5KO Grad-CAM heatmap highlights the interacting cancer cells at a higher rate than the SHKO and RASA2KO Grad-CAM heatmap (Table 3). This suggests that the difference between CUL5KO and RASA2KO behavior is that CUL5KO T cells accelerate the rate of cancer cell-T cell interactions and the formation of T cell aggregates around a cancer cell.

To better understand differences between the full-time and limited-time models, we compared the Grad-CAM visualizations at the same time from the same position. We use to illustrate one input frame from the CUL5KO held-out data at time point 220,880 minutes post

Genetic perturbation	Non-interacting cancer cells	Interacting cancer cells
SHKO	10	5
CUL5KO	2	32
RASA2KO	12	7

Table 3: Number of healthy and interacting cancer cells from ten images between frames 150 and 160 highlighted by Grad-CAM on the full time frame model.

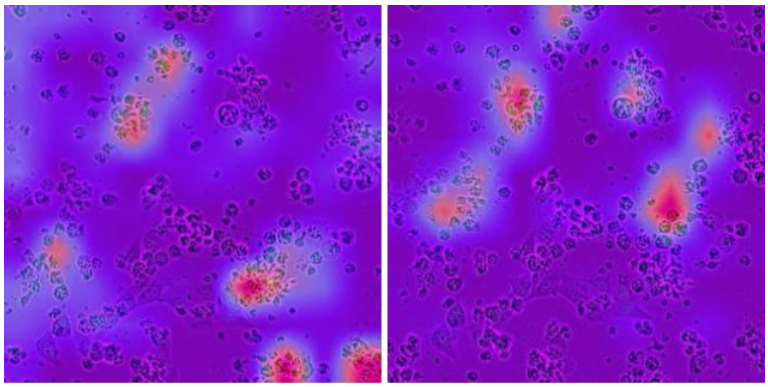


Fig. 4: Grad-CAM visualizations of frame 220 of the held out CUL5 images for the 200-249 frame model (left) and full 350 frame model (right).

culture (Figure 4). Both sets of Grad-CAM images focus on interacting cancer cells—which often appear as large T cell aggregates that hide the seed cancer cell—but the interacting cancer cells they highlight are often different ones (Figure 4). The 200-249 frame model focuses more on regions of the image without aggregates to inform its decision, indicating that overall T cell/cancer cell coverage is an important signature during this 50-frame time window. Although both visualizations appear to focus on similar proportions of the image, the 200-249 frame model's heatmap has a larger area of limited attention across the full image (Figure 4).

To more broadly interrogate the influences on the limited-time frame model, we used Grad-CAM to visualization importance heatmaps across the three different genetic perturbation on held-out frames (Figure 5). Like the full-time model (Figure 3) the limited-time model focuses on interacting cancer cells, which we define as T cells adjacent to or overlapping with cancer cells in the CUL5KO held-out frames. The limited-time model, however, has more diffused highlighted regions of importance for predicting all three conditions than the full time model, capturing most of the cells. This suggests that *total cell coverage*, or the proportion of the area of the image covered by cells, is a more defining signature of the CUL5KO limited-time model than the full-time model. When comparing shorter time intervals, CUL5KO can be characterized by its total cell coverage, but, over longer intervals, the specific interactions between cells proves to be the most important distinguishing feature.

Taken together, our findings indicate that the model trained on later limited time frames takes a larger proportion of the image into account when performing classification, whereas the full model focuses on more limited regions of the image. The greater spread of "attention" and

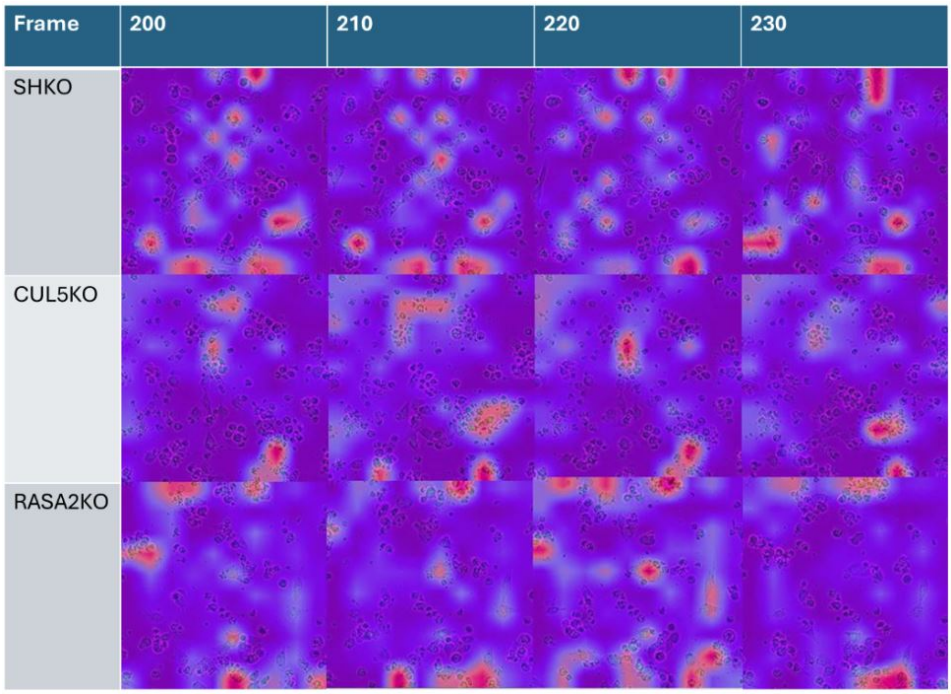


Fig. 5: **Grad-CAM importance scores for the limited-time model across condition and time on held out images.** Frame label indicates both the true label and associated Grad-CAM class label.

the focus on multiple cancer cells and T cells in the limited-time model suggests that the model is effectively counting the number of cells to make a prediction; at limited time points, given the known differences in killing progression, this featurization would be effective for separating the classes as indicated by the limited-time model's accuracy. In contrast, the model trained on the full data cannot rely on the number of cells to differentiate genetic perturbations, and so focuses more on a small number of cell interaction regions to distinguish the knock-outs.

4. Conclusion

Our work analyzes the behaviors of CRISPR-modified TCR T cells interacting with cancer cells in live-cell imaging studies. Most studies count the total area covered by cancer cells across time to characterize the cancer cell killing efficacy of the modified T cells, ignoring the behavior changes in T cell, cancer cells, and their interactions. We identified specific changes in cell behavior across the three experiment types by using Grad-CAM,[10] a visual explainable AI technique, to understand how a deep learning classifier would differentiate the experimental conditions from live-cell images, highlighting the behavioral changes in modified T cells beyond simple cancer cell death rate.

By using Grad-CAM to analyze classification models trained on three types of modified TCR T cells, we found that the amount of T cell or cancer cell coverage differentiates CUL5KO and RASA2KO modified T cells when comparing similar time points. We showed that cell aggregation behavior is a reliable differentiating characteristic to distinguish CUL5KO experiments from the others, and these CUL5KO experiments tend to have consistent T cell

aggregates around cancer cells. We showed that larger empty spaces, suggesting a combination of larger (and fewer) cellular aggregates plus better cancer cell killing, distinguishes RASA2KO experiments from the other two experiments. We found that the safe harbor experiment is defined by no cellular aggregates and substantial coverage of cancer cells, with the T cells both failing to latch on to the cancer cells and furthermore failing to stop their proliferation. We note that coverage plots alone miss these important behavioral signatures.

Our study has a number of limitations, including considering only a single T cell donor and cancer cell line, three genetic modifications, limited replicates, and limited variable titrations of cancer cells to modified T cells. Emerging architectures and pretrained models may improve accuracy relative to the ResNet architecture used here. The importance maps from Grad-CAM are coarse regions over the image, and sometimes the difference between knockouts could be hard to qualitatively observe. Grad-CAM is one of many interpretation approaches, and alternatives such as saliency maps or Shapley Additive Explanations may provide different features of interest. The lack of differences between knockouts may also be an interesting indication of a lack of distinct mechanism changes that would limit downstream efficacy.

However, as a proof of concept, this analysis pipeline for future live-cell imaging experiments will open the door to a more sophisticated interpretation of modified T cell behaviors. We found that existing tools and pretrained image models like ImageNet are effective at classifying biological image samples when fine tuned using live-cell imaging frames. We observed that fine tuning on frames from a wide stretch of time increases the models' attention on individual cellular dynamics, while fine tuning on short time samples later in the experiment will use more characteristic image features for classification.

Overall, we demonstrated that explainable AI techniques are a practical tool for interrogating and understanding biological dynamics from live-cell image, and we developed a framework for studying these dynamics in general live-cell imaging data. Future work pushes our methods towards the clinic. By characterizing the complex behaviors of these possible T cell modifications, we hope to more rapidly identify T cell therapies for broad ranges of cancers, both liquid and solid. Our interpretable classifiers specifically can be used by decision-making AI methods to prioritize specific T cell therapies for new cancer patients by predicting the response of that individual tumor to each type of therapy, and selecting the most effective therapy.

Acknowledgements

M.B., A.V., and B.E.E. were funded by NIH/NCI 5U2CCA233195, NIH/NHGRI R01 HG012967, the Parker Institute for Cancer Immunotherapy (PICI), and NIH/NHGRI R01 HG013736. J.C. was supported by NIH/NCI K08, 1K08CA252605-01, a Burroughs Wellcome Fund Career Award for Medical Scientists, the Lydia Preisler Shorenstein Donor Advised Fund, and the Parker Institute for Cancer Immunotherapy, PICI. B.E.E. is a CIFAR Fellow in the Multiscale Human Program. B.E.E. is on the SAB for ArrePath Inc, Crayon Bio, and Freenome; she consults for Neumora.

References

1. R. C. Sterner and R. M. Sterner, CAR-T cell therapy: current limitations and potential strategies, *Blood Cancer Journal* **11**, p. 69 (2021).

2. A. D. Fesnak, C. H. June and B. L. Levine, Engineered T cells: the promise and challenges of cancer immunotherapy, *Nature Reviews Cancer* **16**, 566 (2016).
3. L. Gaissmaier, M. Elshiaty and P. Christopoulos, Breaking bottlenecks for the TCR therapy of cancer, *Cells* **9**, p. 2095 (2020).
4. J. Carnevale, E. Shifrut, N. Kale, W. A. Nyberg, F. Blaeschke, Y. Y. Chen, Z. Li, S. P. Bapat, M. E. Diolaiti, P. O'Leary *et al.*, RASA2 ablation in T cells boosts antigen sensitivity and long-term function, *Nature* **609**, 174 (2022).
5. X. Liao, W. Li, H. Zhou, B. K. Rajendran, A. Li, J. Ren, Y. Luan, D. A. Calderwood, B. Turk, W. Tang *et al.*, The CUL5 E3 ligase complex negatively regulates central signaling pathways in CD8+ T cells, *Nature Communications* **15**, p. 603 (2024).
6. F. Blaeschke, Y. Y. Chen, R. Apathy, B. Daniel, A. Y. Chen, P. A. Chen, K. Sandor, W. Zhang, Z. Li, C. T. Mowery *et al.*, Modular pooled discovery of synthetic knockin sequences to program durable cell therapies, *Cell* **186**, 4216 (2023).
7. M. Alieva, A. K. Wezenaar, E. J. Wehrens and A. C. Rios, Bridging live-cell imaging and next-generation cancer treatment, *Nature Reviews Cancer* **23**, 731 (2023).
8. J. M. Matthews, B. Schuster, S. S. Kashaf, P. Liu, R. Ben-Yishay, D. Ishay-Ronen, E. Izumchenko, L. Shen, C. R. Weber, M. Bielski *et al.*, OrganoID: A versatile deep learning platform for tracking and analysis of single-organoid dynamics, *PLoS Computational Biology* **18**, p. e1010584 (2022).
9. C. Deben, E. C. De La Hoz, M. L. Compte, P. Van Schil, J. M. Hendriks, P. Lauwers, S. K. Yogeswaran, F. Lardon, P. Pauwels, S. Van Laere *et al.*, OrBITS: label-free and time-lapse monitoring of patient derived organoids for advanced drug screening, *Cellular Oncology* **46**, 299 (2023).
10. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
11. A. Saranya and R. Subhashini, A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends, *Decision Analytics Journal* **7**, p. 100230 (2023).
12. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
13. L. Challenge, Imagenet http://www. image-net. org/challenges, *LSVRC/2012/results. html* (2012).
14. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
15. B. He, Q. Lu, J. Lang, H. Yu, C. Peng, P. Bing, S. Li, Q. Zhou, Y. Liang and G. Tian, A new method for CTC images recognition based on machine learning, *Frontiers in Bioengineering and Biotechnology* **8**, p. 897 (2020).
16. R. Kumar, S. Joshi and A. Dwivedi, CNN-SSPSO: a hybrid and optimized cnn approach for peripheral blood cell image recognition and classification, *International Journal of Pattern Recognition and Artificial Intelligence* **35**, p. 2157004 (2021).

Amyloid, Tau, and APOE in Alzheimer's Disease: Impact on White Matter Tracts

Bramsh Qamar Chandio[1†], Julio E. Villalon-Reina[1], Talia M. Nir[1], Sophia I. Thomopoulos[1], Yixue Feng[1], Sebastian Benavidez[1], Neda Jahanshad[1], Jaroslaw Harezlak[2], Eleftherios Garyfallidis[2], Paul M. Thompson[1], and for the Alzheimer's Disease Neuroimaging Initiative*

[1]*Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Marina del Rey, CA, USA*

[2]*Indiana University Bloomington, IN, USA.* [†]*E-mail: chandio@usc.edu*

Alzheimer's disease (AD) is characterized by cognitive decline and memory loss due to the abnormal accumulation of amyloid-beta (Aβ) plaques and tau tangles in the brain; its onset and progression also depend on genetic factors such as the apolipoprotein E (APOE) genotype. Understanding how these factors affect the brain's neural pathways is important for early diagnostics and interventions. Tractometry is an advanced technique for 3D quantitative assessment of white matter tracts, localizing microstructural abnormalities in diseased populations *in vivo*. In this work, we applied BUAN (Bundle Analytics) tractometry to 3D diffusion MRI data from 730 participants in ADNI3 (phase 3 of the Alzheimer's Disease Neuroimaging Initiative; age range: 55-95 years, 349M/381F, 214 with mild cognitive impairment, 69 with AD, and 447 cognitively healthy controls). Using along-tract statistical analysis, we assessed the localized impact of amyloid, tau, and APOE genetic variants on the brain's neural pathways. BUAN quantifies microstructural properties of white matter tracts, supporting along-tract statistical analyses that identify factors associated with brain microstructure. We visualize the 3D profile of white matter tract associations with tau and amyloid burden in Alzheimer's disease; strong associations near the cortex may support models of disease propagation along neural pathways. Relative to the neutral genotype, APOE $\epsilon3/\epsilon3$, carriers of the AD-risk conferring APOE $\epsilon4$ genotype show microstructural abnormalities, while carriers of the protective $\epsilon2$ genotype also show subtle differences. Of all the microstructural metrics, mean diffusivity (MD) generally shows the strongest associations with AD pathology, followed by axial diffusivity (AxD) and radial diffusivity (RD), while fractional anisotropy (FA) is typically the least sensitive metric. Along-tract microstructural metrics are sensitive to tau and amyloid accumulation, showing the potential of diffusion MRI to track AD pathology and map its impact on neural pathways.

Keywords: diffusion MRI, tractometry, Alzheimer's disease, amyloid, tau

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but most of them did not participate in the analysis or writing of this report. A complete listing of ADNI investigators may be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by progressive cognitive decline and memory loss. Central to its pathology are the abnormal accumulation of amyloid-beta ($A\beta$) plaques and tau tangles in the brain.[1–3] The onset and progression of these pathological processes are influenced by genetic factors such as the apolipoprotein E (APOE) gene.[4] AD pathology not only affects gray matter but also profoundly impacts white matter tracts, which serve as the brain's communication mechanism; these tracts connect different brain regions and facilitate efficient signal transmission. Understanding how amyloid, tau, and APOE influence white matter integrity is crucial for developing early diagnostic tools and monitoring the effects of targeted interventions on the brain.

Amyloid-beta peptides aggregate to form plaques, primarily affecting gray matter[5] but also extending to white matter tracts by disturbing cellular function.[6,7] $A\beta$ deposition leads to myelin degradation, which disrupts the insulating layer around nerve fibers, and axonal injury, which impairs neurons' ability to communicate effectively. Tau is a microtubule-associated protein that stabilizes microtubules in neurons. In AD, tau becomes hyperphosphorylated and forms neurofibrillary tangles,[8] affecting microtubule stability, disrupting axonal transport, and impairing neuronal function.[9,10] The apolipoprotein E (APOE) gene plays a crucial role in lipid metabolism and is a significant genetic factor influencing the risk of developing Alzheimer's disease.[11,12] See footnote for APOE gene types. [a]

Diffusion MRI could offer a less invasive alternative to PET, helping identify affected white matter tracts and leading to personalized therapeutic strategies. Diffusion MRI[15–17] measures water diffusion in the brain, revealing the microstructural properties of the underlying tissue. Tractography, derived from diffusion MRI data,[18–20] maps and visualizes white matter pathways by tracking the directional profiles of water diffusion, providing a detailed picture of brain connectivity. Tractometry enhances this by quantifying specific microstructural properties, such as fractional anisotropy (FA) or mean diffusivity (MD), along the length of individual tracts. This technique maps microstructural alterations in the brain's white matter tracts.[21–25] It analyzes the coherence of neural connections, allowing for precise assessment of characteristic changes in neurological conditions such as Alzheimer's disease or Parkinson's disease.[26]

White matter (WM) microstructure changes with age, and there is a regional variation in the age-dependent trajectories of maturation and decline for the major white matter metrics across the lifespan.[27,28] Several studies of regional microstructure in Alzheimer's disease have used tract-based spatial statistics (TBSS),[29] to link microstructural metrics in specific brain regions to amyloid positivity and clinical dementia severity.[30–32] However, the resolution of

[a]APOE gene has three common variants: $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$. APOE $\epsilon 2$ is the least common, and carriers have a lower risk of developing AD. It may have a protective effect on white matter structure,[13,14] leading to less degeneration, possibly due to enhanced lipid metabolism and repair mechanisms. APOE $\epsilon 3$ is the most common variant and is considered neutral, while the APOE $\epsilon 4$ variant is the greatest known common genetic risk factor for late-onset AD, roughly tripling lifetime risk of AD per allele carried.[11] APOE $\epsilon 4$ is less effective in clearing $A\beta$ from the brain, leading to greater $A\beta$ plaque accumulation and subsequent white matter damage.[12]

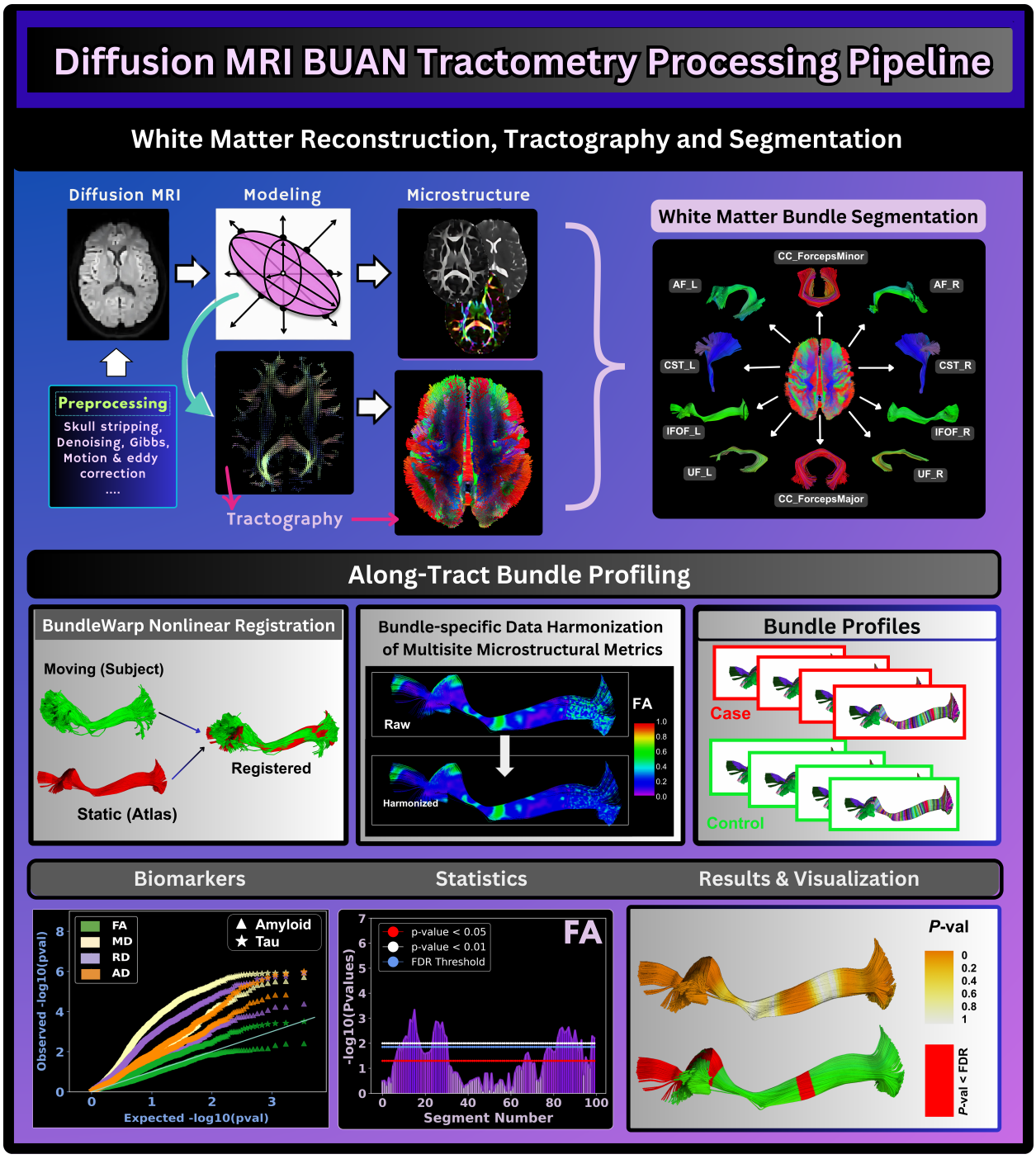


Fig. 1: **BUAN Tractometry Pipeline:** The brain's major neural pathways are digitally reconstructed using diffusion MRI and tractography techniques. Specific white matter tracts are then extracted for visualization and detailed analysis, allowing for localized and focused examination of brain pathways.

TBSS maps is limited by the regions defined in the atlases used.[29] To address this, tractometry methods such as BUAN (Bundle Analytics)[23] map microstructural parameters along the length of white matter tracts, mapping disease effects on neural pathways in 3D and at a finer

anatomical scale.[23,25,26,33,34] Recently, Ba Gari *et al.*[33] used a tractography-based medial tract analysis (MeTa) to enhance the sensitivity for detecting associations of AD, amyloid, and tau with diffusion tensor imaging (DTI) derived microstructural metrics, compared to TBSS.

In this study, we applied our advanced tractometry method, BUAN (Bundle Analytics), to evaluate the impact of amyloid, tau, APOE $\epsilon4$, and APOE $\epsilon2$ on the microstructure of the brain's white matter tracts. BUAN maps the microstructural properties of white matter tracts, and fits along-tract statistical models to detect effects on microstructure that are associated with amyloid plaques, tau tangles, and different APOE genotypes. This is crucial for understanding the effects of AD pathology on brain connectivity. Overall, we found that a range of microstructural metrics were sensitive to tau and amyloid, the two key biomarkers for detecting Alzheimer's disease, supporting the role of diffusion MRI as a non-invasive measure of AD pathology. Relative to APOE $\epsilon3/\epsilon3$ carriers, microstructural alterations were also identified in APOE $\epsilon4$ carriers and, to a lesser extent, in $\epsilon2$ carriers. Mean diffusivity (MD) was most strongly associated with AD pathology, followed by axial diffusivity (AxD) and radial diffusivity (RD). Fractional anisotropy (FA) was the least sensitive metric. The tendency to detect stronger associations in tract regions closer to the cortex may support propagative or "epidemic spreading" models of AD pathology,[35] which argue that AD pathology spreads dynamically along neural pathways or in functionally synchronous networks; future longitudinal studies are needed to verify this.

2. Methods

Data from 730 ADNI3 participants (phase 3 of the Alzheimer's Disease Neuroimaging Initiative; age range: 55-95 years, 349M/381F, 214 with mild cognitive impairment (MCI), 69 with AD, and 447 cognitively healthy controls (CN)) scanned with 7 acquisition protocols (GE36, GE54, P33, P36, S127, S31, S55) were included. Tables 1 and 2 in Fig. 2 detail demographic and acquisition protocol information. Aβ-status, i.e., positive (Aβ+) or negative (Aβ–), was determined by either mean 18F-florbetapir (Aβ+ defined as >1.11)[36,37] or florbetaben (Aβ+ defined as >1.20)[38,39] PET cortical SUVR uptake, normalized by using a whole cerebellum reference region. Tau positivity was defined as a tau SUVR > 1.23.

2.1. *Diffusion MRI Processing*

Raw diffusion MRI (dMRI) were preprocessed using the ADNI3 dMRI protocol.[40,41] Preprocessing of raw diffusion MRI (dMRI) data involved several steps: denoising raw dMRI data using DIPY's principal component analysis (PCA) for GE data, and Marchenko-Pastur PCA for Siemens and Philips data denoising.[42,43] Gibbs artifacts were corrected using MRtrix's *degibbs* tool,[44,45] and extracerebral tissue was removed (skull stripping) with FSL's BET.[46,47] Eddy currents and motion were corrected using FSL's eddy_cuda tool with additional corrections for slice-to-volume and outlier detection.[47,48] Bias field inhomogeneities were corrected using MRtrix's *dwibiascorrection* ANTS function. Preprocessed T1w images from the ADNI database were further processed and aligned to the dMRI data.[45,49] ADNI3 dMRI data lacked reversed phase-encode blips, so echo-planar imaging (EPI) distortion corrections were made using nonlinear registrations to T1-weighted anatomical images. The processed dMRI data were converted back to native space through a series of inversions of the registration matrices, with final outputs visually inspected and manually adjusted as necessary. The DTI model was

Table 1. Participant demographics. 'Site N' denotes the number of sites across North America using the dMRI acquisition protocol specified (protocols are further detailed in **Table 2**). * NB: some sites collected data with more than one protocol, so the total number of sites is less than the sum of the final column.

Protocol	Total N	Diagnosis			M	F	Age (Mean±SD)	CDR-sob (Mean±SD)	Amyloid		Tau		Site N*
		CN	MCI	Dementia					Aβ+	Aβ-	Tau+	Tau-	
S55	270	175	71	24	120	150	74.4±7.9	0.93±2.0	103	117	67	141	18
S31	96	59	27	10	39	57	72.0±8.5	0.96±1.8	26	49	21	56	10
S127	112	68	36	8	53	59	73.7±7.6	0.78±1.5	35	58	21	67	14
P36	46	20	23	3	24	22	73.1±6.9	0.84±1.1	12	27	15	22	5
P33	56	37	14	5	34	22	75.7±7.7	1.03±2.1	19	21	11	26	6
GE36	42	21	17	4	24	18	72.2±7.1	1.29±2.7	13	22	9	26	4
GE54	108	67	26	15	55	53	76.1±8.1	1.16±2.0	37	57	24	63	10
Total	730	447	214	69	349	381	74.1±7.9	0.97±1.9	245	351	168	401	58

Table 2. ADNI3 dMRI acquisition protocols.

Protocol	Scanner		Volumes			Matrix	Acquisition Time (min:sec)
	Vendor	Model	b_0	DWI	Total		
GE36	GE	Basic Widebore 25x	4 b=0 s/mm^2	32 b=1000 s/mm^2	36	256x256	9:52
GE54	GE	Basic 25x	6 b=0 s/mm^2	48 b=1000 s/mm^2	54		7:09
P33	Philips	Basic Widebore	1 b=0 s/mm^2	32 b=1000 s/mm^2	33	128x128	7:32
P36	Philips	Basic Widebore R3	1 b=0 s/mm^2 3 b=2 s/mm^2	32 b=1000 s/mm^2	36		6:54
S31	Siemens	Basic VB17	1 b=0 s/mm^2	30 b=1000 s/mm^2	31	116x116	7:02
S55	Siemens	Basic Skyra E11 & Prisma D13	7 b=0 s/mm^2	48 b=1000 s/mm^2	55		9:18
S127	Siemens	Advanced Prisma VE11C	13 b=0 s/mm^2	6 b=500 s/mm^2 48 b=1000 s/mm^2 60 b=2000 s/mm^2	127		7:25*

Abbreviations

Left/Right (L/R)	Anterior Commissure (AC)
Corticospinal Tract (CST)	Corpus Callosum Mid (CCMid)
Fornix (F)	CC Forceps_Minor
Optic Radiation (OR)	CC Forceps_Major
Frontopontine Tract (FPT)	Posterior Commissure (PC)
Occipitopontine Tract (OPT)	Cerebellum (CB)
Cingulum (C)	Vermis (V)
Arcuate Fasciculus (AF)	Medial Lemniscus (ML)
Extreme Capsule (EMC)	Medial Longitudinal Fasciculus (MLF)
Inferior Fronto Occipital Fasciculus (IFOF)	Spinothalamic Tract (STT)
Inferior Longitudinal Fasciculus (ILF)	Fractional Anisotropy (FA)
Middle Longitudinal Fasciculus (MdLF)	Mean Diffusivity (MD)
Superior Longitudinal Fasciculus (SLF)	Axial Diffusivity (AD)
Uncinate Fasciculus (UF)	Radial Diffusivity (RD)

Fig. 2: Tables 1 and 2 detail demographic and scanner protocol information for the ADNI3 data used in our experiments (data from Thomopoulos et al, 2021). The abbreviation table on the right lists the 38 white matter tracts and four microstructural measures analyzed in this work.

used to extract 4 microstructural measures from processed dMRI: FA, MD, AxD, RD.

2.2. *BUAN Tractometry*

Fig. 7 illustrates the detailed steps of the BUAN tractometry pipeline, along with visualizations of the process. We applied a robust and unbiased model-based spherical deconvolution[50] reconstruction method and a probabilistic particle filtering tracking algorithm that uses tissue partial volume estimation (PVE) to reconstruct [51] whole-brain tractograms. For tracking, the seed mask was created from the white matter (WM) PVE (WM PVE > 0.5), seed density per voxel was set to 2, and step size was set to 0.5. We extracted 38 white matter (WM) tracts from tractograms using RecoBundles [23,52] (see Fig. 2 for full names) using model bundles from the HCP-842 tractography atlas.[53]

After extracting WM bundles, we nonlinearly registered each subject's bundles to model bundles in MNI-space using a streamline-based nonlinear registration method, BundleWarp.[54] Optimal registration of tracts to atlas bundles is crucial for finding accurate segment correspondences among subjects and populations. This enhances the sensitivity of group statistical analyses by eliminating errors due to misalignment across subjects.

BUAN creates the bundle profiles for each bundle using 4 DTI-based microstructural metrics: FA, MD, RD, and AxD calculated in the diffusion native space (see Figure 2 for full bundle names). Bundle profiles are created by dividing the bundles into 100 horizontal segments using the model bundle centroids along the length of the tracts in common space. We cluster our model bundles using the QuickBundles[55] method to obtain a cluster centroid consisting of 100 points per centroid. We calculate Euclidean distances between every point on every streamline of the bundle and 100 points in the model bundle centroid. A segment number is assigned to each point in a bundle based on the shortest distance to the nearest model centroid point. The streamlines are not resampled to have a specific number of points,

and we do not change the distribution of points. Since the assignment of segment numbers is performed in the common space, we establish the segment correspondence among subjects from different groups and populations. Microstructural measures such as FA are then projected onto the points of the bundles in native space. Note that the nonlinearly moved bundles are only used to assign segment numbers to streamlines (and points on the streamlines) in the bundles. Actual statistical analysis always takes place in the native space of the diffusion data. The statistical analysis step uses bundles of the original shape and microstructural measures in the native space using segment labels given during the assignment step for segment-specific group analysis.

Bundle profiles are harmonized using the ComBat method[56,57] to correct for scanner/site effects as described in the harmonized BUAN tractometry pipeline.[58] After data harmonization, we assume each bundle type has its own data distribution, which is considered independent of the rest of the bundles in the brain. For each tract and metric, we pool bundle profiles for a given tract across all subjects from CN, MCI, and AD groups. Pooled bundle profiles consist of 100 segments, and each segment is modeled as a feature. Linear Mixed Models are applied to WM bundles; age and sex are modeled as fixed effects and scanner and subject as a random effect term, the response variable being each DTI metric. Though we harmonized the profiles with ComBat, we further account for scanner and/or site effects by adding it as a random term in the linear mixed models (LMMs)[59] to eliminate any remaining artifacts contributed by scanner/site. We used FURY[60] software to visualize tractometry results in this paper. Fig. 7 provides a comprehensive view of how the bundle-specific nonlinear registration and data harmonization are applied, leading to a focused analysis of specific regions along the tracts.

2.3. *Statistics*

We used LMMs to test the effects of amyloid positivity, tau positivity, and different APOE variants on 38 white matter tracts. In each experiment, age and sex were modeled as fixed effects, and the scanner and subject were modeled as random terms. Multiple testing correction was performed using the False Discovery Rate (FDR)[61] method at P-value < 0.05. See footnote for details on FDR correction applied to WM tracts. [b]

[b]Multiple testing correction is a statistical adjustment process that can control the rate or likelihood of false positives when performing numerous simultaneous tests.[62] In neuroimaging studies, where thousands of brain regions or voxels are analyzed for significant differences or correlations, this adjustment is crucial. It ensures the integrity and reliability of the results by controlling the overall rate of false positives. Common correction methods include the Bonferroni correction,[63] which is stringent and adjusts the significance threshold by dividing it by the number of tests, and the False Discovery Rate (FDR)[61] method, which limits the proportion of false positives among significant findings. These corrections ensure that detected effects are truly significant and not due to random variation. As white matter tracts generated by tractography are not as extensively studied as voxel or ROI-based methods, selecting the appropriate multiple testing correction is challenging. We divided each bundle into 100 segments; for tract-specific FDR correction, we use 100 p-values per bundle to correct for multiple tests using the FDR method. We refer to this bundle-specific FDR corrected threshold as the local threshold, as it only depends on statistics within that bundle. Additionally, we performed multiple test corrections across all bundles in the brain by pooling 100 p-values from each of the 38 tracts, yielding a total of 3,800 p-values to determine the global FDR-corrected threshold.

3. Results

We ran the following five experiments to detect associations of various variables on 38 white matter tracts of the brain. We tested microstructural associations (1) with amyloid positivity; (2) with tau positivity; (3) comparing non $\epsilon 4$ carriers $\epsilon 2\epsilon 3/\epsilon 3\epsilon 3/\epsilon 2\epsilon 2$ with subjects carrying at least one $\epsilon 4$ gene; $\epsilon 2\epsilon 4/\epsilon 3\epsilon 4/\epsilon 4\epsilon 4$, (4) comparing $\epsilon 3\epsilon 3$ with $\epsilon 3\epsilon 4/\epsilon 4\epsilon 4$, and (5) comparing $\epsilon 3\epsilon 3$ with $\epsilon 3\epsilon 2/\epsilon 2\epsilon 2$.

As an overview of the results, quantitative quantile-quantile (QQ) plots (Fig. 3) summarize the overall association signal detected across all 38 white matter bundles between each of the biomarkers (amyloid, tau, and APOE) and each of the DTI metrics (FA, MD, RD, and AxD). These plots visually represent the strength of associations between these biomarkers and DTI metrics, helping to identify which combinations show the most significant relationships. In the visualization layout, the first row of QQ plots highlights which DTI metric exhibits the strongest association with each biomarker. Here, the p-values of the 38 tracts were pooled for each DTI metric and presented in these plots, allowing for a comprehensive assessment of each metric's sensitivity to changes in biomarker levels. The second focuses on the relationship from the opposite perspective: for each DTI metric, it shows which biomarker shows significant associations (the scale of the y-axis varies across the QQ plots to adapt to the observed range of p-values).

Fig. 6 visualizes the p-values along the length of the 34 major tracts (4 tracts with mostly null results were excluded from the plots). In this figure, the x-axis represents 100 segments per tract, displaying the p-value for each segment, while the y-axis corresponds to the individual bundles. Segments highlighted in green indicate p-values less than 0.05, signifying regions of statistically significant associations detected between the metric and Alzheimer's disease biomarkers for that particular bundle and segment. This visualization provides a more detailed view of where significant effects are localized within each tract.

3.1. *Amyloid*

We ran BUAN to assess the effect of amyloid positivity on 38 white matter tracts based on data from 329 amyloid-negative (CN: 235, MCI: 86, Dementia: 8) (156M, 173F) and 277 amyloid-positive (CN: 139, MCI: 87, Dementia: 51) (131M, 146F) participants from the ADNI3 dataset. The following tracts and measures showed significant differences between amyloid negative and amyloid positive: cingulum left (AxD, MD), corpus callosum forceps major and middle sector (MD, RD), extreme capsule left (MD, RD) and right (AxD, MD, RD), frontopontine tract left (AxD, MD) and right (AxD), inferior longitudinal fasciculus right (AxD, MD, RD), middle longitudinal fasciculus left (AxD, MD) and right (AxD, MD, RD), occipito-pontine tract left (MD), optic radiation right (MD), posterior commissure (AxD), and spinothalamic tract left (RD). In significant tracts, diffusivity metrics increase while fractional anisotropy decreases, in those with higher levels of amyloid pathology (this is in the same direction as the known effects of dementia on these metrics).

We consider tract effects to be significant if they pass both local and global FDR thresholds.

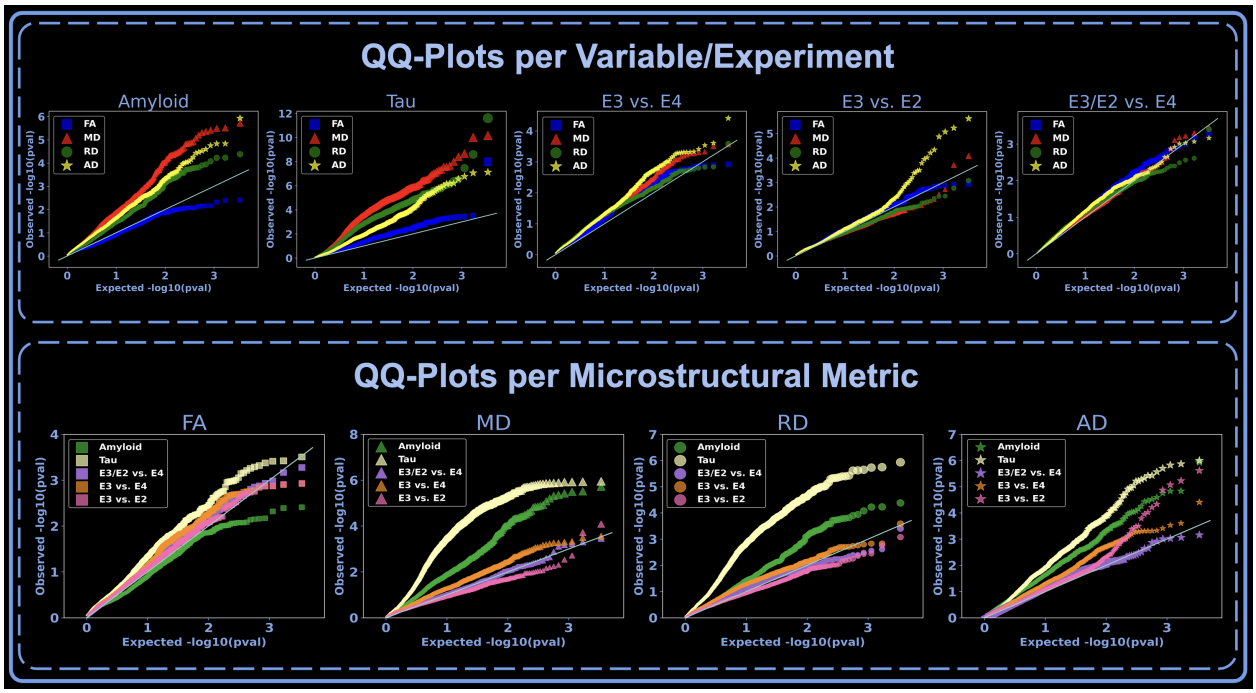


Fig. 3: QQ plots summarize the signal detected by each biomarker (amyloid, tau, and ApoE) and DTI metric (FA, MD, RD, and AxD) across all 38 bundles, indicating which biomarkers and metrics show the strongest associations. In the first row, the plots show which metric shows the strongest association for each biomarker. P-values of the 38 tracts were pooled together for each DTI metric and visualized in QQ plots. In the second row, we analyze for each metric which biomarker shows significant associations. Note the y-axis range varies across the plots depending on the observed range of p-values.

3.2. *Tau*

We ran BUAN to assess the effect of tau positivity on 38 white matter tracts based on data from 401 tau-negative (CN: 293, MCI: 95, and Dementia: 13) (192M, 209F) and 168 tau-positive (CN: 60, MCI: 68, and Dementia: 40) (75M, 93F) participants in the ADNI3 dataset.

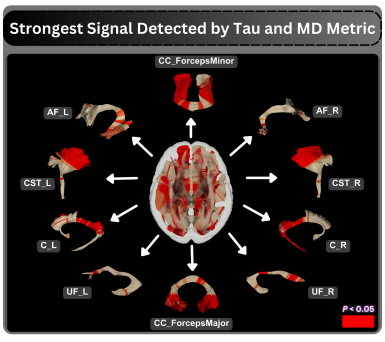


Fig. 4: Tau effects on tracts.

The following tracts and measures showed significant associations between tau positivity and microstructure: Arcuate fasciculus left (MD, RD), cingulum left and right (MD, RD), corpus callosum - forceps major (MD, RD), forceps minor (FA, MD, RD) and mid (AxD, MD, RD), corticospinal Tract left and right (MD, RD), extreme capsule left and right (AxD, MD, RD), frontopontine tract left (MD, RD) and right (FA, AxD, MD, RD), inferior fronto-occipital fasciculus right (RD), inferior longitudinal fasciculus left (MD, RD) and right (AxD, MD, RD), middle longitudinal fasciculus left (AxD, MD, RD) and right (AxD, FA, MD, RD), occipito-pontine tract left (MD, RD) and right (AxD, MD, RD), optic radiation left (RD) and right (AxD, MD, RD), and uncinate fasciculus right (MD, RD). In significant tracts, most diffusivity metrics increase while fractional anisotropy

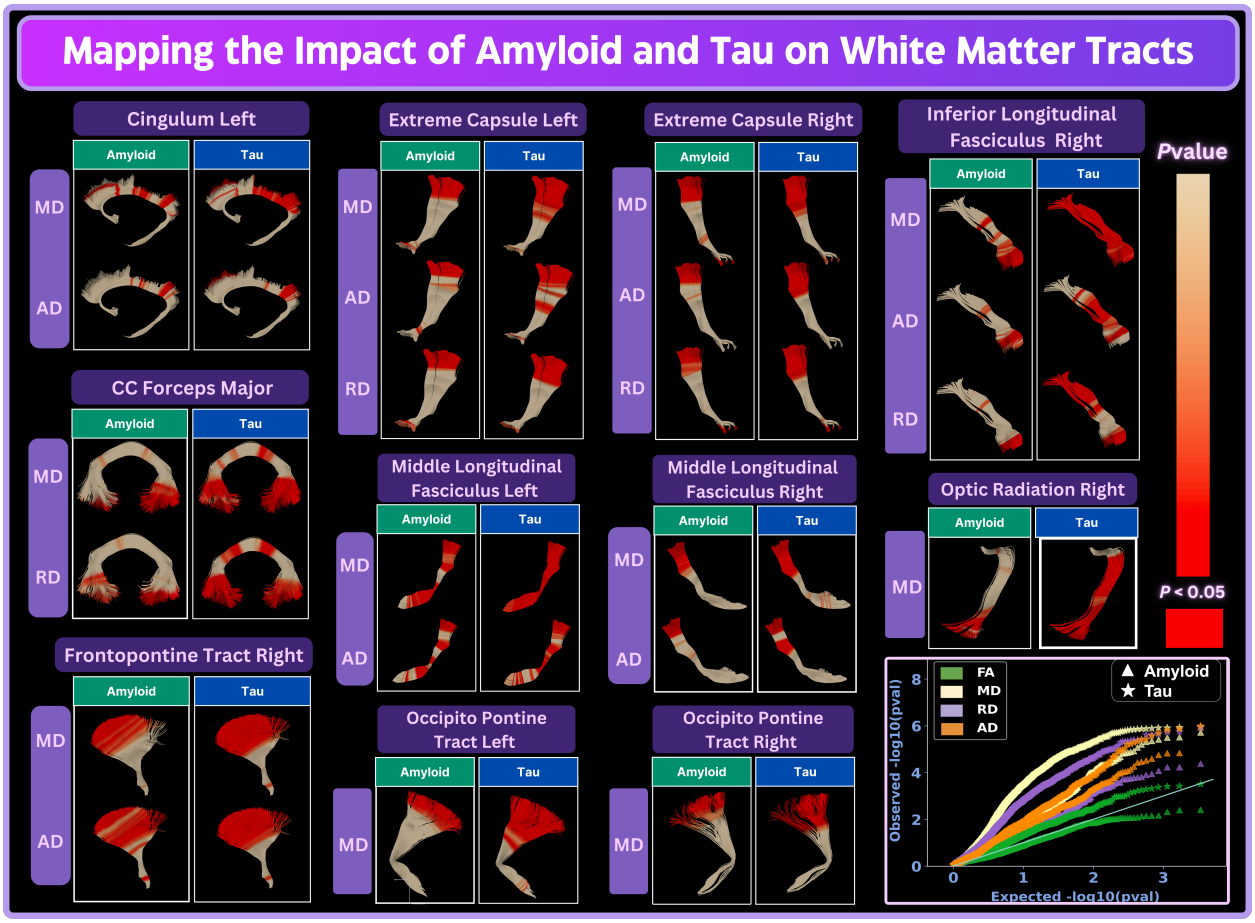


Fig. 5: We compare the effects of amyloid and tau on white matter microstructure along the major white matter tracts. Only tracts showing significant effects, passing both local and global FDR for amyloid and tau, are visualized. *Red* highlights significant associations between the measures of Alzheimer's disease pathology and the microstructural metrics computed with DTI. We consistently observe the strongest associations with tau in various white matter tracts, as seen in the QQ-plot at the right end of the figure. Tau outperforms amyloid in terms of strength of association, for each microstructural metric.

decreases, in line with the expected direction of microstructural abnormalities previously reported in dementia. However, in some tracts, changes in AxD vary along the length of the tracts.

We compare the impact on white matter tracts as influenced by amyloid and tau in Fig. 5. Only tracts that demonstrate significant effects, meeting both local and global false discovery rate (FDR) criteria for amyloid and tau, are included. Significant associations with each biomarker in conjunction with DTI metrics are highlighted in red. We consistently observe stronger associations with tau across various white matter tracts, as illustrated in the QQ-plot at the right end of the figure. For all metrics assessed, tau shows stronger associations compared to amyloid. MD metrics exhibit the strongest association signal for both amyloid and tau. We illustrate the localized effects of tau on MD metrics in Fig. 4. Each tract is color-coded based on p-values, with tracts showing p-values less than 0.05 highlighted in red.

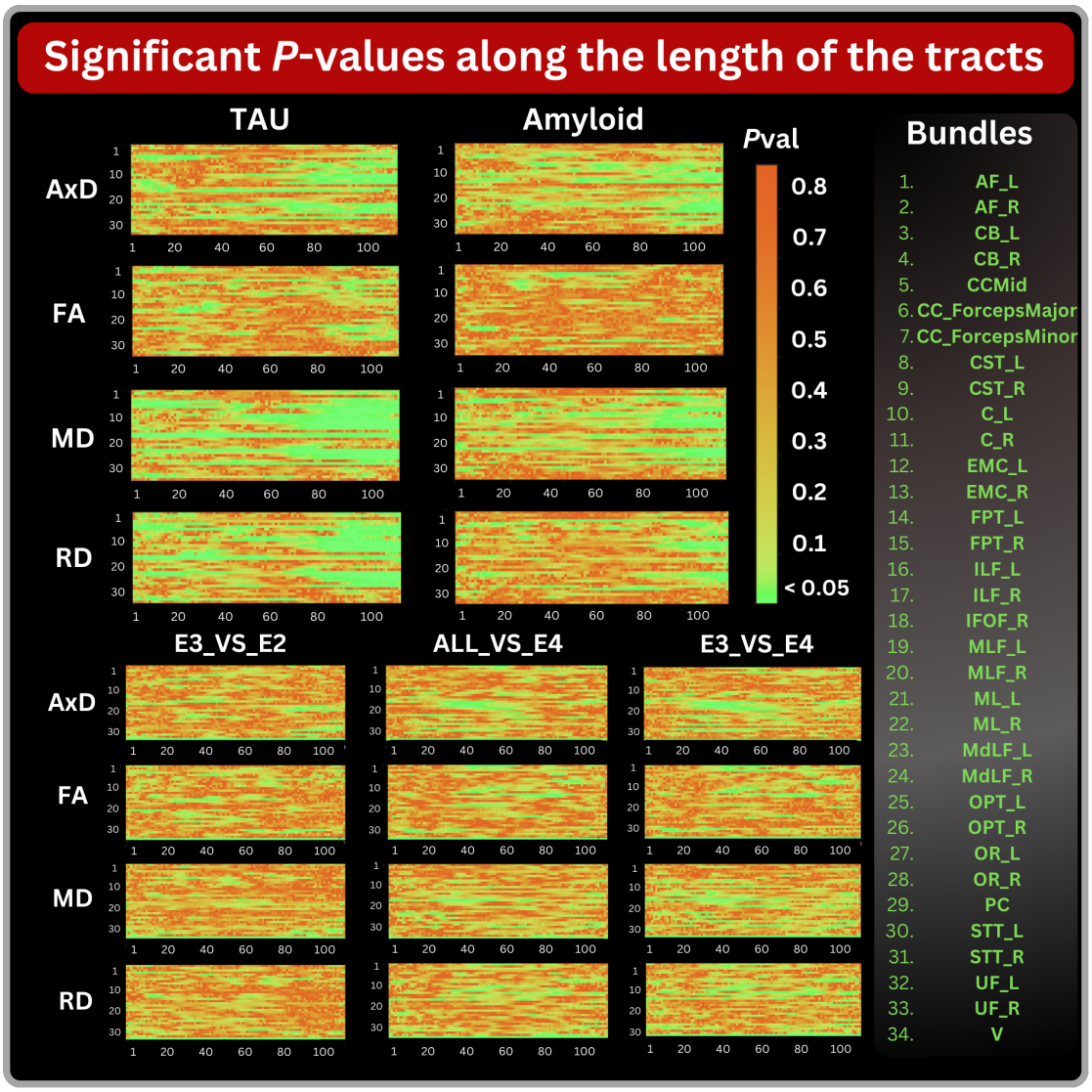


Fig. 6: *P*-values along the length of the 34 major tracts. The x-axis represents 100 segments for each tract, showing the p-value for each segment, while the y-axis corresponds to the different bundles. *Green* pixels indicate segments where p-values<0.05, highlighting regions of higher statistical significance detected by the metric in relation to Alzheimer's disease biomarkers for that specific bundle and segment.

3.3. *APOE $\epsilon 4$ genotype*

We ran BUAN to assess the impact of APOE $\epsilon 4$ - the major common risk gene for late-onset Alzheimer's disease - on 38 major white matter tracts, based on data from 358 non $\epsilon 4$ carriers $\epsilon 2\epsilon 3/\epsilon 3\epsilon 3/\epsilon 2\epsilon 2$ (CN: 224, MCI: 99, Dementia: 35) (168M, 190F) and 203 participants with at least one $\epsilon 4$ gene; $\epsilon 2\epsilon 4/\epsilon 3\epsilon 4/\ \epsilon 4\epsilon 4$ carriers (CN: 136, MCI: 54, and Dementia: 13) (90M, 108F) participants from the ADNI3 dataset. We found the following tracts and measures to be significant: Corticospinal tract left (FA), frontopontine tract left (FA), inferior longitudinal fasciculus right (MD), and middle longitudinal fasciculus right (AxD). MD, RD, and AxD decrease. FA slightly increases.

3.4. *APOE ϵ3 vs. APOE ϵ4*

We ran BUAN to assess the impact of ϵ4 on 38 major white matter tracts using 310 ϵ3ϵ3 (CN:191 MCI:85 Dementia: 34) (140M, 170F) and 192 ϵ3ϵ4/ϵ4ϵ4 (CN:129, MCI:50, and Dementia:13) (88M, 104F) subjects from ADNI3 dataset. We found the following tracts and measures to be significant: Frontopontine Tract left (FA), inferior Longitudinal Fasciculus right (AxD, MD), and Middle Longitudinal Fasciculus right (AxD), and spinothalamic tract left (MD), and right (AxD). MD decreases, AxD changes vary along the length of the tract, with a slight increase in FA.

3.5. *APOE ϵ3 vs. APOE ϵ2*

We ran BUAN to assess the impact of the APOE ϵ2 genotype (which is protective against Alzheimer's disease) on 38 major white matter tracts using 310 ϵ3ϵ3 (CN: 191, MCI: 85, and Dementia: 34) (140M, 170F) and 48 ϵ3ϵ2/ϵ2ϵ2 48, (CN: 33, MCI: 14, and Dementia: 1) (28M, 20F) participants in the ADNI3 dataset. We found the following tracts and measures to be significant: Middle Longitudinal fasciculus right (AxD), spinothalamic tract right (FA, AxD), and uncinate fasciculus right (AxD). FA increases, MD and RD decrease and AxD changes vary along the length of the tracts.

4. Discussion

Our study employed the advanced tractometry method, BUAN (Bundle Analytics), to investigate the effects of amyloid, tau, APOE ϵ4, and APOE ϵ2 on the microstructure of white matter tracts in the brain. The results underscore the significant role of tau and amyloid as biomarkers for Alzheimer's disease (AD), revealing their profound impact on white matter integrity. Tau and amyloid deposition are associated with marked changes in MD, AxD, and RD, with FA being the least sensitive metric. This highlights the critical nature of these biomarkers in the early detection and monitoring of AD progression.

Tau and amyloid significantly alter the microstructural properties of white matter tracts, which are essential for neural communication. APOE ϵ4 carriers showed microstructural changes consistent with poorer white matter integrity, compared to those with the ϵ3/ϵ3 genotype, in line with the heightened genetic risk for AD associated with APOE ϵ4. These alterations are likely due to the reduced efficiency of amyloid clearance and increased inflammation observed in ϵ4 carriers. Conversely, fewer white matter bundles were affected by APOE ϵ2, perhaps in line with its protective role against AD-related white matter degeneration.[13,14]

The findings also revealed that MD is the most affected metric, followed by AxD and RD, whereas FA is the least sensitive. This is consistent with prior literature studying the association of DTI metrics with dementia.[64,65] This differential sensitivity of diffusion metrics highlights the importance of selecting appropriate imaging markers for assessing white matter integrity in AD. MD, in particular, may serve as a more reliable indicator of microstructural damage in the context of AD pathology.

Our results underscore the significant role of the key AD biomarkers in altering the microstructure of key neural pathways, with profound implications for understanding the progression and potential intervention points for AD. Some key tracts - the cingulum bundles and components of the corpus callosum - showed significant alterations in MD and RD in

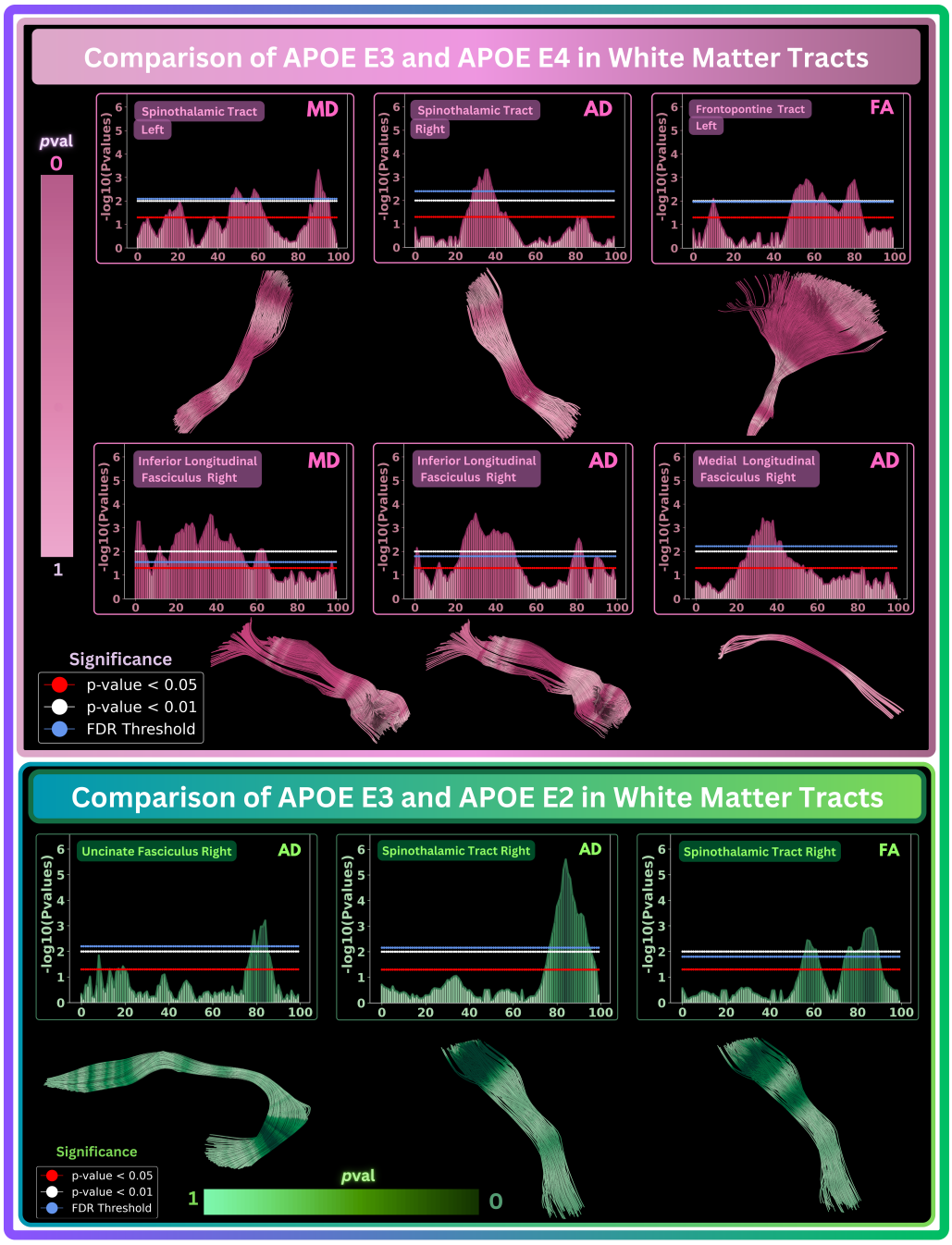


Fig. 7: BUAN results for group differences between $\epsilon 3\epsilon 3$ neutral gene and subjects with either $\epsilon 3\epsilon 4$ or $\epsilon 4\epsilon 4$ gene in white matter tracts. The first and third row shows p-value plots for each tract, where the x-axis represents the segment number along the tract and the y-axis shows a negative logarithm of p-values. The blue horizontal line in the plots represents the FDR corrected threshold. Segments that pass the FDR corrected threshold are considered significant. The second and fourth rows visualize p-values mapped onto the 3D tracts. Where dark *pink* and dark *green* colors imply lower p-values and more significance.

the presence of both amyloid and tau. The increased MD and RD indicate water molecules diffusing more freely in brain tissue - a sign of tissue degeneration and loss of cellular integrity

typical in AD. This diffusion behavior reflects the structural breakdown of neural pathways, which is critical in the progression of Alzheimer's disease.

Amyloid and tau effects on key neural pathways like the cingulum and corpus callosum can impair cognitive function and interhemispheric communication. The cingulum bundle is essential for cognitive and emotional processing, and its disruption weakens connectivity between the frontal lobe and other brain regions, contributing to cognitive decline in AD patients.[66,67] The corpus callosum (CC) is vital for interhemispheric communication, coordinating cognitive and motor functions across both hemispheres.[68] Additional tracts, such as the extreme capsule (EMC), frontopontine tract (FPT), inferior longitudinal fasciculus (ILF), middle longitudinal fasciculus (MLF), optic radiation (OR), and spinothalamic tract (STT), also showed significant changes in diffusivity metrics. The EMC, involved in auditory and language processing, affects communication abilities when damaged.[69] The FPT connects the frontal cortex to the pons, and damage can lead to motor control and executive function issues.[19] The ILF links the temporal and occipital lobes, contributing to visual processing and memory,[70] with disruptions leading to visual memory deficits. The MLF plays a role in language, semantic memory, and integrating auditory and visual information,[71] and its impairment may cause semantic and memory deficits. The OR carries visual information from the thalamus to the visual cortex, and impairment affects visual processing.[72] The STT is critical for pain and temperature sensation,[73] and its impairment affects sensory processing. These findings indicate that AD impacts multiple neural pathways, leading to diverse clinical symptoms.

Moreover, this study highlights the limitations of earlier methods such as TBSS,[29] which, despite identifying significant associations between amyloid positivity, clinical dementia severity, and specific brain regions,[64] suffers from limited resolution due to predefined atlas regions. The BUAN method overcomes these limitations by offering a finer-scale mapping of microstructural changes along the length of white matter tracts, providing a more detailed and accurate assessment of disease-related alterations. The pronounced effects detected in specific bundles reveal the vulnerability of these white matter fiber pathways to Alzheimer's disease pathology, highlighting their potential as biomarkers for early detection and monitoring of disease progression. Future work will integrate microstructural measures derived from sophisticated modeling techniques, such as diffusion kurtosis imaging (DKI),[74] or neurite orientation dispersion and density imaging (NODDI)[75] into BUAN.

4.1. *Conclusion*

In this study, we employ our advanced tractometry method, BUAN (Bundle Analytics), to evaluate the impact of amyloid, tau, APOE $\epsilon 4$, and APOE $\epsilon 2$ on the microstructural properties of white matter tracts in the brain. Among these factors, we find that microstructural alterations in white matter tracts are most significantly associated with tau and amyloid - the two prominent biomarkers of Alzheimer's disease. Fewer bundles are affected by APOE $\epsilon 2$, and comparing APOE $\epsilon 4$ with APOE $\epsilon 3/\epsilon 3$ reveals stronger microstructural alterations than comparing APOE $\epsilon 4$ with $\epsilon 2$ and $\epsilon 3$ variants combined. [c]

[c]Acknowledgement: This research was supported by the NIH (National Institutes of Health) under the AI4AD project grant U01 AG068057, grant numbers P41 EB015922, and RF1 AG057892, the

References

1. M. Goedert, M. Spillantini and R. Crowther, Tau proteins and neurofibrillary degeneration, *Brain pathology* **1**, 279 (1991).
2. J. Sepulcre, M. J. Grothe, F. d'Oleire Uquillas, L. Ortiz-Terán, I. Diez, H.-S. Yang, H. I. Jacobs, B. J. Hanseeuw, Q. Li, G. El-Fakhri *et al.*, Neurogenetic contributions to amyloid beta and tau spreading in the human cortex, *Nature medicine* **24**, 1910 (2018).
3. V. L. Villemagne, V. Doré, S. C. Burnham, C. L. Masters and C. C. Rowe, Imaging tau and amyloid-β proteinopathies in alzheimer disease and other conditions, *Nature Reviews Neurology* **14**, 225 (2018).
4. C. R. Jack Jr, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish *et al.*, Nia-aa research framework: toward a biological definition of alzheimer's disease, *Alzheimer's & dementia* **14**, 535 (2018).
5. C. Haass and D. J. Selkoe, Soluble protein oligomers in neurodegeneration: lessons from the alzheimer's amyloid β-peptide, *Nature reviews Molecular cell biology* **8**, 101 (2007).
6. J. Graff-Radford, E. M. Arenaza-Urquijo, D. S. Knopman, C. G. Schwarz, R. D. Brown Jr, A. A. Rabinstein, J. L. Gunter, M. L. Senjem, S. A. Przybelski, T. Lesnick *et al.*, White matter hyperintensities: relationship to amyloid and tau burden, *Brain* **142**, 2483 (2019).
7. T. Grimmer, M. Faust, F. Auer, P. Alexopoulos, H. Förstl, G. Henriksen, R. Perneczky, C. Sorg, B. H. Yousefi, A. Drzezga *et al.*, White matter hyperintensities predict amyloid increase in alzheimer's disease, *Neurobiology of aging* **33**, 2766 (2012).
8. C. Bancher, C. Brunner, H. Lassmann, H. Budka, K. Jellinger, G. Wiche, F. Seitelberger, I. Grundke-Iqbal, K. Iqbal and H. Wisniewski, Accumulation of abnormally phosphorylated τ precedes the formation of neurofibrillary tangles in alzheimer's disease, *Brain research* **477**, 90 (1989).
9. P. V. Arriagada, J. H. Growdon, E. T. Hedley-Whyte and B. T. Hyman, Neurofibrillary tangles but not senile plaques parallel duration and severity of alzheimer's disease, *Neurology* **42**, 631 (1992).
10. K. Kantarci, M. E. Murray, C. G. Schwarz, R. I. Reid, S. A. Przybelski, T. Lesnick, S. M. Zuk, M. R. Raman, M. L. Senjem, J. L. Gunter *et al.*, White-matter integrity on dti and the pathologic staging of alzheimer's disease, *Neurobiology of aging* **56**, 172 (2017).
11. E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. Small, A. Roses, J. Haines and M. A. Pericak-Vance, Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families, *Science* **261**, 921 (1993).
12. J. Van Bergen, X. Li, J. Hua, S. Schreiner, S. Steininger, F. Quevenco, M. Wyss, A. Gietl, V. Treyer, S. Leh *et al.*, Colocalization of cerebral iron with amyloid beta in mild cognitive impairment, *Scientific reports* **6**, p. 35514 (2016).
13. Z. Li, F. Shue, N. Zhao, M. Shinohara and G. Bu, Apoe2: protective mechanism and therapeutic implications for alzheimer's disease, *Molecular neurodegeneration* **15**, p. 63 (2020).
14. M. J. Grothe, S. Villeneuve, M. Dyrba, D. Bartrés-Faz, M. Wirth, A. D. N. Initiative and A. D. N. Initiative, Multimodal characterization of older apoe2 carriers reveals selective reduction of amyloid load, *Neurology* **88**, 569 (2017).
15. P. J. Basser, J. Mattiello and D. LeBihan, MR diffusion tensor spectroscopy and imaging, *Biophysical Journal* **66**, 259 (1994).
16. D. Le Bihan, J.-F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko and H. Chabriat, Diffusion tensor imaging: concepts and applications, *Journal of Magnetic Resonance Imaging:*

National Institute of Biomedical Imaging and Bioengineering under award numbers R01 EB027585 and R01 EB017230.

An Official Journal of the International Society for Magnetic Resonance in Medicine **13**, 534 (2001).
17. A. L. Alexander, J. E. Lee, M. Lazar and A. S. Field, Diffusion tensor imaging of the brain, *Neurotherapeutics* **4**, 316 (2007).
18. S. Farquharson, J.-D. Tournier, F. Calamante, G. Fabinyi, M. Schneider-Kolsky, G. D. Jackson and A. Connelly, White matter fiber tractography: why we need to move beyond dti, *Journal of neurosurgery* **118**, 1367 (2013).
19. M. Catani and M. T. De Schotten, A diffusion tensor imaging tractography atlas for virtual in vivo dissections, *cortex* **44**, 1105 (2008).
20. G. Gong, Y. He, L. Concha, C. Lebel, D. W. Gross, A. C. Evans and C. Beaulieu, Mapping anatomical connectivity patterns of human cerebral cortex using in vivo diffusion tensor imaging tractography, *Cerebral cortex* **19**, 524 (2008).
21. A. Yendiki, P. Panneck, P. Srinivasan, A. Stevens, L. Zöllei, J. Augustinack, R. Wang, D. Salat, S. Ehrlich, T. Behrens *et al.*, Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy, *Frontiers in neuroinformatics* **5**, p. 23 (2011).
22. M. Cousineau, P.-M. Jodoin, E. Garyfallidis, M.-A. Côté, F. C. Morency, V. Rozanski, M. Grand'Maison, B. J. Bedell and M. Descoteaux, A test-retest study on parkinson's ppmi dataset yields statistically significant white matter fascicles, *NeuroImage: Clinical* **16**, 222 (2017).
23. B. Q. Chandio, S. L. Risacher, F. Pestilli, D. Bullock, F.-C. Yeh, S. Koudoro, A. Rokem, J. Harezlak and E. Garyfallidis, Bundle analytics, a computational framework for investigating the shapes and profiles of brain pathways across populations, *Scientific Reports* **10**, p. 17149 (2020).
24. B. Q. Chandio, T. Chattopadhyay, C. Owens-Walton, J. E. V. Reina, L. Nabulsi, S. I. Thomopoulos, E. Garyfallidis and P. M. Thompson, Fiberneat: Unsupervised white matter tract filtering, in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022.
25. J. D. Yeatman, R. F. Dougherty, N. J. Myall, B. A. Wandell and H. M. Feldman, Tract profiles of white matter properties: automating fiber-tract quantification, *PloS one* **7**, p. e49790 (2012).
26. B. Q. Chandio, C. Owens-Walton, J. E. Villalon-Reina, L. Nabulsi, S. I. Thomopoulos, J. Guaje, E. Garyfallidis and P. M. Thompson, Microstructural changes in the white matter tracts of the brain due to mild cognitive impairment, *Alzheimer's & Dementia* **18**, p. e065339 (2022).
27. A. H. Zhu, T. M. Nir, S. Javid, J. E. Villalon-Reina, A. L. Rodrigue, L. T. Strike, G. I. de Zubicaray, K. L. McMahon, M. J. Wright, S. E. Medland *et al.*, Lifespan reference curves for harmonizing multi-site regional brain white matter metrics from diffusion mri, *bioRxiv* (2024).
28. J. E. Villalón-Reina, A. H. Zhu, T. M. Nir, S. I. Thomopoulos, E. Laltoo, L. Kushan, C. E. Bearden, N. Jahanshad and P. M. Thompson, Large-scale normative modeling of brain microstructure, in *2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM)*, 2023.
29. S. M. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. E. Nichols, C. E. Mackay, K. E. Watkins, O. Ciccarelli, M. Z. Cader, P. M. Matthews *et al.*, Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data, *Neuroimage* **31**, 1487 (2006).
30. S. I. Thomopoulos, T. M. Nir, J. E. Villalon-Reina, A. Zavaliangos-Petropulu, P. Maiti, H. Zheng, E. Nourollahimoghadam, N. Jahanshad and P. M. Thompson, Diffusion MRI metrics and their relation to dementia severity: effects of harmonization approaches, in *17th International Symposium on Medical Information Processing and Analysis*, (SPIE, Campinas, Brazil, December 2021).
31. A. Zavaliangos-Petropulu, T. M. Nir, S. I. Thomopoulos, R. I. Reid, M. A. Bernstein, B. Borowski, C. R. Jack Jr., M. W. Weiner, N. Jahanshad and P. M. Thompson, Diffusion MRI Indices and Their Relation to Cognitive Impairment in Brain Aging: The Updated Multi-

protocol Approach in ADNI3, *Frontiers in Neuroinformatics* **13**, p. 2 (February 2019).
32. A. M. Racine, N. Adluru, A. L. Alexander, B. T. Christian, O. C. Okonkwo, J. Oh, C. A. Cleary, A. Birdsill, A. T. Hillmer, D. Murali *et al.*, Associations between white matter microstructure and amyloid burden in preclinical alzheimer's disease: a multimodal imaging investigation, *NeuroImage: Clinical* **4**, 604 (2014).
33. I. B. Gari, A. Ramesh, S. Javid, S. P. Gadewar, E. Nourollahimoghadam, S. I. Thomopoulos, P. M. Thompson, T. M. Nir, N. Jahanshad, A. D. N. Initiative *et al.*, Medial tractography analysis (meta) for white matter population analyses across datasets, in *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2023.
34. B. Q. Chandio, Advancing white matter tractometry of the brain using diffusion MRI and machine learning, PhD thesis, Indiana University2022.
35. J. W. Vogel, Y. Iturria-Medina, O. T. Strandberg, R. Smith, E. Levitis, A. C. Evans and O. Hansson, Spread of pathological tau proteins through communicating neurons in human alzheimer's disease, *Nature communications* **11**, p. 2612 (2020).
36. S. M. Landau, C. Breault, A. D. Joshi, M. Pontecorvo, C. A. Mathis, W. J. Jagust and M. A. Mintun, Amyloid-β imaging with pittsburgh compound b and florbetapir: comparing radiotracers and quantification methods, *Journal of Nuclear Medicine* **54**, 70 (2013).
37. S. Landau, B. Thomas, L. Thurfjell, M. Schmidt, R. Margolin, M. Mintun, M. Pontecorvo, S. Baker, W. Jagust and A. D. N. Initiative, Amyloid pet imaging in alzheimer's disease: a comparison of three radiotracers, *European journal of nuclear medicine and molecular imaging* **41**, 1398 (2014).
38. S. M. Landau, M. A. Mintun, A. D. Joshi, R. A. Koeppe, R. C. Petersen, P. S. Aisen, M. W. Weiner, W. J. Jagust and A. D. N. Initiative, Amyloid deposition, hypometabolism, and longitudinal cognitive decline, *Annals of neurology* **72**, 578 (2012).
39. S. M. Landau, M. Lu, A. D. Joshi, M. Pontecorvo, M. A. Mintun, J. Q. Trojanowski, L. M. Shaw, W. J. Jagust and A. D. N. Initiative, Comparing positron emission tomography imaging and cerebrospinal fluid measurements of β-amyloid, *Annals of neurology* **74**, 826 (2013).
40. N. Jahanshad, P. V. Kochunov, E. Sprooten, R. C. Mandl, T. E. Nichols, L. Almasy, J. Blangero, R. M. Brouwer, J. E. Curran, G. I. de Zubicaray *et al.*, Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMA–DTI working group, *Neuroimage* **81**, 455 (2013).
41. S. I. Thomopoulos, T. M. Nir, J. E. V. Reina, N. Jahanshad and P. M. Thompson, Diffusion MRI metrics of brain microstructure in alzheimer's disease: Boosting disease sensitivity with multi-shell imaging and advanced pre-processing: Neuroimaging/new imaging methods, *Alzheimer's & Dementia* **16**, p. e046654 (2020).
42. J. V. Manjón, P. Coupé, L. Concha, A. Buades, D. L. Collins and M. Robles, Diffusion weighted image denoising using overcomplete local PCA, *PLoS ONE* **8** (2013).
43. E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. Van Der Walt, M. Descoteaux, I. Nimmo-Smith and D. Contributors, Dipy, a library for the analysis of diffusion MRI data, *Frontiers in Neuroinformatics* **8**, p. 8 (2014).
44. E. Kellner, B. Dhital, V. G. Kiselev and M. Reisert, Gibbs-ringing artifact removal based on local subvoxel-shifts, *Magnetic Resonance in Medicine* **76**, 1574 (2016).
45. J.-D. Tournier, R. Smith, D. Raffelt, R. Tabbara, T. Dhollander, M. Pietsch, D. Christiaens, B. Jeurissen, C.-H. Yeh and A. Connelly, MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation, *Neuroimage* **202**, p. 116137 (2019).
46. S. M. Smith, Fast robust automated brain extraction, *Human Brain Mapping* **17**, 143 (2002).
47. M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich and S. M. Smith, FSL, *Neuroimage* **62**, 782 (2012).
48. J. L. Andersson and S. N. Sotiropoulos, An integrated approach to correction for off-resonance

effects and subject movement in diffusion MR imaging, *Neuroimage* **125**, 1063 (2016).
49. N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich and J. C. Gee, N4itk: improved N3 bias correction, *IEEE transactions on medical imaging* **29**, 1310 (2010).
50. E. J. Canales-Rodríguez, A. Daducci, S. N. Sotiropoulos, E. Caruyer, S. Aja-Fernández, J. Radua, J. M. Yurramendi Mendizabal, Y. Iturria-Medina, L. Melie-García, Y. Alemán-Gómez *et al.*, Spherical deconvolution of multichannel diffusion MRI data with non-Gaussian noise models and spatial regularization, *PLoS ONE* **10**, p. e0138910 (2015).
51. G. Girard, K. Whittingstall, R. Deriche and M. Descoteaux, Towards quantitative connectivity analysis: reducing tractography biases, *Neuroimage* **98**, 266 (2014).
52. E. Garyfallidis, M.-A. Côté, F. Rheault, J. Sidhu, J. Hau, L. Petit, D. Fortin, S. Cunanne and M. Descoteaux, Recognition of white matter bundles using local and global streamline-based registration and clustering, *NeuroImage* **170**, 283 (April 2018).
53. F.-C. Yeh, S. Panesar, D. Fernandes, A. Meola, M. Yoshino, J. C. Fernandez-Miranda, J. M. Vettel and T. Verstynen, Population-averaged atlas of the macroscale human structural connectome and its network topology, *NeuroImage* **178**, 57 (September 2018).
54. B. Q. Chandio, E. Olivetti, D. Romero-Bascones, J. Harezlak and E. Garyfallidis, Bundlewarp, streamline-based nonlinear registration of white matter tracts, *bioRxiv* (2023).
55. E. Garyfallidis, M. Brett, M. M. Correia, G. B. Williams and I. Nimmo-Smith, Quickbundles, a method for tractography simplification, *Frontiers in Neuroscience* **6**, p. 175 (2012).
56. W. E. Johnson, C. Li and A. Rabinovic, Adjusting batch effects in microarray expression data using Empirical Bayes methods, *Biostatistics* **8**, 118 (2007).
57. J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath *et al.*, Harmonization of cortical thickness measurements across scanners and sites, *Neuroimage* **167**, 104 (2018).
58. B. Q. Chandio, J. E. Villalon-Reina, T. M. Nir, S. I. Thomopoulos, Y. Feng, S. Benavidez, N. Jahanshad, J. Harezlak, E. Garyfallidis and P. M. Thompson, Bundle analytics based data harmonization for multi-site diffusion mri tractometry, *bioRxiv* , 2024 (2024).
59. L. V. Hedges, A random effects model for effect sizes., *Psychological Bulletin* **93**, p. 388 (1983).
60. E. Garyfallidis, S. Koudoro, J. Guaje, M.-A. Côté, S. Biswas, D. Reagan, N. Anousheh, F. Silva, G. Fox and F. Contributors, Fury: advanced scientific visualization, *Journal of Open Source Software* **6**, p. 3384 (2021).
61. Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* **57**, 289 (1995).
62. J. Neyman and E. S. Pearson, On the use and interpretation of certain test criteria for purposes of statistical inference part i, *Biometrika* **20**, 175 (1928).
63. O. J. Dunn, Multiple comparisons among means, *Journal of the American statistical association* **56**, 52 (1961).
64. T. M. Nir, N. Jahanshad, J. E. Villalon-Reina, A. W. Toga, C. R. Jack, M. W. Weiner, P. M. Thompson, A. D. N. I. (ADNI *et al.*, Effectiveness of regional dti measures in distinguishing alzheimer's disease, mci, and normal aging, *NeuroImage: clinical* **3**, 180 (2013).
65. Y. Feng, B. Q. Chandio, J. E. Villalon-Reina, S. I. Thomopoulos, T. M. Nir, S. Benavidez, E. Laltoo, T. Chattopadhyay, H. Joshi, G. Venkatasubramanian *et al.*, Microstructural mapping of neural pathways in alzheimer's disease using macrostructure-informed normative tractometry, *bioRxiv* (2024).
66. S. E. Rose, F. Chen, J. B. Chalk, F. O. Zelaya, W. E. Strugnell, M. Benson, J. Semple and D. M. Doddrell, Loss of connectivity in alzheimer's disease: an evaluation of white matter tract integrity with colour coded mr diffusion tensor imaging, *Journal of Neurology, Neurosurgery & Psychiatry* **69**, 528 (2000).

67. Y. Yang, K. Schilling, N. Shashikumar, V. Jasodanand, E. E. Moore, K. R. Pechman, M. Bilgel, L. L. Beason-Held, Y. An, A. Shafer *et al.*, White matter microstructural metrics are sensitively associated with clinical staging in alzheimer's disease, *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **15**, p. e12425 (2023).
68. S. E. Black, S. D. Moffat, C. Y. David, J. Parker, P. Stanchev and M. Bronskill, Callosal atrophy correlates with temporal lobe volume and mental status in alzheimer's disease, *Canadian Journal of Neurological Sciences* **27**, 204 (2000).
69. L. Concha, C. Beaulieu and D. W. Gross, Bilateral limbic diffusion abnormalities in unilateral temporal lobe epilepsy, *Annals of neurology* **57**, 188 (2005).
70. F. Agosta, R. G. Henry, R. Migliaccio, J. Neuhaus, B. L. Miller, N. F. Dronkers, S. M. Brambati, M. Filippi, J. M. Ogar, S. M. Wilson *et al.*, Language networks in semantic dementia, *Brain* **133**, 286 (2010).
71. N. Makris, D. N. Kennedy, S. McInerney, A. G. Sorensen, R. Wang, V. S. Caviness Jr and D. N. Pandya, Segmentation of subcomponents within the superior longitudinal fascicle in humans: a quantitative, in vivo, dt-mri study, *Cerebral cortex* **15**, 854 (2005).
72. H. Bridge, O. Thomas, S. Jbabdi and A. Cowey, Changes in connectivity after visual cortical brain damage underlie altered visual function, *Brain* **131**, 1433 (2008).
73. A. Craig, Pain mechanisms: labeled lines versus convergence in central processing, *Annual review of neuroscience* **26**, 1 (2003).
74. A. J. Steven, J. Zhuo and E. R. Melhem, Diffusion kurtosis imaging: an emerging technique for evaluating the microstructural environment of the brain, *American Journal of Roentgenology* **202**, W26 (2014).
75. H. Zhang, T. Schneider, C. A. Wheeler-Kingshott and D. C. Alexander, NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain, *Neuroimage* **61**, 1000 (2012).

A Pathway-Level Information ExtractoR (PLIER) framework to gain mechanistic insights into obesity in Down syndrome

Sutanu Nandi[1], Yuehua Zhu[2,3], Lucas A Gillenwater[1,4,5,††], Marc Subirana-Granés[6], Haoyu Zhang[6], Negar Janani[6], Casey Greene[4,6,†], Milton Pividori[5,6,*,†††], Maria Chikina[2,*], James C Costello[1,4,5,6,*,†]

[1]*Department of Pharmacology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA*
[2]*Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA*
[3]*School of Medicine, Tsinghua University, Beijing, China*
[4]*Computational Bioscience Program, University of Colorado Anschutz Medical Campus, Aurora, CO, USA*
[5]*Linda Crnic Institute for Down Syndrome, University of Colorado Anschutz Medical Campus, Aurora, CO, USA*
[6]*Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA*

**Co-Senior Authors*
Email: milton.pividori@cuanschutz.edu, mchikina@pitt.edu, james.costello@cuanschutz.edu

Down syndrome (DS), caused by the triplication of chromosome 21 (T21), is a prevalent genetic disorder with a higher incidence of obesity. Traditional approaches have struggled to differentiate T21-specific molecular dysregulation from general obesity-related processes. This study introduces the omni-PLIER framework, combining the Pathway-Level Information ExtractoR (PLIER) with the omnigenic model, to uncover molecular mechanisms underlying obesity in DS. The PLIER framework aligns gene expression data with biological pathways, facilitating the identification of relevant molecular patterns. Using RNA sequencing data from the Human Trisome Project, omni-PLIER identified latent variables (LVs) significantly associated with both T21 and body mass index (BMI). Elastic net regression and causal mediation analysis revealed LVs mediating the effect of karyotype on BMI. Notably, LVs involving glutathione peroxidase-1 (GPX1) and MCL1 apoptosis regulator, BCL2 family members emerged as crucial mediators. These findings provide insights into the molecular interplay between DS and obesity. The omni-PLIER model offers a robust methodological advancement for dissecting complex genetic disorders, with implications for understanding obesity-related processes in both DS and the general population.

Keywords: Down syndrome, obesity, body mass index, matrix factorization, mediation analysis, RNA sequencing, mechanisms of disease, genetic/genomic studies, pathway analysis

1. Introduction

Down syndrome (DS), also known as trisomy 21 (T21), is the result of the triplication of chromosome 21 (chr21) and is the most frequent human aneuploidy[1]. Obesity, the result of

[†]This work is partly supported by HD109765, [††]This work is partly supported by GM007635 and a Blumenthal Fellowship, [†††]This work is partly supported by HG011898

disrupted metabolism leading to excessive adipose accumulation, is associated with increased comorbidities and decreased life expectancy[2]. Obesity, defined by a body mass index (BMI) ≥ 30, is more prevalent in individuals with DS than in the disomic population (D21)[3]. Multiple molecular profiling studies demonstrate systemic dysregulation of obesity-associated processes, including insulin resistance, oxidative phosphorylation, and lipid metabolism, in individuals with T21[4–7]. However, current approaches fail to adequately disentangle T21-specific from general molecular dysregulation in the pathogenesis of obesity. Thus, elucidating molecular mechanisms distinct to obesity in T21 will not only inform DS biology but also provide insights into obesity-related processes in the general population.

Mechanistic insights into DS are complicated by the simultaneous upregulation of most genes on chr21. While the mean overexpression of genes on chr21 is a 1.5X fold change, there is great variability in gene expression across people with DS[5]. Moreover, thousands of genes outside of chr21 are differentially expressed in people with DS. It is helpful to consider T21 in the context of the omnigenic model[8,9], which posits that gene regulatory networks are so highly interconnected that potentially all genes expressed in phenotype-relevant cell types have either a direct or indirect effect. Within this model, there are "core" genes that directly affect the phenotype and "peripheral" genes that indirectly affect the phenotype by regulating these core genes. Integrating gene co-expression modules and genome-wide association studies (GWAS) prioritizes genes missed by standard procedures while aiding interpretation[10]. Thus, the omnigenic framework can help understand the cascading effects in gene regulatory networks, which contribute to the co-occurring conditions in DS, such as obesity, through altered mechanisms compared to the D21.

The Pathway-Level Information ExtractoR (PLIER) is a semi-supervised matrix factorization framework[11]. It transforms an input matrix of high dimensional gene expression data into a relatively small number of latent variables (LVs) and then aligns these LVs with pre-defined pathway/geneset annotations. The LVs aim to maximize the variance within the data and the associated gene loadings are aligned with pathways/gene sets. By leveraging pathway/geneset annotations, PLIER achieves interpretable representations where LVs are more likely to align with independent measurements of biological pathways and processes. The LVs can be plugged into any supervised downstream analysis such as differential expression and eQTL discovery. The PLIER framework has been extensively adapted and reused in various applications[10,12–14].

Merging the omnigenic model with PLIER, we propose the omni-PLIER framework, a methodological advance to gain mechanistic insights into how complex genetic disorders drive the associated conditions. In the omnigenic model, gene co-expression modules impact downstream gene regulatory networks. Here, we use the PLIER model to define LVs as modules. Working with the hypothesis that a causal relationship between a genetic perturbation and a clinical phenotype must be mediated through molecular networks, we combine an elastic net model with causal inference methods to identify LVs derived from molecular data that are mediators in the formal

statistical sense. We apply omni-PLIER to study the link between T21 and obesity and identify known and novel pathway associations providing a foundation for detailed follow-up studies. The omni-PLIER model is available at: https://github.com/CostelloLab/omni-PLIER

2. Methods

2.1. *The omni-PLIER Framework*

The omni-PLIER framework integrates gene expression data with clinical traits to identify latent variables (LVs) associated with both karyotype and BMI. This framework extends the PLIER model by applying elastic net regression and causal mediation analysis to uncover biological pathways mediating the relationship between DS and obesity. The framework allows for the discovery of mechanistic insights into how genetic perturbations drive phenotypic outcomes.

The omni-PLIER workflow, shown in **Figure 1**, proceeds as follows. 1) Input gene expression and sample annotation from the Human Trisome Project. 2) Apply the PLIER model to extract LVs aligned with known biological pathways. 3) Calculate elastic net regression and causal mediation analysis to identify significant LVs that mediate the relationship between T21 and BMI. 4) Output causal networks between LVs and phenotypes for further interpretation.

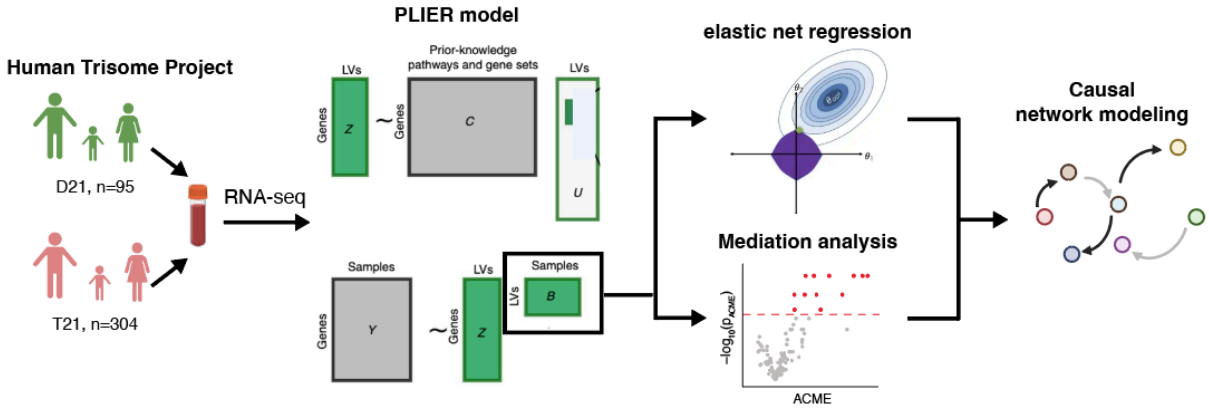


Fig 1. Overview of the omni-PLIER framework

2.2. *Human Trisome Project (HTP) RNA Sequencing Dataset*

Under a study protocol approved by the Colorado Multiple Institutional Review Board (COMIRB #15-2170), the Crnic Institute enrolled participants as part of the Human Trisome Project (HTP; www.trisome.org). Demographic data for study participants were derived from participant and caregiver surveys and the annotation of medical records. Clinical variables relevant to this study include karyotype, age at visit, sex, and body mass index (BMI).

A detailed description of blood processing and molecular quantification for -omic profiling performed by the Human Trisome Project is described by Galbraith et al. and Waugh et al.[5,15]. Briefly, PAXgene RNA Tubes (Qiagen) were used to collect blood samples from 304 T21 and 95

D21 individuals. Whole-blood paired-end RNAseq was performed using Illumina NovaSeq 6000 instrument (Novogene). Reads were filtered for low quality, and adapters were trimmed. Reads were aligned to the human reference genome (assembly GRCh38) using STAR2 and quantified at the gene level to transcripts per million (TPM).

2.3. *Gene Set Enrichment Analysis*

The HTP RNA-seq dataset, along with the sample labels of karyotype or BMI ≥ 30, was input to the 'gseapy' python package (v 1.1.2) for gene set enrichment analysis. We utilized the same pathway information as in the PLIER model: Human Molecular Signatures Database (MSigDB v4.0) collections, C2 (curated gene sets), C6 (oncogenic signature gene sets), C7 (immunologic signature gene sets), bloodCellMarkersIRISDMAP and svmMarkers. These parameters were used: min_size = 5, max_size = 500, method = 'signal_to_noise', and permutation_num = 100,000.

2.4. *PLIER model applied to HTP RNA-seq data*

A gene-by-sample (*g*-by-*s*) matrix is factorized with *k* latent dimensions into $Z_{g \cdot k}\, B_{k \cdot s}$. In addition to the *g*-by-*s* matrix, PLIER considers an additional input of prior knowledge given by a gene-by-geneset binary matrix of pathway/geneset membership, *C* (*g*-by-*p*, where *p* is the number of pathways/genesets). PLIER enforces correspondence between the loadings *Z* and *C* by penalizing the distance between *Z* and its pathway-based prediction $C \cdot U$ (where *U* is a *p*-by-*k* matrix subject to optimization). An elastic-net penalty on the *U* coefficients ensures that each factor utilizes a small fraction of the pathways/genesets. The entire problem is optimized end-to-end using block coordinate minimization.

We determined the number of LVs (*k* parameter) by identifying the number of significant principal components using the *num.pcs* function in the PLIER R package (v0.1.6). The input expression matrix for the PLIER function was the HTP RNA-seq dataset, which was z-score transformed using the *rowNorm* function for genes. We incorporated prior information using the genesets defined in Section 2.2. The default settings of the PLIER function, which automatically configures the L1 and L2 parameters, were used to generate the LV (n=117) by sample matrix.

2.5. *Regression model for latent variable-trait associations*

We integrated gene-trait associations from the PrediXcan family of methods and PLIER LVs through generalized least squares (GLS) regression[10]. The PrediXcan family of methods was utilized for gene-based associations, including S-PrediXcan (for gene-tissue-trait associations) and S-MultiXcan (which combines S-PrediXcan results across tissues and computes gene-trait associations). Our GLS regression model computes an LV-trait association by fitting the model:

$$y = \beta_0 + s\beta_s + \sum_i x_i\beta_i + \epsilon,$$

where y is a vector of S-MultiXcan gene p-values for a trait; s is a binary indicator vector $s_l = 1$ for the top 10% of genes with the largest weights for LV l and zero otherwise; x_i is a gene property used as a covariate (default covariates defined in Pividori et al.[10]); β are effect sizes (with β_0 as the intercept); and $\epsilon \sim MVN(0, \sigma^2\boldsymbol{R})$ are the error terms with a multivariate normal distribution (MVN) where $\boldsymbol{R}$ is the matrix of gene correlations. The model tests whether genes with high weights in an LV are more strongly associated with the phenotype than other genes with small or zero loadings. For more details, see Pividori et al.[10] Consequently, we computed associations for five BMI traits in PhenomeXcan[16] (a large-scale resource with PrediXcan associations across the UK Biobank) across seven omni-PLIER LVs of interest. Due to the limited number of traits, we used nominal significance levels to assess the associations between traits and LVs.

2.6. *Penalized Regression*

The LV-by-sample matrix, *B*, generated from PLIER was batch corrected for the sample source variable using the Combat python package (v 0.20), and then LVs were z-score transformed. Karyotype and clinical variables (age, sex, BMI) were considered for downstream analysis.

We trained elastic net models for two prediction tasks, first to predict D21/T21 using 'LogisticRegressionCV', and second to predict BMI using 'ElasticNetCV', both from the sklearn (v 1.4.2) in python (v 3.11.0). The input dataset was split on the samples into 80% training and 20% testing sets. Using only the training dataset and 5-fold cross-validation, we tuned the α (ratio of L1 to L2 penalization) and λ (penalization weight) parameters using grid search across different ranges. We identified the optimal model parameters using balanced accuracy (T21/D21) or root mean squared error (BMI) over the 5-folds from the training dataset, trained the full model using the training dataset, and evaluated the performance on the testing dataset. Additionally, the contribution of each LV was calculated based on the best performing model coefficients. To establish the model's consistency and reproducibility, we repeated the procedure 1000 times. We then averaged the coefficients of all features from these 1000 models and ranked the features based on these averaged coefficients.

To establish a null model for comparison, we randomly shuffled the target labels (karyotype or BMI) and followed the elastic net model training above to evaluate model performance on the shuffled data. We repeated this procedure 1000 times to establish a null distribution of model performances, then the distribution of the model's performance on the original (unshuffled) data and the shuffled data were compared using a Kruskal-Wallis test.

2.7. *Causal mediation analysis between molecular and clinical variables*

The *mediation* R package (v4.5.0) was used to estimate the causal mediation effects of karyotype on BMI mediated through LVs, with sex and age as covariates. Outcome and mediator models were linear. We performed 100 Monte Carlo draws for quasi-Bayesian approximation (sims=100).

We performed causal analysis on selected LVs, karyotype, BMI, age, and sex. For causal discovery, we utilized the constrained continuous-optimization method PC-NOTEARS with a python implementation of the Peter-Clark algorithm for causal discovery (PC)[17] (causal-learn, v0.1.3.8) and a NOTEARS (Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning)[18] implementation in the *bioCausal* R package (v0.1.0)[19], both with the edge constraints option available. The LVs were z-score transformed for input. PC is a conditional independence testing algorithm that has been extensively benchmarked and shows favorable performance. However, it may not orient all edges, and it does not estimate effects. NOTEARS continuous optimization is applied to maximize the joint multivariate likelihood of the data under the constraint that the inferred relationships form a directed acyclic graph (DAG). Additionally, they provide causal effects. In recent benchmarking work we showed that a combination of PC and NOTEARS is optimal for biological network discovery [20]. In this setting, the continuous optimization is restricted to those edges returned by PC.

We manually set an edge constraint matrix to guarantee that karyotype, age, and sex do not have causal regulators. We ran PC with kernel-based conditional independence tests [21] and α=0.01, followed by NOTEARS with λ=0.01, for a total of 20 times with bootstrapped samples. The output adjacency matrices were averaged to increase the robustness of the output.

The top 10 ranked LVs from the elastic net model based on both predicting karyotype and BMI were used to filter the significant mediating LVs, which were then used for causal graph reconstruction and visualization. To validate the connection between karyotype, LV, and BMI, the correlation between karyotype and BMI was calculated before and after regressing out mediating LVs or non-mediating variables.

3. Results

3.1. *Baseline gene expression and PLIER analysis*

To establish a baseline comparison, we contrasted two alternative workflows for finding associations between phenotypes and pathways using identical input data. In the first workflow, we ran gene set enrichment analysis (GSEA) to identify genesets that were differentially regulated when comparing disomic (D21) to trisomic (T21) individuals. This comparison showed that there were no significant genesets (FDR < 0.25). We additionally compared individuals with a BMI ≥ 30 to those with a BMI < 30 and also found that there were no significant genesets identified.

In the second workflow, we applied PLIER to identify 117 LVs across the HTP cohort. We performed differential analysis using the LV-by-sample matrix and found LVs that were significantly up and down in both karyotype and BMI ≥ 30 (**Figure 2A**). Directly comparing the results from GSEA and PLIER showed that the PLIER LVs were highly enriched for phenotype associations, while the GSEA results were not, suggesting that the PLIER model identifies functionally relevant molecular patterns in the data that are differentially associated with the clinical variables of interest in this study, which are karyotype and BMI (**Figure 2B**).

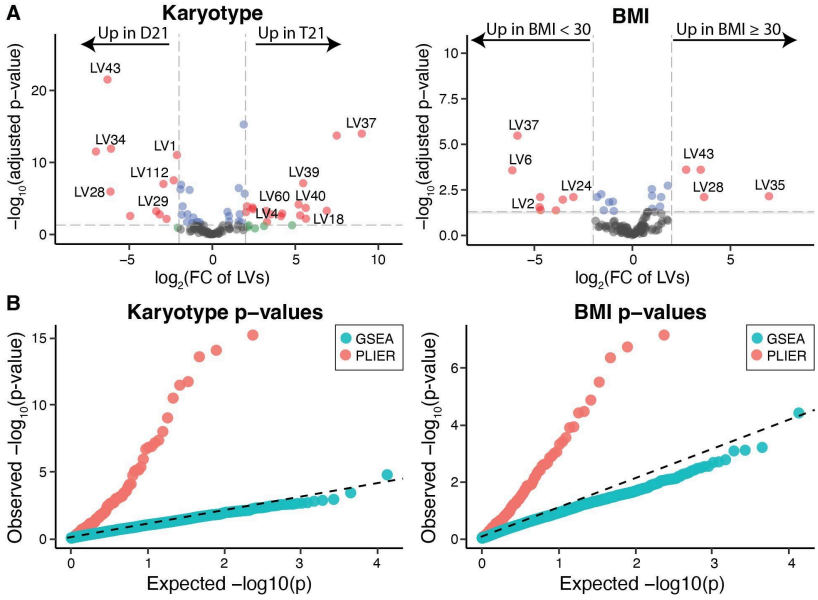


Fig 2. (A) Differential latent variable analysis performed for karyotype (T21 vs. D21) and BMI (BMI ≥ 30 vs. BMI < 30). (B) Comparison between a GSEA pathway analysis p-values and p-values from the differential latent variable analysis. Each individual p-value represents a hypothesized pathway phenotype association.

3.2. *BMI and karyotype associated latent variables*

Given the strong signal found with the PLIER-identified LVs, we performed elastic net regression to first predict karyotype and second to predict BMI using the LVs. In both cases, the elastic net model showed robust predictive performance. We trained the model using 80% of the data and evaluated the held-out 20%. We performed this procedure over 1,000 iterations, randomly sampling to define training and test datasets. To establish the random model, we shuffled the sample labels and performed the same procedure. As shown in **Figure 3A**, the average balanced accuracy is 0.9 for predicting karyotype compared to the expected 0.5 for the random model. For predicting BMI, the RMSE is 6.0 compared to 7.5 for the random model. A significant difference was seen for both prediction tasks (Kruskal–Wallis test, $p<0.001$). We next evaluated the LVs by comparing the average coefficient over the 1,000 iterations of the karyotype and BMI trained models. As shown in **Figure 3B**, 9 out of the top 10 LVs are unique in both conditions (BMI and karyotype), with LV56 being common to both. The top LV annotations are shown in **Figure 3C**.

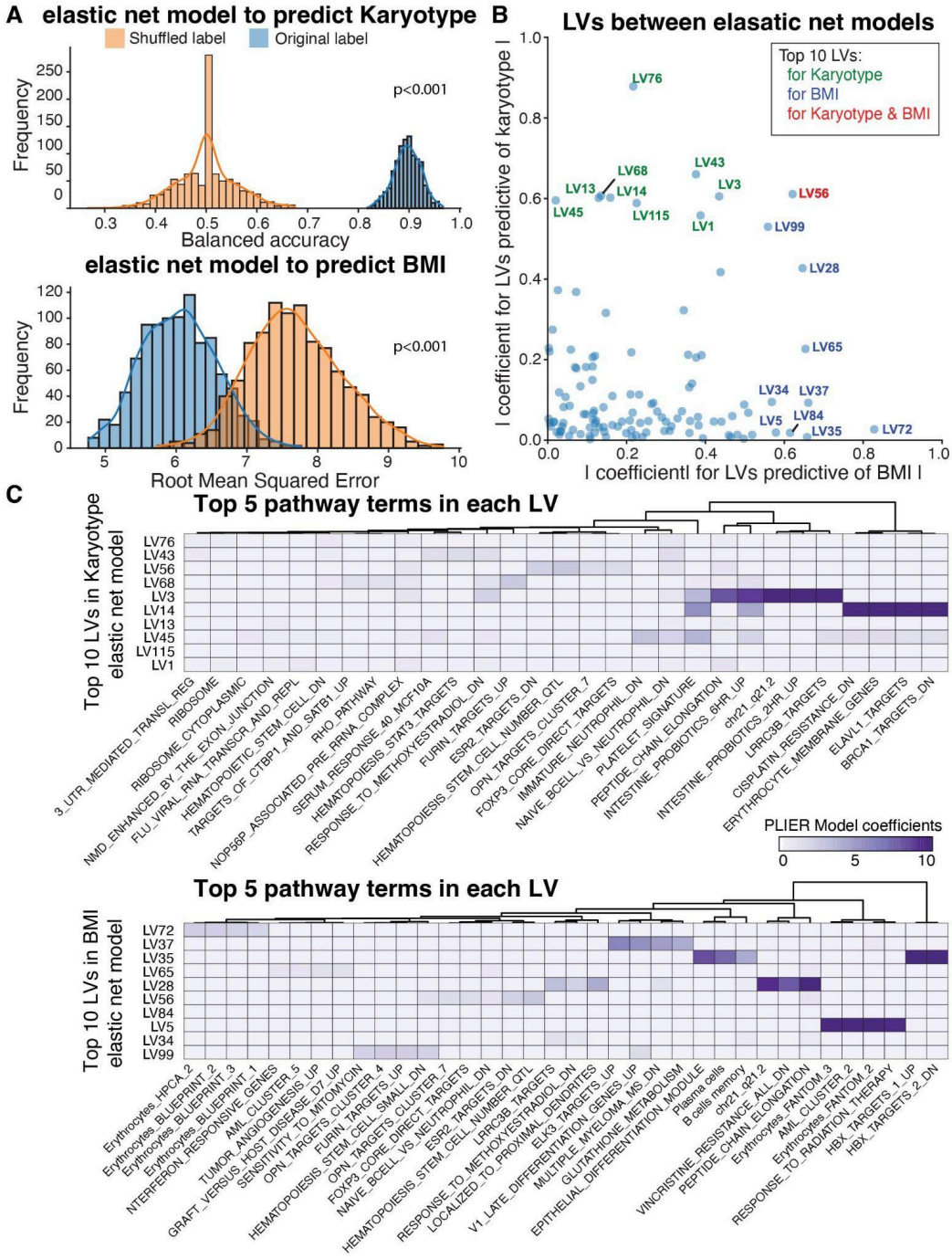


Fig 3. (A) Comparison of model performance of 1000 elastic net models for karyotype and BMI prediction from the PLIER LVs in the HTP cohort. (B) Absolute value of LV coefficients between the karyotype and BMI models. The top 10 LVs are annotated by color. (C) The top 5 pathways associated with the top 10 LVs from the karyotype and BMI models are shown with the color representing model coefficients in the *U* matrix from PLIER.

3.3. *Identifying causal mediators from highly ranked latent variables*

Figure 4A illustrates the workflow for the causal mediation analysis, which was performed to estimate the average causal mediation effects (ACME) of karyotype on BMI mediated through each LV, considering age and sex as covariates. Average direct effect (ADE) and total effect of karyotype on BMI were also reported from each causal mediation test. LVs with p-values < 0.05 from the mediation test were defined as mediator LVs. We then intersected the elastic net model top 10 LVs for both karyotype and BMI models with the mediation analysis to determine which LVs were both causally mediating BMI though karyotype and were predictive of these conditions (**Figure 4B**). We identified 7 top-ranking LVs as causal mediators including LV37, LV76, and LV3. **Figure 4C** shows the causal mediation analysis result for LV37 and LV3.

3.4. *Causal discovery prioritizes key causal mediators for BMI*

To obtain a comprehensive understanding of the relationships between LVs and obesity in the HTP cohort, we performed causal analysis with all 16 mediating LVs using PCnt, a hybrid causal discovery method with no causal regulator constraints on karyotype, age, and sex (**Figure 4A**). A subgraph with seven selected LVs and clinical variables from bootstrapped output was visualized in **Figure 4D**. We found three direct causal regulators for BMI in the subgraph. LV3 and LV37 are the two mediating LVs in the directed path from karyotype to BMI, while age is an independent cause for BMI change. To validate the mediation effect of LV3 and LV37, we adjusted BMI for each of its direct regulators and calculated the correlation between karyotype and BMI (**Figure 4E**). A decrease in the karyotype-BMI correlation after regressing out mediating LVs, as opposed to regressing out age, adds to the confidence in prioritizing LV37 and LV3 as causal mediators.

3.5. *Mechanisms underlying mediating latent variables*

The omnigenic framework allows for the inference of both core genes that affect phenotypes and peripheral genes that propagate their effects across networks or LVs. In this study, we identified LV37 as the primary mediator of BMI through karyotype. Glutathione peroxidase-1 (GPX1) is the top gene in LV37. Interestingly, GPX1 has been previously implicated in obesity[22] and DS[23]. The relationship between karyotype and LV37 is mediated by LV43 (**Figure 4D**). MCL1 Apoptosis Regulator, BCL2 Family Member (MCL1) is the top gene in LV43. MCL1 has known involvement in apoptosis [24] and associations with acute myeloblastic leukemia (AML)[25]. Children with DS have a higher risk than the disomic population for AML[26], and MCL1 was previously identified as a target for treatment in leukemias in DS[27]. Furthermore, glutathione metabolism was previously implicated in modulating the efficacy of BCL2 inhibitors[28].

To interpret these results in terms of known DS biology, we identified which LVs in the causal network contained superoxide dismutase-1 (SOD1) with high loadings. SOD1 is located on chr21 and has been implicated in metabolic regulation of body weight and insulin levels[29].

Moreover, an altered SOD1/GPX1 ratio is observed in T21, contributing to the hallmark oxidative stress observed across DS phenotypes[23]. SOD1 had higher rankings in LV76 and LV3 compared to the other 117 LVs (ranking 5th and 13th, respectively). Interestingly, LV76 and LV3 were on causal pathways distinct from LV37. Furthermore, the position of SOD1 within these LVs suggests a peripheral role in the network, propagating effects that ultimately influence core genes.

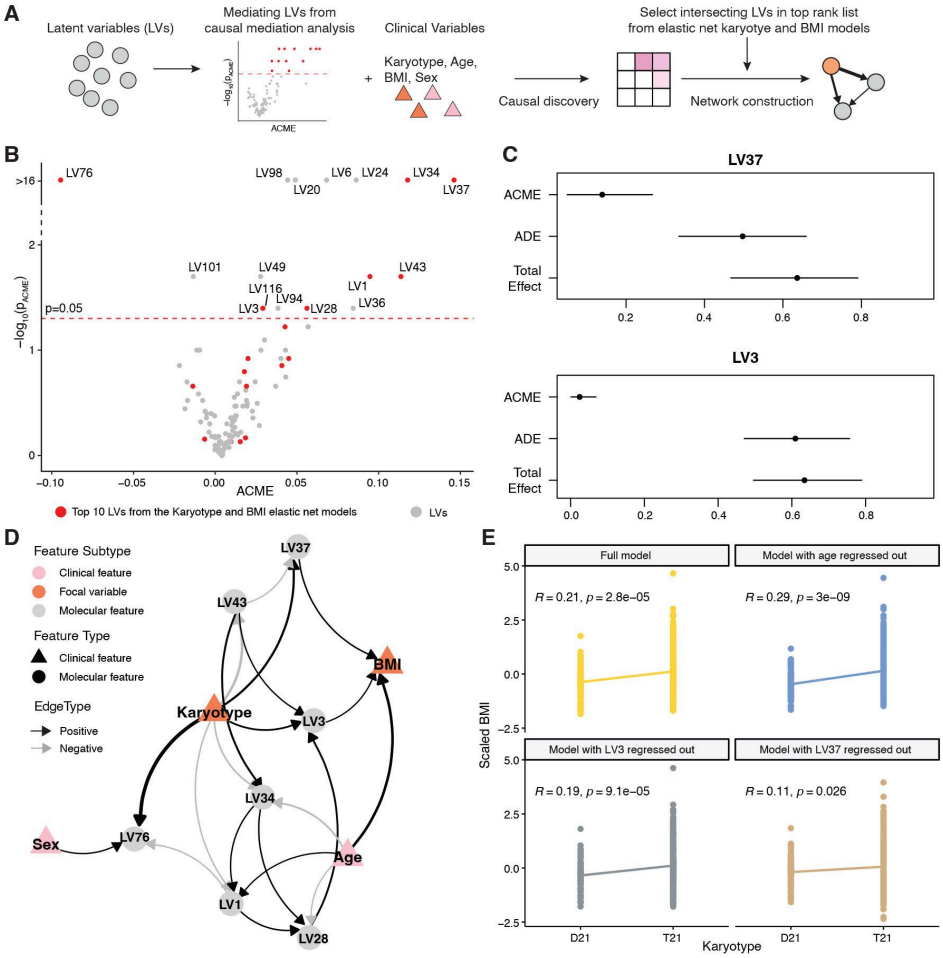


Fig 4. Causal relationship between molecular and clinical variables. A) Causal analysis work flow. B) Volcano plot for average causal mediation effect (ACME) between karyotype and BMI for each latent variable (LV). C) ACME through LV37 and LV3, average direct effect (ADE) and total effect between karyotype and BMI. D) Causal sub-network with key LVs. Directed edges between nodes represent causal directions. Edge weight represents estimated causal effect size. E) Scatter plot for karyotype and BMI before and after regressing out key LVs from BMI.

LV3 was another mediator of BMI through karyotype (**Figure 4C-E**). The gene sets with the highest loadings for LV3 included those implicated in protein translation and a geneset of chr21 genes, supporting the mediation of karyotype (**Figure 2C**). The gene with the highest

loading, *EEF1A1*, encodes an isoform of the alpha subunit of the elongation factor-1 complex responsible for the enzymatic delivery of aminoacyl tRNAs to the ribosome. Ribosomal dysfunction was previously implicated in a study on the impacts of high-fat diets on a DS mouse model (Ts65Dn)[30]**.**

Notably, LV28, which mediates the LV34-LV3 relationship in a causal path between karyotype and BMI (**Figure 4D**), also has *EEF1A1* as the highest loading gene. Since the PLIER methodology optimizes for independent LVs, the high loading of *EEF1A1* implicates a distinct process related to protein translation as mediating the relationship between karyotype and BMI. By integrating PrediXcan gene-trait associations with LVs using a regression model, we identified a significant association between this LV and obesity ($p<0.05$) (**Figure 5**). This finding supports the involvement of LV28 with obesity in an independent analysis.

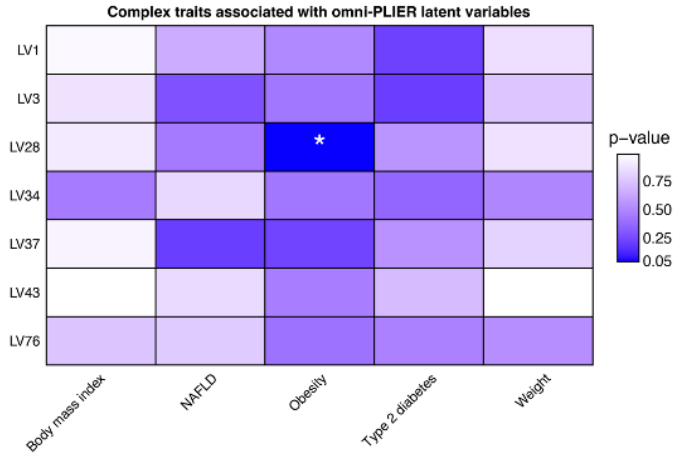


Fig 5. The association between latent variables (LVs) and BMI PhenomeXcan selected traits regressed against the omni-PLIER LVs. The columns represent traits, including Body mass index, Non-Alcoholic Fatty Liver Disease (NAFLD), Obesity, Type 2 diabetes, and Weight. A significant association is marked by an asterisk (*).

4. Discussion

The omni-PLIER framework presented here demonstrates how a complex network of phenotype-genotype/molecular trait interactions can be broken down into interpretable components, facilitating hypothesis formulation and validation. Disentangling disease co-occurrence in the context of T21 demonstrates the benefits of the omni-PLIER workflow.

A critical aspect of the omni-PLIER framework is leveraging the PLIER model to interpret gene expression through its latent components. Standard gene expression analysis performs statistical tests in gene space, such as differential expression (DE), and then conducts pathway level analysis, such as GSEA. This process requires defining groups for comparisons (contrasts) and uses predefined annotations to project gene level measurements into pathway/geneset space. In contrast, the PLIER model first performs sample annotation-agnostic latent variable extraction using pathways/genesets as prior information. Next, LVs are used to test the contrast groups and directly evaluate the LVs, which we have done here by treating the LVs as modules.

The PLIER approach offers several advantages. LVs capture multiple axes of variation while reducing redundant gene expression patterns. For example, a large group of highly correlated genes can dominate the top of a DE list. In the case of PLIER, this group is reduced to a single variable, allowing other less dominant pathways to receive consideration. Additionally, the association between LVs and pathways/genesets is conducted through elastic net regression, so pathways "compete" to explain LVs. This reduces the number of redundant pathways/genesets, as multiple ones are included in the model only if they provide additional information.

The combination of these effects results in a considerable increase in contrast group-to-pathway associations inferred from the dataset, as we showed in **Figure 2**. We additionally combined PLIER LVs with causal modeling, which offers two advantages. First, we find that many LVs are associated with a clinical phenotype of interest. Here, we leverage this observation to gain mechanistic insights into how T21 can drive BMI. Second, causal modeling, a multivariate technique, optimizes a network representation of the data's conditional (in)dependent structure. PLIER analysis alone cannot address independence in the LVs. Similar to non-negative matrix factorization, PLIER LVs are not guaranteed to be orthogonal and are often associated with each other. Causal modeling provides a mathematical framework to organize these associations and infer directionality, though learning causal models from observational data remains challenging [20].

An important aspect of our study design is that one of the variables of interest is genotype, which provides a valuable constraint for learning directed causal models. Specifically, we look for molecular traits (PLIER LVs) that mediate the effect of genotype on BMI. The final result is a simplified network structure that lends itself to interpretation and candidate variable selection. While the number of LVs associated with clinical variables can be large, the combination of univariate mediation analysis and multivariate PC-NOTEARS analysis identified targeted LVs. In our case, three mediating LVs (two direct: LV37 and LV3, and one indirect: LV43).

From the limited number of mediating LVs, we provided evidence that they both support existing knowledge within DS and obesity, and find support for novel mechanisms. It is important to note that, within T21, the ground truth is unknown. Although many findings align with biological knowledge, we cannot directly verify if this approach yields actionable mechanistic insights. Defining suitable benchmarking scenarios that reflect the complexity of a real dataset while providing some notion of ground truth will be the focus of future work.

A particular challenge that is endemic to this area of research is that even well annotated, controlled, and deep molecular datasets are snapshots in time of a complex biological system. We face the same challenge with the HTP dataset. Additionally, the HTP study is based entirely on blood profiling, which lacks important molecular details from other tissues and cell types.

Despite these limitations, the omni-PLIER framework demonstrated a computationally efficient workflow that synthesizes a large number of observations, prior knowledge, and state-of-the-art algorithmic approaches into a unified analytic method.

References

1. Mai CT, Isenburg JL, Canfield MA, et al. National population-based estimates for major birth defects, 2010-2014. *Birth Defects Res*. 2019;111(18):1420-1435.
2. Ofei F. Obesity - a preventable disease. *Ghana Med J*. 2005;39(3):98-101.
3. Bertapelli F, Pitetti K, Agiovlasitis S, Guerra-Junior G. Overweight and obesity in children and adolescents with Down syndrome-prevalence, determinants, consequences, and interventions: A literature review. *Res Dev Disabil*. 2016;57:181-192.
4. Dierssen M, Fructuoso M, Martínez de Lagrán M, Perluigi M, Barone E. Down Syndrome Is a Metabolic Disease: Altered Insulin Signaling Mediates Peripheral and Brain Dysfunctions. *Front Neurosci*. 2020;14:670.
5. Waugh KA, Minter R, Baxter J, et al. Triplication of the interferon receptor locus contributes to hallmarks of Down syndrome in a mouse model. *Nat Genet*. 2023;55(6):1034-1047.
6. Bayona-Bafaluy MP, Garrido-Pérez N, Meade P, et al. Down syndrome is an oxidative phosphorylation disorder. *Redox Biol*. 2021;41:101871.
7. Buonuomo PS, Bartuli A, Mastrogiorgio G, et al. Lipid profiles in a large cohort of Italian children with Down syndrome. *Eur J Med Genet*. 2016;59(8):392-395.
8. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017;169(7):1177-1186.
9. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*. 2019;177(4):1022-1034.e6.
10. Pividori M, Lu S, Li B, et al. Projecting genetic associations through gene expression patterns highlights disease etiology and drug mechanisms. *Nat Commun*. 2023;14(1):5562.
11. Mao W, Zaslavsky E, Hartmann BM, Sealfon SC, Chikina M. Pathway-level information extractor (PLIER) for gene expression data. *Nat Methods*. 2019;16(7):607-610.
12. Ng B, Casazza W, Patrick E, et al. Using Transcriptomic Hidden Variables to Infer Context-Specific Genotype Effects in the Brain. *Am J Hum Genet*. 2019;105(3):562-572.
13. Taroni JN, Grayson PC, Hu Q, et al. MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease. *Cell Syst*. 2019;8(5):380-394.e4.
14. Zhang S, Heil BJ, Mao W, Chikina M, Greene CS, Heller EA. MousiPLIER: A Mouse Pathway-Level Information Extractor Model. *eNeuro*. 2024;11(6). doi:10.1523/ENEURO.0313-23.2024
15. Galbraith MD, Rachubinski AL, Smith KP, et al. Multidimensional definition of the interferonopathy of Down syndrome and its response to JAK inhibition. *Sci Adv*. 2023;9(26):eadg6218.
16. Pividori M, Rajagopal PS, Barbeira A, et al. PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Sci Adv*. 2020;6(37):eaba2083.
17. Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. MIT press; 2001.
18. Zheng X, Aragam B, Ravikumar P, Xing E. DAGs with NO TEARS: Continuous optimization for structure learning. *Adv Neural Inf Process Syst*. March 2018:9492-9503.
19. omni-PLIER. https://github.com/CostelloLab/omni-PLIER.
20. Yuehua Zhu, Panayiotis V. Benos, and Maria Chikina. A Hybrid Constrained Continuous Optimization Approach for Optimal Causal Discovery from Biological Data. In: *ECCB2024*. Bioinformatics; 2024:bioinf - 2024-1389.
21. Zhang K, Peters J, Janzing D, Schölkopf B. Kernel-based Conditional Independence Test and

Application in Causal Discovery. In: ; 01 2011:804-813.
22. Ahmad AA, Rahimi Z, Asadi S, Vaisi-Raygani A, Kohsari M. The GPx-1 Gene Variants (rs1050450) in Obesity: Association with the Risk of Obesity and the GPx Activity in Females. *Rep Biochem Mol Biol*. 2023;12(1):185-194.
23. de Haan JB, Susil B, Pritchard M, Kola I. An altered antioxidant balance occurs in Down syndrome fetal organs: implications for the "gene dosage effect" hypothesis. *J Neural Transm Suppl*. 2003;(67):67-83.
24. Widden H, Placzek WJ. The multiple mechanisms of MCL1 in the regulation of cell fate. *Commun Biol*. 2021;4(1):1029.
25. Escudero S, Zaganjor E, Lee S, et al. Dynamic Regulation of Long-Chain Fatty Acid Oxidation by a Noncanonical Interaction between the MCL-1 BH3 Helix and VLCAD. *Mol Cell*. 2018;69(5):729-743.e7.
26. Mateos MK, Barbaric D, Byatt SA, Sutton R, Marshall GM. Down syndrome and leukemia: insights into leukemogenesis and translational targets. *Transl Pediatr*. 2015;4(2):76-92.
27. Toki T, Kanezaki R, Adachi S, et al. The key role of stem cell factor/KIT signaling in the proliferation of blast cells from Down syndrome-related leukemia. *Leukemia*. 2009;23(1):95-103.
28. Ebner J, Schmoellerl J, Piontek M, et al. ABCC1 and glutathione metabolism limit the efficacy of BCL-2 inhibitors in acute myeloid leukemia. *Nat Commun*. 2023;14(1):5709.
29. Sato A, Shiraishi Y, Kimura T, et al. Resistance to Obesity in SOD1 Deficient Mice with a High-Fat/High-Sucrose Diet. *Antioxidants (Basel)*. 2022;11(7). doi:10.3390/antiox11071403
30. Sarver DC, Xu C, Velez LM, et al. Dysregulated systemic metabolism in a Down syndrome mouse model. *Mol Metab*. 2023;68:101666.

Cross-Species Modeling Identifies Gene Signatures in Type 2 Diabetes Mouse Models Predictive of Inflammatory and Estrogen Signaling Pathways Associated with Alzheimer's Disease Outcomes in Humans

Brendan K. Ball[1], Elizabeth A. Proctor[2-6], Douglas K. Brubaker[7-8†]

[1]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA. [2]Department of Neurosurgery, Penn State College of Medicine, Hershey, PA, USA. [3]Department of Pharmacology, Penn State College of Medicine, PA, USA. [4]Department of Biomedical Engineering, Penn State University, State College, PA, USA. [5]Center for Neuroengineering, Penn State University, State College, PA, USA. [6]Department of Engineering Science & Mechanics, Penn State University, State College, PA, USA. [7]Center for Global Health & Diseases, Department of Pathology, School of Medicine, Case Western Reserve University, Cleveland, OH, USA. [8]Blood Heart Lung Immunology Research Center, University Hospitals, Cleveland, OH, USA.

[†]Email: dkb50@case.edu

Alzheimer's disease (AD), the predominant form of dementia, is influenced by several risk factors, including type 2 diabetes (T2D), a metabolic disorder characterized by the dysregulation of blood sugar levels. Despite mouse and human studies reporting this connection between T2D and AD, the mechanism by which T2D contributes to AD pathobiology is not well understood. A challenge in understanding mechanistic links between these conditions is that evidence between mouse and human experimental models must be synthesized, but translating between these systems is difficult due to evolutionary distance, physiological differences, and human heterogeneity. To address this, we employed a computational framework called translatable components regression (TransComp-R) to overcome discrepancies between pre-clinical and clinical studies using omics data. Here, we developed a novel extension of TransComp-R for multi-disease modeling to analyze transcriptomic data from brain samples of mouse models of AD, T2D, and simultaneous occurrence of both disease (ADxT2D) and postmortem human brain data to identify enriched pathways predictive of human AD status. Our TransComp-R model identified inflammatory and estrogen signaling pathways encoded by mouse principal components derived from models of T2D and ADxT2D, but not AD alone, predicted with human AD outcomes. The same mouse PCs predictive of human AD outcomes were able to capture sex-dependent differences in human AD biology, including significant effects unique to female patients, despite the TransComp-R being derived from data from only male mice. We demonstrated that our approach identifies biological pathways of interest at the intersection of the complex etiologies of AD and T2D which may guide future studies into pathogenesis and therapeutic development for patients with T2D-associated AD.

Keywords: Alzheimer's disease, type 2 diabetes, preclinical translation, cross-species modeling, systems biology

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by memory loss, confusion, and behavioral changes. With more than 6.9 million people living with AD in the

United States[1], $360 billion dollars in health and long-term care costs is expected to be spent in 2024, and projected to rise to $1 trillion by 2050[1]. As the prevalence of AD is expected to increase with the country's aging population, developing effective therapeutics proven to treat or cure AD becomes urgent. Despite the rapid increase of AD cases, studies to develop therapeutics for AD is difficult[2,3]. This difficulty is in part due to the development of AD occurring decades before diagnosis[4] and the multi-factorial nature of the disease[5–8].

In efforts to identify risk factors for AD it was observed that individuals with type 2 diabetes (T2D), a metabolic condition distinguished by chronic hyperglycemia, have an elevated risk in developing AD[9,10]. The development of T2D occurs decades before the diagnosis of AD and is reported to increase the risk of dementia[11]. In the United States, more than 39 million people have T2D, and 116 million have pre-diabetes[12]. This population of people diagnosed with or at risk for developing T2D may face a heightened risk for developing AD in light of the comorbidity of the diseases[13,14]. In clinical studies, common features of both AD and T2D include chronic inflammation[15,16], increased insulin resistance[17], and alterations to mitochondria and energy metabolism[18,19]. Despite multiple studies supporting a link between T2D and AD risk, the biological mechanisms by which this occurs are not well understood.

A critical challenge in understanding the mechanistic links between these conditions is that evidence must be synthesized and translated between experiments in mouse models and human-based clinical studies. Translating information from pre-clinical models to human clinical contexts is difficult due to discrepancies in interspecies physiology[20], timeline of disease development[21], and heterogeneity of the human population[22]. In cases of precision medicine, where complex dependencies between clinical phenotypes are difficult to deconvolute, such as is the case with AD and T2D, there is an important role for computational approaches to resolve this heterogeneity into testable mechanistic hypotheses to guide therapeutic development[23–25].

To overcome this challenge, we developed a computational framework termed translatable components regression (TransComp-R) to identify omics-based signatures in mouse models predictive of AD conditions in human[26–28]. The TransComp-R model works by projecting human omics data into a mouse principal component analysis (PCA) space, followed by linear regression of mouse principal components (PCs) against human disease outcomes to identify translatable mouse PCs. The gene signatures encoded within mouse PCs that best separate conditions between human AD and control outcomes can be interpreted using biological pathway analyses such as gene set enrichment analysis (GSEA). These informed pathways can then be validated through literature and experimental studies.

Here, we aimed to perform a cross-species analysis using publicly available mouse and human transcriptomic data to determine biological pathways by which T2D contributes to AD. We developed a novel extension of TransComp-R that integrated PCs from multiple murine disease models: AD, T2D, and co-occurrence of both diseases (ADxT2D) in a single computational model to compare the predictive power of different murine models of disease and identify mouse-specific features predictive of human AD status. We also modified the existing TransComp-R method by incorporating human demographic variables such as sex and age variables into our model to inform the selection of translatable mouse PCs and better position the insights from the cross-species model to specific human patient subsets, an important goal of precision medicine. Our method synthesizes

mouse models with multiple disease etiologies with human information to prioritize biological pathways affected in disease and prospectively evaluate therapeutic avenues from pre-clinical to clinical contexts with high-throughput omics data.

2. Results

2.1. *Selected mouse and human transcriptomic data were pre-processed for TransComp-R*

Publicly available mouse (GSE152539)[29] and human (GSE48350)[30,31] datasets of microarrayed brain tissue samples were selected from Gene Expression Omnibus (GEO). The mouse dataset uniquely included conditions of solely AD, only T2D, and simultaneous occurrence of both diseases from the hippocampus. The mouse models consisted of six-month-old male $App^{NL\text{-}F/NL\text{-}F}$ knock-ins responsible for heightened amyloid-beta in the brain (Swedish KM670/671NL, Iberian I716F) and wild type (C57BL/6J) mice were fed with either a high-fat diet (custom diet, 40% kcal from fat, and 0.15% from cholesterol) or regular diet (CA-1, 18.8% kcal from fat) for 12 months (n = 3 per condition) for the respective disease groups. The human dataset contains demographic variables of sex and age along with the transcriptomic data of AD (n = 80) and control (n = 173) subjects from four brain regions: hippocampus, entorhinal cortex, superior frontal cortex, and post-central gyrus.

To prepare the data for the TransComp-R framework, both mouse and human transcriptomics datasets were matched for one-to-one homologs. From homolog matching, 13,428 genes were identified, and all other genes that did not have a matching homolog pair were excluded from the analysis. The human data was next filtered for the hippocampal region to account for brain-region variability. Any subjects below the age of 65 were removed from the study to reduce age bias (**Table 1**). Both datasets were individually $\log_2$ transformed and normalized by z-score per gene.

Table 1. Summary of the processed human data across disease condition, age, and sex.

Condition	Age (years) (Mean ± SD)	Sex (%)		Total Sample Size (n)
		Female	Male	
Control	82.7 ± 9.5	11 (46%)	13 (54%)	24
Alzheimer's Disease	84.3 ± 6.6	9 (50%)	9 (50%)	18

2.2. *TransComp-R modeling separates human samples in mouse principal component space*

Here, we applied the TransComp-R methodology, with the incorporation of LASSO to select PCs most predictive of AD outcomes[26]. The TransComp-R model begins with the projection of human data into the mouse PCA space (**Figure 1A**), followed by the evaluation of mouse PC translatability through LASSO and generalized linear model (GLM) regression. The significant mouse PCs that can distinguish between human AD and control are interpreted by GSEA of the gene loading coefficients on each PC (**Figure 1B**). The biological pathways identified from GSEA can provide insight on human biology translated by mouse PCs, which can then be validated through follow-up experiments and literature review.

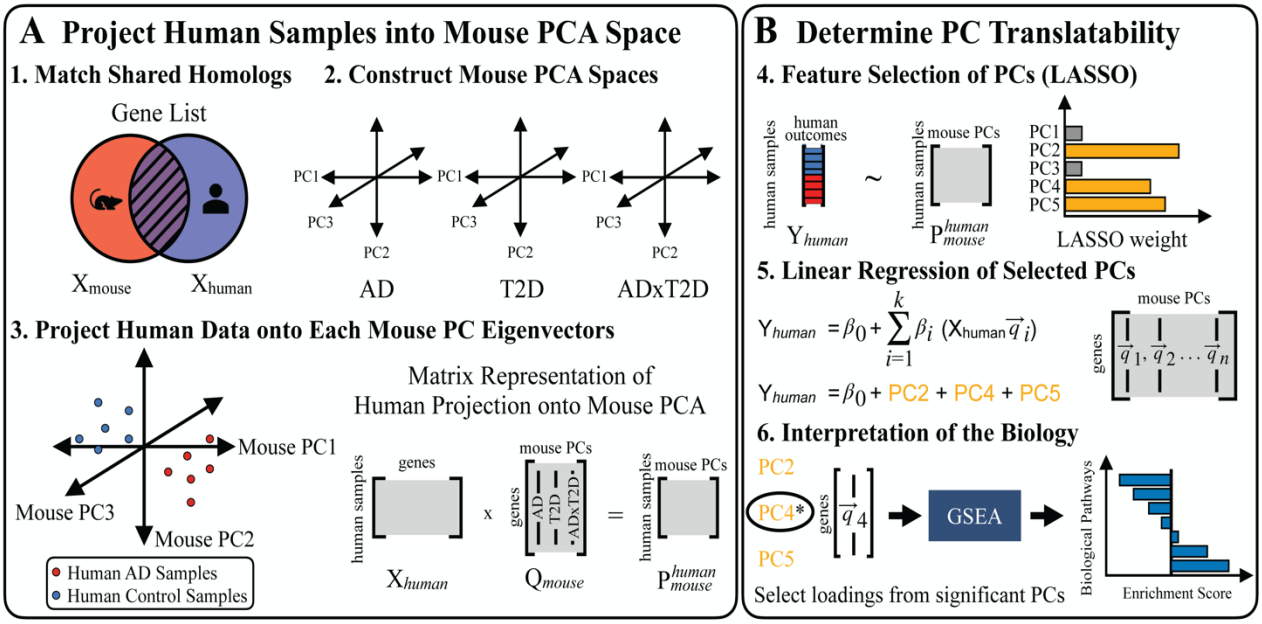


Figure 1. The TransComp-R computational approach. (A) Homolog gene pairs between human and mouse datasets are selected for analysis. Human samples are projected into mouse PCA spaces to combine mouse and human information. **(B)** Principal component translatability from mouse to human is determined by performing a GLM regression against human AD outcomes with PCs selected from LASSO. The loadings from the significant PCs are analyzed via GSEA to identify enriched biological pathways.

Implementing this approach, mouse data were separated into AD, T2D, and ADxT2D with controls prior to constructing separate PCA models, such that three groups of PCs encoded transcriptomic variation between healthy controls and AD, T2D, or ADxT2D mice. To avoid overfitting the mouse data, a threshold of 80% cumulative variance explained was set for each PCA, and as a result, a total of five PCs per disease group were selected (**Supplementary Figure S1**). Next, the human data was projected on the mouse PCA space.

We then trained four separate LASSO models to identify PCs most predictive of binarized human disease outcomes. Using the combined dataset containing rows of human and columns of T2D PCs, we incorporated progressively included human demographic variables associated with the respective human subjects in LASSO such that we examined: models of only mouse PCs, PCs with human sex, PCs with human age, and PCs with both human sex and age main effects. This approach allows us to include human demographic variables in a cross-species translation model, prioritizing not just mouse PCs, but also how mouse PCs capture the heterogeneity of human sex and age when predicting AD outcomes. The PCs were next selected based on 100 rounds of 5-fold cross-validation, where PCs with a significant LASSO coefficient in greater than half of the models were carried forward to the GLM.

From the LASSO models, we found T2D PC2 and ADxT2D PC3 to be consistently selected across all four LASSO models, while T2D PC3 was selected from all models except for the model with only mouse PC main effects (**Figure 2A**). Additionally, AD PC5 and T2D PC5 were selected from the LASSO model with only mouse PCs as main effect variables, but not in other LASSO models that included human demographic variables. The PCs identified by the LASSO model,

which were encoded with transcriptomic variance, were next evaluated for their respective ability to discern between human AD and control status through GLM.

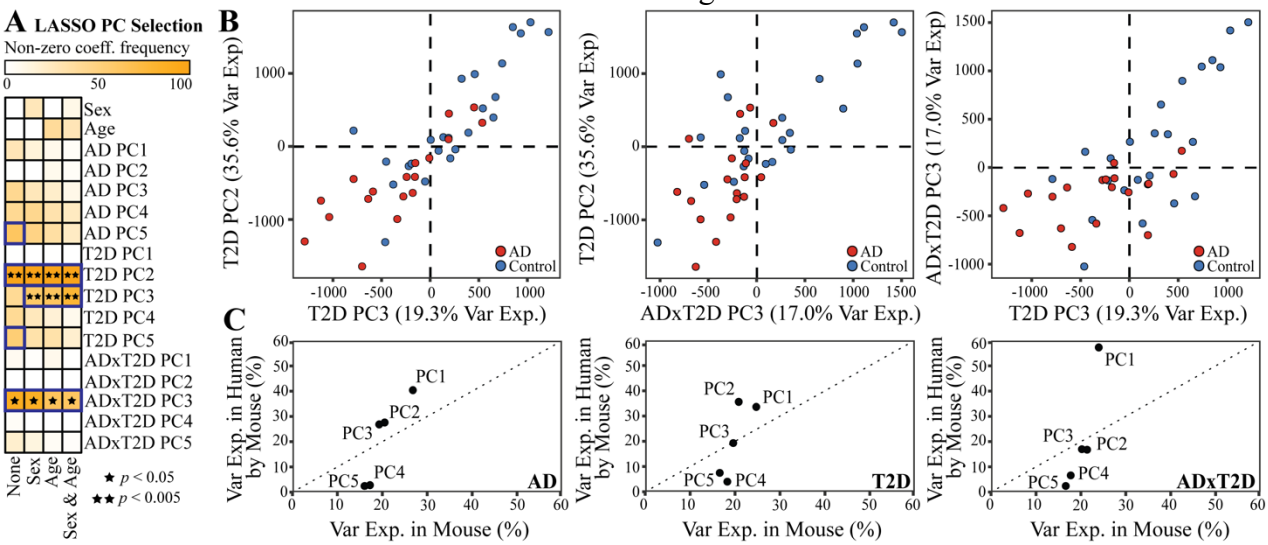


Figure 2. TransComp-R identifies translatable PCs predictive for AD outcomes in human. (A) Selection of PCs using LASSO across 100 rounds of 5-fold cross-validation. The four LASSO models included terms with just mouse PCs, PCs and human sex, PCs and human age, and PCs and human age and sex. PCs with a coefficient frequency greater than 50 rounds of 100 were selected for the GLM and regressed against binarized human disease outcomes (significance defined by simple regression model *p* value). **(B)** A principal component plot of human scores on the selected mouse T2D PC2, T2D PC3, and ADxT2D PC3 separating human control and AD outcomes **(C)** Mouse PCs were separated by disease cohort, comparing the variance explained in mice to the variance in humans explained by mouse respective mouse PCs.

To evaluate the predictability of the selected mouse PCs for human AD , we constructed GLMs with all selected PCs predicting AD status in humans, but these multi-PC models were not significantly predictive due to multi-collinearity (**Supplementary Figure S2**). As a result, we constructed GLMs for each individual PC regressed against human disease outcomes. We found the three mouse PCs consistently selected from LASSO to be predictive of human AD outcomes individually (T2D PC2 $p = 0.0047$, T2D PC3 $p = 0.0042$, and ADxT2D PC3 $p = 0.0130$) (**Figure 2A**). We also note that although AD PC5 and T2D PC5 satisfied the non-zero frequency greater than 50 in the LASSO model with only mouse PCs, the regression against human outcomes was not significant, and was excluded from further analysis (AD PC5 $p = 0.275$, T2D PC5 $p = 0.443$). Consistent LASSO selection of T2D PC2, T2D PC3, and ADxT2D PC3 as significant PCs indicates the importance of including human clinical and demographic variables in the TransComp-R model to detect translatable cross-species biology while controlling for clinical covariates.

We visualized the two T2D mouse PCs and one ADxT2D mouse PC that were identified by TransComp-R as predictive of human AD status (**Figure 2B**). In all three PCs, there was visible separation between the control and AD groups. We next compared the translatability of the selected mouse PCs to their ability to explain the variance in human data (**Figure 2C**). Comparing the proportion of PC variance explained in mouse to the variance explained in human by the same mouse PC, we found that T2D PC3 and ADxT2D PC3 explained a similar ratio, whereas mouse T2D PC2 explained almost double the variance in human by mouse than the mouse PCs alone. This

could imply that certain pathways represented by mouse T2D PC3 and ADxT2D PC3 were conserved consistently across mice and humans, whereas mouse T2D PC2 may had a more pronounced effect in capturing information cross-species.

2.3. *Mouse principal components selected genes contribute to human disease separation*

Having identified three mouse PCs predictive of human AD versus control status from TransComp-R, we were interested in isolating genes that were contributing to the separation between human AD and control subjects. Filtering for human genes ranked with the top and bottom 25 loadings within their respective PCs, we identified genes in the model predictive of AD and control in humans (**Figure 3A-C**). While no genes were shared across the top and bottom 25 ranked on the three mouse PCs, we observed distinct patterns of gene expression among human AD and control groups.

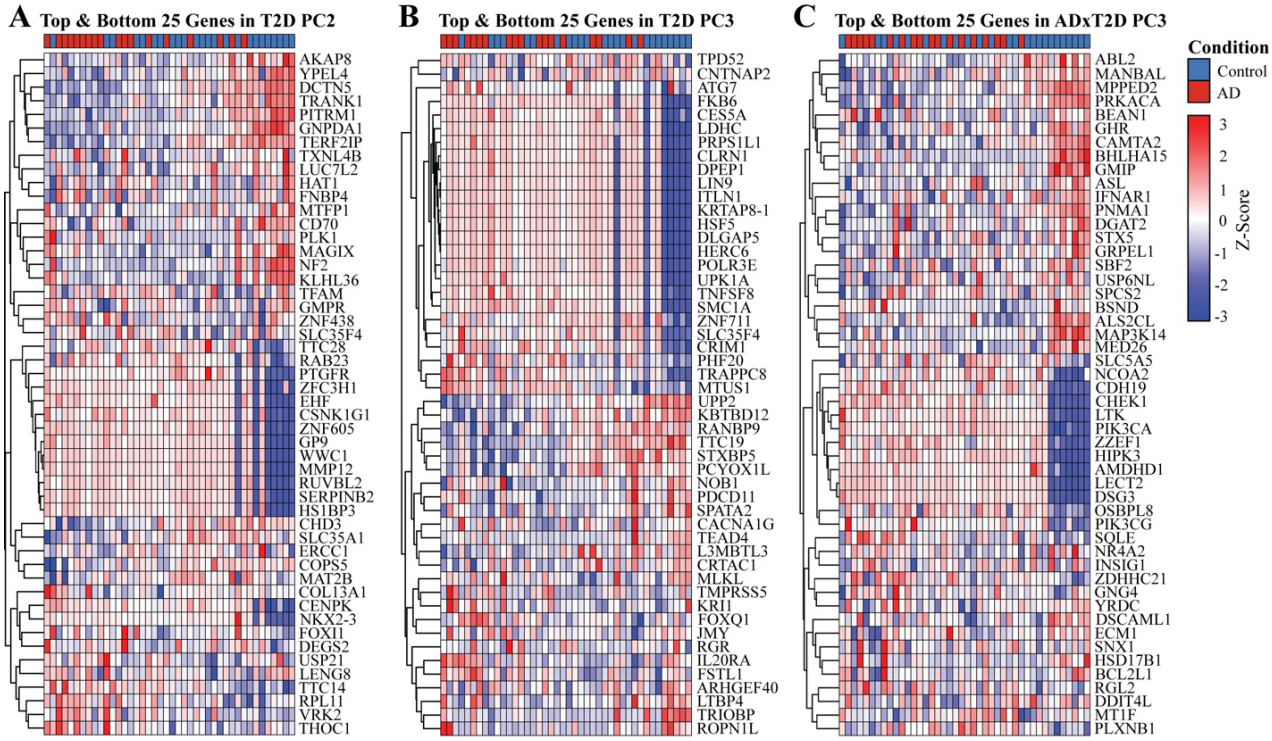


Figure 3. The top and bottom 25 genes of translatable PCs. Z-scored AD human transcriptomic data were filtered by genes with the 25 largest and smallest scores on **(A)** T2D PC2, **(B)** T2D PC3, and **(C)** ADxT2D PC3. Human samples were sorted by their respective PC scores with the most negative (left) to the most positive (right).

2.4. *Gene set enrichment analysis identifies inflammatory and estrogen signaling pathways enriched in human Alzheimer's disease outcome*

We performed GSEA on the selected T2D and ADxT2D PCs and identified pathways associated with inflammatory and estrogen signaling. From the KEGG database, we identified "Complement and Coagulation Cascades" and "Cytokine-Cytokine Receptor Interaction" on T2D PC2 (**Figure 4A**). On T2D PC3, the "Phosphatidylinositol Signaling System" was the only pathway found to be enriched for AD (**Figure 4B**). There were no significant KEGG pathways on T2D PC3. On the Hallmark database, we identified "Interferon Gamma Response," "Interferon Alpha Response,"

"IL6 JAK STAT Signaling," and "Inflammatory Response" to be enriched for AD conditions by T2D PC2 (**Figure 4C**). Interestingly in the Hallmark database, we identified "Estrogen Response Early" was enriched for the control group in T2D PC2 (**Figure 4C**), while "Estrogen Response Late" was enriched for AD in T2D PC3 and ADxT2D PC3 (**Figure 4D-E**).

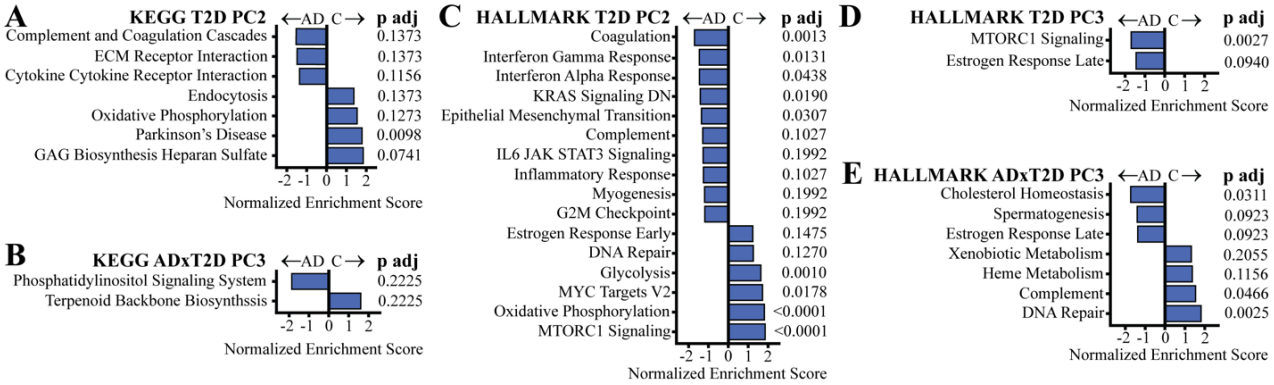


Figure 4. Enriched biological pathways identified from GSEA. Significant KEGG pathways from **(A)** T2D PC2 and **(B)** ADxT2D PC3. No significant pathways were enriched in T2D PC3. Significant Hallmark pathways were identified for **(C)** T2D PC2, **(D)** T2D PC3, and **(E)** ADxT2D PC3. Enriched pathways were defined by a Benjamini-Hochberg adjusted p value < 0.25. Pathways enriched for AD are displayed with a negative normalized enrichment score.

Based on our findings with the estrogen pathways, we were interested in distinguishing the genes that contributed to "Estrogen Response Early" and "Estrogen Response Late." From GSEA, we identified 76, 77, and 49 core enrichment genes contributing to the estrogen-associated pathways in mouse T2D PC2, T2D PC3, and ADxT2D PC3, respectively. Comparing the genes that were contributing to the estrogen response, we found 23 shared genes between T2D PC3 and ADxT2D PC3 (*PDZK1*, *LLGL2*, *KLK11*, *TOP2A*, *PTGES*, *FARP1*, *NAB2*, *CISH*, *MEST*, *KIF20A*, *LTF*, *ISG20*, *IMPA2*, *DUSP2*, *PLAC1*, *PRKAR2B*, *TNNC1*, *OPN3*, *AREG*, *ATP2B4*, *AGR2*, *CALCR*, and *RABEP1*), 2 genes between T2D PC2 and ADxT2D PC3 (*DHCR7* and *MAPT*), 10 genes between T2D PC2 and T2D PC3 (*TPBG*, *FKBP4*, *GLA*, *NXT1*, *CD44*, *PGR*, *RAB31*, *AFF1*, *TFAP2C*, and *TJP3*), and 5 genes shared across all three mouse PCs (*SULT2B1*, *OVOL2*, *SIAH2*, *FDFT1*, and *RBBP8*) (**Supplementary Figure S3**). Additionally, 19 genes enriched in ADxT2D PC3, 39 genes enriched in T2D PC3, and 59 genes enriched in T2D PC2 did not overlap with any other mouse PCs.

2.5. *Male mouse-derived principal components significantly stratify female Alzheimer's disease and control groups in human subjects*

Expanding upon the potential sex-based predictability, we were curious to see if the model was able to distinguish sex and disease status by the PC scores. Here, we separated the scores of each mouse PC by human sex and AD status and found that mouse T2D PC2, T2D PC3, and ADxT2D PC3 significantly stratified human female AD and control groups, and not male AD and control groups, despite the mouse data originating from all male mice (**Figure 5A-C**). The ability of these PCs to distinguish between female AD and control groups shows the model's ability in identifying human sex-based differences in the context of disease development. This is supported by the significance of the separation between the two groups (p value < 0.05).

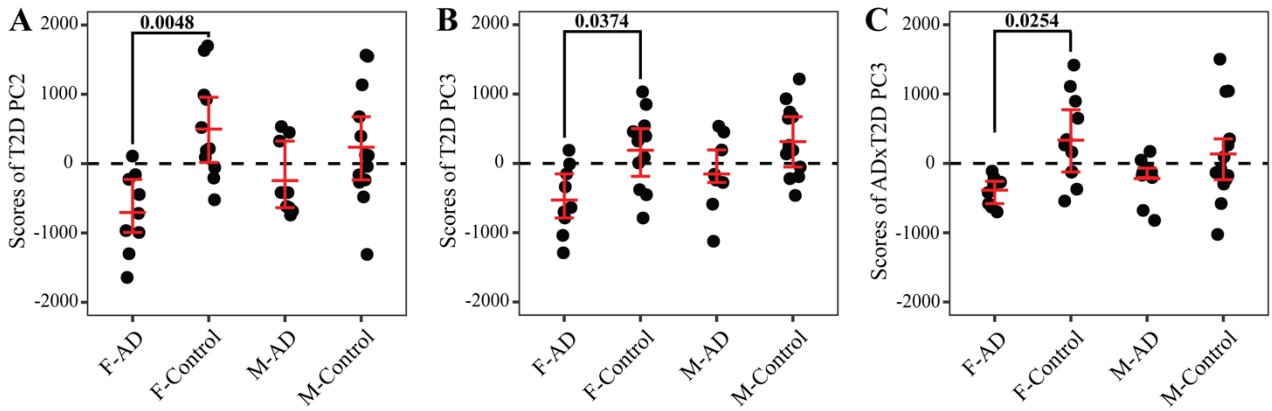


Figure 5. Comparison of sex and disease status among the translatable PCs. Scores of each PC were separated by female (F) AD, female control, male (M) AD, and male control for **(A)** T2D PC2, **(B)** T2D PC3, and **(C)** ADxT2D PC3. A Mann-Whitney pair-wise test corrected by the Benjamini-Hochberg method (FDR q value < 0.05) was used to determine the significance among the groups. The mean of the distribution is labeled with the interquartile range.

3. Discussion

In this work, we aimed to uncover potential biological mechanisms that connected T2D as a risk factor for AD development using mouse and human transcriptomic data. An obstacle in understanding the links between these diseases, in which multifactorial mechanisms interact in humans and biological mechanisms are isolated in animal studies, is that information from mouse models and human-based studies must be synthesized to inform clinical and therapeutic decisions. Currently, translating information from pre-clinical models to patient-specific contexts is often difficult due to discrepancies in interspecies physiology[20], timeline of disease development[21], and heterogeneity of the human population[22]. To overcome these challenges, we innovated on TransComp-R to identify potential biological pathways from mouse PCs that are predictable for AD outcomes. In the TransComp-R workflow, we fused multiple mouse disease models in a single computational model together with human data containing demographic sex and age variables to predict outcomes in AD. With our computational model, we pinpointed potential biological pathways associated with AD, and identified sex-specific differences, despite the mouse disease models being representative of only males.

We identified inflammatory pathways that may link T2D as a risk factor for AD development. These links have the potential translational utility in bridging mouse and human biology to understand and develop therapeutic strategies for AD with T2D exacerbating factors. The mouse T2D PC2 identified several pathways on both KEGG and hallmark databases. From KEGG, “Complement and Coagulation Cascades” and “Cytokine-Cytokine Receptor Interaction” were enriched for AD. From the literature, studies report complement activation to be associated with insulin resistance and T2D[32–34]. Likewise, high complement levels are contributed by neurons and glial cells in AD[35,36]. In both T2D and AD, cytokines are found to actively participate in the progression of disease[37,38].

Using complementary pathway databases, we identified “Interferon Gamma Response,” “Interferon Alpha Response,” “IL6 JAK STAT Signaling,” and “Inflammatory Response” pathways on mouse T2D PC2 enriched in human AD. Interferon gamma[39] and alpha[40], key cytokines in the

innate immune response and response to viral infections, are altered in AD. However, we notice that interferon gamma[41] is more associated with T2D, whereas interferon alpha[42,43] is found to be elevated in subjects with type 1 diabetes instead[44]. IL6 JAK-STAT signaling has been reported to impair the insulin-degrading enzyme, a protein found to be associated with obesity and T2D[45]. In AD, IL6 signaling has been linked with cognitive impairment and metabolic alterations[46]. Collectively, these results may indicate that chronic inflammation could lead to downstream insulin resistance and cognitive deficits[47].

Our results also indicate that estrogen signaling may serve as a potential connection between T2D and AD. From GSEA, ranked genes in ADxT2D PC3 and T2D PC3 both identified "Estrogen Response Late" as pathways enriched for AD, whereas "Estrogen Response Early" was enriched for human control by T2D PC2. Among the three PCs, 49 genes were enriched for ADxT2D PC3, 77 genes were enriched for T2D PC2, and 76 genes were enriched for T2D PC2. Of these, 59 were enriched in T2D PC2, and 23 were shared between ADxT2D PC3 and T2D PC3, in which we compared with previously published literature to potential associations with AD and T2D. Associated with AD in the mouse T2D PC2, we identified *MED13L*[48] and *XBP1*[49] connected to cognitive deficits, changes in mitochondrial metabolism (*PMAIP1*)[50], inflammation (*RASGRP1*)[51], and the expression of *NRIP1*[52] reduced in AD. Similarly, we identified genes associated with insulin resistance (*FASN* and *FKBP5*)[53,54], genetic variances of *RAPGEF1*[55] and increased expression of *AQP3*[56] related to T2D development. Interpreting genes shared across ADxT2D PC3 and T2D PC3, both PCs, we found *MEST*[57] reported to alter Wnt signaling in AD, and *KIF20A*[58], a gene found to be differentially expressed in AD. Likewise in T2D, we found *CISH*[59] to be involved with gluconeogenesis, whereas beta-cells were preserved with upregulated *AGR2*[60].

There were five genes shared across the three mouse PCs identifying estrogen signaling as a potential biological pathway, which included *SULT2B1*, *OVOL2*, *SIAH2*, *FDFT1*, and *RBBP8*. Of the five genes, all but *RBBP8* were reported to have connections to AD or T2D in literature. *SULT2B1*, part of the sulfotransferase family that catalyzes the sulfate conjunction of hormones and neurotransmitters, was found to be upregulated in AD rat models[61]. In a T2D study, *SULT2B1* overexpressed in the liver inhibited hepatic gluconeogenesis in two separate diabetic mouse models: one induced by high-fat diet, and another via leptin-deficiency (ob/ob)[62]. Other genes related to T2D include *OVOL2* and *SIAH2*. The presence of *OVOL2* was found to be linked with beta cell dedifferentiation, a mechanism linked with pancreatic dysfunction[63], and *SIAH2* deficiency improved glucose and insulin tolerance[64]. Related to AD, inhibition of squalene synthase (*FDFT1*) inhibited by squalestatin reduced cellular prion protein in ScN2a, SMB, and ScGT1 (prion-infected cell lines)[65], and protection against amyloid beta-induced synapse damage[66]. Further examination of these genes may be of potential interest to connect biological pathways between T2D and AD.

Interestingly in both diseases, previous studies report that estrogen may play a protective role in AD[67] and T2D[68]. In AD, estrogen provides protection from amyloid-beta toxicity, a hallmark of AD pathology[69,70]. In females that experienced menopause, hormone therapy with estrogen has been found to reduce the risk of T2D onset[68]. Although studies indicate estrogen to be protective, others report that estrogen may be deleterious depending on the timing and onset of T2D[71,72]. These differences could be a result of the varying roles that different genes may have: some genes may contribute to disease when upregulated, while others may serve a protective role that can lead to

disease if downregulated. This variability in genes could further explain the possible observation of estrogen appearing to have both harmful and protective effects. Therefore, further investigations are encouraged to further understand the role of estrogen as a shared pathway between AD and T2D.

Finally, we found that the mouse PCs defined by T2D (PC2 and PC3) and ADxT2D (PC3) were able to distinguish between female AD and control subjects. Despite the mouse groups being entirely male, our model detected sex-based differences in females. This is interesting because females are at a higher risk of developing AD than males[73]. Observing this result, as well as PCs showing enrichment for estrogen, may suggest that despite the widespread lack of female animals in preclinical research, our model is able to detect biological signals in male mice predictive of female human disease biology, thereby enhancing the retrospective utility of prior animal studies that fell short of equitable design. In the specific case of our models, the pathways we identified on the male mouse PC's predictive of human female AD pathology implicate our model's ability to translate transcriptomic signatures across human sex demographics.

There are limitations and opportunities to expand this study. Few research groups have explored the T2D-AD axis, and as a result, there are limited sample sizes available for mouse and human omics data. The incorporation of additional studies that satisfy the criteria of our selection process into the model may improve the confidence of these results. Second, our TransComp-R model only considers homologous gene pairs shared across mice and humans. As a result, we potentially omit genes in pre-processing that may be involved in the development of AD. Additionally, the GLM in our model only regresses against control or AD status without the incorporation of transient phases such as mild cognitive impairment. Finally, the TransComp-R framework has the opportunity to consider other clinical variables that may predict disease outcomes. Some additional factors include information on race, clinical neuropathological scores for AD severity, and current T2D biomarkers. Considering these potential factors may further enhance future cross-species modeling.

Our work expanded upon the existing TransComp-R framework to identify potential biological pathways in which T2D may exacerbate AD development. We show that mouse PCs from T2D and ADxT2D were most predictive of AD outcomes in human. Interestingly, mouse PC's derived from mice with AD alone were not predictive of human AD, which may indicate that metabolic dysfunction encoded on the mouse T2D and T2DxAD PCs plays a more significant role in human AD biology than is typically accounted for. Indeed, these results encourage future applications of TransComp-R to overcome barriers of pre-clinical to human studies and identify affected biological pathways in AD or different diseases. The implications of this work for precision medicine can be expanded to other disease models that may be difficult to synthesize between pre-clinical experiments and clinical studies. This platform could synthesize various omics data from pre-clinical and patient-specific data to rationally select potential pathways to target, which may further enhance clinical studies or possible therapeutic avenues.

4. Materials and Methods

4.1. *Data selection*

Mouse and human datasets were selected with the criteria of matching hippocampal brain region, information containing AD and T2D conditions in the mouse dataset, human sample size greater

than 12 per condition, and at least sex and age information in the human dataset. Additionally, datasets derived from similar sequencing platforms were prioritized. Search terms on GEO included phrases such as “hippocampus Alzheimer’s disease in human,” “mouse Alzheimer’s disease hippocampus,” and “mouse diabetes hippocampus.” Additional searches included the term “gene expression” on the GEO repository.

4.2. *Pre-processing and normalization*

Publicly available transcriptomic human and mouse data were obtained from the GEO repository using Bioconductor tools in R (*GEOquery* 2.70.0, *limma* 3.58.1, and *Biobase* 2.62.0)[74–76]. Before processing, all human subjects with a reported age below 65 years old were removed from the analysis to prevent bias from younger age groups. The imported datasets were $\log_2$ transformed, then human and mouse gene lists were matched for homologous pairs (*orthogene* 1.8.0)[77]. The two datasets were filtered for the hippocampal brain region. The genes were then internally normalized by z-score prior to TransComp-R modeling.

4.3. *Cross-species modeling and variable selection*

We applied TransComp-R by conducting PCA on the mouse data separated in AD, T2D, and ADxT2D groups with controls, such that three groups of PCs encoded transcriptomic variation between healthy controls and AD, T2D, or ADxT2D mice. To avoid overfitting, the number of PCs in its respective group was limited to an 80% cumulative variance explained. Human AD and control subjects were projected into mouse PCA space. Mouse PCs associated with AD outcomes in human were selected by performing LASSO across 100 rounds of 5-fold cross-validation regressing the human positions in mouse PC space against human disease status. Four sets of LASSO models were trained, including main effects of mouse PCs, PCs and human sex, PCs and human age, and PCs and human age and sex. PCs with a coefficient frequency greater than 50 of the 100 rounds were selected for GLMs with individual PCCs and human clinical covariates regressed against human AD outcomes. The significance of the PC was determined if the model p value was less than 0.05.

4.4. *Variance explained in human by mouse principal components*

Human data containing subject information and gene lists, as well as mouse PCs with a matching gene list, was used to calculate the variance explained by mouse in human. Using mouse PCs in the columns of Q, we projected the human data matrix X onto the PCs via matrix multiplication and calculated the percent variance of mouse in X explained by a given column q_i of Q (with T representing the matrix transpose) as:

$$\text{VarExpHuman}(q_i) = \frac{q_i^T [X^T X]\, q_i}{\sum \text{diag}(Q^T X^T X Q)} \quad (1)$$

4.5. *Identifying genes contributing to human separation by mouse principal components*

Genes contributing to the most positive and negative scores were identified by selecting loaded genes with the top 25 and bottom 25 scores in each of the selected PCs. The selected genes were then used to filter the gene list of the human dataset containing z-scored gene expression data. A

heatmap, with the human subjects, clustered by their scores from the TransComp-R model, and the 50 total genes were visualized to compare gene expression between AD and control.

4.6. *Gene set enrichment analysis*

GSEA was performed on the loadings of selected PCs from the GLM in R (*msigdbr* 7.5.1, *fgsea* 1.28.0, and c*lusterProfiler* 4.10.1)[78–80]. From the Molecular Signatures Database, two human collections to perform GSEA included the KEGG and Hallmark databases. The parameters for the minimum gene set size and the maximum gene set size were set to 5 and 500, respectively. The tuning constant, epsilon, was established at 0. For both KEGG and Hallmarks databases, enriched biological pathways were determined significant if the Benjamini-Hochberg adjusted *p* value was less than 0.25.

4.7. *Sex-based comparison across principal component scores*

As an approach to compare predictability across sex, scores of selected PCs were separated by sex and disease categories. A Mann-Whitney pair-wise test was used to determine significance among four groups (AD females, control females, AD males, and control males). To correct for multiple comparisons, *p* values were adjusted with the Benjamini-Hochberg factor. An adjusted *p* value less than 0.05 was considered significant for the analysis.

Acknowledgments

This work is supported by an award from the Good Ventures Foundation and Open Philanthropy (DKB and BKB). BKB is supported by the National Science Foundation through the Graduate Research Fellowship program under grant DGE-1842166. BKB also acknowledges the NIH T32 predoctoral fellowship T32DK101001 from the National Institute of Diabetes and Digestive and Kidney Diseases. This work is also supported by the National Institute on Aging under R01AG072513 (EAP).

Author Contributions

BKB: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, visualization, writing-original draft, writing-review & editing. EAP: Conceptualization, methodology, writing-review & editing. DKB: Conceptualization, funding acquisition, methodology, project administration, resources, writing review & editing.

Competing Interests

The authors declare no competing interests.

Supplementary Material & Code Availability

All supplementary figures, documents, and code for this study are made publicly available at https://github.com/Brubaker-Lab/MouseHuman-TransCompR-T2D-AD.

References

1. Alzheimer's Association. 2024 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* **n/a**, (2024).
2. Asher, S. & Priefer, R. Alzheimer's disease failed clinical trials. *Life Sciences* **306**, 120861 (2022).
3. Cummings, J., Feldman, H. H. & Scheltens, P. The "rights" of precision drug development for Alzheimer's disease. *Alzheimer's Research & Therapy* **11**, 76 (2019).
4. Rajan, K. B., Wilson, R. S., Weuve, J., Barnes, L. L. & Evans, D. A. Cognitive impairment 18 years before clinical diagnosis of Alzheimer disease dementia. *Neurology* **85**, 898–904 (2015).
5. Niu, H., Álvarez-Álvarez, I., Guillén-Grima, F. & Aguinaga-Ontoso, I. Prevalence and incidence of Alzheimer's disease in Europe: A meta-analysis. *Neurología (English Edition)* **32**, 523–532 (2017).
6. Ferretti, M. T. *et al.* Sex differences in Alzheimer disease — the gateway to precision medicine. *Nat Rev Neurol* **14**, 457–469 (2018).
7. Huang, Y.-W. A., Zhou, B., Wernig, M. & Südhof, T. C. ApoE2, ApoE3, and ApoE4 Differentially Stimulate APP Transcription and Aβ Secretion. *Cell* **168**, 427-441.e21 (2017).
8. Nedelec, T. *et al.* Identifying health conditions associated with Alzheimer's disease up to 15 years before diagnosis: an agnostic study of French and British health records. *The Lancet Digital Health* **4**, e169–e178 (2022).
9. Chatterjee, S. *et al.* Type 2 Diabetes as a Risk Factor for Dementia in Women Compared With Men: A Pooled Analysis of 2.3 Million People Comprising More Than 100,000 Cases of Dementia. *Diabetes Care* **39**, 300–307 (2015).
10. Janson, J. *et al.* Increased Risk of Type 2 Diabetes in Alzheimer Disease. *Diabetes* **53**, 474–481 (2004).
11. Zilkens, R. R., Davis, W. A., Spilsbury, K., Semmens, J. B. & Bruce, D. G. Earlier Age of Dementia Onset and Shorter Survival Times in Dementia Patients With Diabetes. *American Journal of Epidemiology* **177**, 1246–1254 (2013).
12. Martin, S. S. *et al.* 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. *Circulation* **149**, e347–e913 (2024).
13. Jash, K. *et al.* Cognitive dysfunction: A growing link between diabetes and Alzheimer's disease. *Drug Development Research* **81**, 144–164 (2020).
14. Kubis-Kubiak, A., Dyba, A. & Piwowar, A. The Interplay between Diabetes and Alzheimer's Disease—In the Hunt for Biomarkers. *International Journal of Molecular Sciences* **21**, 2744 (2020).
15. Donath, M. Y. & Shoelson, S. E. Type 2 diabetes as an inflammatory disease. *Nat Rev Immunol* **11**, 98–107 (2011).
16. Akiyama, H. *et al.* Inflammation and Alzheimer's disease. *Neurobiology of Aging* **21**, 383–421 (2000).
17. Yoon, J. H. *et al.* How Can Insulin Resistance Cause Alzheimer's Disease? *International Journal of Molecular Sciences* **24**, 3506 (2023).
18. Bingham, E. M. *et al.* The Role of Insulin in Human Brain Glucose Metabolism. *Diabetes* **51**, 3384–3390 (2002).
19. Hirvonen, J. *et al.* Effects of Insulin on Brain Glucose Metabolism in Impaired Glucose Tolerance. *Diabetes* **60**, 443–447 (2011).
20. Brubaker, D. K. & Lauffenburger, D. A. Translating preclinical models to humans. *Science* **367**, 742–743 (2020).
21. Saraceno, C., Musardo, S., Marcello, E., Pelucchi, S. & Diluca, M. Modeling Alzheimer's disease: from past to future. *Front. Pharmacol.* **4**, (2013).
22. Gauthier, S. *et al.* Why has therapy development for dementia failed in the last two decades? *Alzheimer's & Dementia* **12**, 60–64 (2016).
23. Brubaker, D. K., Proctor, E. A., Haigis, K. M. & Lauffenburger, D. A. Computational translation of genomic responses from experimental model systems to humans. *PLoS Comput Biol* **15**, e1006286 (2019).
24. Brubaker, D. K. *et al.* Proteogenomic Network Analysis of Context-Specific KRAS Signaling in Mouse-to-Human Cross-Species Translation. *Cell Systems* **9**, 258-270.e6 (2019).
25. Holland, C. H., Szalai, B. & Saez-Rodriguez, J. Transfer of regulatory knowledge from human to mouse for functional genomics analysis. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1863**, 194431 (2020).
26. Brubaker, Douglas. K. *et al.* An interspecies translation model implicates integrin signaling in infliximab-resistant inflammatory bowel disease. *Sci Signal* **13**, eaay3258 (2020).
27. Suarez-Lopez, L. *et al.* Cross-species transcriptomic signatures predict response to MK2 inhibition in mouse models of chronic inflammation. *iScience* **24**, 103406 (2021).
28. Lee, M. J. *et al.* Computational Interspecies Translation Between Alzheimer's Disease Mouse Models and Human Subjects Identifies Innate Immune Complement, TYROBP, and TAM Receptor Agonist Signatures, Distinct From Influences of Aging. *Frontiers in Neuroscience* **15**, (2021).

29. Mazzei, G. *et al.* A high-fat diet exacerbates the Alzheimer's disease pathology in the hippocampus of the App NL−F/NL−F knock-in mouse model. *Aging Cell* **20**, e13429 (2021).
30. Berchtold, N. C. *et al.* Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease. *Neurobiol Aging* **34**, 1653–1661 (2013).
31. Cribbs, D. H. *et al.* Extensive innate immune gene activation accompanies brain aging, increasing vulnerability to cognitive decline and neurodegeneration: a microarray study. *J Neuroinflammation* **9**, 179 (2012).
32. Shim, K., Begum, R., Yang, C. & Wang, H. Complement activation in obesity, insulin resistance, and type 2 diabetes mellitus. *World J Diabetes* **11**, 1–12 (2020).
33. Tan, S. M., Snelson, M., Østergaard, J. A. & Coughlan, M. T. The Complement Pathway: New Insights into Immunometabolic Signaling in Diabetic Kidney Disease. *Antioxid Redox Signal* **37**, 781–801 (2022).
34. Yudkin, J. S., Stehouwer, C. D. A., Emeis, J. J. & Coppack, S. W. C-Reactive Protein in Healthy Subjects: Associations With Obesity, Insulin Resistance, and Endothelial Dysfunction. *Arteriosclerosis, Thrombosis, and Vascular Biology* **19**, 972–978 (1999).
35. Luchena, C., Zuazo-Ibarra, J., Alberdi, E., Matute, C. & Capetillo-Zarate, E. Contribution of Neurons and Glial Cells to Complement-Mediated Synapse Removal during Development, Aging and in Alzheimer's Disease. *Mediators Inflamm* **2018**, 2530414 (2018).
36. Goetzl, E. J., Schwartz, J. B., Abner, E. L., Jicha, G. A. & Kapogiannis, D. High complement levels in astrocyte-derived exosomes of Alzheimer's disease. *Ann Neurol* **83**, 544–552 (2018).
37. Sankar, S. B. *et al.* Amyloid beta and diabetic pathology cooperatively stimulate cytokine expression in an Alzheimer's mouse model. *J Neuroinflammation* **17**, 38 (2020).
38. Ball, B. K., Kuhn, M. K., Fleeman Bechtel, R. M., Proctor, E. A. & Brubaker, D. K. Differential responses of primary neuron-secreted MCP-1 and IL-9 to type 2 diabetes and Alzheimer's disease-associated metabolites. *Sci Rep* **14**, 12743 (2024).
39. Browne, T. C. *et al.* IFN-γ Production by Amyloid β–Specific Th1 Cells Promotes Microglial Activation and Increases Plaque Burden in a Mouse Model of Alzheimer's Disease. *The Journal of Immunology* **190**, 2241–2251 (2013).
40. Taylor, J. M. *et al.* Type-1 interferon signaling mediates neuro-inflammatory events in models of Alzheimer's disease. *Neurobiology of Aging* **35**, 1012–1023 (2014).
41. Taylor, S. *et al.* Suppressing Interferon-γ Stimulates Microglial Responses and Repair of Microbleeds in the Diabetic Brain. *J. Neurosci.* **38**, 8707–8722 (2018).
42. Newby, B. N. & Mathews, C. E. Type I Interferon Is a Catastrophic Feature of the Diabetic Islet Microenvironment. *Front Endocrinol (Lausanne)* **8**, 232 (2017).
43. Marroqui, L. *et al.* Interferon-α mediates human beta cell HLA class I overexpression, endoplasmic reticulum stress and apoptosis, three hallmarks of early human type 1 diabetes. *Diabetologia* **60**, 656–667 (2017).
44. Wang, L. *et al.* Characteristics of Interferon-Associated Diabetes Mellitus in Past 30 Years: A Review. *Horm Metab Res* **54**, 145–152 (2022).
45. Kurauti, M. A. *et al.* Interleukin-6 increases the expression and activity of insulin-degrading enzyme. *Sci Rep* **7**, 46750 (2017).
46. Lyra e Silva, N. M. *et al.* Pro-inflammatory interleukin-6 signaling links cognitive impairments and peripheral metabolic alterations in Alzheimer's disease. *Transl Psychiatry* **11**, 1–15 (2021).
47. Vinuesa, A. *et al.* Inflammation and Insulin Resistance as Risk Factors and Potential Therapeutic Targets for Alzheimer's Disease. *Front Neurosci* **15**, 653651 (2021).
48. Schiano, C., Luongo, L., Maione, S. & Napoli, C. Mediator complex in neurological disease. *Life Sciences* **329**, 121986 (2023).
49. Cissé, M., Duplan, E. & Checler, F. The Transcription Factor XBP1 in Memory and Cognition: implications in Alzheimer's Disease. *Mol Med* **22**, 905–917 (2016).
50. Ling, Y., Hu, L., Chen, J., Zhao, M. & Dai, X. The mechanism of mitochondrial metabolic gene *PMAIP1* involved in Alzheimer's disease process based on bioinformatics analysis and experimental validation. *Clinics* **79**, 100373 (2024).
51. Oh, M. *et al.* Nurr1 performs its anti-inflammatory function by regulating RasGRP1 expression in neuro-inflammation. *Sci Rep* **10**, 10755 (2020).
52. Blondrath, K. *et al.* The nuclear cofactor receptor interacting protein-140 (RIP140) regulates the expression of genes involved in Aβ generation. *Neurobiology of Aging* **47**, 180–191 (2016).
53. Menendez, J. A., Vazquez-Martin, A., Ortega, F. J. & Fernandez-Real, J. M. Fatty Acid Synthase: Association with Insulin Resistance, Type 2 Diabetes, and Cancer. *Clinical Chemistry* **55**, 425–438 (2009).

54. Sidibeh, C. O. *et al.* FKBP5 expression in human adipose tissue: potential role in glucose and lipid metabolism, adipogenesis and type 2 diabetes. *Endocrine* **62**, 116–128 (2018).
55. Hong, K.-W. *et al.* RAPGEF1 gene variants associated with type 2 diabetes in the Korean population. *Diabetes Research and Clinical Practice* **84**, 117–122 (2009).
56. Nejsum, L. N. *et al.* Compensatory increase in AQP2, p-AQP2, and AQP3 expression in rats with diabetes mellitus. *American Journal of Physiology-Renal Physiology* **280**, F715–F726 (2001).
57. Prasad, R. *et al.* Hypermethylation of Mest promoter causes aberrant Wnt signaling in patients with Alzheimer's disease. *Sci Rep* **11**, 20075 (2021).
58. Petralia, M. C. *et al.* Computational Analysis of Pathogenetic Pathways in Alzheimer's Disease and Prediction of Potential Therapeutic Drugs. *Brain Sci* **12**, 827 (2022).
59. Xiao, F. *et al.* Hepatic cytokine-inducible SH2-containing protein (CISH) regulates gluconeogenesis via cAMP-responsive element binding protein (CREB). *The FASEB Journal* **36**, e22541 (2022).
60. Kanno, A. *et al.* Early administration of dapagliflozin preserves pancreatic β-cell mass through a legacy effect in a mouse model of type 2 diabetes. *Journal of Diabetes Investigation* **10**, 577–590 (2019).
61. Chik, M. W., Hazalin, N. A. M. N. & Singh, G. K. S. Regulation of phase I and phase II neurosteroid enzymes in the hippocampus of an Alzheimer's disease rat model: A focus on sulphotransferases and UDP-glucuronosyltransferases. *Steroids* **184**, 109035 (2022).
62. Shi, X. *et al.* Cholesterol Sulfate and Cholesterol Sulfotransferase Inhibit Gluconeogenesis by Targeting Hepatocyte Nuclear Factor 4α. *Molecular and Cellular Biology* **34**, 485–497 (2014).
63. de Jesus, D. S. *et al.* Dysregulation of the *Pdx1/Ovol2/Zeb2* axis in dedifferentiated β-cells triggers the induction of genes associated with epithelial–mesenchymal transition in diabetes. *Molecular Metabolism* **53**, 101248 (2021).
64. Ghosh, S. *et al.* Siah2 modulates sex-dependent metabolic and inflammatory responses in adipose tissue to a high-fat diet challenge. *Biol Sex Differ* **10**, 19 (2019).
65. Bate, C., Salmona, M., Diomede, L. & Williams, A. Squalestatin Cures Prion-infected Neurons and Protects Against Prion Neurotoxicity *. *Journal of Biological Chemistry* **279**, 14983–14990 (2004).
66. Bate, C. & Williams, A. Squalestatin protects neurons and reduces the activation of cytoplasmic phospholipase A2 by Aβ1–42. *Neuropharmacology* **53**, 222–231 (2007).
67. Li, R., Cui, J. & Shen, Y. Brain Sex Matters: estrogen in cognition and Alzheimer's disease. *Mol Cell Endocrinol* **389**, 13–21 (2014).
68. Mauvais-Jarvis, F., Manson, J. E., Stevenson, J. C. & Fonseca, V. A. Menopausal Hormone Therapy and Type 2 Diabetes Prevention: Evidence, Mechanisms, and Clinical Implications. *Endocr Rev* **38**, 173–188 (2017).
69. Nilsen, J., Chen, S., Irwin, R. W., Iwamoto, S. & Brinton, R. D. Estrogen protects neuronal cells from amyloid beta-induced apoptosis via regulation of mitochondrial proteins and function. *BMC Neurosci* **7**, 74 (2006).
70. Marin, R. *et al.* Estradiol prevents amyloid-β peptide-induced cell death in a cholinergic cell line via modulation of a classical estrogen receptor. *Neuroscience* **121**, 917–926 (2003).
71. Carcaillon, L. *et al.* High plasma estradiol interacts with diabetes on risk of dementia in older postmenopausal women. *Neurology* **82**, 504–511 (2014).
72. Pereira, R. I. *et al.* Timing of Estradiol Treatment After Menopause May Determine Benefit or Harm to Insulin Action. *The Journal of Clinical Endocrinology & Metabolism* **100**, 4456–4462 (2015).
73. O'Neal, M. A. Women and the risk of Alzheimer's disease. *Front. Glob. Womens Health* **4**, (2024).
74. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–1847 (2007).
75. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).
76. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115–121 (2015).
77. Schilder, B. & Skene, N. orthogene: an R package for easy mapping of orthologous genes across hundreds of species. Bioconductor (2022).
78. Dolgalev, I. msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format. (2022).
79. Korotkevich, G. *et al.* Fast gene set enrichment analysis. 060012 Preprint at https://doi.org/10.1101/060012 (2021).
80. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).

Biologically Enhanced Machine Learning Model to uncover Novel Gene-Drug Targets for Alzheimer's Disease

Alena Orlenko*[1], Mythreye Venkatesan[1], Li Shen[2], Marylyn D. Ritchie[3], Zhiping Paul Wang[1], Tayo Obafemi-Ajayi[4], Jason H. Moore[1]

[1]*Department of Computational Biomedicine*
Cedars-Sinai Medical Center, Los Angeles, CA, USA

[2]*Department of Biostatistics, Epidemiology and Informatics*
[3]*Department of Genetics and Institute for Biomedical Informatics*
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[4]*Engineering Program*
Missouri State University, Springfield, Missouri, USA

Given the complexity and multifactorial nature of Alzheimer's disease, investigating potential drug-gene targets is imperative for developing effective therapies and advancing our understanding of the underlying mechanisms driving the disease. We present an explainable ML model that integrates the role and impact of gene interactions to drive the genomic variant feature selection. The model leverages both the Alzheimer's knowledge base and the Drug-Gene interaction database (DGIdb) to identify a list of biologically plausible novel gene-drug targets for further investigation. Model validation is performed on an ethnically diverse study sample obtained from the Alzheimer's Disease Sequencing Project (ADSP), a multi-ancestry multi-cohort genomic study. To mitigate population stratification and spurious associations from ML analysis, we implemented novel data curation methods. The study outcomes include a set of possible gene targets for further functional follow-up and drug repurposing.

Keywords: genomics; Alzheimer's disease; feature importance; informatics; epistasis.

1. Introduction

Alzheimer's disease is the most common cause of dementia, and its prevalence is rapidly increasing due to extended lifespans worldwide.[1] With this surge, there is an urgent need to identify therapeutic targets, potential biomarkers, and risk predictive strategies.[2] Lack of success in recent clinical trials confirmed that AD pathology is very complex and a greater understanding of the underlying mechanisms that contribute to aging and neurodegenerative processes is critical.[3] AD is considered to have a large genetic component and is highly heritable.[4] The polygenic nature of AD presents an obstacle to early diagnosis and risk prediction.[2]

*This work was supported by National Institutes of Health (USA) grants R01 AG066833 and R01 LM010098.

Research on AD is a national priority, with 6.5 million Americans affected at an annual cost of more than $250 billion and no definitive cure available.[5–7] This places a significant priority on discovery and approval of therapeutics treatment for AD.[7–9] Drug repurposing involves finding new therapeutic uses for existing drugs that are already on the market.[10] This can lead to significant savings in both time and cost compared to developing new drugs from scratch. Since the safety profiles of these drugs are already well-established, the process can bypass many early-stage trials, speeding up the timeline for reaching patients in need.[10] The Alzheimer's knowledge base (AlzKB) has been developed as a computational AD resource with a particular focus on drug discovery and drug repurposing.[7] It integrates data from 22 diverse sources that spans genes, pathways, drugs, and diseases related to AD to form a specialized open source graph-based knowledge base to aid discovery of complex translational associations for AD drug discovery. The nodes denotes entities (such as genes, pathways, drugs, and diseases) while the edges represent semantic relationships between nodes (entities) such as "*chemical_binds_gene*", "*gene_interacts_with_gene*", "*gene_regulates_gene*", etc. This work leverages the AlzKB's information on gene-gene interaction with known AD genes.

Understanding the role and impact of gene interactions on disease phenotypes is increasingly recognised as an essential aspect of genetic disease research.[11] Most disease-gene association methods do not account for gene-gene interactions, despite their crucial role in complex, polygenic diseases like AD.[2] Exploring the action, function, regulation, and control of proteins can elucidate a clearer understanding of disease processes, cellular functions, and regulatory networks.[12] This is critical in advancing towards precision medicine, given the necessity of anchoring therapeutic targets to a disease mechanism substantiated by genetic evidence.[13] Many of the key functions and life processes in biology are maintained to some extent by different types of protein-protein interactions (PPIs). Knowledge graphs, such as AlzKB, provide a rich heterogeneous network structure that leverages biological and molecular prior knowledge, to uncover possible novel gene-gene interactions that could aid the drug repurposing quest for AD. Drug-gene knowledge sources, such as Drug-Gene interaction database (DGIdb),[14] also provides a rich resource of known interactions between drugs and genes aggregated from multiple sources. This offers additional insights into the molecular mechanisms of drug actions and gene functions, aiding in understanding the underlying biology of diseases and outlining clinically relevant genes.

Machine learning (ML) models in combination with genome-wide association studies (GWAS) have shown promise for identifying novel genes that confer AD risk.[1] To this date, AD GWAS across multiple populations have identified more than 80 loci, with the majority studies conducted in European ancestry cohorts primarily due to large sample sizes.[4] The best known genetic risk factor is the inheritance of the $\epsilon 4$ allele of the apolipoprotein E ($APOE$) gene.[15] Other AD candidate genes have also been identified such as amyloid precursor protein (APP), microtubule-associated protein tau ($MAPT$).[2,15,16] Though ML models have the potential to exploit complex genetic interactions and provide insights into AD pathology, the heterogeneous landscape of AD etiology presents a key challenge.[15] Given the complex biomedical phenotypes that often characterize human diseases, it is becoming increasingly more accepted that epistatic interactions between genes could be more prevalent than previ-

ously assumed.[17,18] Epistatic interactions can be defined as interactions between two or more gene loci where the phenotype cannot be accurately predicted by simply adding the effects of individual gene loci.[19] Epistatic interactions have been detected in multiple GWAS of various disease phenotypes, including AD[20] and other neurological diseases.[21,22] Due to gaps in the current understanding of AD etiology and the complex interactions between genomic and other factors that contribute to its heterogeneity, a multi-modal approach is needed to promote a better mechanistic understanding of the disease.

We present an explainable ML model enhanced by PPI knowledge, specifically epistatic interaction, to identify potential novel non-AD genomic variants with drug targets for further investigation. The underlying hypothesis is that we can leverage the AlzKB and other knowledge sources to pinpoint a set of biologically plausible genes by exploring those with existing drug targets that exhibit a "gene_interacts_with_gene" relationship with known AD genes in the knowledge graph. A key novelty of the ML model is integration of biological knowledge at every level to yield meaningful explanations for model performance and genomic variant (single nucleotide polymorphism (SNP)) feature selection.

2. Methods

We present an ML explainable model, enhanced by the biological knowledge of epistatic interaction, to identify novel genomic variants that could be biomarkers for AD novel gene-drug targets. This framework (see Figure 1) consists of three key phases: (i) druggable gene priority feature selection leveraging AlzKB and DGIdb, (ii) AD study sample data curation, and (iii) ML feature selection and epistasis analysis.

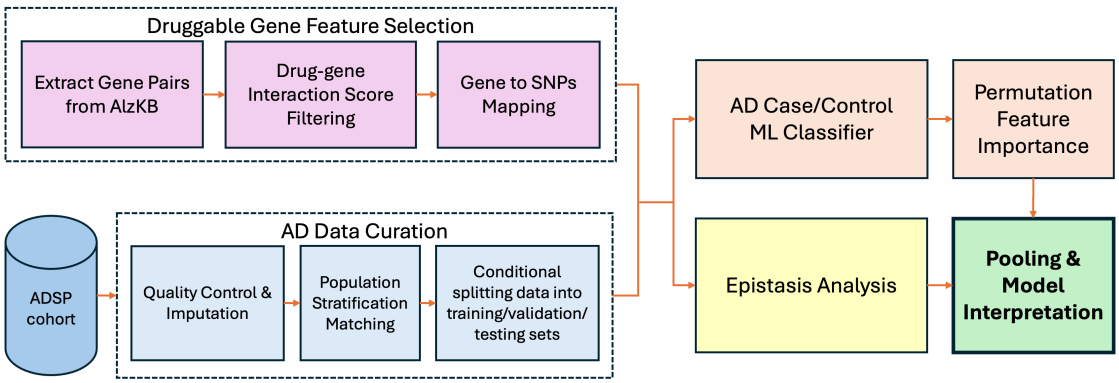


Fig. 1. Flowchart of overall study design.

2.1. *Druggable Gene Feature Selection using AlzKB and DGIdb*

The underlying hypothesis of this study is that nominated gene targets for identification of therapeutic targets for AD can be obtained from the search space of the non-AD genes (i.e. genes not currently known to be implicated for AD) that exhibit a gene-gene interaction with known AD genes in the AlzKB and have at least one drug target. The drug target condition is defined by the edges "*chemical_binds_gene*", "*chemical_upregulates_gene*" and "*chem-*

ical_downregulates_gene". (Note that in this work, AD genes imply all *protein-coding* genes directly linked to the "Alzheimer's Disease" node in AlzKB.) The "*gene_interacts_with_gene*" edges in AlzKB are based on protein-protein interactions.[23] 82 AD genes were connected to 1,805 non-AD protein-coding genes with drug targets, resulting in a total of 2,835 gene pairs. The 82 AD genes served as the baseline model gene list for the subsequent ML model analysis.

2.1.1. *Priority Druggable Gene Selection based on Drug-Gene Interaction Score*

To further prioritize the list of clinically relevant gene selection derived from the AlzKB, we define an additional gene druggability criteria based on the interaction score metric from DGIdb.[14] DGIdb is one of the most comprehensive resources incorporating knowledge about genomic modifications, diseases, and therapeutic targets.[14] The database utilizes experts curation and text-mining of an extensive list of over 40 drug, gene, and interaction sources to extract and rank drug-gene interactions. The interaction score is used to rank the significance and relevance of interactions between drugs and genes. It is calculated based on evidence strength (i.e., the strength of the evidence supporting the interaction from various sources), source credibility, interaction type, the number of supporting sources, and disease relevance. Hence, we retained (AD gene, non-AD gene) pairs from the AlzKB subset, if and only if, the maximum value of the interaction score of the non-AD gene exceeded a 75th percentile threshold (i.e., 11.76 in this work). This yielded a final set of 44 AD genes interacting with 181 non-AD genes, a total of 285 gene pairs (see Figure 2(a) in Results section).

2.1.2. *Gene to SNPs Mapping*

Ensembl REST API is used to obtain the GRCh38 coordinates for the coding regions of each gene.[24] For each gene of interest, we extract all SNPs located within the regulatory regions (100kb upstream and 5kb downstream). The baseline model feature set consists of the union of all the SNPs mapped to each of the 82 AD genes. For the (AD, non-AD) gene-gene interaction datasets, the SNPs feature set is mapped per AD gene, i.e., the SNPs of the AD gene along with all SNPs belonging to each of the non-AD gene interacting with that AD gene.

2.2. *AD Data Sample and Curation*

The AD genotype data utilized in this study is drawn from the Alzheimer's Disease Sequencing Project (ADSP).[25] The ADSP aims to identify genetic variants that influence the risk of AD by sequencing the genomes of individuals (from ethnically diverse populations), focusing primarily on AD case/control phenotypes derived from clinical data. The study sample was extracted from the ADSP R4 v11 2023 release VCF dataset which originally had 346,763,200 variants and 36361 samples.

Pre-filtering quality control was done at two levels: at the variant level, based on sequencing statistics, and at the sample level, based on duplicate samples. To ensure the reliability of genetic analyses and focus on more impactful genetic variants, additional filtering steps were performed to remove singletons and exceedingly rare variants. Singletons imply variants present in only one individual, thus less likely to be relevant to the disease. For exceedingly

rare variants, the total number of counted alleles is very small relative to the number of samples. Thus, subsequent analysis focus on variants with enough occurrences to allow for meaningful statistical analysis. Variants with low call rates (missing call rate > 0.01) and samples with poor genotyping rates (missing call rate > 0.05) were also excluded. Only common variants (minor allele frequency (MAF) >1%) were retained resulting in a final variants count of 9,520,653 and 34971 samples. Imputation of missing values was done using mode-based imputation to avoid false positive signals as a small set of 400 variants had almost no homozygous calls.

2.2.1. *Population Stratification using Propensity Score Matching*

The ADSP R4 v11 2023 release spans 40 study cohorts made up of 5,218 subjects of African ancestry, 2,791 of Asian ancestry, 10,398 of Hispanic ancestry, and 16,191 Non-Hispanic White. Thus, another key consideration of the genomic data preprocessing is to insure that any bias due to population stratification is mitigated before quantitative analysis. Population stratification (PS) refers to the presence of systematic differences in allele frequencies between subpopulations in a population due to different ancestries. These differences can confound genetic association studies if not properly accounted for, leading to false associations or masking true associations between genetic variants and diseases.[26] A commonly used method to address PS is principal components analysis (PCA). This approach uses genotype data (independent loci) to compute the principal components, which are assumed to represent features of genomic ancestry that capture PS. The principal components are then used as covariates in subsequent analyses. However for complex ML analysis, usage of covariates is not applicable. To control for PS in this study, we developed a novel method that adjusts the dataset for ancestral heterogeneity by performing propensity score matching (PSM) on genomic PCA.

To obtain the PCA of the independent genomic loci, we extract a subset of the data based on these parameters: MAF > 0.02, Hardy-Weinberg Equilibrium (HWE) exact test p-value $>$ 1e-7, Linkage Disequilibrium (LD) with a variant window count of 100, a step size 10, and R2 cutoff of 0.1. Subsequently, we apply the PSM procedure using the top eight principal components derived from the PCA computation. The PSM conducts a logistic regression on the 8 PCA covariates to compute the propensity score. The matching is performed using *psmpy* package.[27] A key novelty of the matching process is that it ensures that the individual from the control subset has its closest counterpart in the disease subset based on the computed propensity score using k nearest neighbors matching. The final matched dataset (see Table 1) obtained had 22560 samples equally distributed between AD case and control phenotypes.

Table 1. Demographic summary of cases and controls in the final matched dataset

	Female (%)	Harmonized Age*	Race (%)						Ethnicity (%)		
			White	AA	Asian	Native/Amer. Ind.	Other	N/A	Hispanic/Latino	Non-Hispanic	N/A
Cases	60.48	33 to 90+	66.55	13.95	1.61	0.41	14.73	2.75	26.45	69.34	4.21
Controls	66.33	30 to 90+	47.81	25.61	1.25	0.33	16.11	8.89	36.54	61.99	1.46

*__Harmonized Age__: age at onset for cases, and age at last exam for controls (Age values of 90 or more are coded as "90+").
__Race__: uses NIH Racial Categories. AA denotes Black or African American, Native/Amer. Ind. denotes Native Americans and American Indian/Alaska Native.

2.2.2. *Conditional Splits of AD Data for Robust ML Analysis*

The last phase of the AD data curation involved an intentional split of the derived matched data so that the key fairness characteristics (mitigating population stratification) are not lost during the ML phase. Building a robust ML classifier model requires training and validation datasets as well as a test hold out set, that is not seen by the model during the training and validation phase, to ensure model generalization.[28] The two conditions that had to be preserved and consistent across the splits into three datasets were: (i) Matched case/control pairs and propensity score distributions, (ii) Distributions of significant SNPs reported by recent GWAS studies. The set of significant SNPs is based on the 2023 Lancet meta review[4] studies that listed 101 unique SNPs with a significance threshold p-value < 5e-8. The variant filtering for the matched dataset was based on these parameters: MAF > 0.1, HWE exact test p-value > 1e-7, LD with a variant window count of 100, a step size 10, and R2 cutoff of 0.8. Note that after the filtering phase, only 30 of the 101 SNPs were present in the matched data.

We designed an optimization algorithm using the Optuna platform[29] that satisfied the specified conditions for dataset splits. This entailed running 1000 Optuna trials by sampling different random seeds for training, validation and testing sets splits in equal ratio 1/3:1/3:1/3 for the matched case/control sample pairs. During each trial, 2 objectives with equal weights were evaluated: (i) maximizing the median –log(p) of 30 SNP set across splits; (ii) minimizing the absolute difference between the median –log(p) of 30 SNP. At the end of the optimization procedure, the best trial datasets were selected as the training/validation and test sets for subsequent analyses.

2.3. *ML feature selection and Epistasis Analysis*

2.3.1. *ML AD Case/Control Classifier and Feature Importance*

To identify genomic biomarkers that may indicate potential gene-drug targets, we assessed their predictive power in constructing a ML model for AD case vs control classification using the ADSP matched data. We performed 44 experiments for the gene-gene interaction sets using its corresponding SNPs PPI input feature set. Let $G_{AD} = g_{AD_1}, g_{AD_2}, ..., g_{AD_l}$ be the set of AD genes, where l=44. $G_{nonAD} = g_{AD_1}, g_{AD_2}, ..., g_{AD_m}$ be the set of non-AD genes, where m=181. $I \subseteq G_{AD} \times G_{nonAD}$ denotes the set of interacting (AD, non-AD) gene pairs; $I = \{(g_{\mathrm{AD}_i}, g_{\mathrm{nonAD}_j}) \mid g_{\mathrm{AD}_i} \in G_{\mathrm{AD}}, g_{\mathrm{nonAD}_j} \in G_{\mathrm{nonAD}}\}$, where $|I| = 285$. Let SNP(g) be the set of SNPs for gene g. For each AD gene $g_{AD_i} \in G_{AD}$, its SNPs PPI feature set is the union of g_{AD_i} and the SNPs of all non-AD genes g_{nonAD_j} that interact with g_{AD_i}, $\mathrm{Input\_SNPs}(g_{\mathrm{AD}_i}) = \mathrm{SNP}(g_{\mathrm{AD}_i}) \cup \bigcup_{(g_{\mathrm{AD}_i}, g_{\mathrm{nonAD}_j}) \in A} \mathrm{SNP}(g_{\mathrm{nonAD}_j})$. This is the input data for each of the 44 AD gene PPI experiments. The baseline performance was determined by building the AD case/control with the SNPs derived from all the 82 AD genes, G^{+}_{AD} (see Section 2.1). $\mathrm{Baseline\_SNPs} = \bigcup_{g_{AD} \in G^{+}_{\mathrm{AD}}} \mathrm{SNP}(g_{AD})$.

The AD case/control classifier model was implemented using both automated ML with tree-based pipeline optimizer 2 (TPOT2[30]) platform and the Extreme Gradient Boosting (XGBoost) algorithm.[31] TPOT2 allows for the selections of the best-performing ML model for a given problem in an agnostic manner. The classification pipelines are generated from the sub-

set of ML methods and data pre-processing operators imported from the scikit-learn Python library. During the optimization process, various combinations of pre-processing operator are combined with ML methods into a pipeline in a tree-based manner. XGBoost is a tree-based model implementation of the gradient boosting framework, which combines the predictions of multiple weak learners (usually decision trees) to produce a strong overall model. For fair comparison of both models, the hyperparameters tuning for Xgboost was performed using Optuna method[29] with the objective function set to maximize Receiver Operating Characteristic - Area Under the Curve (ROC AUC) metric across hyperparameters search space of the TPOT2 configuration. For both methods, the training and validation datasets were used for tuning and optimization, and once the final model and hyperparameter set was determined, the final performance was evaluated using testing set.

To identify which variants were driving the predictive power of each model, we performed permutation feature importance (PFI) to compute the univariate contribution of each variant (feature). Note that the PFI coefficient value, which estimates the main effect of each SNP, was calculated exclusively using the testing dataset.

2.3.2. *Epistasis Analysis*

The aim of the epistasis analysis was to compute the level of strength of interactions between SNPs that contribute to the disease, rather than individual SNPs or the additive effects of SNP subsets. Exhaustive searches of epistatic interactions are computationally expensive due to the high dimensionality of genomic datasets, We computed the epistatic interaction using BitEpi, a parallelized bitwise algorithm, which allowed for fast, exhaustive computation of higher-order interactions between SNPs.[32] The genotypes are encoded in bytes (8-bits) with the first 2 bits denoting the combination (e.g. $0/0 \rightarrow 00$, $0/1 \rightarrow 01$) and the remainder bits set as 0. Bitwise operations are subsequently applied to combine genotypes of up to 4 SNPs to create contingency tables and compute the entropy-based metrics (association power (β) and interaction effect size (α)). The β metric reflects the combined association power of the SNPs considered, while the α metric indicates the gain in association power due to the epistatic effect of those SNPs. The α (also known as information gain) metric quantifies the level of strength of interaction of the SNP sets. For all $SNPs \in I$ gene pairs, we computed α for each individual SNP (18778 variants) and its two way interactions (176,297,253 SNP pairs) using the matched testing dataset.

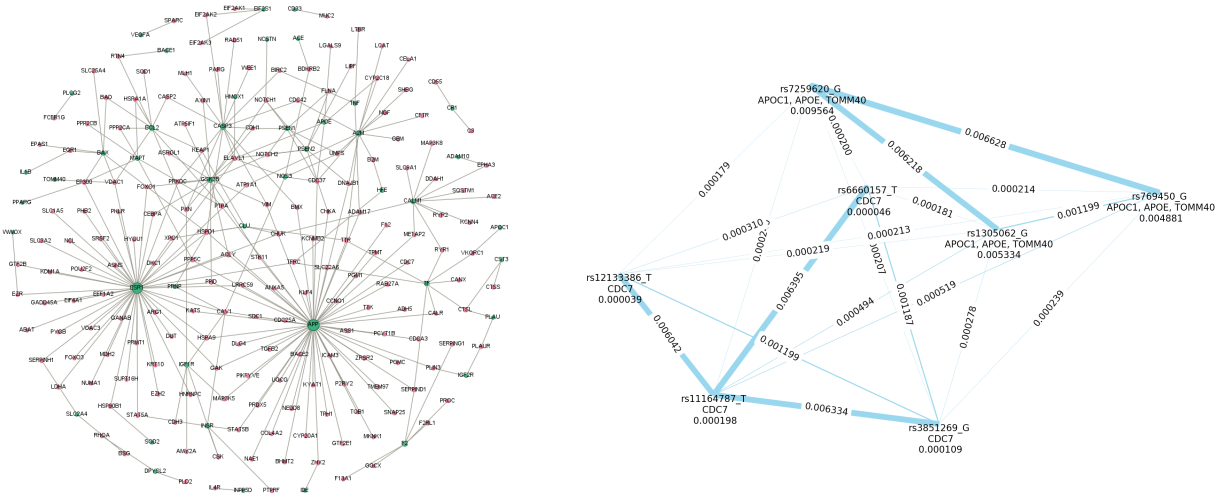
2.3.3. *Pooling & Model Interpretation*

The final phase of the learning framework pools the results obtained from both the epistatic interaction analysis and the ML feature selection to determine a final set of potential genomic biomarkers for AD novel gene-drug targets. Though the ML classification models are able to assess the predictive power of a set of SNP variants to distinguish AD case from control phenotype, it may fall short in the explainability phase of the key drivers. Feature importance scores, as quantified by PFI coefficient scores, are based on main effects of each feature (SNP in this case). The goal is to extract the list of top ranking SNPs exhibiting relatively strong level of interactions and assess their predictive power in distinguishing AD case from control

phenotypes. This will provide additional evidence of their effect on the ML classification models for each AD gene and its set of interacting non-AD gene pairs.

3. Results and Analysis

Figure 2(a) illustrates the outcome of druggable gene feature selection phase which yielded 285 $(g_{\mathrm{AD}_i}, g_{\mathrm{nonAD}_j})$ pairs.



(a) Network of 285 (AD, non-AD) gene pairs (*gene_interacts_with_gene*) filtered by 75th percentile drug interaction score from DGIdb. Green dots denote the 44 AD genes while red, the 181 non-AD genes.

(b) Visualization of network of top 5 epistatic interactions, as quantified by information gain using α values for the main effect and 2-way interactions from BitEPI in ADSP dataset.

Fig. 2. Visualization of (a) 285 (AD, non-AD) gene pairs, (b) SNPs that exhibit high epistatic interactions.

Table 2 illustrates the performance of all the AD case/control classification experiments for both XGBoost and TPOT2 models. Out of the 44 experiments conducted for each of the AD genes (and its set of interacting non-AD genes), only the top 15 best performing experiments (based on XGBoost ROC AUC) are listed in Table 2. See Table S1 in Supplementary file [a] for complete details of all experimental results. From Table 2, we observe that the baseline model (union of all SNPs from 82 AD genes) outperformed all the other models, as expected, with 64.23% for TPOT2 model and 63.65% for XGBoost. While the TPOT2 models seemed to have the better ROC AUC performance overall, the pipelines were very complex and hard to interpret. Hence, based on complexity/performance trade-off, we selected the XGBoost models for further evaluation of the individual contribution of each SNP variant to overall model performance. Among the best performing models, we observe that *APOC1*, *APOE*, and *TOMM40*

[a]Supplementary information is available at: `https://github.com/EpistasisLab/PSB25_ADSP_GIG`

Table 2. AD case/control classification (XGBoost vs TPOT2) performance outcomes for baseline (all AD genes) model, top 15 gene-gene interaction sets, and subset of non-AD genes SNPs that exhibit strong epistatic interaction. Gene-gene interaction sets are sorted by ROC AUC scores from the final XGBoost model.

Gene set	**# non-AD genes**	**# SNPs**	**ROC AUC**		**Recall**		**Precision**		**Accuracy**	
			XGB	TPOT2	XGB	TPOT2	XGB	TPOT2	XGB	TPOT2
All AD genes	-	9539	63.65	64.23	57.53	58.38	59.70	60.17	59.35	59.87
APOC1	1	57	63.15	64.13	53.16	53.94	60.96	61.98	59.56	60.43
APOE	4	389	63.05	64.10	55.53	52.74	60.50	62.10	59.64	60.28
TOMM40	1	141	62.79	64.18	53.40	52.15	60.30	61.98	59.12	60.08
ESR1	53	3399	60.03	52.28	64.63	54.10	56.95	50.05	57.89	50.05
GSK3B	15	1199	59.59	46.17	62.95	0.00	56.60	0.00	57.34	50.00
APP	64	5256	59.29	49.15	64.04	54.10	56.95	50.05	57.82	50.05
CASP3	15	884	59.04	59.80	62.31	70.59	56.39	56.15	57.06	57.73
DPYSL2	2	269	59.02	59.19	63.38	64.92	55.98	56.21	56.77	57.17
BCL2	10	558	58.98	58.94	62.71	70.72	56.18	55.07	56.90	56.52
A2M	14	1112	58.67	59.39	59.76	61.99	56.33	56.70	56.72	57.33
BAX	7	589	58.53	59.04	61.46	45.96	55.53	50.00	56.12	50.00
WWOX	1	1320	58.46	58.67	63.54	64.12	55.79	56.31	56.60	57.18
INSR	5	468	58.43	59.35	61.89	61.09	55.82	56.77	56.45	57.29
CALM1	11	1878	58.42	59.82	61.14	71.14	55.57	55.28	56.13	56.80
TF	8	593	58.36	59.82	61.81	62.23	55.69	57.51	56.32	58.13
non-AD genes*	104	1867	60.06	49.94	65.21	53.67	57.05	50.01	58.06	50.01
High α SNPs†	6	56	62.92	63.65	60.90	59.15	59.19	60.01	59.45	59.87

**non-AD genes SNPs set selected based on $\alpha > 0.003$ from BitEpi.*

†*High α SNPs selected based on top 50 α from BitEpi.*

PPI gene sets had relatively high performance, though the number of their corresponding interacting non-AD genes SNPs was very small. This suggests that model performance could be attributed mainly to SNP(g_{AD}). The gene sets for *ESR1*, *GSK3B*, *APP*, and *CASP3* had a larger number of interacting non-AD genes (and SNPs) and performed relatively well (ROC AUC of $59 - 60\%$).

Figure 3 reports the PFI values for each top performing gene sets based on the XGBoost models. For *APOC1*, *APOE* and *TOMM40* gene sets, SNP rs7259620G has the largest main effect followed by rs769450G and rs449647A, with PFI values of 0.084, 0.057, 0.09 respectively. The rs769450G is a known intronic variant associated with AD risk[33] . There are limited studies demonstrating the association of rs7259620G with AD risk.[33,34] However, the rs449647A (*TOMM40* intronic variant) currently has no GWAS, functional or clinical annotation available. For the remaining 12 gene sets, none of the SNPs exhibited an informative contribution of significant value, with all contributions being less than 1%. Though the performance of these models is 58 - 60% ROC AUC which is relatively close in performance to the top models with quantifiable independent effects (*APOC1*, *APOE*, *TOMM40*). This suggests that the driver for the model performance is likely due to the interaction effect of its variants.

Figure 4(a) presents the PFI values for the baseline (SNPs(G^{+}_{AD})) model. The SNP with the highest PFI score, by a large margin, is the same set of three SNPs from the *APOC1*, *APOE* and *TOMM40* gene sets (Fig. 3). This provides additional evidence that the key driver of model performance for those gene sets were most likely due to the already know AD risk

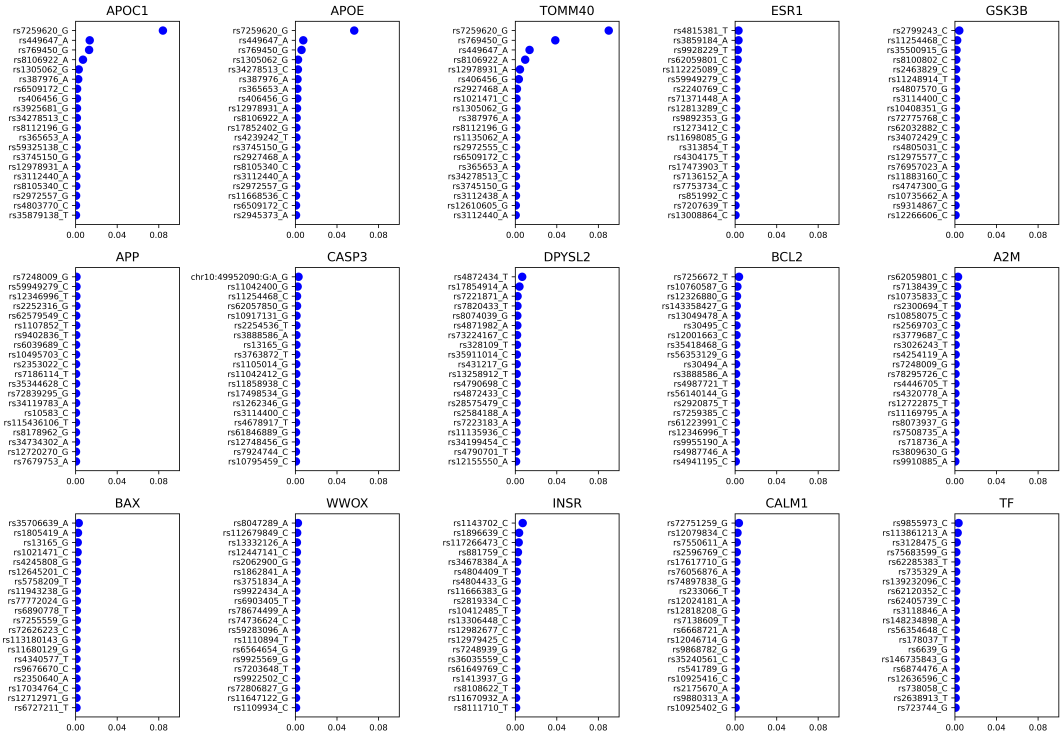
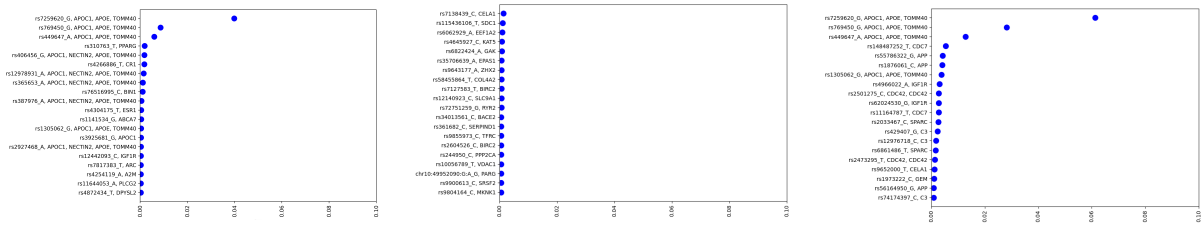


Fig. 3. Permutation feature importance scores for top 15 gene sets from XGBoost model.



(a) Baseline model (all AD) (b) Non-AD SNPs ($\alpha > 0.003$) (c) High α SNPs (Top 50)

Fig. 4. Comparison of XGBoost permutation feature importance scores for selected SNPs sets

genomic variants.

The epistasis interaction analysis outcome is presented in Table 3 for the top 15 SNPs ranked by the information gain of its two-way interaction (α). (see Table S2 in supplementary file for complete list [b]). The effect size of the top two-way SNP combinations, while slightly smaller, was comparable to the top individual SNP effect size indicating that non-additivity could be a contributing factor in explaining AD genomic mechanism. From Table 3, we observe that the strongest pair of interacting SNP variants with α=0.0066 was (rs7259620G, rs769450G). These two SNPs also had largest main effects (α=0.0096, 0.0053) and were two of SNPs driving ML performance for the top 3 gene sets (*APOC1*, *APOE* and

[b]Supplementary information is available at: `https://github.com/EpistasisLab/PSB25_ADSP_GIG`

TOMM40) and baseline model (see Figs. 3, 4(c) and4(a)) (see Table S3 in supplementary file for complete list of PFI values [b]). There were also SNPs from non-AD genes that had strong interaction values: (rs6660157T, rs11164787T $\alpha = 0.0064$), (rs3851269G, rs11164787T $\alpha = 0.0062$), and (rs12133386T, rs11164787T $\alpha = 0.006$). These were all affiliated with the non-coding region of the *CDC7* gene. A visualization of the network of the 7 SNP variants involved in the top 5 interactions (see Figure 2(b)) reveals two groups of interactions: *CDC7* non-coding region, and *APOE* intronic with non-coding *TOMM40* intronic regions. The visualization also reveals slightly weaker interactions (α =0.0012) present for *CDC7* non-coding (rs12133386T, rs3851269G) and (rs3851269G, rs6660157T), and for *TOMM40*/*APOE* (rs1305062G, rs769450G).

Table 3. Top 15 epistatic interaction results from BitEpi

SNP_A	SNP_B	α_{AB}	α_A	α_B	β_{AB}	β_A	β_B	Gene_A	Gene_B
rs7259620_G	rs769450_G	0.0066	0.0096	0.0048	0.5162	0.5096	0.5049	*APOC1*, *APOE*, *TOMM40*	*APOC1*, *APOE*, *TOMM40*
rs6660157_T	rs11164787_T	0.0064	4.60E-05	0.0002	0.5066	0.5000	0.5002	*CDC7*	*CDC7*
rs3851269_G	rs11164787_T	0.0063	0.0001	0.0002	0.5065	0.5001	0.5002	*CDC7*	*CDC7*
rs1305062_G	rs7259620_G	0.0062	0.0053	0.0096	0.5158	0.5053	0.5096	*APOC1*, *APOE*, *TOMM40*	*APOC1*, *APOE*, *TOMM40*
rs12133386_T	rs11164787_T	0.0060	3.90E-05	0.0002	0.5062	0.5000	0.5002	*CDC7*	*CDC7*
rs11166498_G	rs11164787_T	0.0060	6.00E-05	0.0002	0.5062	0.5000	0.5002	*CDC7*	*CDC7*
rs12816187_A	rs9652000_T	0.0060	0.0001	0.0004	0.5064	0.5001	0.5004	CELA1	CELA1
rs1305062_G	rs449647_A	0.0059	0.0053	0.0036	0.5113	0.5053	0.5036	*APOC1*, *APOE*, *TOMM40*	*APOC1*, *APOE*, *TOMM40*
rs2473295_T	rs2501275_C	0.0059	0.0006	0.0002	0.5065	0.5006	0.5002	*CDC42*	*CDC42*
rs2473296_C	rs2501275_C	0.0059	0.0005	0.0002	0.5065	0.5005	0.5002	*CDC42*	*CDC42*
rs7529485_C	rs11164787_T	0.0059	2.70E-05	0.0002	0.5061	0.5000	0.5002	*CDC7*	*CDC7*
rs1883421_C	rs2501275_C	0.0059	0.0005	0.0002	0.5064	0.5005	0.5002	*CDC42*	*CDC42*
rs12116952_G	rs2501275_C	0.0059	0.0006	0.0002	0.5065	0.5006	0.5002	*CDC42*	*CDC42*
rs1063116_A	rs2501275_C	0.0059	0.0005	0.0002	0.5064	0.5005	0.5002	*CDC42*	*CDC42*
rs2501291_G	rs2501275_C	0.0059	0.0006	0.0002	0.5065	0.5006	0.5002	*CDC42*	*CDC42*

Figure 5 illustrates the estimated AD distribution for using contingency table plots for selected 2-way interactions with large α values. These plots display the number of samples for each genotype combinations for the selected SNPs in both case and control cohorts. The plots for the (rs7259620G, rs769450G) pair have substantially increased AD rate when both SNP are homozygous for the alternative allele. More complex associations were observed for (rs6660157T, rs11164787T) pair. Increased AD risk is observed when rs11164787T is homozygous for the reference allele and rs11164787 is homozygous for the alternative allele. When both rs11164787T and rs6660157T are heterozygous, and when rs11164787T is homozygous for the alternative allele and rs11164787 is homozygous for the reference allele. Select genotypes combinations for non-coding region of *CDC42* also associated with increased risk of AD: when both rs760923G and rs760923G are homozygous for the reference allele, when both rs760923G and rs760923G are heterozygous, and when both rs760923G and rs760923G homozygous for the alternative allele.

When the SNPs from non-AD genes of meaningful α values from epistasis analysis are pooled to build the AD case/control classifier, it yields a comparable performance (60.06% ROC AUC) for the XGBoost model (see last row in Table 2). However, the PFI result (Figure 4(b)) which quantifies the univariate contribution of each SNP ($< 2\%$) to model performance still fall shorts in explainability of model performance. This provides additional evidence that the key contributors for model performance is beyond univariate contributions of these SNPs. The non-additive effects of these SNPs could be a factor.

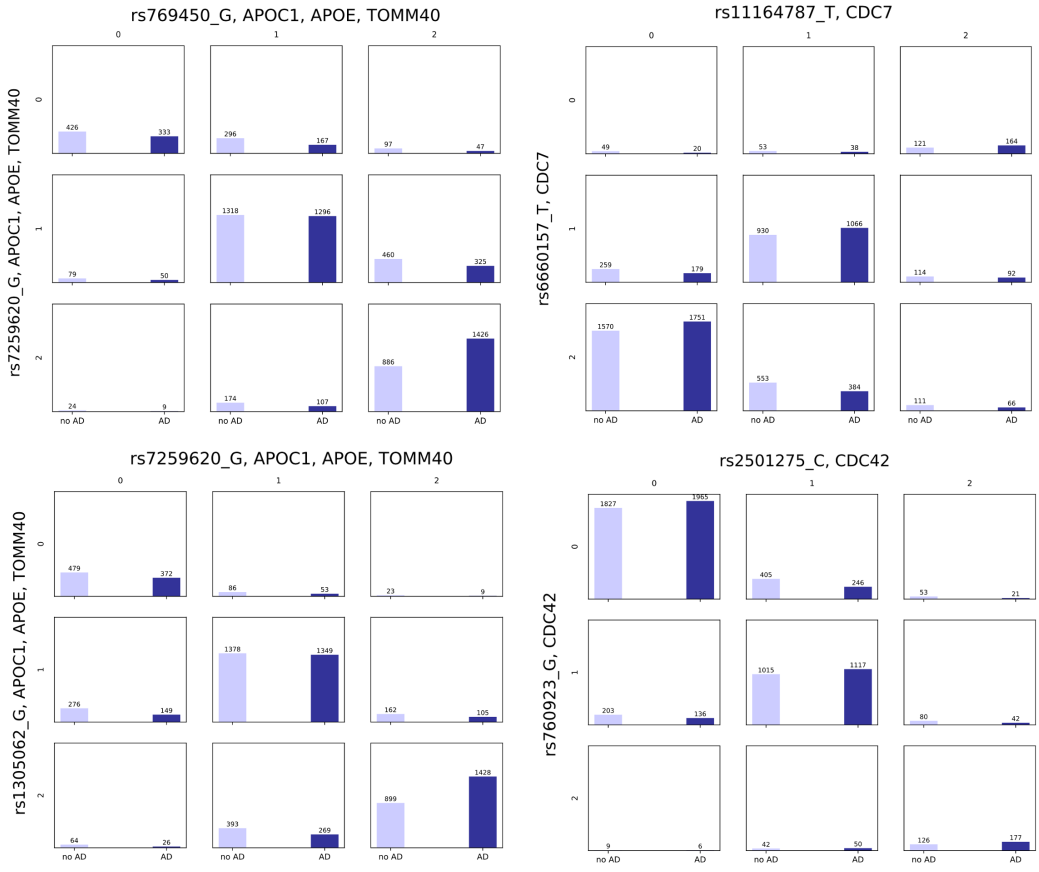


Fig. 5. Genotypes combinations for selected SNP pairs with high information gain (α values).

4. Discussion

In this study, we have implemented an ML SNPs feature selection model that integrates epistatic interaction and leverages the Alzheimer's knowledge base (AlzKB), Drug-Gene interaction database (DGIdb) to identify a list of biologically plausible novel gene-drug targets for further investigation. The prior biological knowledge of gene-gene interactions in AlzKB is based on protein-protein interactions.[23] The model is validated using an ethnically diverse study sample obtained from the Alzheimer's Disease Sequencing Project (ADSP). A primary goal of the ADSP is to further the understanding of the genetic architecture of AD and related dementias and subsequently, turn genetic findings into meaningful therapeutic targets.[25] Given the complexity of the dataset fueled by the multiple ancestry in the sample population, ML analysis directly applied could yield spurious associations that are not related to the disease mechanism but possible ancestry differences. Hence, a key contribution of this work is the extensive novel preprocessing steps applied on the AD case/control ADSP genomic data to mitigate of population stratification. We applied a novel method that combines PCA with propensity score matching. The mitigation of system bias was validated by computing the genomic inflation factor using an external GWS study analysis.

The robustness and generalization of the ML model outcomes is enforced by the conducting conditional splits of the datasets into training, validation and testing sets such that the

matching benefits are not compromised. Conducting the model performance evaluation, feature selection, and epistasis analysis exclusively on the test set, increases the confidence in the generalization and reproduciblity of the results obtained. We utilized two ML methods: TPOT2, an automated ML tool that explores multiple classification algorithms using genetic algorithm and selects the most optimal, and XGBoost, a scalable and highly effective tree-boosting algorithm. Though the TPOT2 performed better overall, subsequent analysis was done based on the XGBoost models, as they were relatively simple and efficient compared to the complexity of the TPOT2 pipelines. Permutation analysis of top models revealed that some SNPs of known AD risk genes are drivers of the performance. Specifically, for the best performing models (all AD genes, *APOE*, *APOC1*, and *TOMM40*) the most informative variant is rs7259620G located 2KB upstream from the *APOE* gene region. GWAS with 17,480 European individuals found an association of the *APOE* rs7259620 G allele with increased AD risk (OR=1.68, p=2x10-23[33]). The second high ranking SNP (rs769450G) is a common intronic variant associated with AD risk. Several large GWAS have also found highly significant associations with various traits including AD.[33] For the rs449647A, a *TOMM40* intronic variant, there was no GWAS, functional or clinical annotation available. Understanding the functional implications of rs449647A could potentially shed light on its contribution to disease risk or progression. However, for other top performing models, the PFI analysis could not quantify any SNP has having substantial univariate contribution to explain model performance.

To identify whether the genotype combinations can better explain phenotype variance in these experiments, we ran an exhaustive pairwise epistasis analysis with BitEpi, a highly scalable and efficient method. Among the combinations of genotypes with strong informative contributions are previously identified SNPs from intronic and noncoding upstream regions of *APOE* and *TOMM40*, and novel SNPs from noncoding regions of *CDC7* (linked to *APP* through gene-gene interactions in AlzKb) and *CDC42* (linked to A2M through gene-gene interactions in AlzKb) (see Table 3, Fig. 5). The non-AD gene SNPs haven't been previously reported in GWAS studies and do not have any functional or clinically relevant affiliations with AD. Epistasis analysis uncovered some novel SNPs, not related to the AD genes, which when pooled into the ML analysis demonstrated comparable predictive power to baseline all AD genes model (see Table 2, Figure 4(b)). This suggests a biological plausible set of genes for further investigation as potential drug-target genes for AD.

This work highlights the limitations of basic ML model interpretation methods, which tend to focus solely on main effects while overlooking impact of epistatic interactions that may contribute to model performance. Evaluating model-based 2- and 3- way PFI is an exhaustive procedure and not scalable for high-dimensional genomic data. We propose that integrating ML analysis with epistasis detection could address this challenge and facilitate advancements in uncovering disease mechanisms and identifying potential therapeutic targets."

Acknowledgements

We wish to thank Yuki Bradford and Rachit Kumar for their help with the data imputation process.

References

1. B. Qorri, M. Tsay, A. Agrawal, R. Au and J. Geraci, Using machine intelligence to uncover alzheimers disease progression heterogeneity, *Exploration of Medicine* **1**, 377 (2020).
2. Y. Lagisetty, T. Bourquard, I. Al-Ramahi, C. G. Mangleburg, S. Mota, S. Soleimani, J. M. Shulman, J. Botas, K. Lee and O. Lichtarge, Identification of risk genes for alzheimer's disease by gene embedding, *Cell genomics* **2** (2022).
3. J. Pleen and R. Townley, Alzheimer's disease clinical trial update 2019–2021, *Journal of Neurology* **269**, 1038 (2022).
4. S. J. Andrews, A. E. Renton, B. Fulton-Howard, A. Podlesny-Drabiniok, E. Marcora and A. M. Goate, The complex genetic architecture of alzheimer's disease: novel insights and future directions, *EBioMedicine* **90** (2023).
5. S. Grueso and R. Viejo-Sobera, Machine learning methods for predicting progression from mild cognitive impairment to alzheimer's disease dementia: a systematic review, *Alzheimer's research & therapy* **13**, 1 (2021).
6. A. Nandi, N. Counts, J. Bröker, S. Malik, S. Chen, R. Han, J. Klusty, B. Seligman, D. Tortorice, D. Vigo *et al.*, Cost of care for alzheimer's disease and related dementias in the united states: 2016 to 2060, *npj Aging* **10**, p. 13 (2024).
7. J. D. Romano, V. Truong, R. Kumar, M. Venkatesan, B. E. Graham, Y. Hao, N. Matsumoto, X. Li, Z. Wang, M. D. Ritchie *et al.*, The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research, *Journal of Medical Internet Research* **26**, p. e46777 (2024).
8. W. K. Self and D. M. Holtzman, Emerging diagnostics and therapeutics for alzheimer disease, *Nature medicine* **29**, 2187 (2023).
9. H. W. Haddad, G. W. Malone, N. J. Comardelle, A. E. Degueure, S. Poliwoda, R. J. Kaye, K. S. Murnane, A. M. Kaye and A. D. Kaye, Aduhelm, a novel anti-amyloid monoclonal antibody, for the treatment of alzheimer's disease: A comprehensive review, *Health Psychology Research* **10** (2022).
10. A. Jiménez, M. J. Merino, J. Parras and S. Zazo, Explainable drug repurposing via path based knowledge graph completion, *Scientific Reports* **14**, p. 16587 (2024).
11. A. Renaux, C. Terwagne, M. Cochez, I. Tiddi, A. Nowé and T. Lenaerts, A knowledge graph approach to predict and interpret disease-causing gene interactions, *BMC bioinformatics* **24**, p. 324 (2023).
12. S. Jin, X. Zeng, F. Xia, W. Huang and X. Liu, Application of deep learning methods in biological networks, *Briefings in bioinformatics* **22**, 1902 (2021).
13. C. X. Alvarado, M. B. Makarious, C. A. Weller, D. Vitale, M. J. Koretsky, S. Bandres-Ciga, H. Iwaki, K. Levine, A. Singleton, F. Faghri *et al.*, omicsynth: An open multi-omic community resource for identifying druggable targets across neurodegenerative diseases, *The American Journal of Human Genetics* **111**, 150 (2024).
14. M. Cannon, J. Stevenson, K. Stahl, R. Basu, A. Coffman, S. Kiwala, J. F. McMichael, K. Kuzma, D. Morrissey, K. Cotto *et al.*, DGIdb 5.0: rebuilding the drug–gene interaction database for precision medicine and drug discovery platforms, *Nucleic acids research* **52**, D1227 (2024).
15. J. I. Castrillo, S. Lista, H. Hampel and C. W. Ritchie, Systems biology methods for alzheimer's disease research toward molecular signatures, subtypes, and stages and precision medicine: application in cohort studies and trials, *Biomarkers for Alzheimer's Disease Drug Development* , 31 (2018).
16. J. Schwartzentruber, S. Cooper, J. Z. Liu, I. Barrio-Hernandez, E. Bello, N. Kumasaka, A. M. Young, R. J. Franklin, T. Johnson, K. Estrada *et al.*, Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new alzheimer's disease risk genes, *Nature genetics* **53**, 392 (2021).

17. J. H. Moore, The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases, *Human Heredity* **56**, 73 (11 2003).
18. T. F. Mackay and J. H. Moore, Why epistasis is important for tackling complex human disease genetics, *Genome Med.* **6**, p. 124 (June 2014).
19. J. H. Moore and S. M. Williams, Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis, *BioEssays* **27**, 637 (2005).
20. T. J. Hohman, W. S. Bush, L. Jiang, K. D. Brown-Gentry, E. S. Torstenson, S. M. Dudek, S. Mukherjee, A. Naj, B. W. Kunkle, M. D. Ritchie, E. R. Martin, G. D. Schellenberg, R. Mayeux, L. A. Farrer, M. A. Pericak-Vance, J. L. Haines and T. A. Thornton-Wells, Discovery of gene-gene interactions across multiple independent data sets of late onset alzheimer disease from the alzheimer disease genetics consortium, *Neurobiology of Aging* **38**, 141 (2016).
21. Q. Sha, Z. Zhang, J. C. Schymick, B. J. Traynor and S. Zhang, Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis, *BMC Med. Genet.* **10**, p. 86 (September 2009).
22. M. Steffens, T. Becker, T. Sander, R. Fimmers, C. Herold, D. A. Holler, C. Leu, S. Herms, S. Cichon, B. Bohn, T. Gerstner, M. Griebel, M. M. Nöthen, T. F. Wienker and M. P. Baur, Feasible and successful: genome-wide interaction analysis involving all 1.9 x 10(11) pair-wise interaction tests, *Hum. Hered.* **69**, 268 (March 2010).
23. D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian and S. E. Baranzini, Systematic integration of biomedical knowledge prioritizes drugs for repurposing, *Elife* **6**, p. e26726 (2017).
24. F. J. Martin, M. R. Amode, A. Aneja, O. Austine-Orimoloye, A. G. Azov, I. Barnes, A. Becker, R. Bennett, A. Berry, J. Bhai *et al.*, Ensembl 2023, *Nucleic acids research* **51**, D933 (2023).
25. Y. Y. Leung, W.-P. Lee, A. B. Kuzma, P. Gangadharan, H. I. Nicaretta, L. Qu, Y. Ren, L. B. Cantwell, O. Valladares, Y. Zhao *et al.*, Adsp whole genome sequencing (wgs) release 4 data update from genome center for alzheimer's disease, *Alzheimer's & Dementia* **19**, p. e077351 (2023).
26. J. N. Hellwege, J. M. Keaton, A. Giri, X. Gao, D. R. Velez Edwards and T. L. Edwards, Population stratification in genetic association studies, *Curr. Protoc. Hum. Genet.* **95**, 1.22.1 (October 2017).
27. A. Kline and Y. Luo, Psmpy: a package for retrospective cohort matching in python, in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022.
28. K. Al-Jabery, T. Obafemi-Ajayi, G. Olbricht and D. Wunsch, *"Computational Learning Approaches to Data Analytics in Biomedical Applications"* (Academic Press, 2019).
29. T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019.
30. P. Ribeiro, A. Saini, J. Moran, N. Matsumoto, H. Choi, M. Hernandez and J. H. Moore, TPOT2: A New Graph-Based Implementation of the Tree-Based Pipeline Optimization Tool for Automated Machine Learning, in *Genetic Programming Theory and Practice XX*, (Springer, 2024) pp. 1–17.
31. T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.
32. A. Bayat, B. Hosking, Y. Jain, C. Hosking, M. Kodikara, D. Reti, N. A. Twine and D. C. Bauer, Fast and accurate exhaustive higher-order epistasis search with BitEpi, *Scientific reports* **11**, p. 15923 (2021).
33. A. Nazarian, A. I. Yashin and A. M. Kulminski, Genome-wide analysis of genetic predisposition to alzheimer's disease and related sex disparities, *Alzheimer's research & therapy* **11**, 1 (2019).

34. N. Sinnott-Armstrong, Y. Tanigawa, D. Amar, N. Mars, C. Benner, M. Aguirre, G. R. Venkataraman, M. Wainberg, H. M. Ollila, T. Kiiskinen *et al.*, Genetics of 35 blood and urine biomarkers in the uk biobank, *Nature genetics* **53**, 185 (2021).

Identifying DNA methylation sites affecting drug response using electronic health record–derived GWAS summary statistics

Delaney A. Smith
Department of Biochemistry, Stanford University Medical School
Stanford, CA 94305, USA
Email: smithdel@stanford.edu

Stephanie A. Arteaga
Department of Biomedical Data Science, Stanford University Medical School
Stanford, CA 94305, USA
Email: sarteaga@stanford.edu

Marie C. Sadler
Department of Computational Biology, University of Lausanne
Lausanne, 1015, Switzerland
Email: marie.sadler@unil.ch

Russ B. Altman
Departments of Genetics and Bioengineering, Stanford University
Stanford, CA 94305, USA
Email: russ.altman@stanford.edu

Adverse drug responses (ADRs) result in over 7,000 deaths annually. Pharmacogenomic studies have shown that many ADRs are partially attributable to genetics. However, emerging data suggest that epigenetic mechanisms, such as DNA methylation (DNAm) also contribute to this variance. Understanding the impact of DNA methylation on drug response may minimize ADRs and improve the personalization of drug regimens. In this work, we identify DNA methylation sites that likely impact drug response phenotypes for anticoagulant and cardiometabolic drugs. We use instrumental variable analysis to integrate genome-wide association study (GWAS) summary statistics derived from electronic health records (EHRs) within the U.K. Biobank (UKBB) with methylation quantitative trait loci (mQTL) data from the Genetics of DNA Methylation Consortium (GoDMC). This approach allows us to achieve a robust sample size using the largest publicly available pharmacogenomic GWAS. For warfarin, we find 71 DNAm sites. Of those, 8 are near the gene *VKORC1* and 48 are on chromosome 6 near the human leukocyte antigen (*HLA*) gene family. We also find 2 warfarin DNAm sites near the genes *CYP2C9* and *CYP2C19*. For statins, we identify 17 DNAm sites. Eight are near the *APOB* gene, which encodes a carrier protein for low-density lipoprotein cholesterol (LDL-C). We find no novel significant epigenetic results for metformin.

Keywords: Pharmacogenomics; Pharmacoepigenetics, Biomarkers, DNA methylation, Electronic Health Records, Biobanks, Personalized Medicine.

1. Introduction

Adverse drug reactions (ADRs) lead to hundreds of thousands of deaths and hospitalizations each year.[1] Pharmacogenomic (PGx) studies show that genetic differences contribute to individual variance in response and are a source of ADRs because metabolic differences lead to higher-than-expected or lower-than-expected drug levels.[2] However, genetics alone do not explain all variance in drug response. Epigenetic modifications, such as DNA methylation (DNAm), have also been implicated.[3] For example, clopidogrel resistance is associated with DNA methylation near the genes BTG anti-proliferation factor 2 (*BTG2*), proteoglycan 2 (*PRG2*), vault RNA 2-1 (*VTRNA2-1*), and Period Circadian Regulator 3 (*PER3*).[4] While our DNAm profile may affect how we respond to many drugs, knowledge of specific interactions that allow prediction of variable drug response is limited.[3] Identifying methylation biomarkers for individual drugs may facilitate the reduction of adverse drug reactions.

PGx Genome-Wide Association Study (GWAS) reports have elucidated which genes and single nucleotide polymorphisms (SNPs) are associated with diverse drug response phenotypes.[5] However, these studies are limited by the fact that they do not account for epigenetic modifications. Pharmacoepigenetic (PEGx) studies, such as epigenome-wide association studies (EWAS) identify associations between DNAm and drug response phenotypes. However, these studies are limited both in number and statistical power. For example, there is currently one EWAS study on statins (linking statin use and type 2 diabetes, N = 6,820) in the EWAS catalog.[6,7] There are no studies on warfarin or metformin response.[7]

Instrument variable (IV) approaches are an alternative method to elucidate likely-causal interactions between an exposure (DNAm) and an outcome (drug response) from observational data.[8] Two sample methods allow researchers to integrate summary statistics from PGx GWAS studies with methylation quantitative trait (mQTL) data from separate sources to elucidate likely causal pharmacoepigenetic effects.[9] Moreover, analysis frameworks that use multiple IVs are less prone to reverse causality and artifacts arising from linkage disequilibrium (LD) patterns.[8] Mendelian randomization (MR) IV methods allow for the detection and elimination of pleiotropic markers while quantifying the direction and magnitude of causal effects (Figure 1).[8] This is key for PEGx studies because DNAm patterns change over time, making it challenging to dissect the cause, consequence, and confounding of PEGx effects.

This approach allows for -omics integration with existing PGx GWAS, identifying causal biomarkers such as DNAm. However, many existing PGx GWAS studies are underpowered (median sample size = 1220) for a robust statistical analysis.[10] While PGx GWAS statistics are more abundant than PGx EWAS reports, they still comprise only 10% of all GWAS entries in the GWAS catalog from 2016 to 2020.[10] A novel alternative method uses Electronic Health Record (EHR) data to generate GWAS summary statistics (Figure 1).[10] Biobank-generated summary statistics can have a large population size (UKBB N ~ 200,000) and have been shown to reflect PGx associations previously reported in traditional GWAS studies, albeit with weaker associations due to nosier phenotypes.[10]

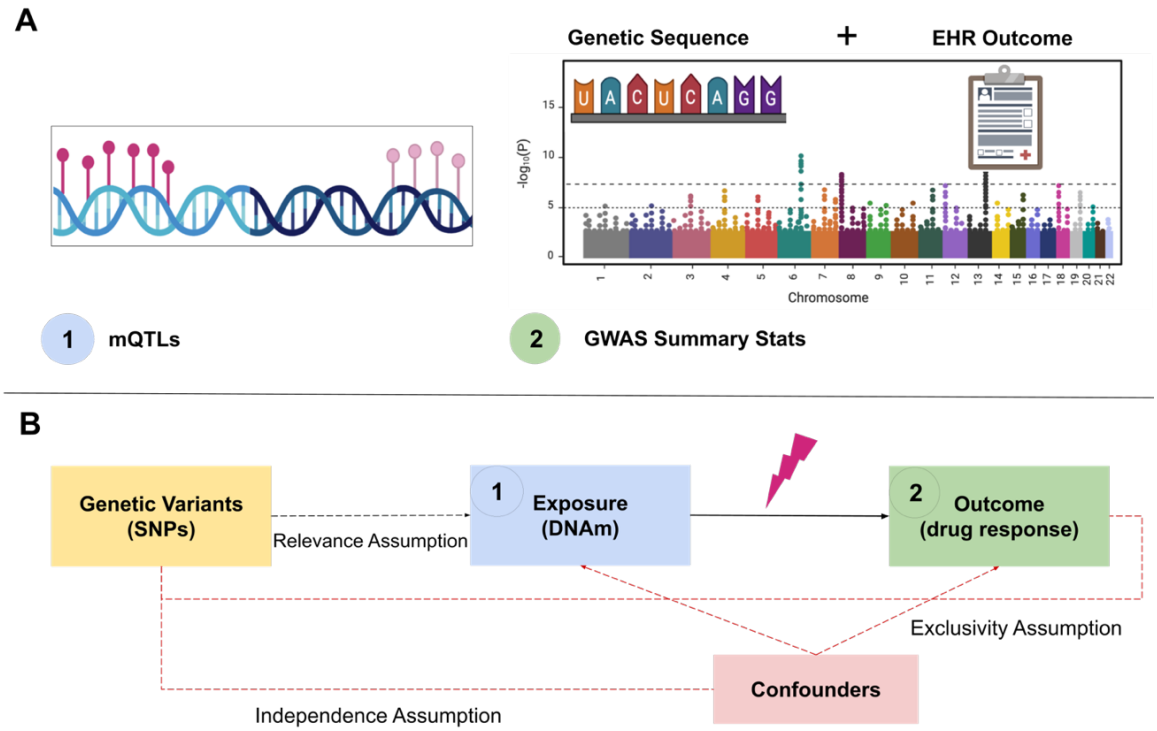


Fig. 1. Schematic of study design. A) mQTLs are taken from GoDMC in source 1 (left) and EHR records are combined with genetic sequences from the UKBB to generate summary statistics for data source 2 (right). B) Sources 1 and 2 are combined in a two-sample MR-IVW framework to determine the effect of DNAm on drug response (lightning bolt). The necessary assumptions are shown in dashed lines.

We demonstrate the efficacy of this approach in identifying DNAm sites that affect individual response to anticoagulant and cardiometabolic drugs. First, we analyze the effect of DNA methylation on warfarin response. Individual genetic differences of several genes, including vitamin K epoxidase reductase complex subunit 1 (*VKORC1*), cytochrome P450 family 2 subfamily C member 9 (*CYP2C9*), and member 19 (*CYP2C19*) are known to affect warfarin response.[11,12] Methylation near *VKORC1* has also been associated with differential warfarin response.[13] We also investigate the effect of DNA methylation on response to cardiometabolic drugs. Specifically, *β*–Hydroxy *β*–methylglutaryl-CoA (HMG-CoA) reductase inhibitors (common name: statins) and metformin. Individual response to these drugs is variable and is measured by low-density lipoprotein cholesterol (LDL-C) and total cholesterol (TC) levels for statins, and hemoglobin A1c (*HbA1c*) for metformin.[14,15] Some of this variance is explained by genetic factors such as variants in the apolipoprotein E (*APOE*) gene for statins and solute carrier family 2 member 2 (*SLC2A2*) for metformin.[16,17] In addition, metformin use is associated with genome-wide changes in DNAm levels, and a recent Swedish twin study revealed several DNAm sites associated with statin use.[18,19]

We report 69 total DNAm sites with an effect on warfarin response. Eight are near the gene *VKORC1*, and 2 are near *CYP2C19* and *CYP2C9*. Most (48) DNAm sites (also called CpGs) are not near known pharmacogenomic genes but are located on chromosome 6 near the *HLA* gene family. In the statin analysis, we find 8 CpGs near the apolipoprotein B (*APOB*) gene in addition to several CpGs near genes previously associated with cholesterol levels such as RING finger protein 39 (*RNF39*).[40,41] We find no novel significant epigenetic results for metformin. These findings allow us to better contextualize the role DNA methylation plays in individual drug responses.

2. Methods

2.1. *Genome-wide association summary statistics from electronic health records*

Genome-wide association summary statistics were generated from the EHRs of ~200,000 participants of the UKBB,[20] as described in Sadler *et al.* 2024.[10] Briefly, longitudinal medication patterns were analyzed to identify drug type, dose regimens, and drug adherence as well as baseline and post-treatment biomarker levels. We used the following pharmacogenetic phenotypes: average warfarin daily dose over the past five prescriptions (N = 4,554; McInnes and Altman), [21] cholesterol response to statins (N = 26,669 for TC, N = 17,063 for LDL-C),[10] and *HbA1c* response to metformin (N = 4,119).[10] GWAS on these quantitative traits were conducted with the REGENIE software (v3.2.4) in a whole-genome regression model for genetic markers with a minor allele frequency (MAF) > 0.05.[22] SNPs in high LD regions were removed along with those not passing LD pruning at $r^2 < 0.9$.[8]

2.2. *Two sample summary statistic instrument variable analysis*

We conducted two sample summary statistic instrument variable analyses using an inverse variance-weighted framework.[8] We used the SMR-IVW software (v1.0) as it allows two-sample IVW analysis with GWAS summary statistics.[8] The settings were: p-value (p) of mQTLs < 1×10^{-6}, LD $r^2 < 0.01$, *cis* window range of 1,000 kilobases (Kb), and the LD matrix was included in causal effect calculations. The tolerated allele frequency difference for each SNP between datasets was 0.1. A Steiger filter was implemented as described in Hemani *et al.* 2017 with a threshold set at -2, equivalent to a one-sided t-test p-value threshold of 0.023.[23] This strict threshold diminished the likelihood of including reverse causal relationships. We used mQTL data from the GoDMC database (N = 32,851), which contains > 170,000 whole blood DNAm sites with at least one significant *cis*-mQTL ($p < 1\times10^{-6}$, < 1 Mb from the DNAm site, N > 5,000).[24] The LD reference panel was from the 1,000 Genomes Project.[25]

2.3. *Multiple hypothesis correction*

To correct for multiple hypothesis testing, we used a false discovery rate ($\alpha = 0.05$) calculated by the Benjamini-Hochberg method from the statsmodels.stats.multitest (v0.14.2) package for Python.[26]

2.4. *Sensitivity to pleiotropy and heterogeneity analyses*

All CpGs that passed the significant threshold were also pruned to ensure a minimum of 3 instrumental variables. The remaining CpGs underwent sensitivity analysis. We first calculated a Cochran's Q statistic using the Metagen R package (v4.9.6),[27] and the corresponding Chi-Squared distribution p-value using the R Stats Chi-Square function (v3.6.2).[28] We removed any CpGs with significant evidence of heterogeneity ($p < 0.05$). [8] Next, we calculated an F-statistic (F) and removed any results with evidence of weak instrument bias ($F < 10$). [8] We tested for evidence of horizontal pleiotropy by analyzing the intercept values of an MR-Egger regression using the 2SMR package (v0.6.6).[23] Any CpGs showing significant evidence of pleiotropy ($p < 0.05$) were removed.

Table 1. Results of warfarin GWAS integration. For brevity, CpGs are displayed together if they are within approximately 1 Mb. When multiple CpGs are grouped, the CpG information represents the signal with the highest absolute value effect size. Full results are available on GitHub: https://github.com/smithdelaney/PGx-MR-from-EHR-GWAS.

CpG	Location	Number of CpGs	β	*SE*
cg06617202	1: 205038787	2	0.289	0.071
cg03935872	2: 17935919	1	0.349	0.086
cg06197503	3: 36422406	1	-0.286	0.063
cg10961486	4: 69959004	1	0.145	0.035
cg27585641	5: 73024506	2	0.336	0.077
cg15601071	6: 30078080	48	-0.768	0.183
cg13455759	9: 119655874	1	0.128	0.032
cg07530925	10: 90564681	1	-0.200	0.050
cg15404570	10: 96943130	1	0.597	0.127
cg03708694	11: 44489577	1	0.407	0.093
cg05555928	11: 63887634	1	0.110	0.027
cg08374890	16: 31117067	8	0.799	0.030
cg04077706	19: 43442484	1	0.128	0.030

3. Results

3.1 Warfarin MR-IVW results

Genome-wide hypothesis correction revealed 76 CpGs which exceeded the significance threshold. Of these, 69 CpGs showed no evidence of pleiotropy, weak instrument bias, or heterogeneity and were considered for further analysis (Figure 2A, Table 1). Eight of these CpGs were *cis* (within 0.5 megabases (Mb)) to the gene *VKORC1* (Figure 2A). Four had a positive beta (β) value (causing a higher warfarin dose) and 4 had a negative β (causing a lower warfarin dose), with absolute value effect sizes ranging from $|\beta| = 0.314$ to $|\beta| = 0.799$. The average absolute-value effect size was $\overline{|\beta|} = 0.554$ with an average standard error ($\overline{SE}$) of 0.046. One CpG (cg15404570) was *cis* to *CYP2C9* and *CYP2C19* and had a positive effect size $\beta = 0.597$ and $SE = 0.127$. Forty-eight CpGs (70% of all significant CpGs) were located on the short arm of chromosome 6, between 28.3 and 31.1 Mb (Figure 2B). These signals are *cis* to genes encoding the tripartite motif (*TRIM*) protein family and the HLA protein family (Figure 2B). Twenty-seven of these CpGs had a positive effect size and 21

had a negative effect size ($|\bar{\beta}| = 0.204$, $\overline{SE}) = 0.0475$). Absolute value effect sizes ranged from $|\beta| = 0.057$ to $|\beta| = 0.768$.

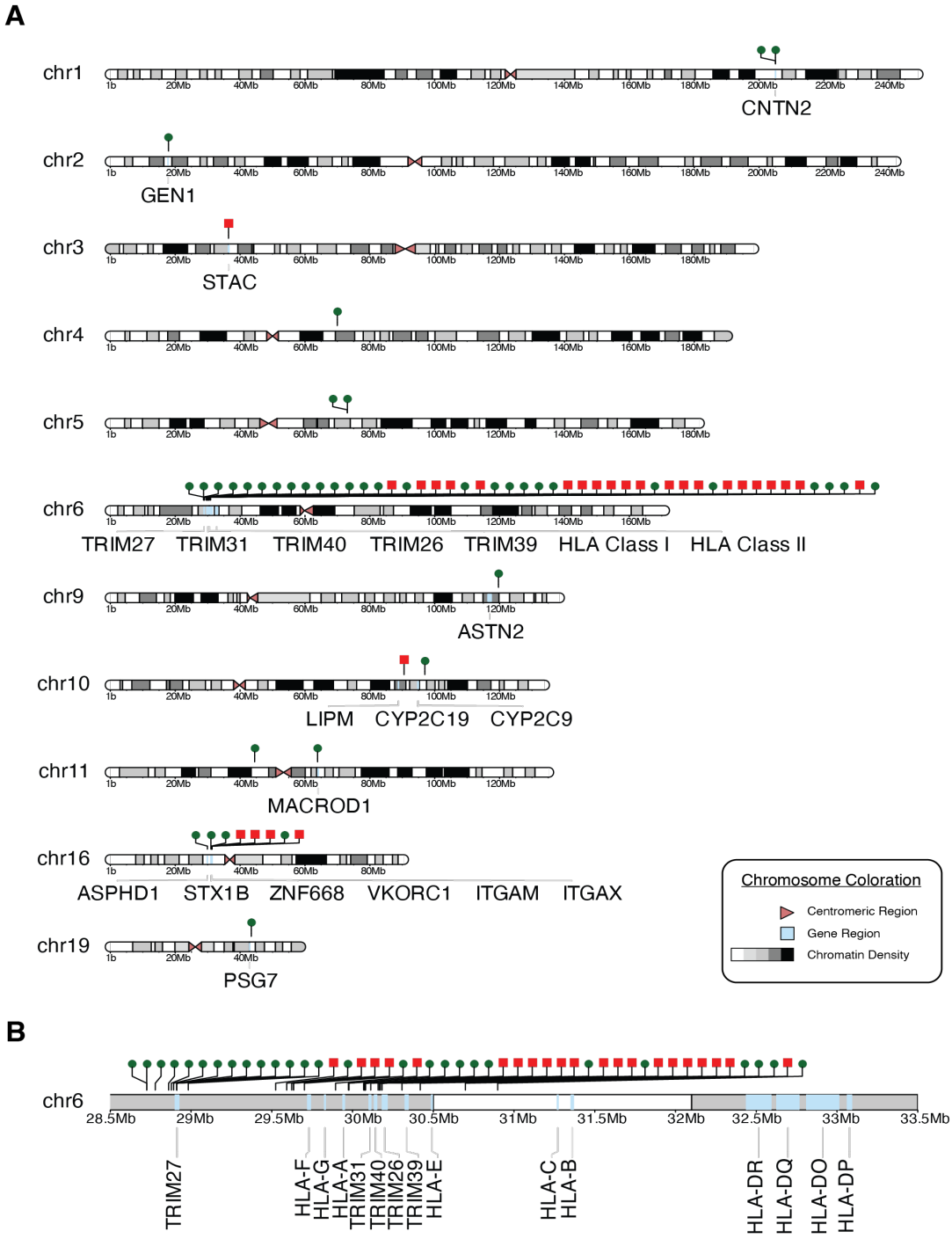


Fig. 2. A) Display of 69 CpGs found in the warfarin analysis. CpGs with a positive effect, or higher warfarin dose, ($\beta > 0$) are shown in green, and those with a negative effect ($\beta < 0$) are shown in red. Nearby genes are also annotated. B) Zoomed-in representation of the short arm of chromosome 6.

3.2 Statin MR-IVW results

GWAS integration results revealed 8 significant CpGs with LDL-C reduction as the outcome and 10 additional CpGs for TC reduction. Following quality control testing, 17 CpGs were further analyzed (Figure 3, Table 2). The 8 CpGs derived from the LDL-C analysis (47% of all CpGs) were *cis* to the *APOB* gene, which encodes an LDL-C carrier protein. All effect sizes for these 8 CpGs were negative, with an average absolute value of $|\overline{\beta}| = 0.088$, and an average standard error of $\overline{SE} = 0.018$. In this case, a negative β means that statin efficacy is increased since the clinical goal of the therapeutic is to reduce cholesterol levels. The absolute value effect size ranged from $|\beta| = 0.079$ to $|\beta| = 0.108$. Two CpGs (cg05337441, cg24309555) were previously annotated for TC or LDL-C in the EWAS catalog.[7] There were 3 CpGs (cg06028875, cg16908633, cg23752348) on the short arm of chromosome 6 which were *cis* to the *RNF39* gene and near the *HLA* gene family (within 1 Mb) (Figure 3). All 3 had a negative effect ($|\overline{\beta}| = 0.075$, $\overline{SE} = 0.016$). Five CpGs were on chromosome 10, four of which had a positive effect size ($|\overline{\beta}| = 0.059$, $\overline{SE} = 0.015$). The magnitude of these effect sizes ranged from $|\beta| = 0.031$ to $|\beta| = 0.158$. Additional genes associated with TC CpGs were DPY30 domain-containing proteins 1 and 2 (*DYDC1/C2*), erythroblast transformation-specific (ETS) proto-oncogene 2 (*ETS2*), tetraspanin 14 (*TSPAN14*), and peroxiredoxin-like 2A (*PRXL2A*) (Figure 3).

Table 2. Results of statin GWAS integration. For brevity, CpGs are displayed together if they are within approximately 1 Mb of each other. When multiple CpGs are grouped, the β is the absolute value average and the CpG name and location represent the signal with the highest absolute value effect size. Full results are available on GitHub: https://github.com/smithdelaney/PGx-MR-from-EHR-GWAS.

CpG	Location	Number of CpGs	β	*SE*
cg00673290	2:21266727	8	-0.108	0.020
cg06028875	6: 30042295	3	-0.087	0.019
cg02750471	10: 82179740	5	0.158	0.031
cg15892280	21:40180000	1	0.088	0.019

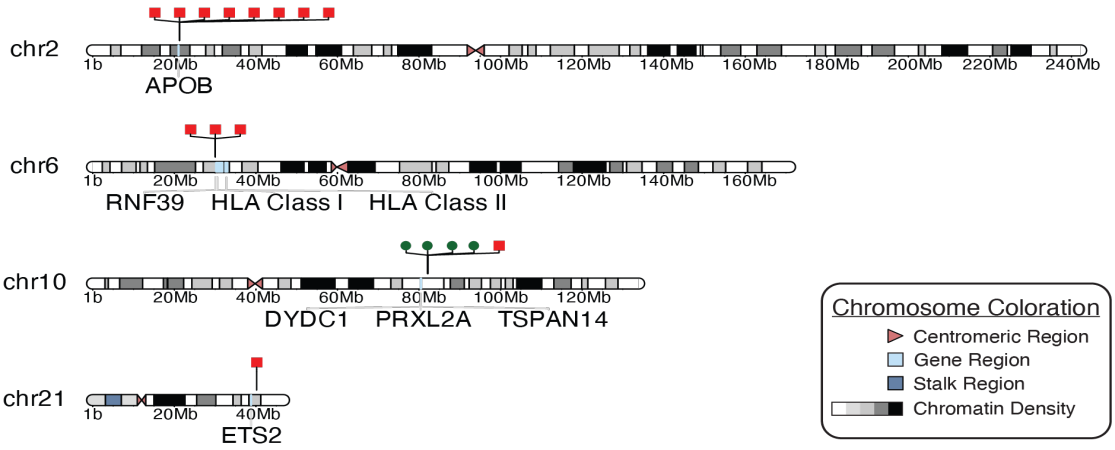


Fig. 3. Display of 17 CpGs found in the statin analysis. CpGs with a positive effect ($\beta > 0$) are shown in green and those with a negative effect ($\beta < 0$) are shown in red. A negative effect in this study means increased statin efficacy. Nearby genes are also annotated.

Discussion

In this study, we provide evidence that DNA methylation plays a causal role in individual response to warfarin and statins. Probing PEGx effects using EWAS studies provides correlative associations between DNAm sites and drug response phenotypes. Our approach uses existing information to infer directional, causal, and quantitative effect estimates. In our warfarin analysis, 8 CpGs were *cis* to *VKORC1*. Warfarin's mechanism of action targets *VKORC1*, and genetic variations in *VKORC1* are known to modulate warfarin's efficacy.[11,12] Recent findings also implicate *cis* DNA methylation near *VKORC1* in warfarin response.[13] In addition, we find 8 CpGs *cis* to *APOB* in the statin analysis. *APOB* encodes an LDL-C carrier and has 187 GWAS associations in the GWAS Catalog with LDL-C and 125 with TC.[29] Two of the CpGs we identified also had previous annotations for LDL-C or TC.[7] Thus, our approach captures the known effects of DNA methylation on drug response. We also show that in the case of metformin, no novel significant DNA methylation effects were detected. This could be because there is no biological effect or because the GWAS is underpowered.[10]

Our method depends on the assumptions underlying IV analysis. The first assumption is that there is a sufficiently strong relationship between the instrumental variables (SNPs) and the exposure (DNAm). The second assumption (independence) is that instrumental variables are independent of confounders. The third assumption (exclusivity) is that any effect the SNP has on the outcome is mediated only through the exposure (no horizontal pleiotropy). The first assumption can be tested by selecting highly significant mQTL effects ($p < 1\times10^{-6}$) and performing a weak instrument bias test (F-statistic).[8] The second and third assumptions are violated when results show evidence of heterogeneity, horizontal pleiotropy, or the presence of invalid instruments. We use Cochran's Q test to detect evidence of heterogeneity, and the presence of invalid instruments.[30] MR-Egger regression intercepts detect the presence of horizontal pleiotropy.[9] In addition, these assumptions may not hold in the presence of LD between the mQTLs and SNPs. The risk of 'LD-hitchhiking' leading to spurious results is managed by selecting CpGs with a minimum of 3 instrumental variables, filtering

out SNPs in high-LD regions, pruning for independence, and using a Steiger filter for directionality.[8] In our analysis, we excluded 7 warfarin CpGs and 1 statin CpG which did not pass all these controls. However, the possibility of horizontal pleiotropy can never be fully excluded.

Our results show DNA methylation CpGs *cis* to warfarin pharmacogenomic genes, *VKORC1*, *CYP2C9*, and *CYP2C19*.[11,12] Therefore, individuals' methylation profiles may account for some of the variability in warfarin response not captured in pharmacogenomic models. The 8 *VKORC1* CpGs had effect sizes in both positive and negative directions, indicating that while the presence of some CpGs reduced the average daily dose of the patient, others likely led to an increase. Seventy percent (70%) of effect CpGs were located near the *HLA* and *TRIM* genes on chromosome 6 (Figure 2B). The CpGs were closest to several *TRIM* genes (*TRIM26, TRIM27, TRIM31,* and *TRIM40).* These genes encode proteins that have varied and widespread functionality. DNAm may regulate the expression of *TRIM* genes, which have many downstream effects, possibly including modulating blood clotting pathways. However, recent work on DNA methylation and gene expression shows that methylated sites can act distally to influence the expression of neighboring genes.[3,24,31,32] Thus, an alternative hypothesis is that these CpGs impact the expression of the *HLA* genes, which are interlaced with *TRIM* genes on chromosome 6 (Figure 2B). Genetic polymorphisms in *HLA* genes have previously been associated with blood disorders including acquired hemophilia A, venous thrombosis, immune thrombotic thrombocytopenic purpura, and sickle cell disease.[33-36] Therefore, the pre-existing association between *HLA* genes and blood diseases may manifest through altered warfarin response, which is affected by DNA methylation near these genes.

Our statin results show that causal CpG identified in this study are not located near known pharmacogenomic genes. Instead, 47% of CpGs are *cis* to *APOB,* which encodes an LDL-C carrier (Figure 3). These CpGs all have an average effect size of $|\bar{\beta}|$ 0.087, all with a negative direction, meaning that methylation in this region causes a decrease in measured LDL-C in response to starting statin treatment. Therefore, the presence of these CpGs causes improved efficacy of statin treatment.

Another apolipoprotein gene, *APOE*, has over 20 variant annotations for statin efficacy in the PharmGKB pharmacogenomic database.[37] Both *APOB* and *APOE* are carriers of LDL-C and other lipoproteins. While *APOB* does not have PharmGKB annotations for statins, there are several genetic variants within the gene associated with LDL-C and TC levels in the GWAS Catalog, as discussed above. Moreover, genetic variation in *APOB* has been associated with familial hypercholesterolemia,[38] and levels of *APOB* are biomarkers of atherogenic particle concentration in the bloodstream (Figure 4).[39] Since DNAm near the *APOB* gene causes decreases in LDL-C in response to statin treatment, these CpGs can be biomarkers of statin response, and the study of these CpGs can increase our understanding of atherogenic disease.

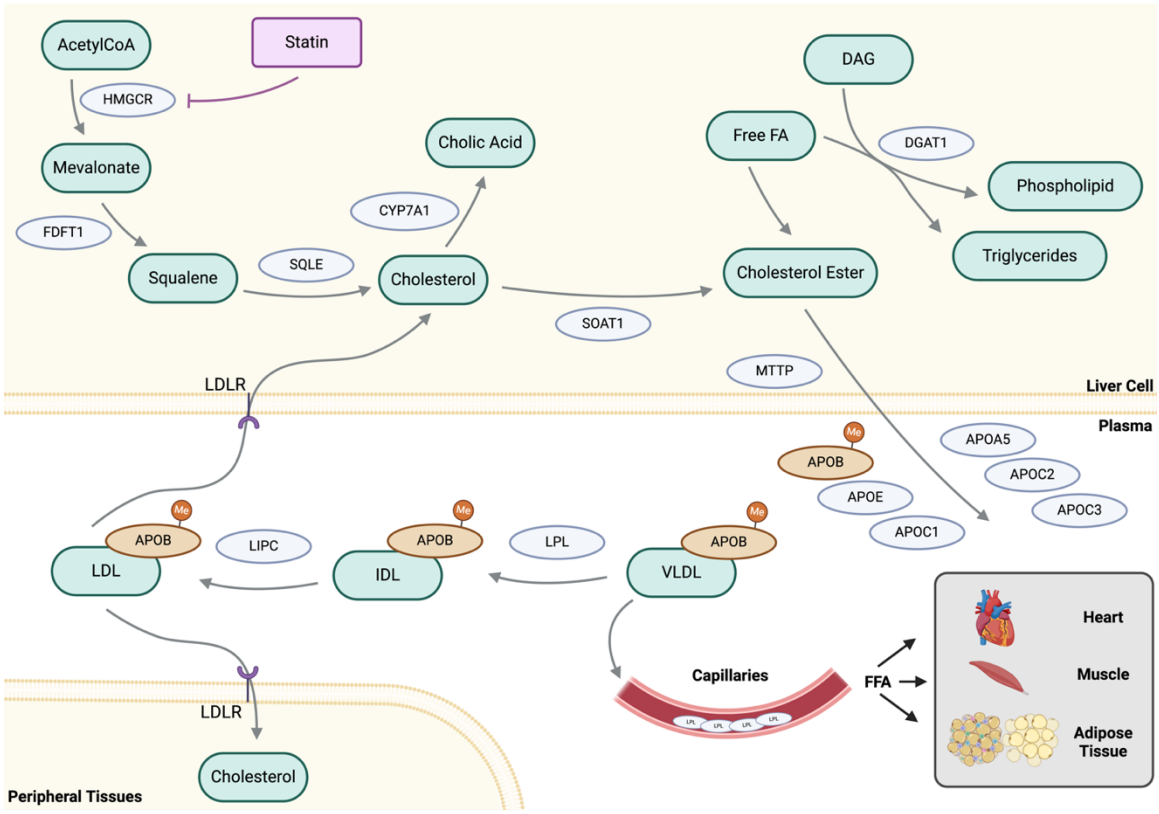


Fig. 4. The statin pharmacodynamic pathway and *APOB*-LDL-C pathway. Purple indicates statin, small molecules are in green, and light blue shapes represent protein-coding genes with HGNC standard names. Orange 'Me' probes represent DNA methylation occurring near the gene coding region (created with BioRender.com; adapted from PharmGKB[37]).

We propose a potential model of how DNAm near *APOB* may affect LDL-C levels in response to statin therapy (Figure 5). The 8 CpGs we identified near *APOB* are in the gene regulatory region. Five CpGs (cg16306978, cg16723488, cg24309555, cg25071744, cg25123895) are in the *APOB* promoter region, one in an enhancer region (cg05337441), and one (cg00673290) in a CpG island within the regulatory region. Generally, DNAm within the regulatory region of a protein-coding gene is associated with decreased expression.[54] Reduced *APOB* can lead to an increased ratio of LDL-C to *APOB*.[55-56] With statin therapy, LDL receptor expression increases as intracellular hepatic cholesterol decreases (Figure 5). Since *APOB* binds to the LDL receptor, more LDL-C is cleared from the plasma per *APOB* particle, leading to a greater decrease in measured LDL-C (Figure 5). It could also be that an individual with reduced *APOB* levels stores more LDL-C in other cholesterol-carrying particles. If these particles are equally reduced with statin therapy, then folks with higher *APOB* levels will have higher post-treatment LDL-C levels. Moreover, cholesterol metabolism is an intricate pathway, and regulatory mechanisms are still being studied, so additional experiments would be required to test these hypotheses.

The statin results also show that the gene *RNF39* had 3 nearby CpG sites. *RNF39* is involved in inflammatory responses throughout the body and has a SNP that has previously been associated with free cholesterol levels.[40,41] We find 2 CpGs near *DYDC1/C2*. These genes are primarily studied for their role in spermiogenesis, but several SNPs in the gene have been previously associated with hypertension.[42,43] One CpG was near the *TSPAN14* gene, which is associated with Niemann-Pick disease, a genetic disorder that leads to the inability to break down fats, such as cholesterol and lipids, inside cells.[14] Another CpG was near *ETS2*, which is a transcription factor. It regulates the transcription of proteasome assembly chaperone 1 (*PSMG1*) which has two SNPs associated with LDL in the GWAS catalog.[7] Finally, the CpG with the largest absolute effect size ($|\beta| = 0.158$) was located near gene *PRXL2A*. This gene interacts with ST3 beta-galactoside alpha-2,3-sialyltransferase (*ST3GLA4*) which has 48 SNPs associated with LDL-C and 32 SNPs associated with TC in the GWAS catalog.[29] These associations provide plausible pathways by which DNA methylation may impact response to statin treatment.

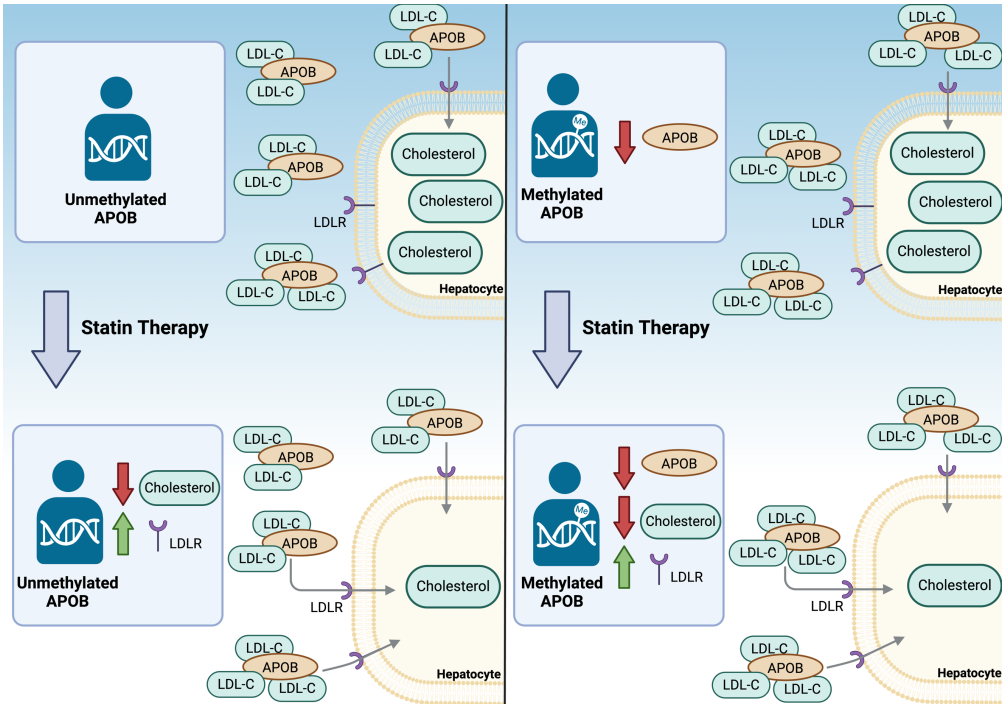


Fig. 5. Lower *APOB* expression may lead to decreased LDL-C levels after statin therapy due to an increased ratio of LDL-C to *APOB* (created with BioRender.com).

We observed that 5 of the LDL-C CpGs (cg16306978, cg24309555, cg25035485, cg25071744, cg25123895) and 3 TC CpGs (cg01528321, cg02750471, cg04043334) identified in the statin analysis (8 total, 47%) had previous annotations for inflammatory disease (inflammatory bowel syndrome (IBD) and Crohn's disease) in the EWAS catalog.[44] A comprehensive EWAS study has published approximately 3,633 CpGs associated with either disease, which make up about 2.1% of all CpGs in GoDMC.[44] The number of overlapping annotations is significantly ($p < 1\text{x } 10^{-8}$) greater than what is expected due to random chance alone. These findings, in combination with the warfarin

CpGs located near *HLA* genes, suggest that there are shared pathways between immune response and response to common cardiovascular and clotting disorder treatments that may be influenced by DNAm patterns. This mirrors other recent findings that are beginning to dissect how these two systems interact outside of the epigenetic space.[45-47] However, since the *HLA* gene region has high genetic diversity, it is possible that the signals detected reflect differences in ancestry and prevalence of *HLA* haplotypes. Thus, we also examined the overlap between the statin DNAm sites, and those annotated for inflammatory diseases, excluding the sites in the *HLA* region (non-*HLA* sites: cg16306978, cg24309555, cg25035485, cg25071744, cg25123895) and found that the overlap remained significant ($p = 0.052$).

This study had several limitations. First, we analyze whole-blood DNA methylomes. DNA methylation is tissue-specific and much of the pharmacokinetic and pharmacodynamic activity occurs in the liver. While some genes have similar DNA methylation patterns across blood and liver, this assumption cannot be generalized to all genes[48]. This means there may be tissue-specific signals we are not detecting. However, blood DNA methylation signal is an accessible diagnostic tool and DNA methylation sites from blood samples remain biologically relevant signals. Moreover, blood DNAm samples have been used to elucidate effects on other phenotypes, such as Alzheimer's and Type 2 Diabetes.[49-51] Another limitation is that both the UKBB and GoDMC sample predominantly European ancestries, which means there may be signals associated with non-European ancestry that are not being detected in this study. Thus, we plan to conduct a replicate analysis using the more genetically diverse biobank, All of Us.[52] Finally, we measure the warfarin average daily dose over the past 5 days, which may be a less robust metric than the patient's clotting time or time in the therapeutic range.[57]

While the longitudinal drug response model presented by Sadler *et al*. minimizes the risk of spurious signals unrelated to drug response,[10] it would still be useful to test whether any of the statin signals are replicated in a study of cholesterol levels alone. In addition, the signals identified in this study are directional from DNA methylation to the outcome of drug response. However, we know that some drugs and diseases induce DNA methylation changes. Therefore, it would be interesting to conduct an explicit bi-directional MR study to identify reverse-causal effects.[53] Moreover, we are learning that DNA methylation does not just regulate the nearest genes but has a more complex regulatory mechanism that may underlie these results.[3,24,31,32] Finally, it is difficult to compare effect sizes generated in this analysis with genetic effects identified through GWAS, because of the different assumptions and experimental set-ups. However, this work does demonstrate that epigenetic considerations are important for advancing our understanding of drug response and ADRs. In summary, we address the problem of insufficient and correlative studies linking DNA methylation and individual drug response with a statistical inference approach.

Acknowledgments

D.A.S. is supported by the NSF Graduate Research Fellowship 2019286895. S.A.A. is supported by the NIH T15 LM007033 from the NLM. M.C.S. was supported by the Swiss National Science Foundation (310030-189147). R.B.A. is supported by Chan-Zuckberg Biohub and NIH GM102365.

References

1. Kommu, S., Carter, C. & Whitfield, P. Adverse Drug Reactions. *StatPearls* (2024).
2. Chenchula, S., Atal, S. & Uppugunduri, C. R. S. A review of real-world evidence on preemptive pharmacogenomic testing for preventing adverse drug reactions: a reality for future health care. *Pharmacogenomics J* **24**, 1–6 (2024).
3. Smith, D. A., Sadler, M. C. & Altman, R. B. Promises and challenges in pharmacoepigenetics. *Camb Prism Precis Med* **1**, e18 (2023).
4. Yang, J. *et al*. Clopidogrel Resistance Is Associated With DNA Methylation of Genes From Whole Blood of Humans. *Front Genet* **11**, 583215 (2021).
5. Swen, J. J. et al. A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomized crossover implementation study. The Lancet 401, 347-356 (2023).
6. Ochoa-Rosales, C. *et al*. Epigenetic Link Between Statin Therapy and Type 2 Diabetes. *Diabetes Care* **43**, 875–884 (2020).
7. Battram T., Yousefi P., Crawford G. *et al*. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res* vol. 7 (2022).
8. Sadler, M. C., Auwerx, C., Porcu, E. & Kutalik, Z. Quantifying Mediation between Omics Layers and Complex Traits. (2021) doi:10.1101/2021.09.29.462396.
9. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512–525 (2015).
10. Sadler, M. C. *et al*. Leveraging large-scale biobank EHRs to enhance pharmacogenetics of cardiometabolic disease medications. 2024.04.06.24305415 Preprint at https://doi.org/10.1101/2024.04.06.24305415 (2024).
11. Li, J., Wang, S., Barone, J. & Malone, B. Warfarin Pharmacogenomics. *P T* **34**, 422–427 (2009).
12. Scordo, M. G. *et al*. Influence of CYP2C9 and CYP2C19 genetic polymorphisms on warfarin maintenance dose and metabolic clearance. *Clin Pharmacol Ther* **72**, 702–710 (2002).
13. He, S. *et al*. Methylation of CYP1A1 and VKORC1 promoter associated with stable dosage of warfarin in Chinese patients. *PeerJ* **9**, 11549 (2021).
14. Zineh, I. Pharmacogenetics of Response to Statins. *Curr Atheroscler Rep* **9**, 187–194 (2007).
15. Florez, J. C. The pharmacogenetics of metformin. *Diabetologia* **60**, 1648–1655 (2017).
16. Cai, C., Wen, Z. & Li, L. The relationship between ApoE gene polymorphism and the efficacy of statins controlling hyperlipidemia. *Am J Transl Res* **13**, 6772–6777 (2021).
17. Zhou, K. *et al*. Variation in the glucose transporter gene SLC2A2 is associated with glycemic response to metformin. *Nat Genet* **48**, 1055–1059 (2016).
18. Marra, P. S. *et al.* Metformin use history and genome-wide DNA methylation profile: potential molecular mechanism for aging and longevity. *Aging (Albany NY)* **15**, 601–616 (2023).
19. Qin, X., Wang, Y., Pedersen, N. L., Tang, B. & Hägg, S. Dynamic patterns of blood lipids and DNA methylation in response to statin therapy. *Clin Epigenetics* **14**, 153 (2022).

20. Bycroft, C. *et al*. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
21. McInnes, G. & Altman, R. B. Drug Response Pharmacogenetics for 200,000 UK Biobank Participants. 2020.08.09.243311 Preprint at https://doi.org/10.1101/2020.08.09.243311 (2020).
22. Mbatchou, J. *et al*. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097–1103 (2021).
23. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* **13**, 1007081 (2017).
24. Min, J. L. *et al*. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet* **53**, 1311–1321 (2021).
25. Sudmant, P. H. *et al*. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
26. Perktold, J. *et al.* statsmodels/statsmodels: Release 0.14.2. Zenodo https://doi.org/10.5281/zenodo.10984387 (2024).
27. Metagen R package (v4.9.6): Release 4.9.6. RDRR https://rdrr.io/cran/meta/src/R/metagen.R (2024).
28. R: The R Stats Package. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html (2024).
29. Sollis, E. *et al*. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, 977–985 (2023).
30. Bowden, J. *et al*. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International Journal of Epidemiology* **48**, 728–742 (2019).
31. Zaghlool, S. B. *et al*. Mendelian inheritance of trimodal CpG methylation sites suggests distal cis-acting genetic effects. *Clin Epigenetics* 8, 124 (2016).
32. Tong, Y. *et al*. MICMIC: identification of DNA methylation of distal regulatory regions with causal effects on tumorigenesis. *Genome Biol* 19, 73, (2018).
33. Pavlova, A., Zeitler, H., Scharrer, I., Brackmann, H.-H. & Oldenburg, J. HLA genotype in patients with acquired haemophilia A. *Haemophilia* **16**, 107–112 (2010).
34. Ivasková, E. *et al*. HLA and venous thrombosis: a prospective study. *Cor Vasa* **33**, 424–427 (1991).
35. Laghmouchi, A., Graça, N. A. G. & Voorberg, J. Emerging Concepts in Immune Thrombotic Thrombocytopenic Purpura. *Front Immunol* **12**, 757192 (2021).
36. Wong, K., Lai, W. K. & Jackson, D. E. HLA Class II regulation of immune response in sickle cell disease patients: Susceptibility to red blood cell alloimmunization (systematic review and meta-analysis. *Vox Sang* **117**, 1251–1261 (2022).
37. Whirl-Carrillo, M. *et al*. An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clin Pharmacol Ther* **110**, 563–572 (2021).
38. Zorzo, R. A. *et al*. LDLR gene's promoter region hypermethylation in patients with familial hypercholesterolemia. *Sci Rep* **13**, 9241 (2023).

39. Behbodikhah, J. *et al*. Apolipoprotein B and Cardiovascular Disease: Biomarker and Potential Therapeutic Target. *Metabolites* **11**, 690 (2021).
40. Richardson, T. G. *et al*. Characterising metabolomic signatures of lipid-modifying therapies through drug target mendelian randomisation. *PLoS Biol* **20**, 3001547 (2022).
41. Jia, X., Zhao, C. & Zhao, W. Emerging Roles of MHC Class I Region-Encoded E3 Ubiquitin Ligases in Innate Immunity. *Front. Immunol* **12**, (2021).
42. Sakaue, S. *et al*. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415–1424 (2021).
43. Li, S. *et al*. Interaction of SH3P13 and DYDC1 protein: a germ cell component that regulates acrosome biogenesis during spermiogenesis. *Eur J Cell Biol* **88**, 509–520 (2009).
44. Ventham, N. T. *et al*. Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nat Commun* **7**, 13507 (2016).
45. Boyalla, V., Gallego-Colon, E. & Spartalis, M. Immunity and inflammation in cardiovascular disorders. *BMC Cardiovasc Disord* **23**, 148 (2023).
46. Dal Lin, C., Tona, F. & Osto, E. The crosstalk between the cardiovascular and the immune system. *Vasc Biol* **1**, 83–88 (2019).
47. Wielscher, M. *et al*. DNA methylation signature of chronic low-grade inflammation and its role in cardio-respiratory diseases. *Nat Commun* **13**, 2408 (2022).
48. Olsson Lindvall, M. *et al*. Comparison of DNA Methylation Profiles of Hemostatic Genes between Liver Tissue and Peripheral Blood within Individuals. *Thromb Haemost* **121**, 573–583 (2021).
49. Konki, M. *et al*. Peripheral blood DNA methylation differences in twin pairs discordant for Alzheimer's disease. *Clinical Epigenetics* **11**, 130 (2019).
50. Sun, Y. *et al*. Identification of candidate DNA methylation biomarkers related to Alzheimer's disease risk by integrating genome and blood methylome data. *Transl Psychiatry* **13**, 1–10 (2023).
51. Juvinao-Quintero, D. L., Sharp, G. C., Sanderson, E. C. M., Relton, C. L. & Elliott, H. R. Investigating causality in the association between DNA methylation and type 2 diabetes using bidirectional two-sample Mendelian randomisation. *Diabetologia* **66**, 1247–1259 (2023).
52. Mayer, C. S. & Huser, V. Learning important common data elements from shared study data: The All of Us program analysis. *PLoS One* **18**, 0283601 (2023).
53. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, 89–98 (2014).
54. Stadler, M. B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480, 490–495 (2011).
55. Glavinovic, T. et al. Physiological Bases for the Superiority of Apolipoprotein B Over Low-Density Lipoprotein Cholesterol and Non–High-Density Lipoprotein Cholesterol as a Marker of Cardiovascular Risk. J Am Heart Assoc 11, e025858 (2022).
56. Vega, G. L. & Grundy, S. M. Hypercholesterolemia with cholesterol-enriched LDL and normal levels of LDL-apolipoprotein B. Effects of the step I diet and bile acid sequestrants on the cholesterol content of LDL. Arterioscler Thromb Vasc Biol 16, 517–522 (1996).

57. Kuruvilla, M. & Gurk-Turner, C. A review of warfarin dosing and monitoring. Proc (Bayl Univ Med Cent) 14, 305–306 (2001).

Spherical Manifolds Capture Drug-Induced Changes in Tumor Cell Cycle Behavior

Olivia Wen[1,2], Samuel C. Wolff[2,3], Wayne Stallaert[6], Didong Li[4], Jeremy E. Purvis[2,3,†] & Tarek M. Zikry[2,4,5,†]

[1]Department of Biology, [2]Computational Medicine Program, [3]Department of Genetics, [4]Department of Biostatistics, [5]School of Data Science and Society, University of North Carolina at Chapel Hill, NC
[6]Department of Computational and Systems Biology, University of Pittsburgh, PA
[†]Corresponding Authors: jeremy_purvis@med.unc.edu, tarek@unc.edu

CDK4/6 inhibitors such as palbociclib block cell cycle progression and improve outcomes for many ER+/HER2- breast cancer patients. Unfortunately, many patients are initially resistant to the drug or develop resistance over time in part due to heterogeneity among individual tumor cells. To better understand these mechanisms of resistance, we used multiplex, single-cell imaging to profile cell cycle proteins in ER+ breast tumor cells under increasing palbociclib concentrations. We then applied spherical principal component analysis (SPCA), a dimensionality reduction method that leverages the inherently cyclical nature of the high-dimensional imaging data, to look for changes in cell cycle behavior in resistant cells. SPCA characterizes data as a hypersphere and provides a framework for visualizing and quantifying differences in cell cycles across treatment-induced perturbations. The hypersphere representations revealed shifts in the mean cell state and population heterogeneity. SPCA validated expected trends of CDK4/6 inhibitor response such as decreased expression of proliferation markers (Ki67, pRB), but also revealed potential mechanisms of resistance including increased expression of cyclin D1 and CDK2. Understanding the molecular mechanisms that allow treated tumor cells to evade arrest is critical for identifying targets of future therapies. Ultimately, we seek to further SPCA as a tool of precision medicine, targeting treatments by individual tumors, and extending this computational framework to interpret other cyclical biological processes represented by high-dimensional data.

Keywords: Manifold learning; Dimensionality reduction; ER+/HER2- Cancer.

1. Introduction

Despite promising results of CDK4/6 inhibitors for treating ER+/HER2- breast cancer, 10-20% of patients show initial drug resistance, and all patients develop resistance over time.[1] Resistance is thought to arise from the heterogeneity of molecular states in individual tumor cells, and one potential source of this cell-to-cell heterogeneity is the cell cycle. In recent years, single-cell studies have revealed that the cell cycle can show remarkable flexibility.[2] For example, individual tumor cells may progress through cell cycle phases with variable durations, or show altered expression levels of core cell cycle regulators.[3–5] The ability of cells to upregulate or downregulate certain protein signaling pathways is referred to as cell cycle plasticity.

Prior studies of ER+/HER2- cells —both in a cell culture model and from a primary tumor sample— have demonstrated how molecular differences allow individual tumor cells to evade CDK4/6 inhibitor therapy through alternative cell cycle paths.[6] Therefore, to mechanistically understand how resistance develops, it is important to develop robust analytical methods for characterizing the underlying manifolds along which cell cycle trajectories proceed.

Commonly, to more easily detect trends in increasingly high-dimensional single-cell data, dimensionality reduction methods are used. Dimensionality reduction techniques transform high-dimensional data into a low-dimensional space such that the most valuable information, or original structure, of the data is preserved. Manifold learning, a nonlinear approach to dimensionality reduction, is often applied to high-dimensional data as a tool for visualization, data exploration, and statistical analysis. In this study, we collected and analyzed single-cell data of T47D, a model human ER+/HER2- breast cancer cell line,[7] that we introduced to varying doses of palbociclib, one of three FDA-approved CDK4/6 inhibitors (Fig. 1). Complex molecular signatures were obtained for each cell by performing iterative indirect immunofluorescence imaging (4i)[8] using 20 cellular features relevant to proliferation. The result is a high-dimensional dataset that we can broadly interpret as a representation of the cell cycle. We seek to reduce the high-dimensional representation and visualize it in a more interpretable space. The selection of an appropriate method for characterizing cell cycle data is a necessary and crucial step for both biological and statistical interpretations. However, it is difficult to assess the performance of these methods as no fixed statistics exist to directly compare the effectiveness of one method over another. Thus, we must biologically interpret the results using known cell cycle markers and trends.

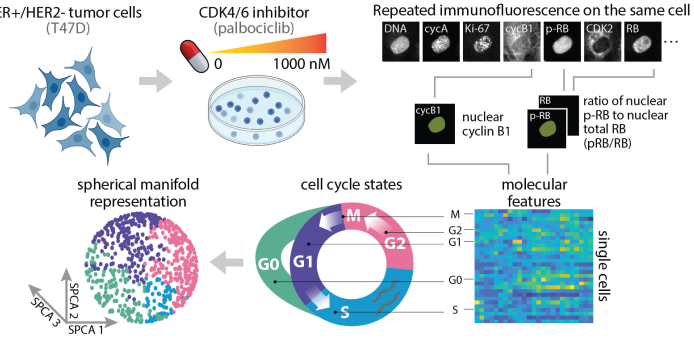


Fig. 1. **Pipeline for generating cell cycle manifold from single-cell images**. T47D tumor cells were treated with increasing concentrations of palbociclib. 4i was then performed using a panel of cell cycle-specific markers, resulting in a tabular dataset after raw image processing and segmentation. Because individual cells are not synchronized, individual cells span a range of cell cycle states. SPCA was then applied to estimate a hypersphere manifold representation of the cell cycle in a lower dimensional space. Each dot represents an individual tumor cell in a specific cell cycle state.

Here, we present spherical principal component analysis (SPCA) as an effective tool for modeling the underlying cell cycle structure of single-cell ER+/HER2- breast cancer cell line data (Fig. 1).[a] By assuming the data lie on a reduced spherical space, SPCA helps preserve gradual cell state transitions and cell-to-cell heterogeneity. Using trajectory inference approaches, we demonstrate how SPCA captures cyclical patterns of cell cycle regulators not

[a]All code, additional experimental details, and full supplemental feature plots can be found at `https://github.com/purvislab/SingleCell_HyperSphere`. Data are available at `https://doi.org/10.5281/zenodo.13621367`.

found in potential of heat-diffusion for affinity-based transition embedding (PHATE) or principal component analysis (PCA) models. Structural differences in spherical manifolds across treatment conditions also point to driving factors of CDK4/6 inhibitor response, which can help identify downstream clinical targets involved in treatment-resistant pathways.

2. Related Work

Manifold Estimation

Manifold learning is often a necessary step of high-dimensional data analysis. One branch of manifold learning techniques is manifold estimation. Manifold estimation approaches identify a low-dimensional embedding that preserves local and global structures without imposing assumptions about the structure of the data. PHATE is an example of a manifold estimation approach that has been found to be successful in producing clean, denoised models of biological data and preserving continuous trajectories.[9] PHATE captures local and global relationships by computing neighborhood relationships between cells, performing diffusion using local affinities, and projecting diffusion distances to create a two- or three-dimensional embedding. These embeddings can be used for hypothesis generation and visual comparison of cell cycle progressions.[6,10] However, because PHATE and other manifold estimation approaches do not assume an underlying structure, no statistical inferences can be made on these embeddings or between manifolds produced by different datasets.

Manifold Approximation

Other techniques, called manifold approximation, assume data to lie on an underlying structure. Thus, fitted values and error metrics can be computed. The most commonly used method for manifold approximation[11] is PCA.[12] PCA identifies features that are responsible for the most variance and projects the linearly transformed data onto a subspace of fewer dimensions. PCA has a long history of use in the biological field and requires low computational power, but is sensitive to noise, making it suboptimal for use with heterogeneous data such as single-cell data.[13,14]

Other manifold approximation methods assume data to lie on a more complex surface. SPCA, a variant of PCA, is one such method that assumes data lie on a sphere in a lower dimensional space.[15] For a dataset X reduced using SPCA, a spherical manifold M is parameterized by a radius r, a center c, and an affine subspace V. First, the subspace V where the optimal sphere lies is estimated from the input dataset X. A loss function is minimized to identify an optimal sphere by reducing the number of points that lie outside or inside the surface of the sphere. The optimal center and radius are estimated from the minimization of the loss function. From the parameters c (center), r (radius), and V (subspace), a projection for X onto the sphere is defined. SPCA has previously been applied to cell cycle data of retinal pigmented epithelial (RPE) cells and has been found to fit the data better than other methods.[10,14] However, a deep dive of the exact cell cycle trends was not explored, nor were SPCA manifolds of different datasets, such as cells of different treatment conditions, compared.

Trajectory Inference

Trajectory inference methods can be applied to high-dimensional, single-cell datasets to quantify the progression of dynamic cellular processes.[16] Revelio, a method that leverages PCA,

revealed single-cell transcriptomic data to follow a 2D circular trajectory.[17] However, this method seeks to remove cell cycle effects whereas we aim to study them and the response of proteomic states to forms of perturbation. Revelio also orders cells according to gene markers of cell state transitions and is suboptimal when applied to cells that follow different dynamics.[18]

One robust method for capturing dynamic cellular processes and representing noisy, single-cell data is Slingshot, a curve-based trajectory inference method.[19] Slingshot identifies single or multiple branched trajectories using two main steps: (1) constructing a minimum-spanning tree between clusters of data to identify a global lineage structure and (2) fitting smooth principal curves to each lineage. Orthogonal projections of each data point onto the curve assign a pseudotime, representing cell cycle progression, for each cell.

Slingshot also allows for various levels of supervision and flexibility in the choice of upstream data analysis methods. At a minimum, Slingshot requires data that has been clustered and reduced, a list of cluster labels, and specification of the dimensionality reduction method performed. Additional supervision can be achieved by specifying a start cluster, an end cluster, or the number of lineages to infer. Previously, Slingshot has been found to identify smooth cell cycle trends in PHATE embeddings.[6] The flexibility of Slingshot and its success with cell cycle data makes it an ideal method for comparing cell cycle paths inferred using different manifold learning approaches.

3. Methods

Experimental Details

T47D ER+/HER2- breast cancer cells were obtained from the ATCC (catalog number HTB-133) and maintained at 37°C with 5% CO_2 in RPMI-1640 media supplemented with 10% fetal bovine serum (FBS). Cells were plated on a glass 96-well plate coated with poly-L lysine at 25,000 cells per well. Cells were allowed to adhere for 24 hours at 37°C with 5% CO_2 in RPMI-1640 media with 10% FBS. After 24 hours, media and non-adherent cells were removed. RPMI-1640 media with 10% FBS was added containing vehicle, or palbociclib at 0, 1, 10, 100, or 1,000 nM. Cells were incubated at 37°C with 5% CO_2. After 24 hours of treatment, cells were fixed with PFA, and iterative indirect immunofluorescence imaging (4i) was performed as described below. Single-cell proteomic measurements for samples were obtained using 4i by adapting the protocol previously described in Refs. 6,8. Following image and data preprocessing, cell cycle phases were annotated using a three component Gaussian Mixture Model (sklearn v0.24.1) on the log-transformed measurements of DNA content, cyclin A, and cyclin B1, as these features were previously shown to minimally represent the cell cycle.[20] The full and close-up 4i images used for this study can be seen in Fig. S1, S2[a].

Manifold Approximation with SPCA

SPCA[15] was implemented in Python to identify the c_i, r_i, and V_i of the sphere that characterizes cells from each treatment condition i. Three dimensions were chosen to aid in visual comparison, but other methods, such as the identification of an elbow plot,[15] exist to identify the optimal lower dimension of a dataset. To identify a shared subspace V_G for comparison of all treatment conditions on a uniform scale, we applied SPCA using the complete dataset.

20-feature cell signatures from each treatment condition i were projected onto spheres with center c_i and radius r_i in subspace V_G. Orientations of plots were selected based on the best visual separation of phases or treatment conditions.

Manifold Comparisons with PHATE and SPCA

We visually and statistically compared the performance of the spherical manifolds approximated by SPCA to three-dimensional manifolds produced by PHATE and PCA. PHATE[9] was performed in Python (phate v1.0.11) on the complete dataset of all treatment conditions. A k-nearest neighbor graph was constructed to create the three-dimensional PHATE structure using the following hyperparameters: $n\_components = 4$, n_jobs = -1, $knn = 200$, and $t = 12$. The hyperparameters were tuned according to hyperparameters selected for previous PHATE models of cell cycle data.[6,20,21]

Python was also used to perform PCA[12] (scikit-learn v1.3.2). PCA was run using the complete dataset such that all treatment conditions can be evaluated in the same space. To produce a three-dimensional visualization, $n\_components$, the number of features to extract in the reduced dataset, was set to three.

Cell Cycle Trajectory Inference Using Slingshot

To assess the recapitulation of temporal trends, we applied trajectory inference to infer cell cycle paths. Slingshot was performed in R (slingshot v2.6.0) using each of the manifold learning approaches (PHATE, PCA, and SPCA) as the upstream dimensionality reduction method. Slingshot trajectories were inferred through cells from each of the treatment conditions. We provided cell cycle phase annotations (G0, G1, S, G2/M) as cluster labels and specified G0 as the start cluster. Pseudotimes were normalized to a scale of 0 to 1 to allow for the comparison of lineages on a uniform scale. To identify feature expression trends over pseudotime, locally estimated scatterplot smoothing (LOESS)[22] curves were fit using Python (v2.1.2).

4. Results

Recapitulating the Cell Cycle

From the 20-feature proteomic signatures and cell cycle phase labels (G0, G1, S, G2/M) of 64,502 T47D cells, we generated tabular datasets of cells from each treatment condition (n_0=10,366, n_1=10,675, n_{10}=13,051, n_{100}=15,688, n_{1000}=14,722). Each row describes a cell's unique molecular state, thus providing a complete representation of the cell cycle altogether. To identify a lower dimensional manifold that preserves the cyclical nature of the cell cycle, we performed SPCA[15] for each palbociclib dose resulting in five three-dimensional hyperspheres characterized by unique centers and radii projected onto a shared global reduced space.

We expect neighborhood relationships to be preserved such that cells in similar states, and thus with similar molecular signatures, are located near each other on a lower-dimensional manifold. Similarly, cells with different proteomic profiles are located far apart in 20-dimensional space and should remain further away on a three-dimensional manifold. To assess the ability of SPCA to capture differences in cell states, we visualized the distribution of cell cycle phases across the SPCA hyperspheres (Fig. 2A). For each of the treatment conditions, we obtained a spherical manifold that successfully captured differences between phases and the canonical progression of cells through the cell cycle, from G0, G1, S, to G2/M. Across

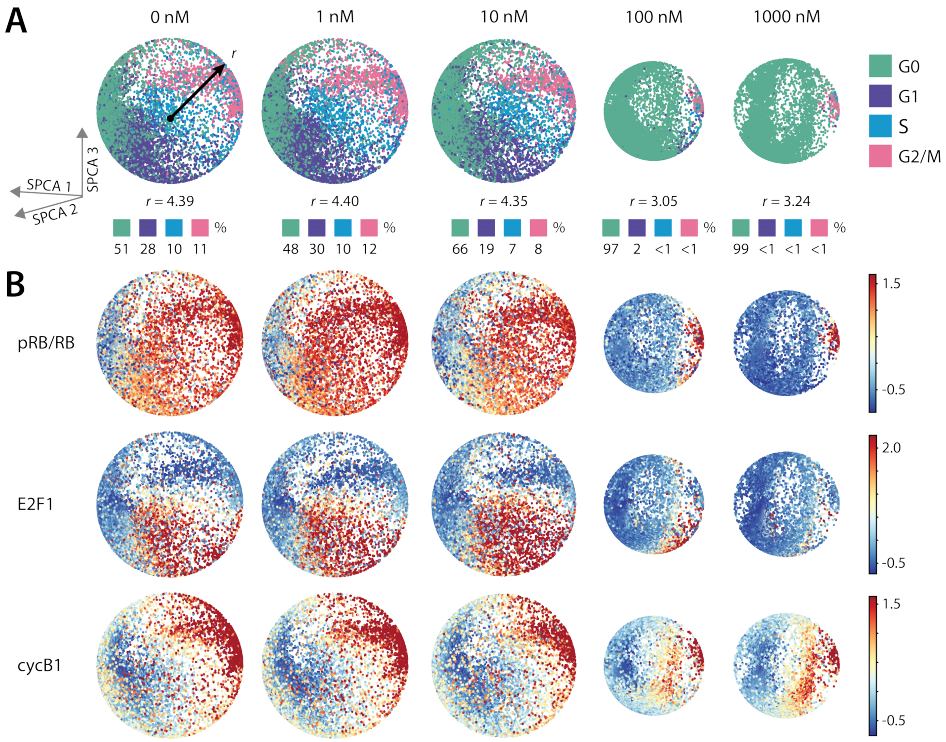


Fig. 2. **SPCA captures shifts in cell cycle phases and regulators across treatment conditions**. Data points from each treatment condition were projected onto three-dimensional hyperspheres identified by SPCA in a shared subspace. Points are colored according to (A) their cell cycle phase label or (B) normalized expression level of pRB/RB, E2F1, or cycB1.

all conditions, cells belonging to the same cell cycle phase were located near each other and within distinct regions along the surface of the spheres. We observed the cell cycle phases to be evenly distributed and occupy the same regions in the 0 nM, 1 nM, and 10 nM projections. At 100 nM and 1,000 nM, we saw an increase in the proportion of G0 cells concentrated mainly in the western hemisphere and along the vertical center axis, notably in the direction of cells in proliferative cell states. The small proportion of proliferative (G1, S, G2/M) cells was visible in a small region on the eastern hemisphere of the two manifolds. Additionally, the hyperspheres representative of cells treated with 100 nM and 1,000 nM had smaller radius sizes compared to the hyperspheres of lower palbociclib doses. Thus, we observed a delineation between the lower ($\leq$10 nM) and higher ($\geq$100 nM) treatment conditions. For all figures, plots for all features are available in the supplement[a].

We next investigated the ability of SPCA to capture more gradual cell-to-cell transitions by inspecting changes in the expression of each of the 20 cell cycle regulators (Fig. 2B). The resulting plots visually recapitulated known trends in protein expression levels for every feature. A high ratio of pRB to RB (pRB/RB) is needed to transition past the restriction point in late G1 to S phase. RB, or retinoblastoma protein, is hypophosphorylated by CDK4/6 and cyclin D1 complexes and hyperphosphorylated by cyclin E-CDK2.[23] Therefore, we expected cells in G0 and early G1 to have relatively lower pRB/RB values. In the G0 and G1 regions of the $\leq$10 nM palbociclib-treated cells, we observed an increasing gradient of pRB/RB values.

The cells with the highest pRB/RB ratios aligned with cells in the G2/M regions (Fig. 2A, B). Known trends were also observed for E2F1 and cyclin B1.[24,25] The highest values of E2F1 were located in S and nearby G1 regions while G2/M and bordering S phase cells expressed the highest values of cyclin B1. The 100 nM and 1,000 nM hyperspheres revealed more nuanced results. Compared to the lower treatment conditions, the G0 cells in the two highest treatment conditions expressed the lowest amounts of pRB/RB and E2F1, visible by the contrast in the color intensities of the G0 regions. Although the expression of pRB/RB and E2F1 decreased to a more extreme state, cyclin B1 followed a different trend. A subset of G0 cells demonstrated low expression of cyclin B1 while another group, most notably under the 1,000 nM palbociclib dose, had higher levels of cyclin B1 nearing those characteristic of proliferative G2/M cells.

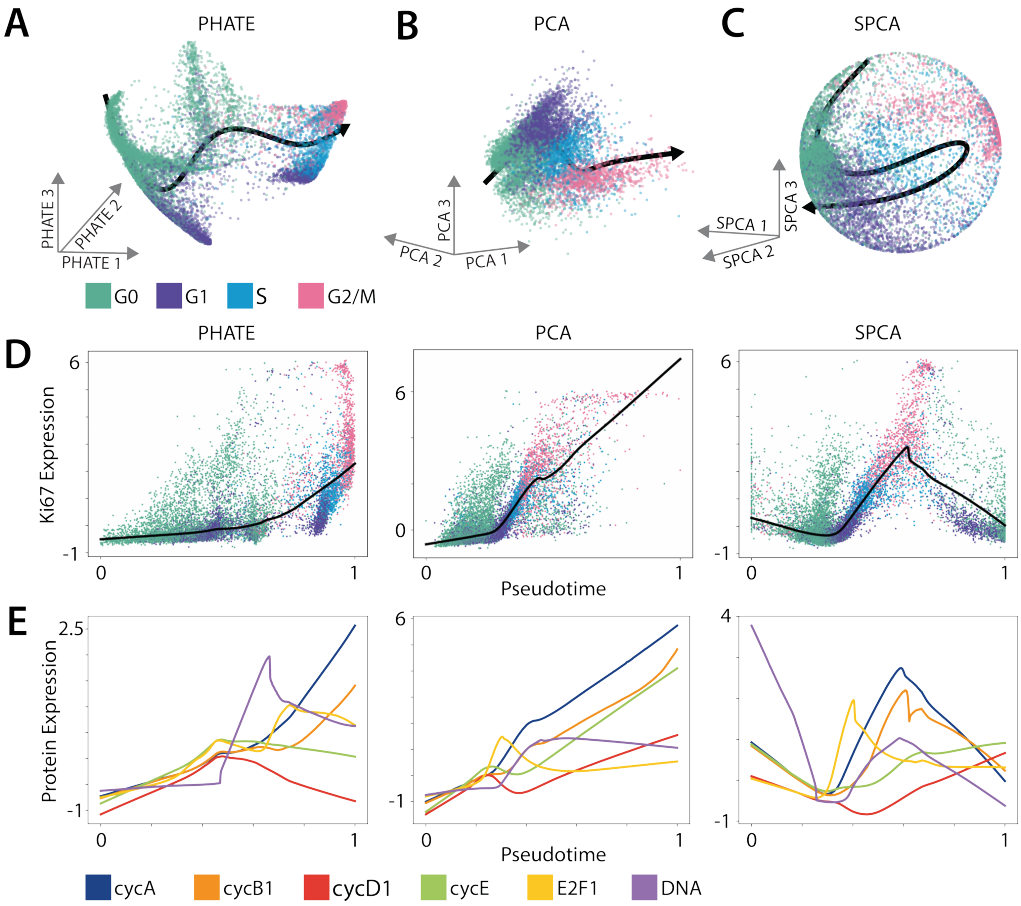


Fig. 3. **SPCA recapitulates cyclical protein level trends**. (A) PHATE, (B) PCA, and (C) SPCA were performed on untreated cells (0 nM palbociclib). Data points from each manifold learning method were plotted in three dimensions and colored according to their cell cycle phase annotation. Trajectories identified by Slingshot (black line) were overlaid onto their respective plots. (D) Ki67 expression of each cell was plotted according to the cell's normalized Slingshot pseudotime. A LOESS curve (black line) was fit through the points for each method. (E) LOESS curves fit through points plotted according to Slingshot pseudotime and median levels of core cell cycle regulators (cyclin A, cyclin B1, cyclin D1, cyclin E1, E2F1, and DNA content) were overlaid for PHATE, PCA, and SPCA.

Comparing Cell Cycle Structures from PHATE, PCA, and SPCA

SPCA successfully captured the overall structure of the cell cycle and known proteomic trends. To validate the effectiveness of SPCA as a representative tool for modeling the cell cycle, we compared SPCA to two other manifold learning methods, PHATE and PCA. Unlike SPCA, PHATE does not allow for the projection of multiple datasets to a shared space. PCA does have this capability but some uninterpretable alignment of principal component spaces is required. Due to these limitations, both PHATE and PCA were performed on the entire dataset such that all treatment conditions could be compared on a uniform scale. First, we examined the distribution of cell cycle phases in untreated cells using all three methods (Fig. 3A-C). Overall, we found cells belonging to the same cell cycle phase to be concentrated in the same region. However, the separation of phases varied. The least visual separation of phases was observed in the structure for PCA. In the projections produced by PHATE and SPCA, we saw greater separation between phase regions. We also observed a separation within phases in the PHATE manifold, specifically in G0 and G1, showing a discontinuous progression of phases. G0 cells occupied three main arms in one region of the PHATE structure while G1 cells were clustered in one of two regions on opposite sides of the manifold. One G1 cluster was located along an arm of the structure shared with G0 cells and the other group bordered the S phase region. Upon visual inspection, all manifolds suggested a canonical ordering of cell cycle phases. To assess how well these structures represented temporal trends, we performed Slingshot,[19] a trajectory inference method, to infer cell cycle paths. For each method, Slingshot identified a single trajectory through the canonical ordering of cell cycle phases - G0, G1, S, and G2/M - when provided a starting phase of G0. However, while the trajectories identified using PHATE and PCA proceeded in one direction from G0 to G2/M, the trajectory found from SPCA returned to the G0 and G1 regions, indicating a cyclic pattern (Fig. 3D).

Using the normalized pseudotime assigned to each cell, we next examined how expression levels of each feature fluctuated throughout the identified cell cycle paths (Fig. 3D). Ki67 is a key proliferative marker that accumulates over the course of the cell cycle reaching a peak in G2 and M.[26,27] Temporal orderings of cells identified for PHATE, PCA, and SPCA all followed an increasing trend of Ki67 expression (Fig. 3B). While each method ordered cells from G0, G1, S, to G2/M, cells in the SPCA pseudotime ordering returned to a state of G0 or G1 following the G2/M phase. Similarly, Ki67 expression decreased to a level consistent with that of the initial G0 cells. Cells were also more evenly distributed over pseudotime time for SPCA compared to PHATE, which had a separation between arrested and proliferating cells, and PCA, which had a majority of cells concentrated in the first half of the trajectory. Thus, SPCA successfully captured the gradient of protein accumulation we observed for key cell cycle regulators (Fig. 2, S3)[a] whereas PHATE and PCA identified less continuous trends.

We next asked how overall feature trends followed known accumulation patterns of key cell cycle regulators, specifically cyclins, E2F1, and DNA (Fig. 3E).[24,25,28] Expression of these cell cycle markers follows a cyclical pattern and aligns with key molecular events. Only cyclin A and cyclin B1 trends for PHATE and PCA as well as E2F1 trends for PCA aligned with expected points of accumulation, whereas all trends, except for cyclin E, identified using SPCA followed known expression patterns. For PHATE and PCA, the majority of feature trends

followed strictly increasing patterns. Cyclin D1 for PHATE and E2F1 for PCA experienced a peak in expression and a decrease, returning near initial expression levels. DNA content trends for the two methods, and cyclin E for PHATE peaked and revealed a more subtle decrease. All features for SPCA demonstrated a cyclic pattern such that final expression levels nearly matched initial levels, except for DNA content which had a lower final expression than the G0 cells identified to be at the beginning of the cell cycle path.

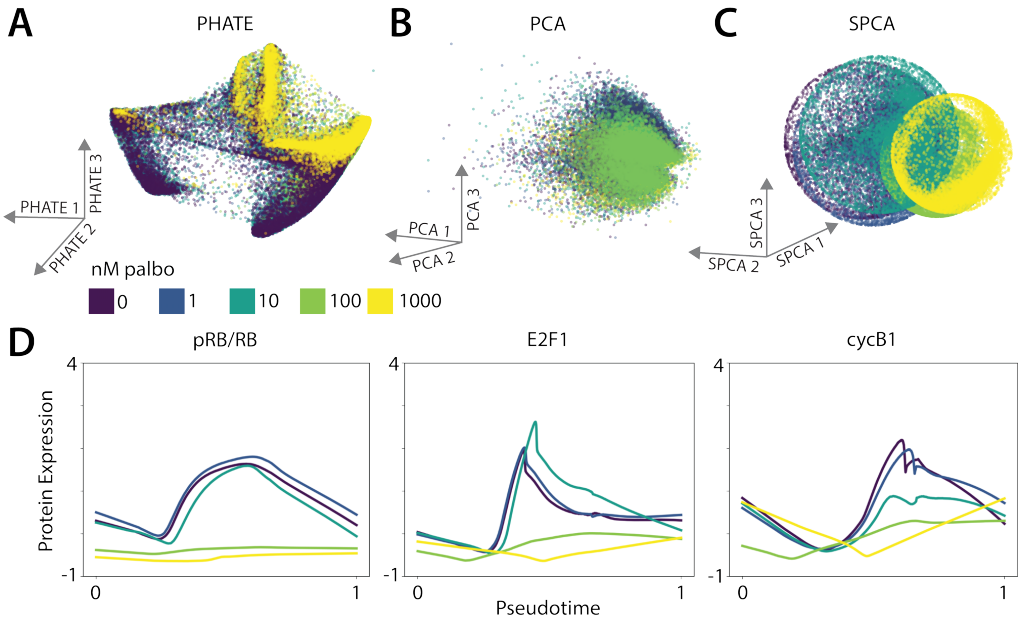


Fig. 4. **SPCA captures dose-dependent shifts in cell cycle manifold**. (A) PHATE and (B) PCA were performed using 20-feature single-cell signatures from all treatment conditions. (C) SPCA was performed for individual treatment conditions and the data points were projected onto their respective hyperspheres in a shared space identified by performing SPCA using all cells. Points are colored according to palbociclib dose for each individual cell. (D) LOESS curves were fit through points plotted according to Slingshot pseudotime generated using SPCA and median protein expression levels (pRB/RB, E2F1, cycB1) across five treatment conditions.

In response to palbociclib treatment, a greater proportion of cells become arrested (Fig. 2A). Therefore, we expect cells treated with different doses of palbociclib to reflect differences in the makeup of cell states and behaviors. PCA showed minimal delineation between treatment conditions (Fig. 4B) whereas PHATE and SPCA structures (Fig. 4A, C) captured differences between $\leq$10 nM and $\geq$100 nM palbociclib-treated cells. Cells belonging to the 100 nM and 1,000 nM treatment conditions concentrated along the arms of G0 cells of lower treatment conditions in the PHATE structure (Fig. 3A). Interestingly, cells of higher treatment conditions did not concentrate in the areas occupied by G0 cells in spheres of $\leq$10 nM doses. Instead, in addition to shrinking in size, SPCA spheres for 100 nM and 1,000 nM migrated in the direction of proliferative cell states. When we compared feature expression trends over pseudotime across treatment conditions for each method, SPCA more accurately captured cell cycle trends (Fig. 2B) and characterized behaviors expected of a dose response (Fig. 4D).

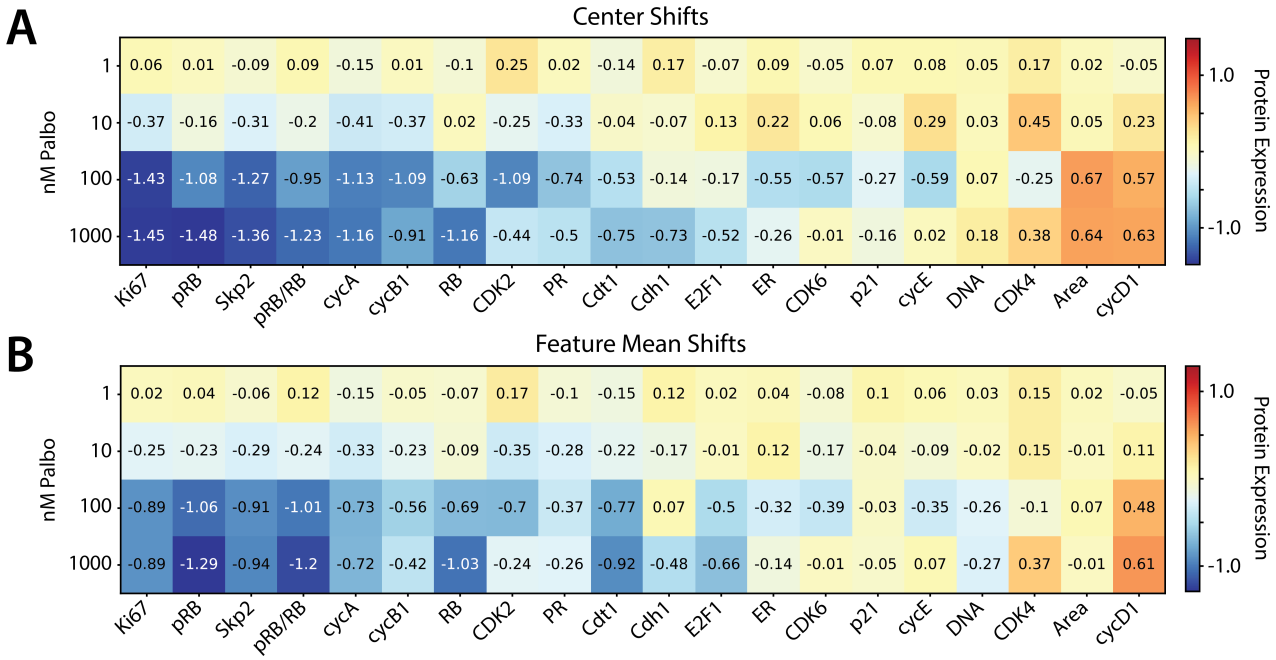


Fig. 5. **Shifts in centers of SPCA hyperspheres reveal changes in cell cycle regulation across treatment conditions**. (A) Shifts in the centers and (B) mean protein abundance of three-dimensional hyperspheres identified by SPCA for each treatment condition were calculated from each dose response to the untreated condition. Results from each pairwise comparison are represented in each row of the heatmaps.

SPCA Elucidates Mechanisms of CDK4/6 Inhibitor Resistance

To identify which specific factors were driving shifts in cell cycles across treatment conditions, evident by visual observations of feature expression differences (Fig. 2B) and the shift in positions of the 100 nM and 1,000 nM SPCA structures from the lower treatment conditions (Fig. 4C), we compared centers of the spheres. We quantified center shifts by subtracting the 20-feature center for 0 nM from the centers of each treatment condition. There was a clear distinction in protein levels between cells treated with lower (1 nM and 10 nM) and higher (100 nM and 1,000 nM) doses of palbociclib (Fig. 5A). Notably, we found a more significant depletion of proteins including Ki67, pRB, Skp2, cyclin A, cyclin B1, and RB, and enrichment of CDK4, cell area, and cyclin D1 in higher treatment conditions. Overall, the same trends were identified by comparing differences in feature means (Fig. 5B). However, the differences between the treatment groups were not as substantial, specifically for Ki67, Skp2, cyclin A, cyclin B1, and cell area which showed almost no change in mean expression across treatment conditions. The greatest depletion was found in pRB expression while cyclin D1 accumulation was the highest among all cell cycle regulators according to mean expression shifts from 100 nM and 1,000 nM to untreated cells. CDK4 expression was the second most elevated protein according to mean expression. Although CDK4 was also enriched according to center shifts between 1,000 nM and 0 nM treatment groups, CDK4 expression peaked in the 10 nM center shift as opposed to in the 1,000 nM mean expression shift. A similar pattern was found for cyclin E which was also elevated in the 10 nM to 0 nM comparison of centers, suggesting

an increase in cyclin E expression in 10 nM palbociclib-treated cells, while mean expression values revealed the opposite.

Because a majority of cells under 100 nM and 1,000 nM palbociclib treatment were arrested in G0 (Fig. 2A) and Slingshot identified a cyclical return to G0 cells in SPCA's cell cycle trajectory (Fig. 3C, D), we wanted to determine if there were differences between these groups of G0 cells. G0 cells were partitioned according to the median pseudotime of cells in a treatment condition. We will refer to the group of G0 cells with a pseudotime less than the median pseudotime value as 'early G0' and the remaining G0 cells bordering G2/M phase as 'late G0'. When we compared the proteomic signatures of early and late G0 cells, we observed notable differences between the groups (Fig. 6). For ≤100 nM doses, we observed higher expression of cell cycle regulators Ki67, Skp2, RB, CDK2, PR, Cdt1, Cdh1, ER, CDK6, p21, and cyclin E, and a decrease in cell area in late G0 cells. The greatest contrast was observed between early and late G0 cells treated with ≤10 nM palbociclib. An opposite trend was observed for the 1,000 nM early G0 cells which had higher expression of cell cycle markers including cyclin D1, CDK4, and cyclin E compared to the late G0 cells for 1,000 nM and other early G0 cells.

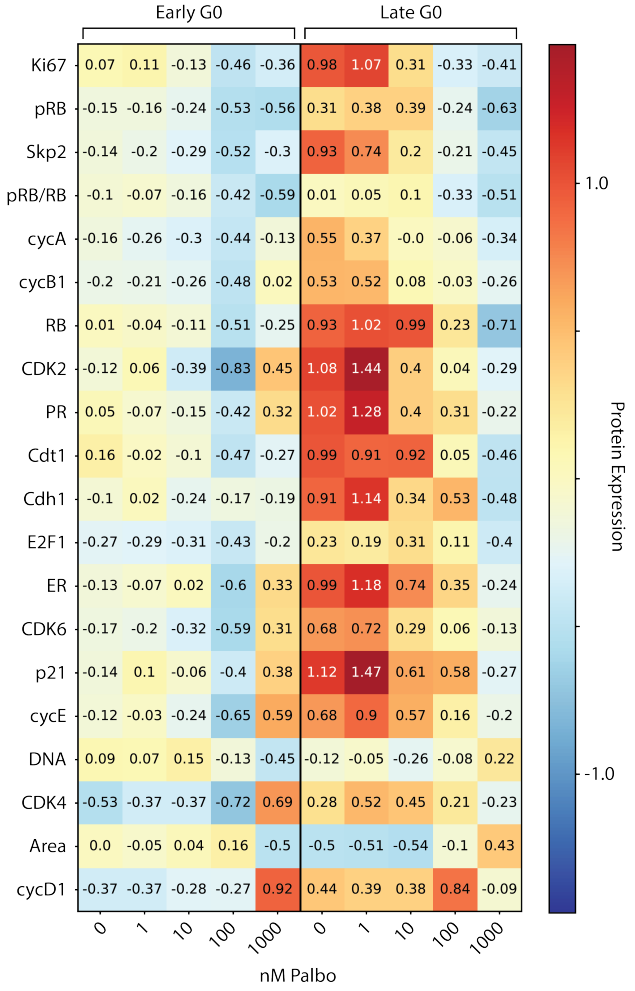


Fig. 6. **SPCA and Slingshot identify differences in G0 cells**. G0 cells were separated according to median normalized pseudotime. Mean protein levels for each cell cycle feature are represented in each row of the heatmap.

5. Discussion

We validated SPCA as a tool for characterizing cell cycle plasticity of breast tumor cells in response to palbociclib treatment. SPCA recapitulated the underlying cyclical structure of multiplex, single-cell breast tumor data and enabled direct visual and quantitative comparisons across treatment conditions. SPCA captures heterogeneity of molecular states by preserving fundamental differences between stages of the cell cycle, shown by the delineation of each cell cycle phase, while revealing gradual transitions in protein expression patterns. In addition to the continuous progression of cell states, the even distribution of phases and cells across the spherical manifolds suggests that the structure is representative of cell cycle data. Other

methods such as PHATE and PCA do not allow for as much flexibility in quantitative analysis or comparison of treatment conditions compared to SPCA. These methods failed to capture the cyclical nature of phase progression and protein expression.

Furthermore, we note that the spheres characterizing cells in the higher treatment group reveal differences from lower treatment conditions that are not observed by PHATE or PCA. The decrease in radius size, paired with a shift in center location, and skewed distribution of phases suggests that cells experience a fundamental shift in their cell cycles at the 10 nM and 100 nM transition. The greater proportion of G0 cells and smaller radius size of the 100 nM compared to the 10 nM sphere indicate less heterogeneity in cell cycle states. The overall shifts in the positions of the hyperspheres and the migration of 100 nM and 1,000 nM palbociclib-treated G0 cells towards proliferative regions in lower treatment conditions also suggest that cells under higher dosage have different mean states and, thus, traverse alternative paths through the cell cycle. The dichotomy between $\leq$10 nM and $\geq$100 nM hyperspheres aligns with prior knowledge that the IC50 for palbociclib lies within this range.[29]

SPCA allows us to quantitatively assess this difference between low and high treatment conditions via a comparison of each hypersphere's 20-feature center coordinates. These structural characteristics of SPCA manifolds can reveal trends that cannot be realized by looking at feature expression alone. A decrease in the hypersphere radius size along dose increases (Fig. 2A), indicates a reduction in heterogeneity of cell states. Center shifts pointed to further depletion of cell markers (Ki67, Skp2, cyclin A, cyclin B1) compared to mean expression, but also an increase in cell area and DNA which were found to remain consistent (cell area) or be downregulated (DNA) according to mean expression under increasing palbociclib treatment. Conversely, Cdt1 was found to be one of the highest-ranking features with decreased mean expression in higher treatment groups, but this difference was not as prominent when examining center shifts. These differences highlighted by radius and center shifts may indicate which cell cycle regulators are most responsible for driving changes in cell cycle behavior, but future experiments will need to be done to validate this hypothesis.

Structural differences that allowed for the identification of cyclical cell cycle trajectories with SPCA, but not PHATE or PCA, are also worth further investigation. Differences in early and late G0 cells suggest greater heterogeneity of multiple molecular states within cells categorized as G0. These differences may suggest that improved methods of cell cycle phase annotation need to be performed and that our framework of using SPCA and Slingshot could be used as a tool for differentiating between cell states. For example, $\leq$10 nM late G0 cells with high expression of proliferative markers, but low cell area could suggest that these are new daughter cells. However, differences in early and late G0 cells may indicate true differences in G0 cells or CDK4/6 inhibitor resistance. Cells with low proliferation markers, such as $\leq$100 nM early and 1,000 nM late G0 cells, may indicate varying depths of quiescence.[21,30–32] Overexpression of cyclin D1 and elevated levels of CDK2, shown in late 1,000 nM G0 cells in Fig. 6 across doses, has previously been found to be a potential mechanism of CDK4/6 inhibitor resistance via formation of cyclin D1-CDK2 complexes.[6,33,34] Cyclin E overexpression and constitutive activation is another characteristic of breast tumor cell behavior and an indicator of CDK4/6 inhibitor response.[35–37] Thus, these cell profiles can be used to characterize cells

and identify potential mechanisms involved in treatment-resistant pathways.

In future work, this pipeline can be utilized for multi-modal precision medicine. Rather than estimating hyperspheres for each treatment dose to compare, we may estimate hyperspheres for individual tumors across a patient population. In this way, we can compare an individual's cancer progression and resistance, and find personalized biomarkers for clinical targeting. Furthermore, though we have demonstrated the use of these innovative computational and statistical techniques on a single-cell breast tumor dataset, this framework can be extended to other biological contexts. SPCA can be generalized to study not only disease responses along the cell cycle, but single-cell responses to other forms of perturbation as well, including stem cell differentiation pathways. Other cyclical biological processes such as circadian rhythm and weather patterns can also be studied, leveraging the inherent underlying structures of these data, although prior knowledge or assessment that the data is spherical, which was established for cell cycle data based on extensive study,[211417] is needed. This novel framework for modeling cyclical biological data can allow for the rapid identification and quantification of novel trends in responses to forms of perturbations to biological systems.

Acknowledgments D.L. was supported by NIH grants R01 AG079291, R56 LM013784, R01 HL149683, and UM1 TR004406. T.M.Z. was supported by NIH F31HL156464. This work was also supported by grants NSF-2242980 (J.E.P.), R01-GM138834 (J.E.P.), and R01-CA280482 (J.E.P.).

References

1. I. Gomes, C. Abreu, L. Costa and S. Casimiro, The evolving pathways of the efficacy of and resistance to cdk4/6 inhibitors in breast cancer, *Cancers* **15**, p. 4835 (2023).
2. S. J. Altschuler and L. F. Wu, Cellular heterogeneity: do differences make a difference?, *Cell* **141**, 559 (2010).
3. H. X. Chao, R. I. Fakhreddin, H. K. Shimerov, K. M. Kedziora, R. J. Kumar, J. Perez, J. C. Limas, G. D. Grant, J. G. Cook, G. P. Gupta *et al.*, Evidence that the human cell cycle is a series of uncoupled, memoryless phases, *Molecular systems biology* **15**, p. e8604 (2019).
4. C. Liu, Y. Konagaya, M. Chung, L. H. Daigh, Y. Fan, H. W. Yang, K. Terai, M. Matsuda and T. Meyer, Altered g1 signaling order and commitment point in cells proliferating without cdk4/6 activity, *Nature Communications* **11**, p. 5305 (2020).
5. H. W. Yang, S. D. Cappell, A. Jaimovich, C. Liu, M. Chung, L. H. Daigh, L. R. Pack, Y. Fan, S. Regot, M. Covert *et al.*, Stress-mediated exit to quiescence restricted by increasing persistence in cdk4/6 activation, *Elife* **9**, p. e44571 (2020).
6. T. M. Zikry, S. C. Wolff, J. S. Ranek, H. M. Davis, A. Naugle, N. Luthra, A. A. Whitman, K. M. Kedziora, W. Stallaert, M. R. Kosorok *et al.*, Cell cycle plasticity underlies fractional resistance to palbociclib in er+/her2- breast tumor cells, *Proceedings of the National Academy of Sciences* **121**, p. e2309261121 (2024).
7. S. Yu, T. Kim, K. H. Yoo and K. Kang, The t47d cell line is an ideal experimental model to elucidate the progesterone-specific effects of a luminal a subtype of breast cancer, *Biochemical and Biophysical Research Communications* **486**, 752 (2017).
8. G. Gut, M. D. Herrmann and L. Pelkmans, Multiplexed protein maps link subcellular organization to cellular states, *Science* **361**, p. eaar7042 (2018).
9. K. R. Moon, D. Van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. v. d. Elzen, M. J. Hirn, R. R. Coifman *et al.*, Visualizing structure and transitions in high-dimensional biological data, *Nature biotechnology* **37**, 1482 (2019).
10. W. Stallaert, K. M. Kedziora, C. D. Taylor, T. M. Zikry, J. S. Ranek, H. K. Sobon, S. R. Taylor, C. L. Young, J. G. Cook and J. E. Purvis, The structure of the human cell cycle, *Cell systems* **13**, 230 (2022).
11. I. T. Jolliffe, *Principal component analysis for special types of data* (Springer, 2002).
12. H. Hotelling, Analysis of a complex of statistical variables into principal components., *Journal of educational psychology* **24**, p. 417 (1933).
13. R. Xiang, W. Wang, L. Yang, S. Wang, C. Xu and X. Chen, A comparison for dimensionality reduction methods of single-cell rna-seq data, *Frontiers in genetics* **12**, p. 646936 (2021).
14. H. Luo, J. E. Purvis and D. Li, Spherical rotation dimension reduction with geometric loss functions, *Journal of Machine Learning Research* **25**, 1 (2024).
15. D. Li, M. Mukhopadhyay and D. B. Dunson, Efficient manifold approximation with spherelets, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**, 1129 (2022).
16. L. Deconinck, R. Cannoodt, W. Saelens, B. Deplancke and Y. Saeys, Recent advances in trajectory inference from single-cell omics data, *Current Opinion in Systems Biology* **27**, p. 100344 (2021).
17. D. Schwabe, S. Formichetti, J. P. Junker, M. Falcke and N. Rajewsky, The transcriptome dynamics of single cells during the cell cycle, *Molecular systems biology* **16**, p. e9946 (2020).
18. A. Riba, A. Oravecz, M. Durik, S. Jiménez, V. Alunni, M. Cerciat, M. Jung, C. Keime, W. M. Keyes and N. Molina, Cell cycle gene regulation dynamics revealed by rna velocity and deep-learning, *Nature communications* **13**, p. 2865 (2022).
19. K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom and S. Dudoit, Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics, *BMC genomics*

19, 1 (2018).
20. J. S. Ranek, W. Stallaert, J. J. Milner, M. Redick, S. C. Wolff, A. S. Beltran, N. Stanley and J. E. Purvis, Delve: feature selection for preserving biological trajectories in single-cell data, *Nature Communications* **15**, p. 2765 (2024).
21. W. Stallaert, S. R. Taylor, K. M. Kedziora, C. D. Taylor, H. K. Sobon, C. L. Young, J. C. Limas, J. Varblow Holloway, M. S. Johnson, J. G. Cook *et al.*, The molecular architecture of cell cycle arrest, *Molecular Systems Biology* **18**, p. e11087 (2022).
22. W. G. Jacoby, Loess:: a nonparametric, graphical tool for depicting relationships between variables, *Electoral studies* **19**, 577 (2000).
23. S. Kim, A. Leong, M. Kim and H. W. Yang, Cdk4/6 initiates rb inactivation and cdk2 activity coordinates cell-cycle commitment and g1/s transition, *Scientific reports* **12**, p. 16810 (2022).
24. J. R. Nevins, The rb/e2f pathway and cancer, *Human molecular genetics* **10**, 699 (2001).
25. C. Norbury and P. Nurse, Animal cell cycles and their control, *Annual review of biochemistry* **61**, 441 (1992).
26. E. Endl and J. Gerdes, The ki-67 protein: fascinating forms and an unknown function, *Experimental cell research* **257**, 231 (2000).
27. M. Sobecki, K. Mrouj, J. Colinge, F. Gerbe, P. Jay, L. Krasinska, V. Dulic and D. Fisher, Cell-cycle regulation accounts for variability in ki-67 expression levels, *Cancer research* **77**, 2722 (2017).
28. T. Evans, E. T. Rosenthal, J. Youngblom, D. Distel and T. Hunt, Cyclin: a protein specified by maternal mrna in sea urchin eggs that is destroyed at each cleavage division, *Cell* **33**, 389 (1983).
29. J. Bollard, V. Miguela, M. R. De Galarreta, A. Venkatesh, C. B. Bian, M. P. Roberto, V. Tovar, D. Sia, P. Molina-Sánchez, C. B. Nguyen *et al.*, Palbociclib (pd-0332991), a selective cdk4/6 inhibitor, restricts tumour growth in preclinical models of hepatocellular carcinoma, *Gut* **66**, 1286 (2017).
30. J. S. Kwon, N. J. Everetts, X. Wang, W. Wang, K. Della Croce, J. Xing and G. Yao, Controlling depth of cellular quiescence by an rb-e2f network switch, *Cell reports* **20**, 3223 (2017).
31. J. M. Lemons, X.-J. Feng, B. D. Bennett, A. Legesse-Miller, E. L. Johnson, I. Raitman, E. A. Pollina, H. A. Rabitz, J. D. Rabinowitz and H. A. Coller, Quiescent fibroblasts exhibit high metabolic activity, *PLoS biology* **8**, p. e1000514 (2010).
32. K. J. Soprano, Wi-38 cell long-term quiescence model system: A valuable tool to study molecular events that regulate growth, *Journal of cellular biochemistry* **54**, 405 (1994).
33. M. T. Herrera-Abreu, M. Palafox, U. Asghar, M. A. Rivas, R. J. Cutts, I. Garcia-Murillas, A. Pearson, M. Guzman, O. Rodriguez, J. Grueso *et al.*, Early adaptation and acquired resistance to cdk4/6 inhibition in estrogen receptor–positive breast cancer, *Cancer research* **76**, 2301 (2016).
34. M. Palafox, L. Monserrat, M. Bellet, G. Villacampa, A. Gonzalez-Perez, M. Oliveira, F. Brasó-Maristany, N. Ibrahimi, S. Kannan, L. Mina *et al.*, High p16 expression and heterozygous rb1 loss are biomarkers for cdk4/6 inhibitor resistance in er+ breast cancer, *Nature communications* **13**, p. 5258 (2022).
35. J. Gray-Bablin, J. Zalvide, M. P. Fox, C. J. Knickerbocker, J. A. DeCaprio and K. Keyomarsi, Cyclin e, a redundant cyclin in breast cancer, *Proceedings of the National Academy of Sciences* **93**, 15215 (1996).
36. P. W. Hinds, S. Mittnacht, V. Dulic, A. Arnold, S. I. Reed and R. A. Weinberg, Regulation of retinoblastoma protein functions by ectopic expression of human cyclins, *Cell* **70**, 993 (1992).
37. C. E. Caldon, C. M. Sergio, J. Kang, A. Muthukaruppan, M. N. Boersma, A. Stone, J. Barraclough, C. S. Lee, M. A. Black, L. D. Miller *et al.*, Cyclin e2 overexpression is associated with endocrine resistance but not insensitivity to cdk2 inhibition in human breast cancer cells, *Molecular cancer therapeutics* **11**, 1488 (2012).

Astrocyte Reactivity Polygenic Risk Score May Predict Cognitive Decline in Alzheimer's Disease

Jared M Phillips[1,2], Julie A Schneider[3], David A Bennett[3], Paul K Crane[4], Shannon L Risacher[5,6], Andrew J Saykin[5,6], Logan C Dumitrescu[1,7], Timothy J Hohman[1,7 †]; the Alzheimer's Disease Neuroimaging Initiative*

[1] Vanderbilt Memory and Alzheimer's Center, Vanderbilt University Medical Center, Nashville, TN, USA
[2] Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, TN, USA
[3] Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA
[4] Department of Medicine, University of Washington, Seattle, WA, USA
[5] Indiana Alzheimer's Disease Research Center, Indiana University School of Medicine, Indianapolis, IN, USA
[6] Stark Neurosciences Research Institute, Indiana University School of Medicine, Indianapolis, IN, USA
[7] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA
† Email: timothy.j.hohman@vumc.org
† Address: 3319 West End Ave, Suite 847
Nashville, TN 37203

Alzheimer's disease (AD) is a polygenic disorder with a prolonged prodromal phase, complicating early diagnosis. Recent research indicates that increased astrocyte reactivity is associated with a higher risk of pathogenic tau accumulation, particularly in amyloid-positive individuals. However, few clinical tools are available to predict which individuals are likely to exhibit elevated astrocyte activation and, consequently, be susceptible to hyperphosphorylated tau-induced neurodegeneration. Polygenic risk scores (PRS) aggregate the effects of multiple genetic loci to provide a single, continuous metric representing an individual's genetic risk for a specific phenotype. We hypothesized that an astrocyte activation PRS could aid in the early detection of faster clinical decline. Therefore, we constructed an astrocyte activation PRS and assessed its predictive value for cognitive decline and AD biomarkers (i.e., cerebrospinal fluid [CSF] levels of Aβ1-42, total tau, and p-tau181) in a cohort of 791 elderly individuals. The astrocyte activation PRS showed significant main effects on cross-sectional memory ($\beta = -0.07$, $p = 0.03$) and longitudinal executive function ($\beta = -0.01$, $p = 0.03$). Additionally, the PRS interacted with amyloid positivity (p.intx = 0.02), whereby indicating that amyloid burden modifies the association between the PRS and annual rate of language decline. Furthermore, the PRS was negatively associated with CSF Aβ1-42 levels ($\beta = -3.4$, $p = 0.07$) and interacted with amyloid status, such that amyloid burden modifies the association between the PRS and CSF phosphorylated tau levels (p.intx = 0.08). These findings suggest that an astrocyte activation PRS could be a valuable tool for early disease risk prediction, potentially enabling intervention during the interval between pathogenic amyloid and tau accumulation.

Keywords: Alzheimer's disease, polygenic risk, astrocyte reactivity, cognition, biomarkers

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but most of them did not participate in the analysis or writing of this report. A complete listing of ADNI investigators may be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1. Introduction

Alzheimer's disease (AD) is a highly polygenic condition characterized by a neuropathological sequence of extracellular amyloid-beta plaques and intracellular neurofibrillary tangles that leads to neurodegeneration and cognitive decline [12]. A distinguishing feature of AD is its prolonged prodromal phase, during which pathology accumulates well before clinical symptoms manifest [2, 14]. This prodromal period spans decades of pathological changes prior to the onset of noticeable cognitive deficits, making early diagnosis of clinical dementia both challenging and crucial in developing precision interventions. Polygenic risk scores (PRS) of AD have displayed some utility in predicting the global genetic risk of developing AD [5] yet demonstrate mixed success clinically [8, 10, 22, 26]. This may be partly due to the case-control genome-wide association study (GWAS) designs used to generate summary statistics that enable PRS calculation, which lack the phenotypic specificity needed to move towards precision interventions.

Astrocyte activation plays a varied and complex role in AD, with numerous detrimental functions that may contribute to disease pathogenesis including induction of tau hyperphosphorylation, impairment of glutamate and ion buffering abilities, and weakening of the neurovascular unit [13, 15, 16, 28]. Recent evidence has emerged that highlights astrocyte activation as an important cellular event linking initial amyloid pathology with subsequent phosphorylated tau accumulation [3]. Most notably, recent findings leveraging *in vivo* measurements of peripheral glial fibrillary acidic protein (GFAP), a strong correlate of astrocyte activation, found that high plasma GFAP expression, representing a greater degree of astrocyte reactivity, relates to higher AD neuropathological burden [3, 29]. This association was most pronounced in amyloid-positive individuals [3]. In acute brain injury, astrocyte reactivity is both beneficial and detrimental, contributing significantly to post-traumatic tissue repair and synaptic remodeling in conditions such as traumatic brain injury and stroke [4] while also facilitating release of pro-inflammatory factors that may exacerbate cognitive decline [19]. As such, the level of chronic astrocyte activation, particularly in the presence of amyloid pathology, may influence an individual's risk of subsequently developing tau pathology and dementia. Heterogeneity in astrocyte responses to trauma, whether acute or chronic, points to genetic factors that may influence the molecular response of astrocytes to insult [4, 24]. Consequently, investigating the genetic architecture of astrocyte activation in the context of AD may yield insights beneficial in advancing targeted interventions for individuals at risk of developing the detrimental effects of long-term reactive states.

In this study, we sought to accomplish three main aims: 1) to elucidate the genetic architecture of an astrocyte activation phenotype, 2) to build a PRS of astrocyte activation, and 3) to test its ability to predict cognitive decline and associations with AD biomarker levels. Using post-mortem measures of mRNA sequencing from the dorsolateral prefrontal cortex, we calculated an established astrocyte activation transcript signature [33]. Then, we employed this transcript

signature as an outcome in GWAS to identify genetic signals associated with the astrocyte activation phenotype. Finally, we built a PRS in an independent dataset to test its associations with cognitive performance in multiple domains and AD biomarker burden.

2. Methods

2.1. Participants

Participants were sourced from two well-characterized cohort studies of aging, including the Religious Orders Study/Rush Memory and Aging Project (ROS/MAP) and the Alzheimer's Disease Neuroimaging Initiative (ADNI). Data collection commenced in 1994 for ROS and in 1997 for MAP, resulting in extensive longitudinal clinical-pathologic data on aging and AD risk factors. ROS includes religious clergy members from across the United States, while MAP includes individuals from northeastern Illinois. Initiated in 2003, ADNI encompasses over 1800 individuals between 55 to 90 years old, through four study phases, with the principal objective of validating biomarkers for Alzheimer's disease clinical trial applications (http://adni.loni.usc.edu/). All participants provided informed consent and the studies were carried out in accordance with Institutional Review Board-approved protocols. The Vanderbilt University Medical Center Institutional Review Board authorized secondary analyses of the data. Data were accessed and harmonized as part of the Alzheimer's Disease Sequencing Project Phenotype Harmonization Consortium (https://adsp.niagads.org/). Please see **Table 1** for an overview of each cohort's participant demographics.

Table 1. Participant Demographics

ROS/MAP	
Sample Size	598
Age at death (years)	81.1 +/- 6.97
Education (years)	16.53 +/- 3.5
Astrocyte Activation Z Score	0 +/- 0.61
Female, no. (%)	391 (65%)
Amyloid Positive at Autopsy, no. (%)	383 (64%)
Tau Positive at Autopsy, no. (%)	340 (57%)
AD diagnosis at last visit, no. (%)	252 (42%)
ADNI	
Sample Size	791
Age at baseline (years)	75.31 +/- 7.39
Education (years)	16.03 +/- 2.84
Total number of visits	6.32 +/- 2.93
Longitudinal follow-up (years)	4.89 +/- 3.51
Female, no. (%)	342 (43%)
Amyloid Positive at baseline, no. (%)	527 (67%)
Tau Positive at baseline, no. (%)	385 (49%)
AD diagnosis at baseline, no. (%)	196 (25%)

2.2. Cerebrospinal fluid biomarker measures

Lumbar puncture was performed as described in the ADNI procedures manual (http://www.adni-info.org/). CSF measures of β-amyloid(1–42) were obtained using the xMAP platform and CSF measures of total tau and p-tau 181 were obtained using the Elecsys platform. Amyloid positivity was defined as CSF β-amyloid(1–42) concentrations lower than 192 pg/mL as outlined previously [31].

2.3. Neuropsychological composites

Harmonized scores representing composite memory, executive function, and language were used in the present analyses and have been previously described in detail [25]. Briefly, this harmonization process involved experts assigning individual test item-level data into memory, executive function, language, visuospatial, or "none of" domains. Investigators ensured identical scoring of anchor items across studies and a confirmatory factor analysis was conducted to choose the best single factor or bi-factor model. Anchor items were items identified as having been administered and scored precisely the same way in two or more cohorts. All items had freely estimated parameters, with anchor items forced to have the same parameters across studies. We used these co-calibrated parameters for anchor and study-specific items to generate cognitive scores that were on the same scale across cohorts.

2.4. Genetic data quality control and imputation

For ADNI, genetic data were collected with four arrays (Illumina Human610-Quad, Illumina HumanOmniExpress, Illumina Omni 2.5 M, and Illumina Global Screening Array v2). For ROSMAP, genetic data were collected with three arrays (Global Screening Array-24 v3.0, Affymetrix GeneChip 6.0, Illumina HumanOmniExpress). All genetic data were processed using a standardized quality control and imputation pipeline [7]. First, variants which had a low genotype rate (<95%), low minor allele frequency (MAF<1%) or were outside of Hardy-Weinberg equilibrium ($p<1\times10^{-6}$) were removed. Participants were excluded if the reported and genotypic sex differed, if there was poor genotyping efficiency (missing>1% of variants), or cryptic relatedness was present (*PIHAT*>0.25). Imputation was performed on the University of Michigan Imputation Server using the TOPMed reference panel (hg38) with SHAPEIT phasing [6, 11, 32]. Following imputation, datasets were filtered to exclude variants with low imputation quality (R^2<0.8), duplicated/multi-allelic variants, and MAF<1%. Within the self-identified non-Hispanic White racial group, principal component analysis was conducted and genetic ancestry outliers relative to a 1000 Genomes reference population (eg. Utah residents with Northern and Western European Ancestry [CEU]) were excluded.

2.5. Autopsy measures of DLPFC bulk mRNA expression

A standardized protocol for post-mortem biological specimens was used consistently across centers performing autopsies, as previously described [1]. RNA extraction from specific brain regions was conducted using a Qiagen miRNeasy mini kit along with a RNase-free DNase Set for quantification on a Nanodrop. The integrity and purity of the RNA were assessed using an Agilent Bioanalyzer. Samples with a RIN score greater than five were included for bulk next-generation RNA sequencing.

Sequencing was performed in multiple phases. Phase one focused on the dorsolateral prefrontal cortex (dlPFC). Phase two added more dlPFC samples and included samples from the posterior cingulate cortex (PCC) and the head of the caudate nucleus (CN). Phase three included additional

participant samples from the dlPFC. Detailed information on RNA processing and sequencing is available on Synapse (syn3388564). In summary, phase one employed poly-A selection, strand-specific dUTP library preparation, and Illumina HiSeq with 101 bp paired-end reads, achieving a coverage of 150 million reads for the first 12 reference samples. These deeply sequenced reference samples included 2 males and 2 females from non-impaired, mild cognitive impairment, and Alzheimer's disease cases. The remaining samples were sequenced with a coverage of 50 million reads. Phase two used the KAPA Stranded RNA-Seq Kit with RiboErase (kapabiosystems) for ribosomal depletion and fragmentation. Sequencing for this phase was performed on an Illumina NovaSeq6000 with 2 × 100 bp cycles, targeting 30 million reads per sample. In phase three, RNA was extracted with a Chemagic RNA tissue kit (Perkin Elmer, CMG-1212) using a Chemagic 360 instrument, and ribosomal RNA was depleted using RiboGold (Illumina, 20,020,599). Sequencing for phase three was carried out on an Illumina NovaSeq6000 with 40-50 million 2 × 150 bp paired-end reads.

Data processing and QC of RNA sequencing runs was performed by the Vanderbilt Memory and Alzheimer's Center Computational Neurogenomics Team using an automated pipeline and is described in detail elsewhere [30]. Samples whose last visit was >5 years before death or who had non-AD dementia were excluded.

2.6. Statistical analyses

See **Figure 1** for an overview of analytical activities.

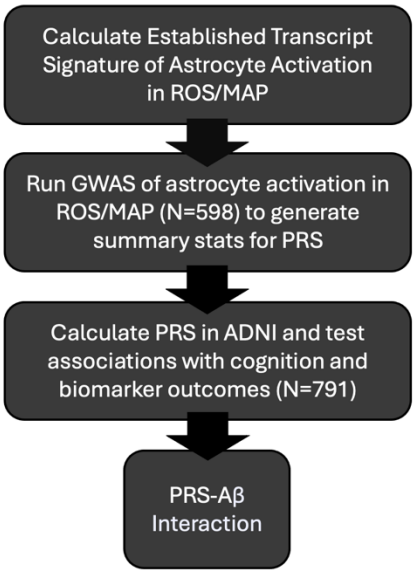


Figure 1. Workflow outlining analytical activities.

2.6.1. Astrocyte reactivity z-score calculation

Methods for generating an astrocyte reactivity z-score were derived from procedures reported by Wu et al [33]. Briefly, single-nucleus RNA sequencing measures from the dorsolateral prefrontal cortices of 24 participants, representing 162,562 individual nuclei, were clustered into transcriptionally similar clusters using a k-nearest neighbor graph. Further dimensionality reduction occurred through t-SNE and expression of canonical genes, including *AQP4* for astrocytes, was used to identify cell type clusters. This analysis was then repeated within the astrocyte cluster, resulting in ten astrocyte sub-clusters. Next, the expression of genes characteristic of reactive astrocytes as reported in Zamanian et al [34]., including *GFAP*, *CD44*, *OSMR*, and *CHI3L1*, was surveyed, resulting in the identification of three sub-clusters that displayed high expression of all four genes. Differential gene expression was assessed using Seurat to obtain marker genes for these activated astrocyte clusters. Genes were required to be expressed in at least 10% of nuclei in the given cluster and at least log(0.25)-fold difference between the clusters.

Genes that were significantly over-expressed in reactive astrocytes compared to both other astrocyte clusters and all other cells were preserved in the marker gene-set (n=25).
Next, we obtained normalized bulk mRNA sequence counts from the ROS/MAP dorsolateral prefrontal cortex dataset, which did not overlap with the snRNA sequencing dataset used to identify reactive astrocyte markers. Four genes were unavailable due to quality control filtering, resulting in 21 genes in the final gene set. Participants with values for all 21 genes were included, leading to a sample size of 843 individuals. Finally, a summary z-score representing higher or lower-than-average reactive astrocyte gene expression was calculated to leverage as an outcome in downstream analyses.

2.6.2. Genome-wide association study of astrocyte reactivity
Following generation of the astrocyte reactivity z-score, we conducted a GWAS to assess the effect of genetic variants on astrocyte reactivity. GWAS were performed with PLINK linear association models (v1.90b5.2, https://www.cog-genomics.org/plink/1.9). 646 participants in ROS/MAP had both genetic data and an astrocyte reactivity z-score. We excluded a random sample of 48 participants from GWAS to later assess the correlation of the astrocyte reactivity z-score and PRS in these individuals, resulting in a final sample size of 598 participants in GWAS. GWAS covariates included RNA-sequencing batch, RNA sequencing sample collection phase, age at death, sex, and the first five principal components of genetic ancestry.

2.6.3. Polygenic risk score generation
No participants in ADNI were included in the astrocyte reactivity GWAS. First, GWAS variants were compared to the ADNI genetic data. Any ambiguous, palindromic variants were filtered out. Then overlapping variants between the GWAS and the ADNI genetic data were retained and subsequently compared for variants on opposite strands between the GWAS and the genetic data, and strand differences were resolved. Then, linkage disequilibrium (LD) clumping was performed with PLINK in the ADNI genetic data (r^2=0.5, window=250kb), to choose the variant with the most significant phenotypic association within each genetically-linked genomic region. Each PRS was built with three different P-value thresholds: P=0.01, P=0.001, and P=0.00001, wherein variants were included in the PRS only if their phenotypic association was less than the given threshold. The LD-clumped genetic data were then leveraged to calculate each PRS with PLINK's profile function which calculates scores as follows: Weights were retrieved from the variant associations with AD or with resilience from the respective GWAS. For each variant the given weight was multiplied by 0, 1, or 2, based on how many risk alleles an individual possessed. The summation of this process results in a summary score for an individual.
Since *APOE* polymorphism is a robust risk factor for AD, PRS were calculated with and without the *APOE* region, defined by a 1Mb region up and downstream of the *APOE* gene.

2.6.4. Baseline and longitudinal linear association models
We performed a series of linear fixed and linear mixed effects models in R (v. 4.1.2) for each PRS. Fixed effects in our models included baseline age, sex, and the given PRS. Longitudinal linear

mixed effects models included a PRS-by-interval term, where interval was determined by the difference between a participant's age at each cognitive visit and their baseline age. Additionally, linear mixed effects models allowed slope and intercept to vary for each participant. In addition, we performed identical sets of models with the addition of a PRS-by-amyloid term in linear models and a PRS-by-amyloid-by-interval term for linear mixed effects models, with amyloid measured by the CSF Aβ1–42 assay outlined above. Biomarker-based outcomes of our models were cross-sectional CSF Aβ1–42, CSF total tau, CSF p-tau 181. Cognition-based outcomes of our models were baseline memory, executive function, and language, or longitudinal decline in memory, executive function, and language, using linear and linear mixed effects models, respectively. We re-ran significant or near-significant interaction models as amyloid-stratified models to obtain main effect statistics for amyloid positive (N=527) and amyloid negative (N=257) individuals. We also conducted sensitivity analyses using data-driven cutpoints determined by Gaussian mixture modeling (GMM) to reevaluate amyloid positivity within our sample (amyloid positivity defined as CSF β-amyloid(1–42) concentrations lower than 195 pg/mL; amyloid positive N = 520, amyloid negative N = 264).

3. Results

The 21 genes included in the astrocyte activation gene module were positively correlated with one another, with the exceptions of *ARGHEF3* and *ZFYVE28* (**Supplemental Figure 1**). We subsequently ran GWAS to generate summary statistics to be leveraged in the PRS calculation. GWAS results highlighted loci on chromosomes 2, 6, 7, and 11 with an acceptable genomic inflation factor of 1.0 (**Supplemental Figure 2**). To evaluate the correlation of each PRS with the astrocyte reactivity Z-score, we built the PRS with a variety of p-value cutoffs in a subset of 48 random participants in ROS/MAP who possessed astrocyte reactivity Z-scores but were excluded from GWAS. The correlation was by far the strongest in the PRS with p-value cutoff < 0.01 (0.98; see **Supplemental Figure 3**). As such, subsequent analyses focused only on PRS with this p-value cutoff. The correlation between the PRS and astrocyte activation z-score did not differ when excluding the *APOE* region, and no strong loci were observed in the *APOE* region at the GWAS level (**Supplemental Figure 2** and **Supplemental Figure 3**). Consequently, we leveraged PRS which included the *APOE* region in proximate analyses.

Table 2. PRS Main Effect Model Results

Outcome	β	p
Memory at baseline	-0.07	**0.03**
Executive function at baseline	-0.02	0.43
Language at baseline	-0.03	0.22
Longitudinal memory	-4.6E-3	0.43
Longitudinal executive function	-0.01	**0.03**
Longitudinal language	-2.3E-3	0.67
CSF Aβ1-42 at baseline	-3.4	0.07
CSF total tau at baseline	-0.29	0.87
CSF pTau at baseline	0.68	0.43

We then built the PRS in an independent dataset and evaluated its associations with cross-sectional and longitudinal cognition as well as cross-sectional AD biomarker levels, including CSF Aβ1–42, total tau, and phosphorylated tau. All main effects on cognition and biomarker outcomes are presented in **Table 2** and/or **Figure 2**. The astrocyte activation PRS had significant effects on both cross-sectional memory (**Figure 2A**) and longitudinal executive function (**Figure 2B**), such that a higher PRS was associated with worse cross-sectional memory performance and a faster rate of executive function decline. In addition, the PRS was negatively associated with the CSF Aβ1-42 level (**Figure 2C**), although this result was just below the significance threshold.

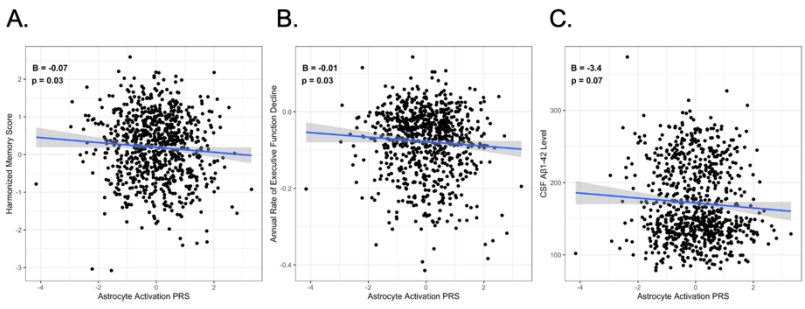


Figure 2. PRS associations with cross-sectional memory, annual rate of executive function decline, and CSF Aβ1–42 level.

Table 3. PRS-Aβ1-42 Interaction Model Results

Outcome	β	p
Memory at baseline	3.7E-4	0.46
Executive function at baseline	2.8E-4	0.54
Language at baseline	3.7E-4	0.37
Longitudinal memory	-7.9E-6	0.93
Longitudinal executive function	1.2E-4	0.13

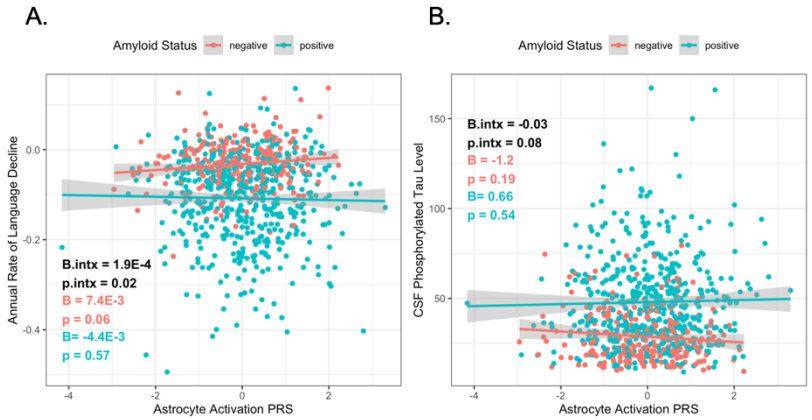


Figure 3. PRS-Aβ42 interactions on annual rate of language decline and CSF phosphorylated tau. Interaction model statistical results are shown in black while amyloid-stratified main effect statistics are shown in colors corresponding to each stratification on the plot.

Next, we performed a series of interaction models to determine if amyloid status modified the effect of the PRS on each outcome (**Table 3** and **Figure 3**). Effects of the PRS on annual rate of language decline differed across amyloid status, and amyloid-negative individuals largely drove the significant interaction (**Figure 3A**). Effects of the PRS on CSF phosphorylated tau level also differed across amyloid status, with the near-significant interaction being driven by deviations between amyloid-negative and amyloid-positive individuals with higher PRS (**Figure 3B**). Results were consistent across both the predefined amyloid positivity threshold and the threshold generated through GMM (**Supplemental Figure 4**). Together, these results suggest a differential effect of the PRS when stratified by amyloid status.

4. Discussion

The findings from our study underscore the potential of an astrocyte activation polygenic risk score (PRS) in the preclinical detection and risk stratification of Alzheimer's disease (AD). Together, our results highlight several critical points that add to the growing body of literature on the role of astrocytes in AD pathology and suggest practical applications for astrocyte activation PRS in clinical settings.

4.1. Genetic architecture of astrocyte activation

We leveraged an established transcript signature of astrocyte activation to serve as a single, continuous outcome in GWAS. Interestingly, the top locus, rs17416058, located on chromosome 11, is an expression quantitative trait locus in brain for *ARNTL* (alias: *BMAL1*), a circadian clock gene (Sources: Braineac and BrainSeq databases). Astrocyte-specific deletion of *BMAL1* has been shown to induce astrocyte activation, indicating a crucial role of circadian rhythm in regulating astrocytic gene expression [18]. Furthermore, astrocytes deficient in *BMAL1* display an enhanced response to amyloid-beta pathology, signaling disease-relevant changes in the face of altered gene expression [23]. Carriage of the minor allele is associated with decreased expression of *BMAL1* in the BrainSeq hippocampus dataset and a higher astrocyte activation transcript signature ($\beta = 0.25$, $p = 1.3E\text{-}7$), which is in line with the observed direction of effect in the aforementioned biological literature. As such, *BMAL1* may represent an important genomic locus influencing an individual's degree of astrocyte reactivity, though this finding requires validation in a well-powered dataset.

4.2. Predictive utility of astrocyte activation PRS

The constructed astrocyte activation PRS demonstrated predictive value for cognitive decline, providing a potential genetic tool to anticipate AD progression. The significant associations between higher PRS and both cross-sectional memory ($\beta = -0.07$, $p = 0.03$; **Figure 2A**) and longitudinal executive function decline ($\beta = -0.01$, $p = 0.03$; **Figure 2B**) suggest that individuals with a higher genetic predisposition for astrocyte activation exhibit worse cognitive performance cross-sectionally and over time. These findings align with previous research indicating that astrocyte reactivity exacerbates neurodegeneration and cognitive impairment [9, 17, 27, 29]. Furthermore, the negative associations between the astrocyte activation PRS and CSF amyloid-beta 1-42 levels ($\beta = -3.4$, $p = 0.07$; **Figure 2C**) provide additional insights into the biological underpinnings of AD. Although the result was marginally below the significance threshold, it suggests that higher genetic risk for astrocyte activation is associated with lower CSF amyloid-beta 1-42 levels, potentially reflecting greater amyloid plaque burden in the brain. This association aligns with the hypothesis that astrocyte activation is linked to amyloid pathology and subsequent neurodegenerative processes [3].

4.3. Interaction with amyloid positivity

The interaction between the astrocyte activation PRS and amyloid positivity highlights a nuanced understanding of AD pathology. In the case of annual rate of language decline, the significant

interaction appears to largely be driven by the effect in amyloid-negative individuals, such that higher PRS relates to a slower rate of language decline (**Figure 3A**). We observed a smaller effect in amyloid-positive individuals, though both stratifications aligned with the anticipated directions of effect. In the case of CSF phosphorylated tau levels, a stronger effect was also observed in amyloid-negative individuals (**Figure 3B**). However, the difference in the directions of effect between amyloid-negative and amyloid-positive individuals drives the near-significant interaction. This suggests that the astrocyte activation PRS may identify individuals who are more susceptible to tau pathology in the presence of amyloid accumulation and a potential protective effect of astrocyte activation in the absence of amyloid pathology. It is plausible that increased astrocyte reactivity in the absence of amyloid pathology may lead to decreased neurodegeneration and subsequent cognitive decline, as reactive astrocytes are known to excrete various growth factors that maintain neuronal and synaptic integrity [20]. However, further interrogating this effect would require more precise transcriptional and morphological profiling of reactive astrocytes in the presence and absence of amyloid pathology, an area ripe for future investigation.

4.4. Clinical implications and future directions

The astrocyte activation PRS holds promise as a clinical tool for early AD risk stratification and intervention. By identifying individuals at higher genetic risk for astrocyte activation, clinicians can better predict the trajectory of cognitive decline and tailor preventive strategies accordingly. Furthermore, the PRS can aid in the selection of candidates for clinical trials targeting astrocyte-mediated pathways, thereby enhancing the precision and efficacy of therapeutic interventions. Future research should focus on refining the astrocyte activation PRS by genetically surveying the astrocyte activation transcript signature in larger, harmonized datasets to increase statistical power at the GWAS level. Validation of its predictive power in large, diverse cohorts would also be greatly beneficial. Additionally, exploring the mechanistic pathways linking astrocyte activation to amyloid and tau pathology will deepen our understanding of AD etiology and to what extent astrocyte activation is genetically regulated. Finally, newer tools allowing for more robust quantification of astrocyte activation *in vivo* using positron emission tomography tracers could serve as a complementary approach to the transcript signature leveraged here and increase statistical power in future studies [21].

4.5. Strengths and weaknesses

Our study has numerous strengths. We leveraged multiple well-characterized, deeply phenotyped cohort studies of aging to first determine the genetic architecture of astrocyte activation and then validate a PRS in predicting clinically relevant outcomes. Incorporating longitudinal measures of cognition and both amyloid and tau biomarker outcomes in our analyses allowed us to survey associations across the amyloid/tau/neurodegeneration framework. Despite its strengths, our study has notable weaknesses. Primarily, we were underpowered at the GWAS level due to the nature of building the astrocyte activation transcript signature from mRNA transcript sequencing from post-

mortem brain tissue. Harmonization of brain transcriptomics across cohorts will enable higher-powered analyses in the future. Our study was also limited to individuals of Western European ancestry, limiting the generalizability of our findings to more diverse populations. We will be better equipped to investigate the utility of an astrocyte activation PRS in diverse populations as more data becomes available. In addition, we chose to employ a data-driven approach leveraging a previously published transcript signature of astrocyte activation [33]. However, a theory-driven approach could provide additional opportunities for discovery. Notably, key astrocyte genes known to be upregulated in reactive states were excluded from the transcript signature we used in our analyses. Potential candidates include: *GFAP*, *Serpina3n*, *VIM*, *AQP4*, and *Lcn2*, which are commonly upregulated in reactive astrocytes [34]. Future analyses incorporating such genes into the gene module will allow us to evaluate whether the inclusion of additional genes captures more of the polygenic architecture of astrocyte reactivity and improves the predictive ability of the PRS. Furthermore, the p-value cutoff used for PRS, though strongly correlated with the astrocyte activation transcript signature itself, was selected somewhat arbitrarily. This less-restrictive cutoff likely includes variants with smaller effects, which collectively may explain a large portion of variance in the phenotype. On the other hand, this may increase the risk of overfitting through the inclusion of more SNPs. Newer tools that enable fine-tuning of p-value cutoff selection for PRS will improve statistical power and predictive ability in future analyses. Furthermore, Since LD structure in the dataset used to build the PRS is likely playing a critical role in the relationship between the PRS and the astrocyte activation phenotype, assessing different R^2 thresholds when using meta-analysis results leveraging multiple cohorts will be an important part of future work. Finally, none of the observed associations survived correction for multiple comparisons, potentially due to the GWAS's power and sample size constraints. This will also be aided by the ever-increasing availability of brain transcriptomic measures and genetic data.

4.6. Conclusions

In summary, our study supports the potential role of an astrocyte activation PRS in predicting cognitive decline and AD biomarker burden. These findings emphasize the importance of astrocyte reactivity in AD progression and highlight the potential of genetic tools in early disease detection and personalized medicine. Further research and validation in well-powered datasets are needed to fully characterize the clinical utility of an astrocyte activation PRS in treating AD.

5. Acknowledgments

Study data were obtained from the Religious Orders Study/Rush Memory and Aging Project (ROS/MAP) and the Alzheimer's Disease Neuroimaging Initiative (ADNI). ROSMAP data are available at www.radc.rush.edu. ADNI data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging

and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. The present work was further supported by the National Institutes of Health under award numbers F31 AG085980, U24 AG074855, R01 AG059716, and R01 AG073439. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

6. Supplemental Material

Supplemental material is available online at https://doi.org/10.6084/m9.figshare.27181179.v1

References

1. A. Bennett D, A. Schneider J, S. Buchman A, L. Barnes L, A. Boyle P, S. Wilson R (2012) Overview and Findings from the Rush Memory and Aging Project. CAR 9:646–663. doi: 10.2174/156720512801322663

2. Amieva H, Jacqmin-Gadda H, Orgogozo J-M, Le Carret N, Helmer C, Letenneur L, Barberger-Gateau P, Fabrigoule C, Dartigues J-F (2005) The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. Brain 128:1093–1101. doi: 10.1093/brain/awh451

3. Bellaver B, Povala G, Ferreira PCL, Ferrari-Souza JP, Leffa DT, Lussier FZ, Benedet AL, Ashton NJ, Triana-Baltzer G, Kolb HC, Tissot C, Therriault J, Servaes S, Stevenson J, Rahmouni N, Lopez OL, Tudorascu DL, Villemagne VL, Ikonomovic MD, Gauthier S, Zimmer ER, Zetterberg H, Blennow K, Aizenstein HJ, Klunk WE, Snitz BE, Maki P, Thurston RC, Cohen AD, Ganguli M, Karikari TK, Rosa-Neto P, Pascoal TA (2023) Astrocyte reactivity influences amyloid-β effects on tau pathology in preclinical Alzheimer's disease. Nat Med 29:1775–1781. doi: 10.1038/s41591-023-02380-x

4. Burda JE, Bernstein AM, Sofroniew MV (2016) Astrocyte roles in traumatic brain injury. Experimental Neurology 275:305–315. doi: 10.1016/j.expneurol.2015.03.020

5. Clark K, Leung YY, Lee W-P, Voight B, Wang L-S (2022) Polygenic Risk Scores in Alzheimer's Disease Genetics: Methodology, Applications, Inclusion, and Diversity. JAD 89:1–12. doi: 10.3233/JAD-220025

6. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh P-R, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C (2016) Next-generation genotype imputation service and methods. Nat Genet 48:1284–1287. doi: 10.1038/ng.3656

7. Eissman JM, Dumitrescu L, Mahoney ER, Smith AN, Mukherjee S, Lee ML, Scollard P, Choi SE, Bush WS, Engelman CD, Lu Q, Fardo DW, Trittschuh EH, Mez J, Kaczorowski CC, Hernandez Saucedo H, Widaman KF, Buckley RF, Properzi MJ, Mormino EC, Yang HS, Harrison TM, Hedden T, Nho K, Andrews SJ, Tommet D, Hadad N, Sanders RE, Ruderfer DM, Gifford KA, Zhong X, Raghavan NS, Vardarajan BN, Pericak-Vance MA, Farrer LA, Wang LS, Cruchaga C, Schellenberg GD, Cox NJ, Haines JL, Keene CD, Saykin AJ, Larson EB, Sperling RA, Mayeux R, Cuccaro ML, Bennett DA, Schneider JA, Crane PK, Jefferson AL, Hohman TJ (2022) Sex differences in the genetic architecture of cognitive resilience to Alzheimer's disease. Brain 145:2541–2554. doi: 10.1093/brain/awac177

8. Euesden J, Gowrisankar S, Qu AX, St. Jean P, Hughes AR, Pulford DJ (2020) Cognitive Decline in Alzheimer's Disease: Limited Clinical Utility for GWAS or Polygenic Risk Scores in a Clinical Trial Setting. Genes 11:501. doi: 10.3390/genes11050501

9. Ferrari-Souza JP, Ferreira PCL, Bellaver B, Tissot C, Wang Y-T, Leffa DT, Brum WS, Benedet AL, Ashton NJ, De Bastiani MA, Rocha A, Therriault J, Lussier FZ, Chamoun M, Servaes S, Bezgin G, Kang MS, Stevenson J, Rahmouni N, Pallen V, Poltronetti NM, Klunk WE, Tudorascu DL, Cohen AD, Villemagne VL, Gauthier S, Blennow K, Zetterberg H, Souza DO, Karikari TK, Zimmer ER, Rosa-Neto P, Pascoal TA (2022) Astrocyte biomarker signatures of amyloid-β and tau pathologies in Alzheimer's disease. Mol Psychiatry 27:4781–4789. doi: 10.1038/s41380-022-01716-2

10. for the Alzheimer's Disease Neuroimaging Initiative, Schork NJ, Elman JA (2023) Pathway-Specific Polygenic Risk Scores Correlate with Clinical Status and Alzheimer's Disease-Related Biomarkers. JAD 95:915–929. doi: 10.3233/JAD-230548

11. Fuchsberger C, Abecasis GR, Hinds DA (2015) minimac2: faster genotype imputation. Bioinformatics 31:782–784. doi: 10.1093/bioinformatics/btu704

12. Jack CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ (2010) Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. The Lancet Neurology 9:119–128. doi: 10.1016/S1474-4422(09)70299-6

13. Jiwaji Z, Tiwari SS, Avilés-Reyes RX, Hooley M, Hampton D, Torvell M, Johnson DA, McQueen J, Baxter P, Sabari-Sankar K, Qiu J, He X, Fowler J, Febery J, Gregory J, Rose J, Tulloch J, Loan J, Story D, McDade K, Smith AM, Greer P, Ball M, Kind PC, Matthews PM,

Smith C, Dando O, Spires-Jones TL, Johnson JA, Chandran S, Hardingham GE (2022) Reactive astrocytes acquire neuroprotective as well as deleterious signatures in response to Tau and Aß pathology. Nat Commun 13:135. doi: 10.1038/s41467-021-27702-w

14. Johnson DK, Storandt M, Morris JC, Galvin JE (2009) Longitudinal Study of the Transition From Healthy Aging to Alzheimer Disease. Arch Neurol 66. doi: 10.1001/archneurol.2009.158

15. Kim H, Leng K, Park J, Sorets AG, Kim S, Shostak A, Embalabala RJ, Mlouk K, Katdare KA, Rose IVL, Sturgeon SM, Neal EH, Ao Y, Wang S, Sofroniew MV, Brunger JM, McMahon DG, Schrag MS, Kampmann M, Lippmann ES (2022) Reactive astrocytes transduce inflammation in a blood-brain barrier model through a TNF-STAT3 signaling axis and secretion of alpha 1-antichymotrypsin. Nat Commun 13:6581. doi: 10.1038/s41467-022-34412-4

16. Kim J, Yoo ID, Lim J, Moon J-S (2024) Pathological phenotypes of astrocytes in Alzheimer's disease. Exp Mol Med. doi: 10.1038/s12276-023-01148-0

17. Kim J-H, Michiko N, Choi I-S, Kim Y, Jeong J-Y, Lee M-G, Jang I-S, Suk K (2024) Aberrant activation of hippocampal astrocytes causes neuroinflammation and cognitive decline in mice. PLoS Biol 22:e3002687. doi: 10.1371/journal.pbio.3002687

18. Lananna BV, Nadarajah CJ, Izumo M, Cedeño MR, Xiong DD, Dimitry J, Tso CF, McKee CA, Griffin P, Sheehan PW, Haspel JA, Barres BA, Liddelow SA, Takahashi JS, Karatsoreos IN, Musiek ES (2018) Cell-Autonomous Regulation of Astrocyte Activation by the Circadian Clock Protein BMAL1. Cell Reports 25:1-9.e5. doi: 10.1016/j.celrep.2018.09.015

19. Li L, Zhou J, Han L, Wu X, Shi Y, Cui W, Zhang S, Hu Q, Wang J, Bai H, Liu H, Guo W, Feng D, Qu Y (2022) The Specific Role of Reactive Astrocytes in Stroke. Front Cell Neurosci 16:850866. doi: 10.3389/fncel.2022.850866

20. Linnerbauer M, Rothhammer V (2020) Protective Functions of Reactive Astrocytes Following Central Nervous System Insult. Front Immunol 11:573256. doi: 10.3389/fimmu.2020.573256

21. Liu Y, Jiang H, Qin X, Tian M, Zhang H (2022) PET imaging of reactive astrocytes in neurological disorders. Eur J Nucl Med Mol Imaging 49:1275–1287. doi: 10.1007/s00259-021-05640-5

22. Logue MW, Panizzon MS, Elman JA, Gillespie NA, Hatton SN, Gustavson DE, Andreassen OA, Dale AM, Franz CE, Lyons MJ, Neale MC, Reynolds CA, Tu X, Kremen WS (2019) Use of an Alzheimer's disease polygenic risk score to identify mild cognitive impairment in adults in their 50s. Mol Psychiatry 24:421–430. doi: 10.1038/s41380-018-0030-8

23. McKee CA, Lee J, Cai Y, Saito T, Saido T, Musiek ES (2022) Astrocytes deficient in circadian clock gene Bmal1 show enhanced activation responses to amyloid-beta pathology without changing plaque burden. Sci Rep 12:1796. doi: 10.1038/s41598-022-05862-z

24. Monterey MD, Wei H, Wu X, Wu JQ (2021) The Many Faces of Astrocytes in Alzheimer's Disease. Front Neurol 12:619626. doi: 10.3389/fneur.2021.619626

25. Mukherjee S, Choi S-E, Lee ML, Scollard P, Trittschuh EH, Mez J, Saykin AJ, Gibbons LE, Sanders RE, Zaman AF, Teylan MA, Kukull WA, Barnes LL, Bennett DA, Lacroix AZ, Larson EB, Cuccaro M, Mercado S, Dumitrescu L, Hohman TJ, Crane PK (2023) Cognitive domain harmonization and cocalibration in studies of older adults. Neuropsychology 37:409–423. doi: 10.1037/neu0000835

26. Oetjens MT, Kelly MA, Sturm AC, Martin CL, Ledbetter DH (2019) Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. Nat Commun 10:4897. doi: 10.1038/s41467-019-12869-0

27. Pelkmans W, Shekari M, Brugulat-Serrat A, Sánchez-Benavides G, Minguillón C, Fauria K, Molinuevo JL, Grau-Rivera O, González Escalante A, Kollmorgen G, Carboni M, Ashton NJ, Zetterberg H, Blennow K, Suarez-Calvet M, Gispert JD, for the ALFA study (2024) Astrocyte biomarkers GFAP and YKL-40 mediate early Alzheimer's disease progression. Alzheimer's & Dementia 20:483–493. doi: 10.1002/alz.13450

28. Rodríguez-Giraldo M, González-Reyes RE, Ramírez-Guerrero S, Bonilla-Trilleras CE, Guardo-Maya S, Nava-Mesa MO (2022) Astrocytes as a Therapeutic Target in Alzheimer's Disease–Comprehensive Review and Recent Developments. IJMS 23:13630. doi: 10.3390/ijms232113630

29. Sánchez-Juan P, Valeriano-Lorenzo E, Ruiz-González A, Pastor AB, Rodrigo Lara H, López-González F, Zea-Sevilla MA, Valentí M, Frades B, Ruiz P, Saiz L, Burgueño-García I, Calero M, Del Ser T, Rábano A (2024) Serum GFAP levels correlate with astrocyte reactivity, post-mortem brain atrophy and neurofibrillary tangles. Brain 147:1667–1679. doi: 10.1093/brain/awae035

30. Seto M, Weiner RL, Dumitrescu L, Mahoney ER, Hansen SL, Janve V, Khan OA, Liu D, Wang Y, Menon V, De Jager PL, Schneider JA, Bennett DA, Gifford KA, Jefferson AL, Hohman TJ (2022) RNASE6 is a novel modifier of APOE-ε4 effects on cognition. Neurobiology of Aging 118:66–76. doi: 10.1016/j.neurobiolaging.2022.06.011

31. Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, Blennow K, Soares H, Simon A, Lewczuk P, Dean R, Siemers E, Potter W, Lee VM -Y., Trojanowski JQ, Alzheimer's Disease Neuroimaging Initiative (2009) Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Annals of Neurology 65:403–413. doi: 10.1002/ana.21610

32. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, Pitsillides AN, LeFaive J, Lee S-B, Tian X, Browning BL, Das S, Emde A-K, Clarke WE, Loesch DP, Shetty AC, Blackwell TW, Smith AV, Wong Q, Liu X, Conomos MP, Bobo DM, Aguet F, Albert C, Alonso A, Ardlie KG, Arking DE, Aslibekyan S, Auer PL, Barnard J, Barr RG, Barwick L, Becker LC, Beer RL, Benjamin EJ, Bielak LF,

Blangero J, Boehnke M, Bowden DW, Brody JA, Burchard EG, Cade BE, Casella JF, Chalazan B, Chasman DI, Chen Y-DI, Cho MH, Choi SH, Chung MK, Clish CB, Correa A, Curran JE, Custer B, Darbar D, Daya M, de Andrade M, DeMeo DL, Dutcher SK, Ellinor PT, Emery LS, Eng C, Fatkin D, Fingerlin T, Forer L, Fornage M, Franceschini N, Fuchsberger C, Fullerton SM, Germer S, Gladwin MT, Gottlieb DJ, Guo X, Hall ME, He J, Heard-Costa NL, Heckbert SR, Irvin MR, Johnsen JM, Johnson AD, Kaplan R, Kardia SLR, Kelly T, Kelly S, Kenny EE, Kiel DP, Klemmer R, Konkle BA, Kooperberg C, Köttgen A, Lange LA, Lasky-Su J, Levy D, Lin X, Lin K-H, Liu C, Loos RJF, Garman L, Gerszten R, Lubitz SA, Lunetta KL, Mak ACY, Manichaikul A, Manning AK, Mathias RA, McManus DD, McGarvey ST, Meigs JB, Meyers DA, Mikulla JL, Minear MA, Mitchell BD, Mohanty S, Montasser ME, Montgomery C, Morrison AC, Murabito JM, Natale A, Natarajan P, Nelson SC, North KE, O'Connell JR, Palmer ND, Pankratz N, Peloso GM, Peyser PA, Pleiness J, Post WS, Psaty BM, Rao DC, Redline S, Reiner AP, Roden D, Rotter JI, Ruczinski I, Sarnowski C, Schoenherr S, Schwartz DA, Seo J-S, Seshadri S, Sheehan VA, Sheu WH, Shoemaker MB, Smith NL, Smith JA, Sotoodehnia N, Stilp AM, Tang W, Taylor KD, Telen M, Thornton TA, Tracy RP, Van Den Berg DJ, Vasan RS, Viaud-Martinez KA, Vrieze S, Weeks DE, Weir BS, Weiss ST, Weng L-C, Willer CJ, Zhang Y, Zhao X, Arnett DK, Ashley-Koch AE, Barnes KC, Boerwinkle E, Gabriel S, Gibbs R, Rice KM, Rich SS, Silverman EK, Qasba P, Gan W, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Papanicolaou GJ, Nickerson DA, Browning SR, Zody MC, Zöllner S, Wilson JG, Cupples LA, Laurie CC, Jaquish CE, Hernandez RD, O'Connor TD, Abecasis GR (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590:290–299. doi: 10.1038/s41586-021-03205-y

33. Wu R, Tripathy S, Menon V, Yu L, Buchman AS, Bennett DA, De Jager PL, Lim ASP (2023) Fragmentation of rest periods, astrocyte activation, and cognitive decline in older adults with and without Alzheimer's disease. Alzheimer's & Dementia 19:1888–1900. doi: 10.1002/alz.12817

34. Zamanian JL, Xu L, Foo LC, Nouri N, Zhou L, Giffard RG, Barres BA (2012) Genomic Analysis of Reactive Astrogliosis. J Neurosci 32:6391–6410. doi: 10.1523/JNEUROSCI.6221-11.2012

Connecting intermediate phenotypes to disease using multi-omics in heart failure

Anni Moore [*1], Rasika Venkatesh[*1], Michael G. Levin[2], Scott M. Damrauer[3,4], Nosheen Reza[2], Thomas P. Cappola[1,2], Marylyn D. Ritchie[1,4]**

[1]*Genomics and Computational Biology, University of Pennsylvania Perelman School of Medicine, 3700 Hamilton Walk Philadelphia, PA, 19104, USA*

[2]*Division of Cardiovascular Medicine, University of Pennsylvania Perelman School of Medicine, 3400 Civic Center Blvd Philadelphia, PA, 19104, USA*

[3]*Department of Surgery, University of Pennsylvania Perelman School of Medicine, 3400 Spruce St Philadelphia, PA 19104*

[4]*Department of Genetics, University of Pennsylvania Perelman School of Medicine, 3700 Hamilton Walk Philadelphia, PA, 19104, USA*

***Email:* marylyn@pennmedicine.upenn.edu*

**Authors contributed equally to the manuscript*
***Corresponding author*

Heart failure (HF) is one of the most common, complex, heterogeneous diseases in the world, with over 1-3% of the global population living with the condition. Progression of HF can be tracked via MRI measures of structural and functional changes to the heart, namely left ventricle (LV), including ejection fraction, mass, end-diastolic volume, and LV end-systolic volume. Moreover, while genome-wide association studies (GWAS) have been a useful tool to identify candidate variants involved in HF risk, they lack crucial tissue-specific and mechanistic information which can be gained from incorporating additional data modalities. This study addresses this gap by incorporating transcriptome-wide and proteome-wide association studies (TWAS and PWAS) to gain insights into genetically-regulated changes in gene expression and protein abundance in precursors to HF measured using MRI-derived cardiac measures as well as full-stage all-cause HF. We identified several gene and protein overlaps between LV ejection fraction and end-systolic volume measures. Many of the overlaps identified in MRI-derived measurements through TWAS and PWAS appear to be shared with all-cause HF. We implicate many putative pathways relevant in HF associated with these genes and proteins via gene-set enrichment and protein-protein interaction network approaches. The results of this study (1) highlight the benefit of using multi-omics to better understand genetics and (2) provide novel insights as to how changes in heart structure and function may relate to HF.

Keywords: Heart Failure; Cardiovascular; Multi-omics; Gene Transcription; Proteome; Network

1. Introduction

1.1. *Heart failure has a high disease burden.*

Heart failure (HF) carries one of the highest disease burdens in the world, with 1-3% of the global population estimated to be living with HF. This includes 6.7 million people in the United States (US) alone, and does not include the 33% of the US population at-risk of developing HF[1]. The etiology of HF is heterogeneous and complex, but has ultimately been defined as a clinical syndrome with symptoms and signs caused by structural and functional cardiac abnormalities[2]. Its

risk is promoted by increasing age and by the presence of comorbidities such as myocardial infarction, diabetes, hypertension, obesity, arrhythmias, infiltrative and inflammatory disorders, and exposure to drugs or environmental toxins[3–5]. Despite the complexity of HF, it has been demonstrated that risk is impacted by genetic predisposition to some degree[6]. While an exact consensus of heritability for HF has not been reached, some have estimated that the heritability of HF is around 26%[7].

The overall progression of HF can be broken up into four stages:

Stage A: risk of HF but patients have no symptoms or structural heart changes
Stage B: no symptoms of HF or asymptomatic but patients do have structural heart changes
Stage C: patients experiencing symptoms of HF
Stage D: advanced heart failure requiring specialized interventions[8,9].

As noted in Stage A and B, HF is often preceded by a phase of undetected progression, highlighting the need for better insight into the changes, such as structural heart changes[9,10]. These changes often appear specifically at the left ventricle (LV), and include decreased left ventricular ejection fraction (LVEF), LV dilation and/or hypertrophy, and valvular disease in which the heart cannot pump as effectively, losing function. LV mass (LVM) has been shown to be an independent predictor of HF, with risk for HF increasing by 1% for every 1% increase in excess LV mass[11]. Likewise, LVEF, which measures LV contractile function (the percentage of blood leaving the heart with each contraction) has been shown to be associated with HF prognosis[12,13]. Abnormal measurements of several of these parameters measuring both structural and functional changes together are reliable markers of cardiovascular risk and eventual HF diagnosis.

To quantify changes in the volume of blood in the heart before and after contraction, we can use LV end-diastolic volume (LVEDV) and LV end-systolic volume (LVESV) respectively. Together these four measures of heart structure and function (LVM, LVEF, LVEDV, and LVESV) can provide an overall characterization of progression towards potential HF and represent an intermediate phenotype or endophenotype. Identifying overlaps in changes seen in both intermediate MRI trait measures and HF could give us a better idea of vital aspects that lead towards full, advanced HF.

1.2. *Using multi-omics to increase knowledge gained from GWAS.*

Given the known genetic contribution to HF and the prevalence of patients with the disease, many groups have performed genome-wide association studies (GWAS) to identify genetic variants associated with HF[14–21]. While this approach allows us to gain valuable insights into potential genetic variation that contributes to the disease, it still leaves a crucial gap in connecting how these variants are actually resulting in mechanistic change, and in which specific tissues. This is especially relevant in quantitative phenotypes, where GWAS is insufficient to capture the full heterogeneity measured by the trait. Transcriptome-wide association studies (TWAS) use GWAS summary statistics along with reference gene expression from specific tissues to predict how genetic variants affect gene expression within those tissues. TWAS and PWAS also provide a boost in overall statistical power, as they are less affected by multiple test corrections due to being

a gene-based test of association[22,23]. These methods are also more portable than GWAS; they are less impacted by population structure in datasets as they operate on a gene and protein level[24].

In this study, we make use of TWAS and PWAS methods to investigate genetic-derived gene and protein changes among cardiovascular related-tissues using the largest published GWAS summary statistics of HF and MRI measures of LV structure and function to date[25–27]. Our goals are: 1) to integrate multi-omics data in the form of reference gene expression and protein expression datasets to identify novel HF and related trait associated genes, 2) to evaluate whether TWAS and PWAS approaches uncover the same association signals or provide novel gene-based associations, and 3) determine whether these genes associated with HF and related traits are part of shared pathways and/or networks between traits. This study is also, to our knowledge, one of the first times that both TWAS *and* PWAS have been performed simultaneously on quantitative traits.

2. Methods

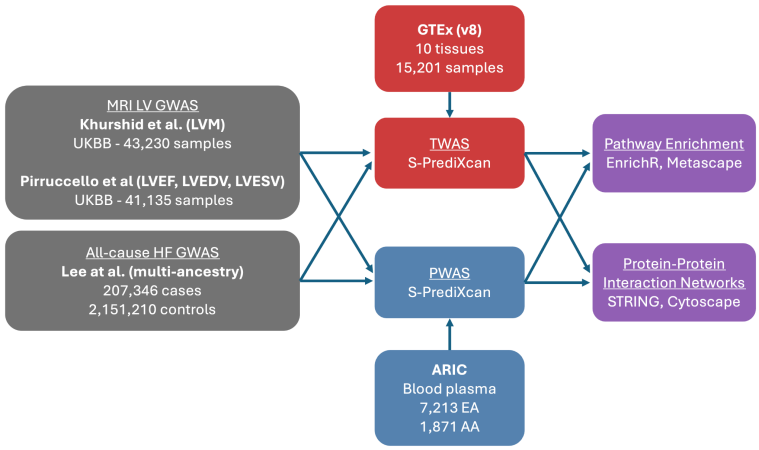


Figure 1. Overview of study analysis to identify genes, proteins, and related interactions between MRI-derived cardiac intermediate traits and heart failure GWAS. (UKBB: UK Biobank, LVM: Left ventricular mass, LVEF: Left ventricular ejection fraction, LVEDV: Left ventricular end-diastolic volume, LVESV: Left ventricular end-systolic volume, EA: European American, AA: African American)

2.1. *Cardiovascular data*

2.1.1. *MRI traits*

Of all cardiac chambers, dysfunction of the left ventricle is the most common structural abnormality in HF cases. We chose four measurements taken from the left ventricle derived from MRI imaging with previously published GWAS data to characterize potential associations with HF: LVM indexed to body surface area, LVEDV, LVESV, and LVEF[26,27]. LVM measurements were taken from Khurshid et al, and includes 43,230 samples (91% European ancestry) with MRI imaging and genotype data from the UK Biobank[26]. LVEDV, LVESV, and LVEF association studies were performed on 41,135 samples also from the UK Biobank with MRI imaging and genotype data by Pirruccello et al[27].

2.1.2. *Heart failure (HF)*

We identified the largest all-cause HF GWAS study to date including 207,346 non-overlapping samples of cases and 2,151,210 controls meta-analyzed from HERMES, the Million Veterans Project (MVP), FinnGen, Mount Sinai BioMe (BIOME), Global Biobank Meta-analysis Initiative (GBMI), eMERGE, Geisinger DiscovEHR, and Penn Medicine BioBank (PMBB)[25]. This included an overall sample of 81.1% European ancestry, 9.7% African American, 6.5% East Asian, and 2.6% Admixed American.

2.2. *Transcriptome-wide association study (TWAS)*

To provide tissue-specific context to GWAS results from the selected MRI traits and HF studies we conducted transcriptome-wide association studies (TWASs) using S-PrediXcan[28] and multivariate adaptive shrinkage (MASHR) eQTL models from the Genotype-Tissue Expression (GTEx) Project v8, available in PredictDB[29,30] GTEx eQTLs were derived from a sample group of mostly European ancestry (84.6% European ancestry, 12.9% African American, 1.3% Asian and 1.1% unknown) that closely parallels the composition of the HF multi-ancestry cohort. Using this reference, we imputed genetically regulated gene expression (GReX) for all genes in ten tissues known to be relevant in the cardiovascular system and heart failure (aorta, coronary artery, tibial artery, atrial appendage, left ventricle, whole blood, visceral adipose, subcutaneous adipose, liver, and kidney)[31–37]. Associations for each of the intermediate MRI traits from the UK Biobank, as well as multi-ancestry and EUR populations from all-cause HF were calculated independently for each of these ten tissues. Significant genes were determined using a Bonferroni threshold of ($p<0.05/$(# genes x 10 tissues tested) per trait. All code is available at https://github.com/RitchieLab/HFmultiomics_PSB2025.

2.3. *Proteome-wide association study (PWAS)*

We performed a proteome-wide association study (PWAS) using S-PrediXcan[28] with the GWAS summary statistics for MRI traits from UK Biobank and for the multi-ancestry and European (EUR) population all-cause HF studies. PWAS identifies genetic associations that may influence complex traits, such as all-cause HF and MRI traits, by regulating protein abundance in tissue[38]. Blood plasma-derived protein quantitative trait loci (cis-pQTLs) from the Atherosclerosis Risk in Communities (ARIC)[39] study were used to construct the models. This large bi-ethnic study was made up of 9,084 participants, consisting of 7,213 European Americans (EA) and 1,871 African Americans (AA). S-PrediXCan PWAS EA and AA models were identified in PredictDB and were constructed using ARIC consortium data by utilizing PEERS covariates, expression information from eQTL associations, gene and SNP annotations[40,41]. PWAS was conducted on multi-ancestry and EUR studies of all-cause HF, as well as on traits from UK Biobank with the intermediate MRI traits using the EA cohort information, and additionally the AA cohort for the multi-ancestry HF study. The resulting PWAS associations were assessed for statistical significance using a Bonferroni significance threshold ($p<0.05/$# proteins tested) for each trait.

2.4. *Network and pathway analyses*

2.4.1. *Pathway enrichment analysis*

Gene set enrichment was performed using EnrichR[42,43] for the significant results from TWAS and PWAS for each MRI trait and HF phenotype, respectively. Enrichment analysis explored the specific pathways and processes associated with the statistically significant genes and proteins from the TWAS and/or PWAS. Pathway results were annotated with KEGG 2021, Reactome 2022, and Gene Ontology (GO) Biological Process 2023 pathways. The significant pathways were identified as having Fisher's exact test p-value < 0.05[38,44].

2.4.2. *Network analysis and identification of hub genes and proteins*

The statistically significant genes and proteins identified via TWAS and PWAS were used to construct a protein-protein interaction (PPI) network using the online Search Tool for the Retrieval of Interacting Genes (STRING v11)[45], where the number of interactions present was assessed for significance. Network interactions were thresholded by a minimum confidence score of > 0.4, as calculated by STRING[45]. The networks were then visualized using Cytoscape 3.10.2[46], and degree centrality analysis was performed using the cytoHubba module to identify and visualize the hub genes and proteins[47,48].

2.4.3. *Classification of of sub-clusters*

Additionally, the Molecular Complex Detection (MCODE)[49] module in Cytoscape was used to screen modules of the larger PPI networks and construct clusters by identifying densely-connected regions of the network[50]. The networks were thresholded to have an MCODE degree cutoff of 3, node density cutoff of 0.1, node score cutoff of 0.2, number of nodes > 3[47,51]. Gene set enrichment analysis using KEGG 2021[52–54], Reactome 2022[55,56], and Gene Ontology (GO) Biological Process 2023[57,58] of each cluster was then conducted using Metascape[59], using the default parameters of minimum overlap of 3, p-value cutoff of 0.01, and minimum enrichment score of 1.5.

3. Results

3.1. *TWAS and PWAS Association Analyses*

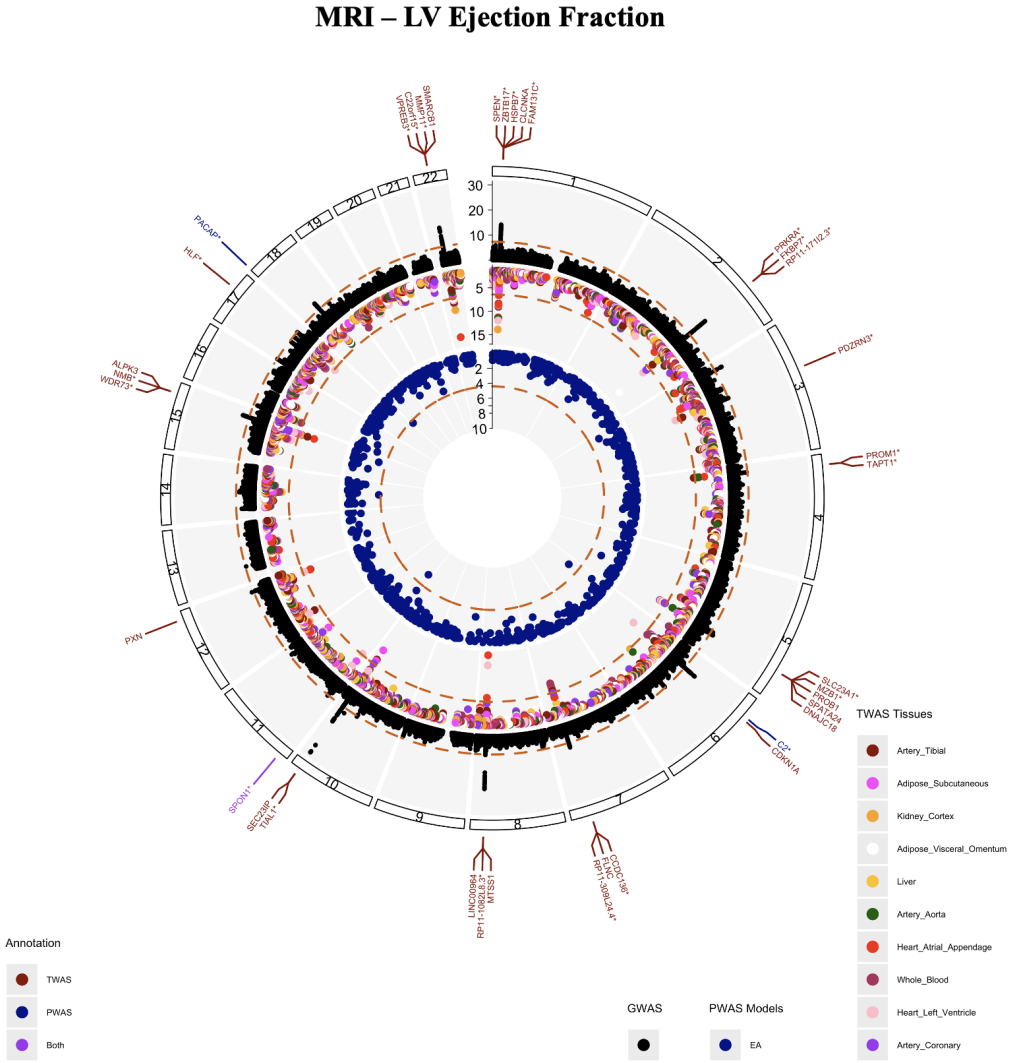
3.1.1 *MRI trait gene and protein associations*

After fine-mapping, 35 unique genes within ten tissues and three proteins from blood plasma (SPON1, C2, PACAP) were significant for LVEF based on a Bonferroni threshold (TWAS:p<3.814E-07, PWAS:3.75E-05) (**Figure 2A**). 16 of these genes were significant in three or more tissues, and one gene, *SPON1*, replicated in both TWAS and PWAS for LVEF. 35 genes and one protein (THBS4) appeared significantly associated with LVM measures (TWAS:p<3.842E-07, PWAS:3.79E-05). Five of these significant genes (*FKBP7*, *WNT3*, *HSPQ4*, *PSMC3*, and *PRKRA*) appeared in three or more tissues tested. Finally, amongst the ten tissues tested, 33 genes and three proteins (ENG, QPCTL, SPON1) were significant for LVEDV (TWAS:p<3.815E-07, PWAS:3.79E-05) and 48 genes along with four proteins (RAB5A, SRL, PACAP, SPON1) for LVESV (TWAS:p<3.815E-07), protein:3.79E-05). *SPON1* was also

significantly associated with LVESV for both TWAS and PWAS. Figures for LVM, LVEDV, LVESV are available in **Supplemental Figure 1A-C**. The full significant results of the TWAS and PWAS for MRI traits are available in **Supplemental Table 1 and 2**.

Between MRI traits, several genes appeared significant between measures. Genes *FKBP7*, *PRKRA*, and *RP11-171I2.3* were associated with all four MRI-based traits in at least one tissue. 15 genes overlapped between LVEDV and LVESV, four genes were shared between LVEDV and LVM, four between LVEDV and LVM, five genes between LVESV and LMV, and finally 30 genes between LVESV and LVEF. Amongst protein results, SPON1 was the only protein shared between traits and was significantly associated with LVEF, LVEDV, and LVESV.

A



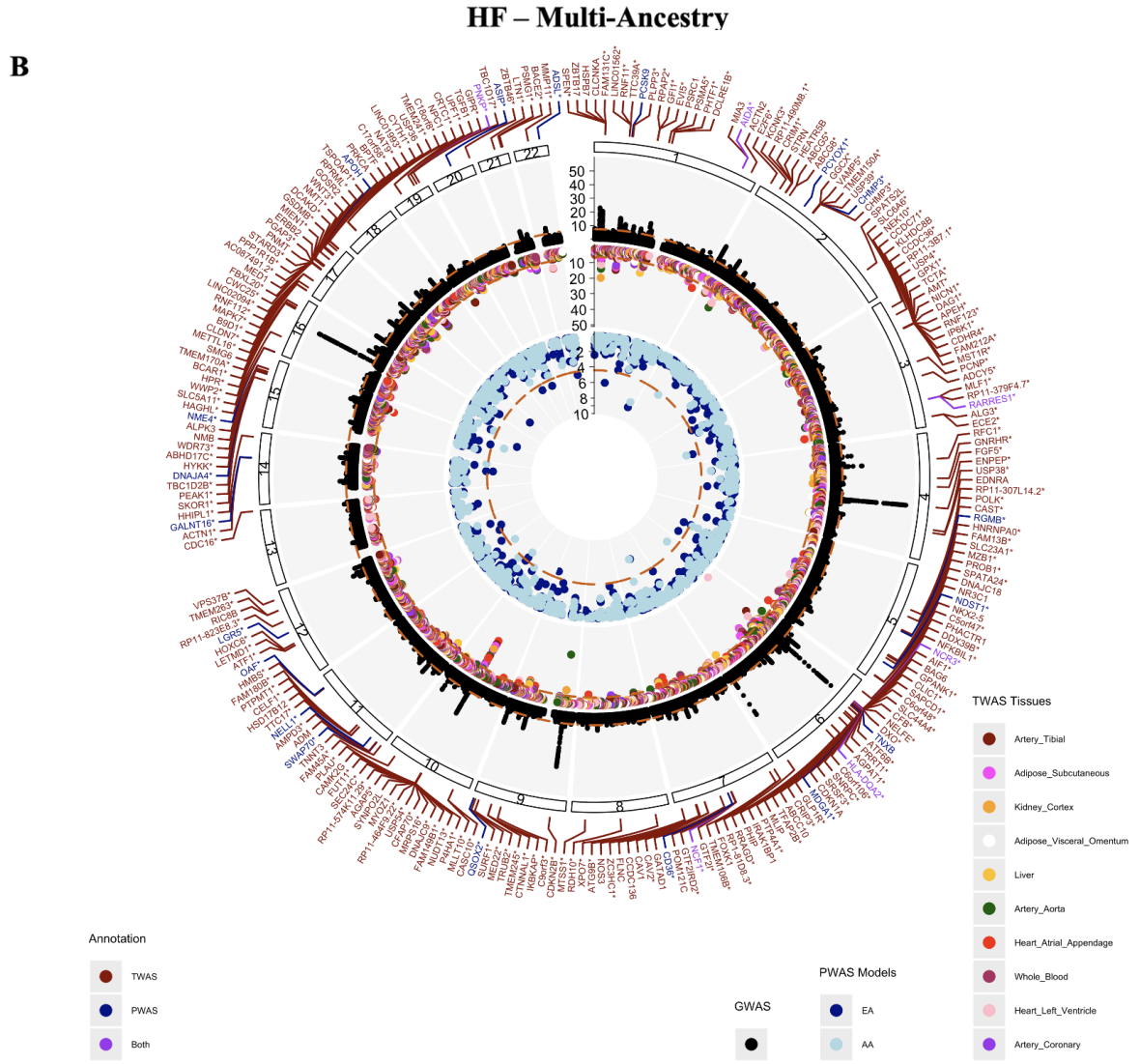


Figure 2. Circos plots for A) MRI LV Ejection Fraction and B) HF - Multi-ancestry representing identified associations through GWAS (black), TWAS (red), and PWAS (blue). The outermost track of annotations depicts genes and proteins identified through TWAS (red) and PWAS (blue), or both analyses (purple), with the asterisk denoting novel results not previously reported from the source GWAS, GWAS Catalog, or NCBI.

3.1.2 *HF gene and protein associations*

231 unique genes and 29 proteins significantly associated with HF in a multi-ancestry population (TWAS:p<3.806E-07, PWAS:p<3.79E-05) (**Figure 2B**). Six genes replicated across genes and proteins, including *RARRES1*, *NCF1*, *AIDA*, *HLA-DQA2*, *PNKP*, and *NCR3*. 185 of the total 231 associating genes were significant in at least one vascular tissue (heart atrial appendage, heart left ventricle, tibial artery, coronary artery, whole blood). 90 genes were significant in at least one vascular tissue and one peripheral tissue (liver, kidney, subcutaneous adipose, visceral adipose). Of the ten tissues tested, heart atrial appendage and heart left ventricle tissues had the largest number of genes significantly associating with HF. Genes *CRIP3* and *USP54* were significant in

all ten tissues tested. Similar associations with HF were noted in the European population (**Supplemental Figure 1D**).

3.2. *Network and Pathway analyses*

3.2.1. *MRI trait gene-set enrichment*

In order to identify the known biologically relevant pathways associated with the statistically significant TWAS and PWAS genes for each phenotype, gene-set enrichment analysis was performed using EnrichR for Reactome 2022, KEGG 2021, and Gene ontology (GO) 2023 pathways. Significant pathways were identified at a p-value < 0.05; the full set of significant pathways for each phenotype are available in **Supplemental Table 3**. For LVEF, the most significant pathway by p-value was positive regulation of actin filament bundle assembly (p = 4.23E-03), made up of genes *PXN* and *MTSS1* (**Figure 3A**). Several pathways involved in kidney development function were also identified to be significant, such as renal cell filtration differentiation (p-value = 9.47E-03), and nephron tubule development (p = 9.47E-03).

LVM (**Supplemental Figure 2A**) was enriched for a variety of cell proliferation and differentiation pathways such as positive regulation of endothelial cell proliferation (p = 1.11E-03) and negative regulation of muscle cell differentiation (p = 4.54E-03), identifying the gene *IGF2* and protein THBS4 as important. Similarly developmentally important cardiovascular pathways were identified for the LVEDV **(Supplemental Figure 2B**), including embryonic hemopoiesis (p = 2.43E-04), megakaryocyte differentiation (p = 4.211E-04), and cardiac atrium morphogenesis (p = 4.73E-04). The most significant pathway for LVESV (**Supplemental Figure 2C**) was modulation by host of symbiont process (p = 5.34E-03), in addition to developmental pathways - glomerular epithelial cell differentiation (p = 1.29E-02), and renal filtration cell differentiation (p = 1.29E-02).

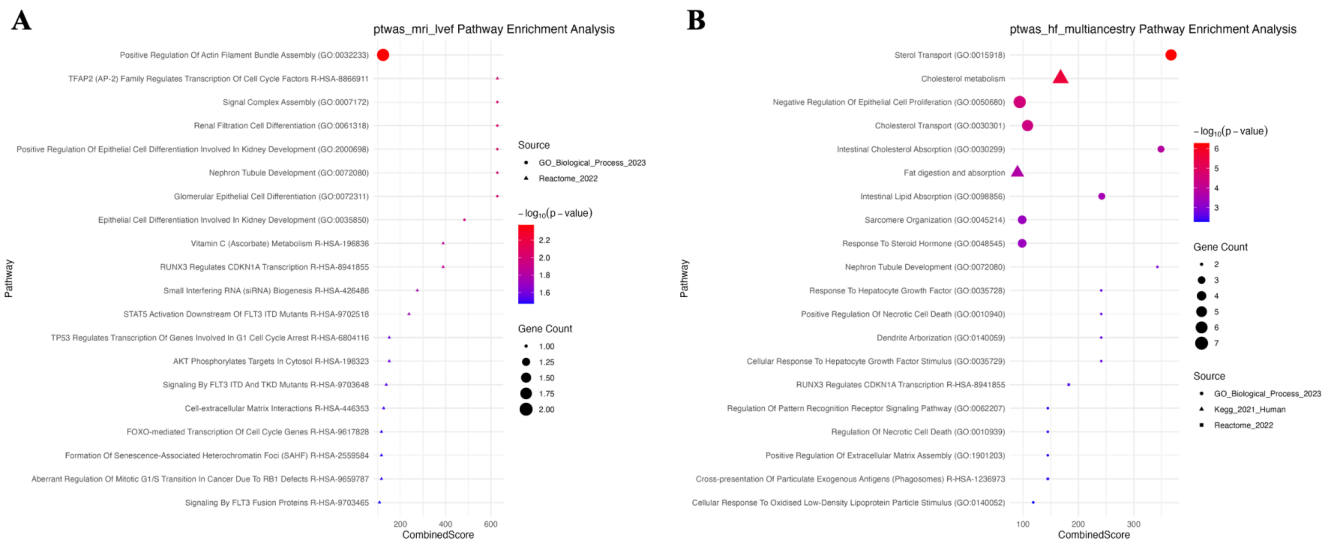


Figure 3. Gene-set enrichment results of TWAS and PWAS significant hits for A) MRI LVEF and B) multi-ancestry HF phenotypes.

3.2.2. *HF gene-set enrichment*

A variety of relevant gene-sets were found to have overrepresented pathways previously identified as important in all-cause HF[60]. For the HF multi-ancestry cohort (**Figure 3B**), the most significant pathways include sterol transport (p = 5.34E-07) and cholesterol metabolism (p = 2.84E-06), which are known to be impacted in a variety of cardiovascular disease states, including heart failure[61,62]. Similar pathways were enriched in the EUR population for HF (**Supplemental Figure 2D**). Genes implicated in these pathways include *ABCG8, STARD3, ABCG5, NPC1, CAV1, APOH, PCSK9,* and *CD36*.

3.2.3. *PPI network analysis of MRI trait genes and proteins*

To evaluate the association of candidate genes and proteins identified by TWAS and PWAS, PPI networks were constructed for each MRI trait phenotype using the STRING database. The PPI network for LVEF contained 32 nodes, 16 of which were connected, and 14 edges at a confidence threshold of > 0.4, with a PPI enrichment p-value = 5.24E-07, indicating that there were significantly more interactions observed than expected by random chance. The hub nodes identified via the cytohubba plugin by degree centrality were *FLNC*, *ALPK3*, *SPATA24*, and *HSPB7* (**Figure 4A**). MCODE identified 1 cluster in the network, with nodes *FLNC*, *HSPB7*, and *ALPK3*, at a score of 1.5, as computed by multiplying node density by the number of members.

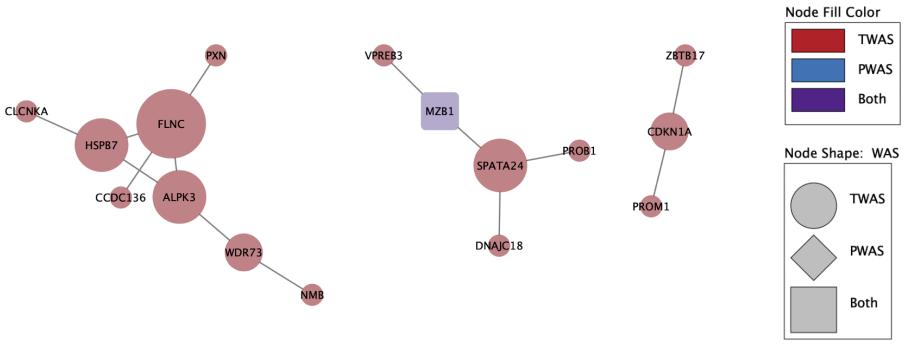
The network for LVM (**Supplemental Figure 3A**) contained 10 nodes (3 connected) and 2 edges, all corresponding to hits from TWAS. The network had a PPI enrichment score of p = 0.303. The hub node identified was FKBP7, a protein which functions as a molecular chaperone to accelerate protein folding. MCODE did not identify clusters in this network.

The network for LVESV (**Supplemental Figure 3B**) consisted of 45 nodes (19 connected) and 16 edges, with a PPI enrichment p-value = 3.14E-06. Similarly to LVEF, the genes *FLNC*, *ALPK3*, *HSPB7*, and *BHMG1* were identified as hub nodes. MCODE once again identified 1 cluster in the network, with nodes *FLNC*, *HSPB7*, and *ALPK3*, at a score of 3.

The LVEDV network (**Supplemental Figure 3C**) contained 34 nodes (13 connected) and 11 edges. Cytohubba identified BHMG1 as a hub node. MCODE did not identify clusters in this network. Metascape pathway enrichment of clusters for the MRI phenotypes did not yield additional enriched terms; the full cluster and pathway enrichment results are available in **Supplemental Table 4**.

MRI – LV Ejection Fraction PPI Network

A



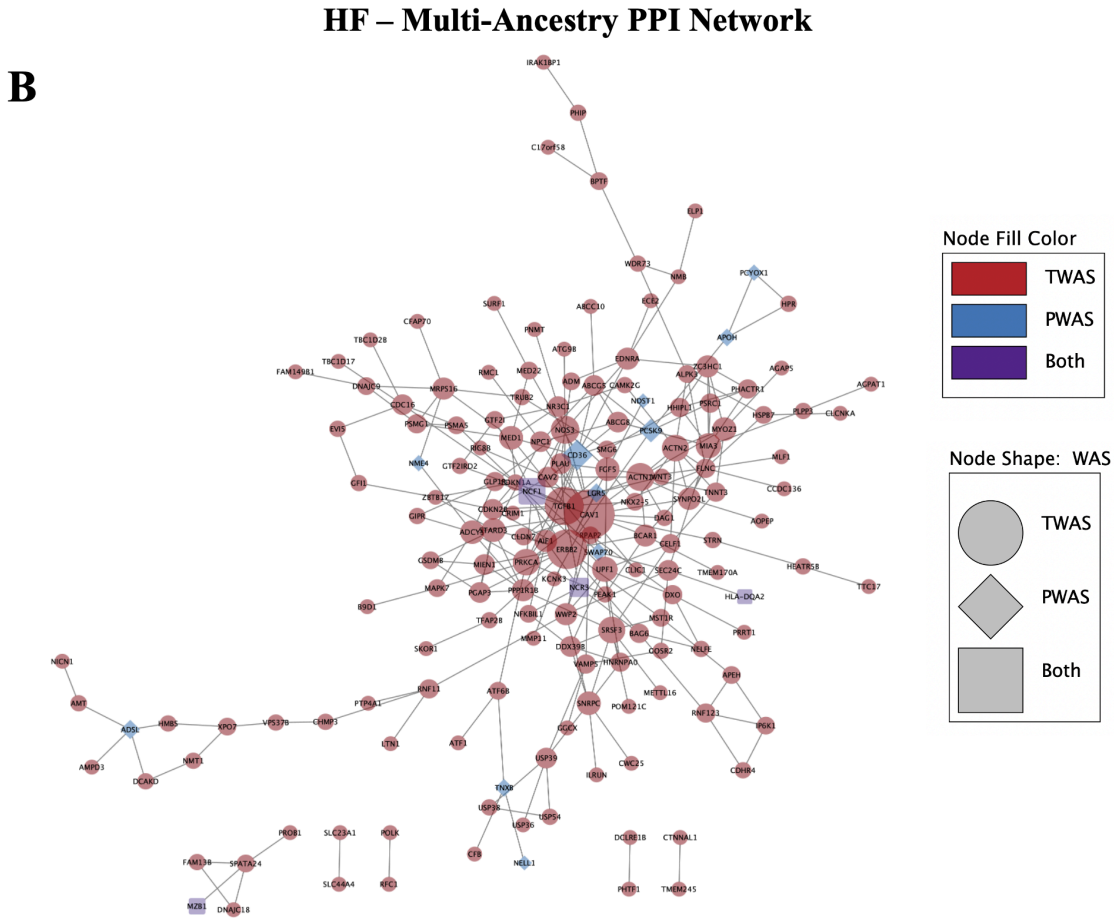


Figure 4. PPI networks constructed using TWAS and PWAS significant hits for A) MRI LVEF and B) multi-ancestry HF phenotypes. Size of nodes denotes degree centrality, with largest nodes identified as hub nodes.

3.2.4. *PPI network analysis of HF genes and proteins*

The PPI network for the HF multi-ancestry cohort consisted of 234 nodes and 294 edges at the medium confidence score > 0.4 in STRING, with an average local clustering coefficient of 0.377 (**Figure 4B**). STRING found that the network had a PPI enrichment p-value = 1.0E-16, suggesting that the network had significantly more edges than expected by random chance. In Cytoscape, the hub nodes identified by degree centrality were *CAV1*, *ERBB2*, *TGFB1*, and *CD36*. Additionally, 5 non-overlapping clusters were identified in MCODE, with a max cluster score of 4.889 and minimum score of 3.0, where a higher score denotes a greater number of nodes in the cluster. Metascape was used to evaluate pathways for the genes present in each cluster, identifying pathways in sarcomere organization ($LOG_{10}P$=-9.09), Hemostasis ($LOG_{10}P$=-7.84), VEGFA-VEGFR2 pathway ($LOG_{10}P$=-7.65), and heart development ($LOG_{10}P$=-3.14), among several others. The PPI network for HF in the EUR population (**Supplemental Figure 3D**) yielded similar findings.

4. Discussion

We performed the first ever intermediate cardiac imaging trait TWAS at the gene level and first-ever protein imputation and largest for all-cause HF, followed by enriched gene sets and constructed interaction networks to contextualize our findings. Current established cardiomyopathy (CM) genes have been found studying familial forms of disease, however, here we focused on identifying associations based on structural and functional cardiac changes and expect to see differences. Overall we demonstrated (1) an increase in information gain using TWAS and PWAS in addition to GWAS in connecting intermediate cardiac phenotypes and full-stage HF, (2) the value in using quantitative intermediate measures for interpreting the changes that occur during disease progression, and (3) how these intermediate measures may relate to changes seen in full-stage HF.

4.1. *Information gain from multi-omics approaches*

While GWAS has value in identifying singular variants associated with traits, variant results are unable to give clues toward mechanistic effects as it relates to the traits of interest. Genes deemed significant from GWAS are usually only identified based on their proximity to significant variants, but we know that proximity actually is not always a good indicator of effect on gene expression and often varies per tissue[63,64]. However, our results are derived from published variant to gene expression values (eQTLs). Our TWAS showed an overall increase in the number of genes associated for every MRI-trait tested compared to GWAS. For example, from Pirruccello et al., 19 genes were identified as proximal to significant variants from GWAS for LVEF. We replicated 8 of those in our TWAS and PWAS, and also identified 29 additional genes and proteins (4 within GWAS Catalog[65] and NCBI[66], 25 novel) that were significant in a tissue specific context (**Figure 2A**)[27]. This trend continued for LVEDV, LVESV, and LMV from Khurshid et al (**Supplemental Figures 1A-C**)[26]. For all-cause multi-ancestry HF GWAS, we replicated 50 genes from the source GWAS in our TWAS results. 199 genes and proteins were not identified in GWAS, indicating that GWAS variants may not be affecting closest proximity genes. Of the remaining 199 genes and proteins identified in our TWAS, 179 are considered novel associations and are not found in previous HF GWAS within GWAS Catalog (**Figure 2B**).

Additionally, to our knowledge this is the first study looking at the protein level based on quantitative cardiac traits, making all protein results novel. By adding proteins to genes in our gene-protein networks, these results 1) give higher confidence associations to those replicating at the gene and protein level, and 2) connect gene clusters with missing information. For example, SPON1 replicated at the gene and protein level for LVEF and LVESV and has been implicated in cardiomyopathies[67–69], blood pressure[70], cognitive decline[71–73], cancers[74,75], as well as EF in a recent study[76]. NCF1, replicated at the gene and protein level for HF and does not appear to have previously been associated with HF in literature, but has been reported for Williams Syndrome which can cause structural vascular changes[77]. NCF1 also has ties to immune response and is a key regulator of reactive oxygen species[78,79]. Other replicating proteins for HF (APOH, TNXB, PCSK9, and RGMB), have previously been associated with HF[16,80–82].

4.2. *Relationships between intermediate cardiac measures and HF*

The four MRI derived cardiac measures explored in this study are often collectively used to mark structural and functional changes in the heart, and are reliable indicators of HF risk and eventual

diagnosis. Therefore the relationship between these measures as well as the overlap between each and HF at the gene, protein, and pathway level is of interest.

Between the MRI cardiac measures we see overlap at the gene and protein level (**Figure 2A** and **Supplemental Figures 1A-C**). Three genes, *FKBP7*, *RP11-171I2.3*, and *PRKRA* were significant in at least one tissue for each of the four MRI cardiac traits. Both *FKBP7* and *PRKRA* have been discussed in relation to cardiac traits before. *PRKRA* has been implicated in studies impacting heart structure, and while *FKBP7* has been less discussed, was linked to atrial fibrillation in one study[83–87]. *PRKRA* and *FKBP7* are both involved in stress response and could be indicative of problematic changes beginning to occur. It is worth noting that *FKBP7*, *PRKRA*, and third gene (*PLEKHA3,* significant with LVM, LVEDV, and LVESV), all lie within a region of chromosome 2 that also includes the gene *TTN*, which is well established heritable cause of dilated cardiomyopathy, a leading cause of heart failure[88–90].

Of the four traits LVEF and LVESV had the most overall overlaps, with 30 different genes and one protein, SPON1 appearing for both traits. When we clustered the results, one PPI cluster for LVEF is dominated by genes previously implicated with dilated cardiomyopathy and heart failure risk (*HSPB7*, *FLNC*, *ALPK3*, *CLCNKA*) **(Figure 4A)**, as well as links to the brain via *WDR73*[91–97]. We also see a cluster composed of much of the same genes for LVESV, and overlapping enriched pathways for cardiac cell development (*ALPK3*), and transepithelial chloride transport (*CLCNKA*). Renal transport also appears as a significantly enriched pathway for LVEF as a result of *CLCNKA,* as well as renal filtration cell differentiation and nephron tubule as top pathway results. Renal impairment is common among HF patients and is documented to increase mortality risk[98,99]. The significance of structural LV genes contributing to these pathways could indicate that these genes impact both disorders, or that there may be some interplay between HF and renal conditions that contribute to progression.

Within our multi-ancestry HF results there appears to be substantial overlap in the MRI derived cardiac measures. 30 genes that appeared significant in one of the MRI cardiac traits also appeared to be associated with all-cause HF. LVEF and LVESV had the largest number of overlapping genes with the disease respectively (18 genes for LVEF and 18 genes for LVESV), further supporting their use as HF indicators. AIDA was the top significant protein from blood plasma, significant with both the European American and African American pQTL reference panels (**Supplemental Table 2**), a gene previously discussed as part of inflammatory response that also promotes atherosclerosis and coronary artery disease[100]. Overall, HF associated gene and protein results tend to enrich lipid and specifically cholesterol related pathways (**Figure 3B**). Nephron tubule development also appears as a top pathway for GF, implicating the kidneys again. We did not see top CM genes such as *MYBPC3*, *MYH7*, *MYL2/3*, *DSP*, but these also were not significant in our source GWASs and not surprising given that our studies were focused on cardiac structure and functional changes rather than familial types. However, we did find some known CM genes: *FLNC* (LVEF, LVESV, HF), *PLN* (LVEDV), and *ACTIN2* (HF), as well as genes related to known CM genes like *TNNT2* (*TNNT3* - HF) and *TMEM43* (*TMEM170A*, *TMEM241*, *TMEM150A*, *TMEM245* - HF).

4.3. *Limitations and future directions*

Limitations exist within this study. First, this analysis only encompasses the genetic factors of HF and intermediate cardiac measures. Genetics play a significant role in the development of HF;

however, it has been shown that environment and comorbidities play a large role in increasing risk as well[101,102]. We did not consider social factors or health records of the patients used in the source GWASs, and therefore may not have fully been able to characterize all potential sources of disease progression. Future studies attempting to characterize or stratify individual-level risk of developing HF would benefit from including these data modalities, in addition to those used in this study.

Additionally, we used a GWAS study with an all-case HF multi-ancestry cohort to represent HF in our study. While this is beneficial for increasing sample size and increasing power, this might also muddle distinct signals within known phenotypic subgroups of HF or specific ancestries given the complexity of the disease as earlier stated. Future studies may find improvement by stratifying HF by subgroups, such as diastolic vs systolic dysfunction, to achieve more informed results. Beyond our cohort limitations, we also acknowledge that the imputation of gene and protein expression is influenced by the ancestry and completeness of the reference eQTL and pQTL sets we used (GTEx v8 and ARIC). For example, known cardiomyopathy genes *LMNA*, *FHL1*, and *CALR3* were not in the GTEx eQTL set used[25]. These QTL reference sets were also not disease specific. The creation of disease specific models using disease-specific RNAseq and proteomic data may be an avenue for further improvement. Lastly, here we only considered imputed gene and protein data modalities, based on multi-omics data from well characterized reference populations. The use of additional modalities, such as RNAseq, protein abundance, or methylation information would provide stronger evidence for our conclusions.

5. Acknowledgments

This work was supported through the following grants from the National Institutes of Health (NIH): AG066833, HL169458. We also thank Tess Cherlin from the University of Pennsylvania for her assistance with visualizations in this manuscript.

6. Appendix

Supplemental figures and tables are available at:
https://ritchielab.org/publications/supplementary-data/psb-2025/hfmultiomics
Code is available at: https://github.com/RitchieLab/HFmultiomics_PSB2025

References

1. Bozkurt, B. *et al.* Heart Failure Epidemiology and Outcomes Statistics: A Report of the Heart Failure Society of America. *J. Card. Fail.* **29**, 1412–1451 (2023).
2. Bozkurt, B., Coats, A. & Tsutsui, H. Universal definition and classification of heart failure. *J. Card. Fail.* **27**, 387–413 (2021).
3. Ziaeian, B. & Fonarow, G. C. Epidemiology and aetiology of heart failure. *Nat. Rev. Cardiol.* **13**, 368–378 (2016).
4. Kenchaiah, S. *et al.* Obesity and the risk of heart failure. *N. Engl. J. Med.* **347**, 305–313 (2002).
5. Roger, V. L. Epidemiology of Heart Failure: A Contemporary Perspective. *Circ. Res.* **128**, 1421–1434 (2021).
6. Lopes, L. R. & Elliott, P. M. Genetics of heart failure. *Biochim. Biophys. Acta* **1832**, 2451–2461 (2013).

7. Lindgren, M. P. *et al.* A Swedish Nationwide Adoption Study of the Heritability of Heart Failure. *JAMA Cardiol* **3**, 703–710 (2018).
8. Ammar, K. A. *et al.* Prevalence and prognostic significance of heart failure stages: application of the American College of Cardiology/American Heart Association heart failure staging criteria in the community. *Circulation* **115**, 1563–1570 (2007).
9. Goldberg, L. R. & Jessup, M. Stage B heart failure: management of asymptomatic left ventricular systolic dysfunction. *Circulation* **113**, 2851–2860 (2006).
10. Kosmala, W. & Marwick, T. H. Asymptomatic Left Ventricular Diastolic Dysfunction: Predicting Progression to Symptomatic Heart Failure. *JACC Cardiovasc. Imaging* **13**, 215–227 (2020).
11. de Simone, G., Gottdiener, J. S., Chinali, M. & Maurer, M. S. Left ventricular mass predicts heart failure not related to previous myocardial infarction: the Cardiovascular Health Study. *Eur. Heart J.* **29**, 741–747 (2008).
12. Aurigemma, G. P., Zile, M. R. & Gaasch, W. H. Contractile behavior of the left ventricle in diastolic heart failure: with emphasis on regional systolic function. *Circulation* **113**, 296–304 (2006).
13. Solomon, S. D. *et al.* Influence of ejection fraction on cardiovascular outcomes in a broad spectrum of heart failure patients. *Circulation* **112**, 3738–3744 (2005).
14. Smith, J. G. *et al.* Discovery of Genetic Variation on Chromosome 5q22 Associated with Mortality in Heart Failure. *PLoS Genet.* **12**, e1006034 (2016).
15. Villard, E. *et al.* A genome-wide association study identifies two loci associated with heart failure due to dilated cardiomyopathy. *Eur. Heart J.* **32**, 1065–1076 (2011).
16. Rasooly, D. *et al.* Genome-wide association analysis and Mendelian randomization proteomics identify drug targets for heart failure. *Nat. Commun.* **14**, 3826 (2023).
17. Levin, M. G. *et al.* Genome-wide association and multi-trait analyses characterize the common genetic architecture of heart failure. *Nat. Commun.* **13**, 6914 (2022).
18. Yu, B. *et al.* Genome-wide association study of a heart failure related metabolomic profile among African Americans in the Atherosclerosis Risk in Communities (ARIC) study. *Genet. Epidemiol.* **37**, 840–845 (2013).
19. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat. Commun.* **11**, 163 (2020).
20. Smith, N. L. *et al.* Association of genome-wide variation with the risk of incident heart failure in adults of European and African ancestry: a prospective meta-analysis from the cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium. *Circ. Cardiovasc. Genet.* **3**, 256–266 (2010).
21. Morrison, A. C. *et al.* Genomic variation associated with mortality among adults of European and African ancestry with heart failure: the cohorts for heart and aging research in genomic epidemiology consortium. *Circ. Cardiovasc. Genet.* **3**, 248–255 (2010).
22. Li, B. & Ritchie, M. D. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Front. Genet.* **12**, 713230 (2021).
23. Cao, C. *et al.* Power analysis of transcriptome-wide association study: Implications for practical protocol choice. *PLoS Genet.* **17**, e1009405 (2021).
24. Li, D., Liu, Q. & Schnable, P. S. TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. *Plant Physiol.* **186**, 1800–1811 (2021).

25. Lee, D. S. M. *et al.* Common- and rare-variant genetic architecture of heart failure across the allele frequency spectrum. *medRxiv* (2023) doi:10.1101/2023.07.16.23292724.
26. Khurshid, S. *et al.* Clinical and genetic associations of deep learning-derived cardiac magnetic resonance-based left ventricular mass. *Nat. Commun.* **14**, 1558 (2023).
27. Pirruccello, J. P. *et al.* Genetic analysis of right heart structure and function in 40,000 people. *Nat. Genet.* **54**, 792–803 (2022).
28. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
29. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).
30. Araujo, D. S. *et al.* Multivariate adaptive shrinkage improves cross-population transcriptome prediction and association studies in underrepresented populations. *HGG Adv* **4**, 100216 (2023).
31. Jahng, J. W. S., Song, E. & Sweeney, G. Crosstalk between the heart and peripheral organs in heart failure. *Exp. Mol. Med.* **48**, e217 (2016).
32. Oikonomou, E. K. & Antoniades, C. The role of adipose tissue in cardiovascular health and disease. *Nat. Rev. Cardiol.* **16**, 83–99 (2019).
33. Schefold, J. C., Filippatos, G., Hasenfuss, G., Anker, S. D. & von Haehling, S. Heart failure and kidney dysfunction: epidemiology, mechanisms and management. *Nat. Rev. Nephrol.* **12**, 610–623 (2016).
34. Poelzl, G. *et al.* Liver dysfunction in chronic heart failure: prevalence, characteristics and prognostic significance. *Eur. J. Clin. Invest.* **42**, 153–163 (2012).
35. Xanthopoulos, A., Starling, R. C., Kitai, T. & Triposkiadis, F. Heart Failure and Liver Disease: Cardiohepatic Interactions. *JACC Heart Fail* **7**, 87–97 (2019).
36. Tanai, E. & Frantz, S. Pathophysiology of Heart Failure. *Compr. Physiol.* **6**, 187–214 (2015).
37. Vanburen, P. *et al.* Blood gene expression signatures associate with heart failure outcomes. *Physiol. Genomics* **43**, 392–397 (2011).
38. Li, S.-J. *et al.* Identifying causal genes for migraine by integrating the proteome and transcriptome. *J. Headache Pain* **24**, 1–11 (2023).
39. Wang, J.-J., Zheng, Z. & Zhang, Y. Association of Hematological Biomarkers of Inflammation with 10-Year Major Adverse Cardiovascular Events and All-Cause Mortality in Patients with Metabolic Dysfunction-Associated Steatotic Liver Disease: The ARIC Study. *J. Inflamm. Res.* **17**, 4247–4256 (2024).
40. Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat. Genet.* **54**, 593–602 (2022).
41. Mi, S. Protein prediction models - ARIC. *PredictDB* (2022).
42. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, (2013).
43. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, (2016).
44. Akçimen, F. *et al.* Transcriptome-wide association study for restless legs syndrome identifies new susceptibility genes. *Communications Biology* **3**, 1–5 (2020).
45. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic*

Acids Res. **47**, D607–D613 (2018).
46. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
47. Gong, W., Guo, P., Liu, L., Guan, Q. & Yuan, Z. Integrative Analysis of Transcriptome-Wide Association Study and mRNA Expression Profiles Identifies Candidate Genes Associated With Idiopathic Pulmonary Fibrosis. *Front. Genet.* **11**, (2020).
48. Chin, C. H. *et al.* cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **8 Suppl 4**, (2014).
49. Bader, G. D. & Hogue, C. W. V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 1–27 (2003).
50. Sun, P. G., Gao, L. & Han, S. Prediction of Human Disease-Related Gene Clusters by Clustering Analysis. *Int. J. Biol. Sci.* **7**, 61 (2011).
51. Song, J. *et al.* Integrative transcriptome-wide analysis of atopic dermatitis for drug repositioning. *Communications Biology* **5**, 1–13 (2022).
52. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, (2000).
53. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, (2019).
54. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, (2023).
55. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, (2022).
56. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, (2020).
57. The Gene Ontology Consortium *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000).
58. Aleksander, S. A. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
59. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, (2019).
60. Dutka, M., Zimmer, K., Ćwiertnia, M., Ilczak, T. & Bobiński, R. The role of PCSK9 in heart failure and other cardiovascular diseases-mechanisms of action beyond its effect on LDL cholesterol. *Heart Fail. Rev.* (2024) doi:10.1007/s10741-024-10409-7.
61. Levine, G. N., Keaney, J. F., Jr & Vita, J. A. Cholesterol reduction in cardiovascular disease. Clinical benefits and possible mechanisms. *N. Engl. J. Med.* **332**, 512–521 (1995).
62. Daniels, T. F., Killinger, K. M., Michal, J. J., Wright, R. W., Jr & Jiang, Z. Lipoproteins, cholesterol homeostasis and cardiac health. *Int. J. Biol. Sci.* **5**, 474–488 (2009).
63. Gaffney, D. J. Global properties and functional complexity of human gene regulatory variation. *PLoS Genet.* **9**, e1003501 (2013).
64. Gutierrez-Arcelus, M. *et al.* Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet.* **11**, e1004958 (2015).
65. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
66. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32**, D35–40 (2004).
67. Gao, Y. *et al.* Genome-wide analysis of DNA methylation and risk of cardiovascular disease

in a Chinese population. *BMC Cardiovasc. Disord.* **21**, 240 (2021).
68. Girerd, N. *et al.* Inflammation and remodeling pathways and risk of cardiovascular events in patients with ischemic heart failure and reduced ejection fraction. *Sci. Rep.* **12**, 8574 (2022).
69. Liu, T. *et al.* Identification of Circular RNA-MicroRNA-Messenger RNA Regulatory Network in Atrial Fibrillation by Integrated Analysis. *Biomed Res. Int.* **2020**, 8037273 (2020).
70. Clemitson, J.-R. *et al.* Genetic dissection of a blood pressure quantitative trait locus on rat chromosome 1 and gene expression analysis identifies SPON1 as a novel candidate hypertension gene. *Circ. Res.* **100**, 992–999 (2007).
71. Jahanshad, N. *et al.* Genome-wide scan of healthy human connectome discovers SPON1 gene variant influencing dementia severity. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 4768–4773 (2013).
72. Fernandez, S. *et al.* Is Associated with Amyloid-β and ε4-Related Cognitive Decline in Cognitively Normal Adults. *J Alzheimers Dis Rep* **5**, 111–120 (2021).
73. Liu, Z. *et al.* APOE influences working memory in non-demented elderly through an interaction with SPON1 rs2618516. *Hum. Brain Mapp.* **39**, 2859–2867 (2018).
74. Tamjidifar, R. *et al.* Prognostic and Diagnostic Values of miR-506 and SPON 1 in Colorectal Cancer with Clinicopathological Considerations. *J. Gastrointest. Cancer* **52**, 125–129 (2021).
75. Miyakawa, R. *et al.* SPON1 is an independent prognostic biomarker for ovarian cancer. *J. Ovarian Res.* **16**, 95 (2023).
76. Shah, A. M. *et al.* Large scale plasma proteomics identifies novel proteins and protein networks associated with heart failure development. *Nat. Commun.* **15**, 528 (2024).
77. Kozel, B. A. *et al.* Williams syndrome predisposes to vascular stiffness modified by antihypertensive use and copy number changes in NCF1. *Hypertension* **63**, 74–79 (2014).
78. Liu, H. *et al.* Ncf1 knockout in smooth muscle cells exacerbates angiotensin II-induced aortic aneurysm and dissection by activating the STING pathway. *Cardiovasc. Res.* **120**, 1081–1096 (2024).
79. Holmdahl, R., Sareila, O., Olsson, L. M., Bäckdahl, L. & Wing, K. Ncf1 polymorphism reveals oxidative regulation of autoimmune chronic inflammation. *Immunol. Rev.* **269**, 228–247 (2016).
80. Da Dalt, L. *et al.* PCSK9 deficiency rewires heart metabolism and drives heart failure with preserved ejection fraction. *Eur. Heart J.* **42**, 3078–3090 (2021).
81. Xu, Q. *et al.* PCSK9: A emerging participant in heart failure. *Biomed. Pharmacother.* **158**, 114106 (2023).
82. Nayor, M. *et al.* Aptamer-Based Proteomic Platform Identifies Novel Protein Predictors of Incident Heart Failure and Echocardiographic Traits. *Circ. Heart Fail.* **13**, e006749 (2020).
83. Hsu, J. *et al.* Genetic Control of Left Atrial Gene Expression Yields Insights into the Genetic Susceptibility for Atrial Fibrillation. *Circ Genom Precis Med* **11**, e002107 (2018).
84. Wang, Q. *et al.* Epistasis regulates genetic control of cardiac hypertrophy. *Res Sq* (2023) doi:10.21203/rs.3.rs-3509208/v1.
85. Wang, X. *et al.* Genetic and phenotypic relationships between coronary atherosclerotic heart disease and electrocardiographic traits. in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE, 2022). doi:10.1109/bibm55620.2022.9995557.
86. Gao, H., Liu, Y., Sheng, W., Shou, W. & Huang, G. Progresses in genetic testing in congenital heart disease. *Medicine Plus* **1**, 100028 (2024).

87. Shah, S. *et al.* Four genetic loci influencing electrocardiographic indices of left ventricular hypertrophy. *Circ. Cardiovasc. Genet.* **4**, 626–635 (2011).
88. Akhtar, M. M. *et al.* Clinical Phenotypes and Prognosis of Dilated Cardiomyopathy Caused by Truncating Variants in the Gene. *Circ. Heart Fail.* **13**, e006832 (2020).
89. Chauveau, C. *et al.* Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. *Hum. Mol. Genet.* **23**, 980–991 (2014).
90. Tharp, C. A., Haywood, M. E., Sbaizero, O., Taylor, M. R. G. & Mestroni, L. The Giant Protein Titin's Role in Cardiomyopathy: Genetic, Transcriptional, and Post-translational Modifications of TTN and Their Contribution to Cardiac Disease. *Front. Physiol.* **10**, 1436 (2019).
91. Stark, K. *et al.* Genetic association study identifies HSPB7 as a risk gene for idiopathic dilated cardiomyopathy. *PLoS Genet.* **6**, e1001167 (2010).
92. Cappola, T. P. *et al.* Common variants in HSPB7 and FRMD4B associated with advanced heart failure. *Circ. Cardiovasc. Genet.* **3**, 147–154 (2010).
93. Begay, R. L. *et al.* Gene Splice Mutations Cause Dilated Cardiomyopathy. *JACC Basic Transl Sci* **1**, 344–359 (2016).
94. Cui, H. *et al.* Mutation profile of FLNC gene and its prognostic relevance in patients with hypertrophic cardiomyopathy. *Mol Genet Genomic Med* **6**, 1104–1113 (2018).
95. Van Sligtenhorst, I. *et al.* Cardiomyopathy in α-kinase 3 (ALPK3)-deficient mice. *Vet. Pathol.* **49**, 131–141 (2012).
96. Almomani, R. *et al.* Biallelic Truncating Mutations in ALPK3 Cause Severe Pediatric Cardiomyopathy. *J. Am. Coll. Cardiol.* **67**, 515–525 (2016).
97. Cappola, T. P. *et al.* Loss-of-function DNA sequence variant in the CLCNKA chloride channel implicates the cardio-renal axis in interindividual heart failure risk variation. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 2456–2461 (2011).
98. Smith, G. L. *et al.* Renal impairment and outcomes in heart failure: systematic review and meta-analysis. *J. Am. Coll. Cardiol.* **47**, 1987–1996 (2006).
99. Damman, K. *et al.* Worsening renal function and prognosis in heart failure: systematic review and meta-analysis. *J. Card. Fail.* **13**, 599–608 (2007).
100. Lalonde, S. *et al.* Integrative analysis of vascular endothelial cell genomic features identifies AIDA as a coronary artery disease candidate gene. *Genome Biol.* **20**, 133 (2019).
101. Kannel, W. B. *et al.* Profile for estimating risk of heart failure. *Arch. Intern. Med.* **159**, 1197–1204 (1999).
102. Nayak, A., Hicks, A. J. & Morris, A. A. Understanding the Complexity of Heart Failure Risk and Treatment in Black Patients. *Circ. Heart Fail.* **13**, e007264 (2020).

Plasma protein-based and polygenic risk scores serve complementary roles in predicting inflammatory bowel disease

Jakob Woerner[1†], Thomas Westbrook[1†], Seokho Jeong[2], Manu Shivakumar[1], Allison R. Greenplate[3], Sokratis A. Apostolidis[4], Seunggeun Lee[5], Yonghyun Nam[2], Dokyoon Kim[2]

[1]*Genomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, PA, USA*
[2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA*
[3]*Institute for Immunology and Immune Health, University of Pennsylvania, Philadelphia, PA, USA*
[4]*Division of Rheumatology, Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[5]*Graduate School of Data Science, Seoul National University, Seoul, South Korea*
[†]*Equal Contribution*
Email: Yonghyun.Nam@pennmedicine.upenn.edu, Dokyoon.Kim@pennmedicine.upenn.edu

Inflammatory bowel disease (IBD), encompassing Crohn's disease (CD) and ulcerative colitis (UC), has a significant genetic component and is increasingly prevalent due to environmental factors. Current polygenic risk scores (PRS) have limited predictive power and cannot inform time of symptom onset. Circulating proteomics profiling offers a novel, non-invasive approach for understanding the inflammatory state of complex diseases, enabling the creation of proteomic risk scores (ProRS). This study utilizes data from 51,772 individuals in the UK Biobank to evaluate the unique and combined contributions of PRS and ProRS to IBD risk prediction. We developed ProRS models for CD and UC, assessed their predictive performance over time, and examined the benefits of integrating PRS and ProRS for enhanced risk stratification. Our findings are the first to demonstrate that combining genetic and proteomic data improves IBD incidence prediction, with ProRS providing time-sensitive predictions and PRS offering additional long-term predictive value. We also show that the ProRS achieves better predictive performance among individuals with high PRS. This integrated approach highlights the potential for multi-omic data in precision medicine for IBD.

Keywords: plasma proteomics; polygenic risk score; autoimmunity; multi-omics; inflammatory bowel disease.

1. Introduction

Inflammatory bowel disease (IBD) represents a chronic inflammatory condition of the gastrointestinal tract. Its subtypes, Crohn's disease (CD) and ulcerative colitis (UC) are related but unique conditions with differing properties, symptoms, and risk factors.[1] IBD affects approximately 2.4 to 3.1 million people in the United States, with most diagnoses occurring in adulthood.[2–4] Epidemiologic and genetic studies have demonstrated that these inflammatory conditions are driven by a complex interplay between genetic susceptibility and environmental factors. Genome-wide association studies (GWASs) have identified over 200 significant genetic loci,[5] and family history of the disease is the strongest risk factor.[6] Multiple lifestyle factors,[7,8] including smoking and

psychological stress, as well as environmental factors[9] such as urbanization, industrialization, and westernization are also associated with the onset and progression of IBD.

Patients with IBD often develop severe complications, including strictures or fistulas in the intestine, and in extreme cases, colorectal cancer. Therefore, identifying high-risk individuals before the onset of IBD symptoms is crucial to potentially preventing or delaying irreversible bowel damage and disease progression.[10] Many studies have developed models to stratify high-risk and low-risk individuals for CD and UC using polygenic risk scores (PRSs) that incorporate GWAS summary statistics and individual genotype data.[11,12] PRSs use genetic variants to estimate an individual's susceptibility to developing a disease. However, since IBD is also influenced by non-genetic factors like lifestyle and environmental influences, accurately assessing IBD risk using models based solely on genetic data is challenging.

IBD is an autoimmune condition, so the current state of an individual's immune system provides valuable information about symptom onset.[13] While genetic data provide insights into susceptibility, they cannot predict when symptoms will appear or how the disease will progress. A PRS can identify individuals with high genetic risk for IBD, but these individuals may not necessarily develop the disease if they effectively manage factors that influence their immune system and overall health. This highlights the importance of integrating both genetic predisposition and variable non-genetic factors for a comprehensive assessment of IBD risk.

Recently, high-dimensional circulating plasma proteomics profiling has been used as a non-invasive tool to understand complex diseases on a large scale and act as endophenotypes related to disease pathogenesis and progression. Plasma proteomics provide a snapshot of an individual's current immune status, including many health-related processes and pathways. Studies have found proteins associated with the prevalence of a range of complex diseases,[14,15] including IBD.[16,17] Additionally, protein levels prior to diagnosis have been linked with subsequent disease onset,[18,19] including in IBD,[20] further motivating their use as a predictive tool. Consequently, these developments produced proteomic risk scores (ProRS), where protein signatures are consolidated into a score for the current risk of developing a disease.[21–23] Proteomic signatures are broadly more predictive of complex disease incidence and prevalence than PRS.[15,24] However, many diseases have both genetic and non-genetic components predictive of disease onset, so efforts have been made to combine scores through multi-omic integration of PRS and ProRS. Evidence suggests this combination improves the prediction of coronary artery disease,[25] coronary plaques,[26] and type 2 diabetes;[22] however, this has not been explored in IBD.

We used data from 51,772 patients in the UK Biobank (UKB) to characterize the unique contributions of polygenic risk and proteomic risk to IBD onset prediction in the largest available proteomics dataset (Figure 1). Despite the superior performance of ProRS compared to PRS for IBD risk assessment, we combined circulating plasma proteomics with genetics in two ways to leverage their gene-environment (GxE) interactions[27] and provide a more comprehensive risk assessment. We directly integrated proteomics and genetics as predictors in the same model, as well as stratified patients by PRS before assessing ProRS, showcasing the interactions between the two modalities and identifying disease-associated protein biomarkers. Since proteomics data can be obtained during a routine clinical blood test, we tested the accuracy of a personalized medicine approach through omics integration for IBD risk.

2. Methods

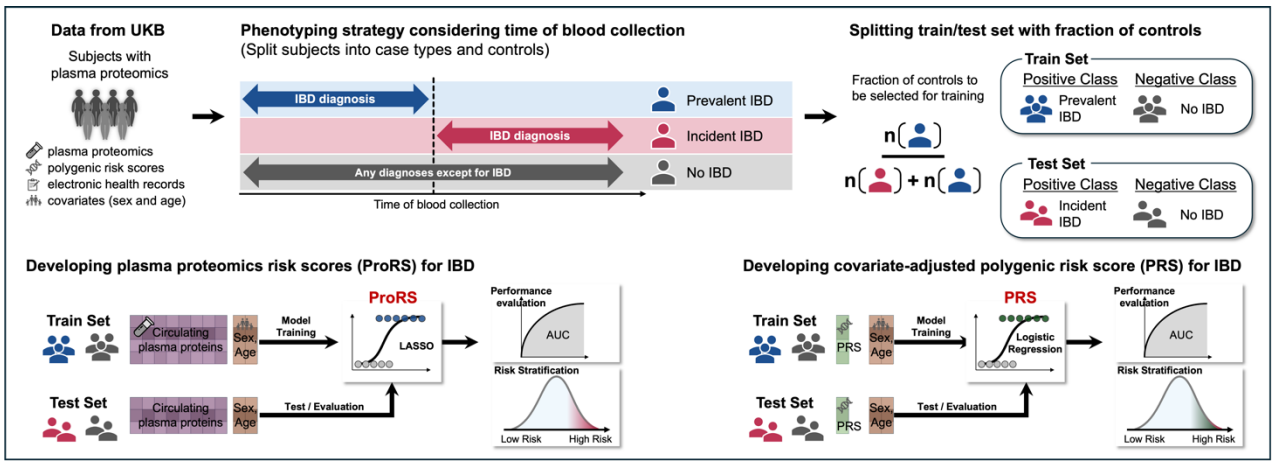


Fig. 1. Study Overview. Data was collected from UKB including plasma protein levels, disease-specific PRS models, age, sex, and ICD-10 codes. Cases were separated into prevalent and incident groups, and the fraction of cases that were prevalent determined the fraction of controls assigned to the training set. The rest were assigned to the testing set. The training set was used to create logistic regression models for the covariate-only model and covariate-adjusted PRS as well as LASSO models for the ProRS and combined score.

2.1. *Data and study participants*

2.1.1. *UK Biobank*

We used data from the UKB, a large-scale biomedical database that provides an extensive collection of genetic, health, and lifestyle information from half a million participants from the UK aged 40-69 at recruitment. With genotype information, International Classification of Diseases (ICD) codes from electronic health records, and biological samples saved for later analysis, the biobank provides the largest resource to study IBD. The breadth of data collected by the UKB and its large sample sizes enabled this project to analyze multi-omic data in a substantial sample with an adequate number of disease cases.

2.1.2. *Circulating proteomics*

In October 2023, the UKB released plasma protein levels of 53,018 blood samples from participants collected at recruitment between 2006 and 2010 as part of the Pharma Proteomics Project (UKB-PPP).[28] The circulating levels of 2,923 proteins were recorded using the Olink Explore 3072 proximity extension assay. The data had about 17.5% missingness. To preserve as many samples as possible, the missing values were imputed using the *k*-nearest neighbors imputation method with $k = 10$.[29] Before imputation, individuals with greater than 54% missingness ($n = 698$) and proteins with greater than 30% missingness ($n = 3$) were excluded, reducing the total missingness to 9.5%.

2.1.3. *Phenotyping*

Binary phenotypes for each IBD subtype were established using the ICD diagnosis codes K50* for CD and K51* for UC. If an individual's date of first report of disease occurrence was before their blood draw, from which circulating proteomics were profiled, they were labeled as a prevalent case. If their date of first disease occurrence was after their initial blood draw, they were labeled as an incident case. Otherwise, they were considered controls (Figure 1). Additionally, Hospital Episode Statistics were used to identify the specific ICD code for each case. A rheumatologist classified each code within K50* and K51* as an autoimmune disease or other rheumatic condition. 35 individuals with codes in the UC block (K51) that had non-autoimmune diseases (K51.4, K51.5) but no other autoimmune disease in the block, were removed from the analysis. 55 individuals had both CD and UC codes at baseline, and so were considered prevalent cases in both models. For survival analysis, individuals were considered to have the event at their date of first occurrence of the disease. Individuals were censored at their date of death if they appeared in the central death registry. To generalize the findings as much as possible, our analyses included all individuals, regardless of ancestral background. However, the vast majority of the study population self-identified as white British (n = 43,047, 83.1%).

2.2. *Risk Scoring*

We developed a ProRS for each of UC and CD to quantify the likelihood of disease onset in undiagnosed individuals using proteomics data. To differentiate protein levels between healthy subjects and IBD patients, we stratified cases by time of disease onset (see Phenotyping) and used the prevalent cases for model development (training set). Since ProRS aims to predict future IBD onset after blood collection, the incident cases were used for model evaluation (testing set). Due to limited follow-up time in the UKB, there are fewer incident IBD cases compared to prevalent cases. This discrepancy results in an imbalance between the number of training and testing cases, which could potentially affect the accuracy and evaluation of our models by introducing bias and reducing generalizability. To address this imbalance, we randomly split controls into the training and testing sets with the same ratio as prevalent to incident cases in the data.

The train/test split can be described as follows. Let $\boldsymbol{S}^{\text{case}}_{\text{train}}(\cdot)$ and $\boldsymbol{S}^{\text{case}}_{\text{test}}(\cdot)$ be the case set of patients who were diagnosed with the disease before and after blood collection respectively, where the parentheses represent the disease of interest. Given the index disease, the control set is defined as $\boldsymbol{S}^{\text{control}}(\cdot) = \{\boldsymbol{S}^{\text{case}}_{\text{train}}(\cdot) \cup \boldsymbol{S}^{\text{case}}_{\text{test}}(\cdot)\}^{\text{c}}$. We then randomly selected the training control set $\boldsymbol{S}^{\text{control}}_{\text{train}}(\cdot)$ from $\boldsymbol{S}^{\text{control}}(\cdot)$ such that the proportion of all controls that are in $\boldsymbol{S}^{\text{control}}_{\text{train}}(\cdot)$ equaled the proportion of the total number of cases ($|\boldsymbol{S}^{\text{case}}_{\text{train}}(\cdot) + \boldsymbol{S}^{\text{case}}_{\text{test}}(\cdot)|$) that are prevalent cases ($|\boldsymbol{S}^{\text{case}}_{\text{train}}(\cdot)|$). The testing control set is then the remaining set of controls:

$$\begin{cases} \boldsymbol{S}^{\text{control}}_{\text{train}}(\cdot) \subset \boldsymbol{S}^{\text{control}}(\cdot) \\ \boldsymbol{S}^{\text{control}}_{\text{test}}(\cdot) = \boldsymbol{S}^{\text{control}}(\cdot) \setminus \boldsymbol{S}^{\text{control}}_{\text{train}}(\cdot) \end{cases}$$

This approach ensures that the model is trained and evaluated on disjoint datasets with balanced case-control ratios so that the ProRS's performance can be accurately assessed despite the

differences in the numbers of prevalent and incident IBD cases. The resultant case and control counts in each set for both diseases are shown in Table 1.

To evaluate the contribution of each omic level to risk prediction, four models were created for CD and UC separately: a covariate-only model, PRS, ProRS, and a combined model. The covariate model used only the age at plasma protein measurement and sex in an unpenalized logistic regression, acting as a baseline prediction. The PRS model was based on scores from Thompson et al.,[30] which were added as predictors to an unpenalized logistic regression with the covariates. After removing individuals missing a PRS, we analyzed an overall sample size of 51,772 for CD and 51,737 for UC.

Since not all 2,920 proteins are expected to be informative of disease status, we applied covariate-adjusted Least Absolute Shrinkage and Selection Operator (LASSO) models to develop the ProRS while adjusting for potential confounders (sex and age).[31] The method allows for simultaneous protein marker selection and regularization, defined by the equation:

$$\hat{\beta} = \text{argmin}_{\beta} \sum_{i=1}^{n} \log\left(\exp\left(-y_i(X_i^T\beta)\right) + 1\right) + \lambda \sum_{j=1}^{p} |\beta_j| \quad (1)$$

where n is the sample size, y_i is the class information (case or control) for individual i, X_i are the values of the predictors (circulating protein levels and covariates), β_j is the regression coefficient for predictor j, p is the number of predictors, and λ is the regularization parameter controlling the strength of the penalty. Protein features with non-zero coefficient ($\beta_j \neq 0$) in the trained model were considered significant proteins associated with IBD. The chosen λ was the minimum λ from 5-fold cross-validation. The combined model was created in the same fashion, with all protein values, the PRS, age, and sex as predictors in a LASSO model. Prior to LASSO, each predictor was standardized to a mean of 0 and standard deviation of 1 so that the coefficient magnitudes would be comparable. Scores for all four models were computed for each individual in the disease's testing set for further analysis and performance evaluation of disease onset prediction.

Table 1. Case-control counts and covariates by disease. The number of individuals in the case and control groups for the train and test sets for CD and UC, along with sex, mean age, and the number of individuals self-identified as white-British in each subgroup. The p-values for "Female" and "White-British" were calculated by the chi-square test, and the p-values for "Age" were calculated by the Wilcoxon signed rank test for difference between the case and control groups.

Phenotypes		Train				Test			
		Total	Case	Control	p-value	Total	Case	Control	p-value
CD	N	31,879	242	31,637	-	19,893	151	19,742	-
	Female	17,188 (54%)	124 (51%)	17,064 (54%)	0.439	10,738 (54%)	78 (52%)	10,660 (54%)	0.622
	Age	56.8 (56.7-56.9)	56.3 (55.3-57.4)	56.8 (56.7-56.9)	0.3374	56.9 (56.8-57.0)	58.5 (57.2-59.8)	56.9 (56.8-57.0)	0.01195
	White-British	26,481 (83%)	204 (84%)	26,277 (83%)	0.6699	16,566 (83%)	129 (85%)	16,437 (83%)	0.5466
UC	N	34,567	453	34,114	-	17,170	225	16,945	-
	Female	18,627 (54%)	228 (50%)	18,399 (54%)	0.1387	9,280 (54%)	118 (52%)	9,162 (54%)	0.6756
	Age	56.8 (56.7-56.9)	58.7 (58.0-59.4)	56.8 (56.7-56.9)	1.2e-6	56.8 (56.7-56.9)	57.0 (55.9-58.0)	56.8 (56.7-56.9)	0.8897
	White-British	28,740 (83%)	386 (85%)	28,354 (83%)	0.2629	14,276 (83%)	195 (87%)	14,081 (83%)	0.1833

2.3. *Statistical Analyses*

2.3.1. *Risk prediction evaluation*

All data analyses were performed in R 4.4.0. All models were adjusted for age and sex. Area under the receiver operating characteristic curve (AUC) and Nagelkerke's R^2 were used as evaluation metrics to assess the classification ability of each quantitative score.[32] DeLong's test was used to compare AUCs and establish confidence intervals[33] with the pROC R package.[34] This nonparametric approach is suitable for comparing AUCs of two correlated receiver operating characteristic curves, especially when the models are built from the same samples. The CD and UC ProRS models had more proteins with non-zero coefficients than the combined models. In order to evaluate their genetic backing, SNP-based heritability estimates were established for the circulating levels of each protein from pQTL summary statistics of European ancestry individuals[28] using LD score regression of roughly 1.2 million HapMap3 SNPs.[35] Gene set enrichment analysis was then used to test if the heritability estimates were higher in the sets of removed proteins than expected by chance. This analysis was run with the clusterProfiler R package[36] using the heritability estimates of all 2,923 proteins as the background set. Kaplan-Meier cumulative incidence curves were constructed to visualize and test the cumulative incidence of each disease using the survminer R package.[37]

2.3.2. *Longitudinal Analyses*

As protein levels in an individual are dynamic while genotypes are static, performances of the PRS and ProRS models were evaluated in the short term (5 years) and in the long term (10 years) after the blood draw. In these experiments, individuals were only considered incident cases if they were diagnosed with the disease within that time frame (five or ten years). Otherwise, they were considered controls.

To test the relationship between the ProRS and time to diagnosis, mean scores were calculated on a backward timescale for each year leading up to the diagnosis date. Those who would go on to develop IBD were tested against those who did not. Using the approach described in Guo, You, Zhang et al., a nested case-control study was implemented to match individuals with incident diagnosis events to healthy controls.[38] Individuals were matched based on age and sex, with a 1:5 case-control ratio. The event date for matched controls was set to their corresponding case, and incident cases past 14 years were set to have an event date of 14 years. Mean values at each time point were fitted using locally weighted smoothing curves ($\alpha = 0.8$). The Mann-Kendall trend test was used to compare differences in ProRS between cases and controls longitudinally.

3. Results

3.1. *Genomics and proteomics uniquely predict IBD incidence*

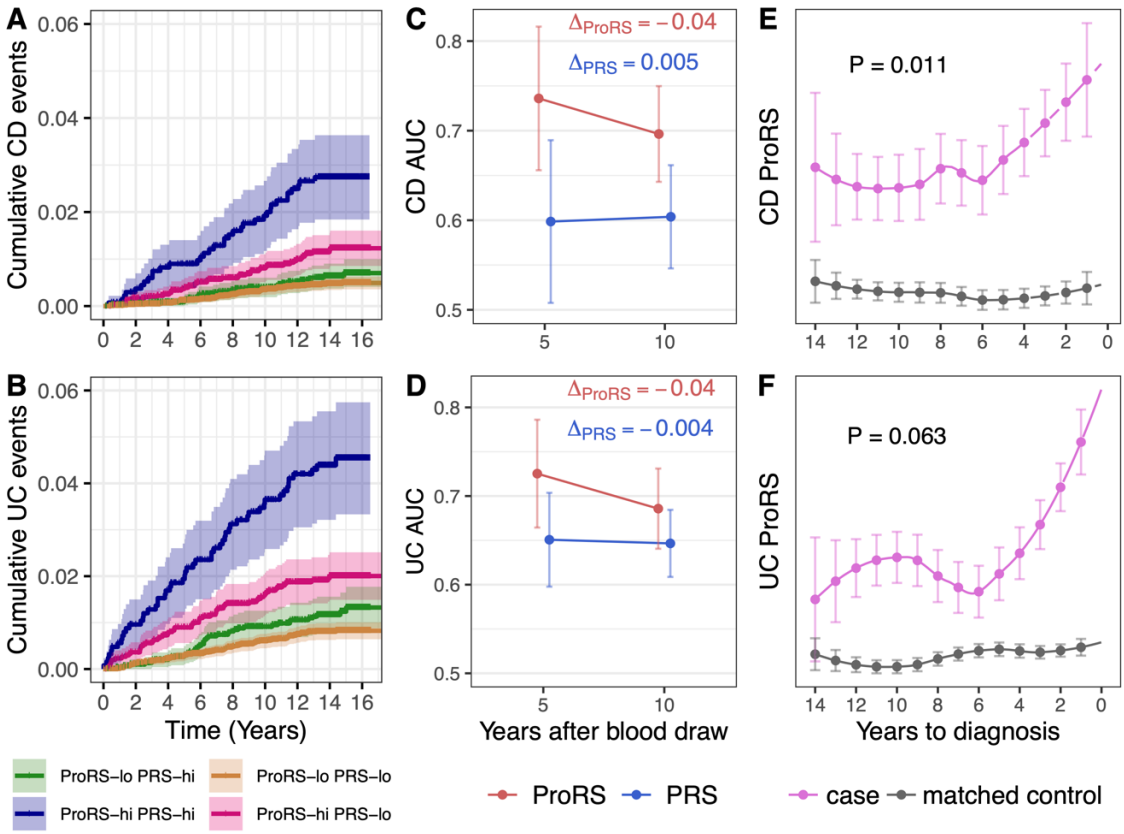


Fig. 2. Longitudinal Analysis. (**A** + **B**) Kaplan Meier curves of time to disease onset, stratified by PRS and ProRS. Individuals are considered high risk when at the 75th percentile or higher. (**C** + **D**) AUC of risk score models at five and ten years after blood draw. (**E** + **F**) ProRS of disease cases compared to age- and sex-matched controls using locally weighted smoothing curves, where the x-axis represents the time after blood draw that individuals were diagnosed. (**A**,**C**,**E**) for Crohn's disease, (**B**,**D**,**F**) for ulcerative colitis.

3.1.1. *PRS and ProRS both effectively stratify individuals at risk for IBD*

ProRS models for CD and UC were developed using LASSO, selecting 216 proteins and 338 proteins, respectively, to predict disease onset. Although age and sex were included as input variables, neither the CD nor the UC model included these covariates as significant features, aligning with the known lack of a sex bias in these diseases.[39] Consistent with other studies, both PRS and ProRS effectively stratified individuals at high risk for disease (Supplemental Figure 1). We also observed that high ProRS was more distinguishing than high PRS. To assess their combined predictive utility, we stratified individuals based on both polygenic risk and proteomic risk. This yielded a cumulative incidence curve with four strata (Figure 2A-B), where high risk was defined as greater than the 75th percentile for each score, and low risk as all others. Interestingly, polygenic risk further stratified individuals within the proteomic risk categories, suggesting PRS can offer additional information on time to disease onset beyond what ProRS can provide.

3.1.2. *ProRS are time-sensitive and reduce in predictive ability over time*

Since circulating protein signatures indicate current health status, we hypothesized that the ProRS predictive accuracy is higher closer to disease onset, while stable for PRS. We tested the models at 5 years and 10 years post blood draw, finding that the ProRS model had an AUC reduction of 0.04 in both CD and UC (CD: 0.74→0.70, UC: 0.73→0.69). The PRS AUC, however, remained similar in both diseases (Figure 2C-D). This reduction may be explained by the observation that the ProRS for both CD and UC increased dramatically in the ~5 years preceding disease diagnosis, whereas matched controls demonstrated little difference in risk over time (Figure 2E-F). The increasing difference in mean ProRS between cases and controls at each time point indicates a likely increase in IBD protein signatures in the years leading up to disease onset (CD: $p = 0.011$, UC: $p = 0.063$).

3.2. *Genomics and proteomics in combination improve IBD prediction*

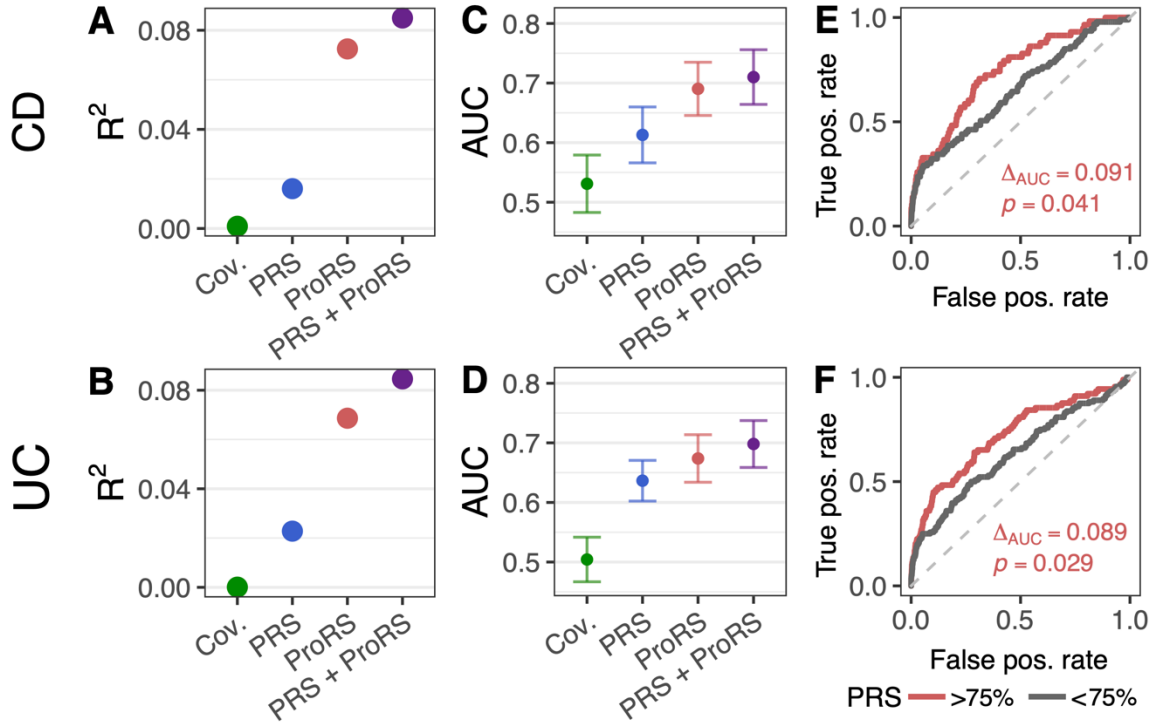


Fig. 3. Combining PRS and ProRS. (A + B) R^2 estimate for disease incidence variance in the covariate-only model and the adjusted PRS, ProRS, and combined models. **(C + D)** AUC comparison of risk score models to predict disease incidence. **(E + F)** Performance of the ProRS model in high disease-risk individuals (>75 percentile) and low disease-risk individuals (<75 percentile). **(A,C,E)** for Crohn's disease, **(B,D,F)** for ulcerative colitis.

3.2.1. *PRS adds complementary predictive information to ProRS*

We evaluated each risk score individually and in combination to test their unique and combined contributions to IBD risk prediction. As previously observed, the ProRS had a much higher R^2 (Figure 3A-B) and AUC (Figure 3C-D) than the PRS for predicting IBD subtype incidence, indicating that ProRS more meaningfully stratifies patients at high risk for the disease. The

combined model, however, outperforms either omic modality alone. Compared to the ProRS model, the combined model's R^2 increased by 0.012 in CD and 0.016 in UC, while AUC increased by 0.020 in CD and 0.024 in UC. This emphasizes the importance of both genetic and proteomic screening in the clinic to identify patients likely to develop CD or UC soon.

3.2.2. *Adding PRS to ProRS removes more heritable proteins*

In the construction of the ProRS, LASSO selected 216 predictors for CD and 338 for UC. When the PRS was added as a predictor, it became a significant predictor with the 9th largest coefficient in CD and the 4th largest in UC. The number of predictors with non-zero coefficients decreased to 203 in CD and 284 in UC. We hypothesized that the PRS might replace proteins whose levels are influenced by genetics. To test this, we used LD score regression to estimate the heritability for each protein and performed gene set enrichment analysis to see if the heritabilities for the removed proteins were significantly higher than expected by chance. These sets consisted of 24 proteins for CD and 66 proteins for UC. With a p-value of 0.004 for CD and 0.08 for UC, there is evidence that the PRS accounts for heritable differences in protein levels.

3.2.3. *High PRS for CD and UC is associated with better incident disease prediction accuracy*

It is thought that genetically susceptible individuals develop IBD due to specific environmental or lifestyle triggers. We hypothesize that protein measurements can reflect when such conditions are met. To test this, we stratified individuals into high (>75 percentile) and low (<75 percentile) PRS groups and evaluated the accuracy of ProRS (Figure 3E-F). Compared to the low PRS group, we observed that the AUC in the high PRS group is 0.091 higher in CD (p = 0.041), and 0.089 higher in UC (p = 0.029). This suggests that an IBD-related inflammatory state from the ProRS model is more predictive in those already known to be at higher risk. This substantial difference in ProRS classification may be explained by higher false positive rates in the low PRS group, resulting from inflammatory states not caused by IBD.

4. Discussion

We evaluated the predictive ability of circulating plasma proteins and genetics for IBD risk and their interactions. Our study highlights three novel findings with implications for their clinical utility. Firstly, combining proteomic and genomic information enabled more precise patient stratification into risk groups. This approach yielded better predictive performance, as indicated by higher AUC and R^2 values, and improved survival analysis for predicting time-to-disease onset. Secondly, stratifying patients by PRS revealed substantial differences in the ProRS model performance for predicting later onset of CD and UC. This may indicate that the inflammatory protein signature is more likely to be an accurate marker of the disease in individuals with high PRS, as opposed to being a confounding condition in low PRS individuals. Thirdly, we found that ProRS prediction accuracy decreases over time, whereas the performance of PRS remained stable. This is likely because ProRS, based on dynamic circulating plasma protein levels, becomes less distinguishing over time, while the static nature of PRS maintains its predictive power.

IBD is a highly polygenic and heritable disease with a significant environmental component. A leading theory of IBD pathogenesis is that environmental exposures in life may trigger inflammatory

bowel disease in genetically susceptible individuals.[40] Although this exposure component is difficult to measure, the genetic component is increasingly measurable. Additionally, circulating proteins can act as an early endophenotype to indicate whether the exposure has happened, and autoimmunity initiated. Our demonstration that the performance of ProRS to predict onset of IBD subtypes is increased in high PRS individuals provides further support to this theory.

Advancements in proteomic technologies have enabled biobanks to generate large-scale data for analyzing the circulating proteome, with many new projects already underway.[41,42] Thus, the utility of risk scoring for precision medicine in both clinical and research settings is becoming more realistic. With increasingly affordable genotyping technologies, it is plausible that lifetime polygenic risk for diseases could be part of a patient's health history available to clinicians. If circulating plasma proteomics were measured in a patient and a ProRS developed, the additional insight from a PRS could help refine this risk. For example, higher PRS could indicate higher confidence in the estimated probability of developing IBD. Additionally, the falling accuracy of ProRS over time suggests scores from older data should be analyzed with skepticism. Given the difference in the cost of genotyping a patient and generating proteomics panels, we suggest an initial assessment with a cheaper genomics approach may be more efficient. If a patient is at high genetic risk for IBD, regularly generating proteomics panels may be necessary.

There are several limitations in our study motivating future work. We used a simple linear model with an L1 penalty to generate the ProRS, but such models may oversimplify the complex biological interactions between circulating proteins and genetic factors. Although preliminary evidence suggests that ensemble methods for proteomic scores perform equally to linear methods when predicting cardiovascular events[31], linear models inherently cannot capture higher-order interactions that might be important for predicting disease risk. In future studies, more sophisticated computational methodologies should be explored for predictive capacity, such as graph machine learning algorithms that might better represent the relationships between biological entities. Another limitation is that this study was only performed in one biobank, with no external validation. Given the uniqueness of the UKB proteomics dataset, it is not possible to replicate the results on a large scale, but more datasets will soon be available for validation. This single biobank also means that results can only be interpreted for a British population. The effect of ancestry could not be sufficiently evaluated in this study due to power constraints. However, protein risk scores have been reported to be transferable across populations with no heterogeneity in effect, even with models trained on much smaller sample sizes.[31] Nonetheless, we acknowledge the need for more diverse cohorts in multi-omic studies. A further limitation of the UKB is the well documented challenge of using mapped ICD-10 codes for phenotyping.[43] Studies suggest positive predictive values of >70% for mapping electronic health records to stroke[44] and acute myocardial infarction,[45] however further work is needed to evaluate their accuracy in phenotyping IBD onset.

This study demonstrates the predictive nature of genetic risk scores, proteomic risk scores, and especially their combination, on IBD incidence. Future work involves using large biobank proteomics to predict IBD progression and prognosis, as shown in smaller studies.[46,47] There is also evidence that proteomics[48] and genomics[49] can be employed to subtype IBD, and their integration may be useful to further distinguish disease types to inform the best clinical care. Our approach is appropriate to analyze any heritable condition that can arise throughout life and would be valuable to apply to more autoimmune and neurodegenerative diseases. These results offer hope for successfully integrating biological data to improve risk prediction.

5. Acknowledgments

This work was supported by NIGMS R01 GM138597 and has been conducted using the UK Biobank Resource under Application Number 45227.

6. Code Availability

No previously unreported algorithm was used to generate results central to the conclusions. Any additional information required to re-analyze the data reported in this work paper is available from the lead contact upon request.

7. Supplemental Materials

All supplemental materials are available at https://s3.amazonaws.com/biomedinfolab/supp/ibd.pdf.

References

1. Seyedian, S. S., Nokhostin, F. & Malamir, M. D. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. *J. Med. Life* **12**, 113–122 (2019).
2. Lewis, J. D. *et al.* Incidence, Prevalence, and Racial and Ethnic Distribution of Inflammatory Bowel Disease in the United States. *Gastroenterology* **165**, 1197-1205.e2 (2023).
3. Xu, F. Health-Risk Behaviors and Chronic Conditions Among Adults with Inflammatory Bowel Disease — United States, 2015 and 2016. *MMWR Morb. Mortal. Wkly. Rep.* **67**, (2018).
4. Dahlhamer, J. M. Prevalence of Inflammatory Bowel Disease Among Adults Aged ≥18 Years — United States, 2015. *MMWR Morb. Mortal. Wkly. Rep.* **65**, (2016).
5. El Hadad, J., Schreiner, P., Vavricka, S. R. & Greuter, T. The Genetics of Inflammatory Bowel Disease. *Mol. Diagn. Ther.* **28**, 27–35 (2024).
6. Santos, M. P. C., Gomes, C. & Torres, J. Familial and ethnic risk in inflammatory bowel disease. *Ann. Gastroenterol.* **31**, 14–23 (2018).
7. Mawdsley, J. E. & Rampton, D. S. Psychological stress in IBD: new insights into pathogenic and therapeutic implications. *Gut* **54**, 1481–1491 (2005).
8. Parkes, G. C., Whelan, K. & Lindsay, J. O. Smoking in inflammatory bowel disease: Impact on disease course and insights into the aetiology of its effect. *J. Crohns Colitis* **8**, 717–725 (2014).
9. Ng, S. C. *et al.* Geographical variability and environmental risk factors in inflammatory bowel disease. *Gut* **62**, 630–649 (2013).
10. Noor, N. M., Sousa, P., Paul, S. & Roblin, X. Early Diagnosis, Early Stratification, and Early Intervention to Deliver Precision Medicine in IBD. *Inflamm. Bowel Dis.* **28**, 1254–1264 (2022).
11. Khunsriraksakul, C. *et al.* Construction and Application of Polygenic Risk Scores in Autoimmune Diseases. *Front. Immunol.* **13**, 889296 (2022).
12. Gettler, K. *et al.* Common and Rare Variant Prediction and Penetrance of IBD in a Large, Multi-ethnic, Health System-based Biobank Cohort. *Gastroenterology* **160**, 1546–1557 (2021).
13. de Souza, H. S. P. & Fiocchi, C. Immunopathogenesis of IBD: current state of the art. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 13–27 (2016).
14. Jacobs, B. M. *et al.* Plasma proteomic profiles of UK Biobank participants with multiple sclerosis. *Ann. Clin. Transl. Neurol.* **11**, 698–709 (2024).
15. Smelik, M. *et al.* An interactive atlas of genomic, proteomic, and metabolomic biomarkers promotes the potential of proteins to predict complex diseases. *Sci. Rep.* **14**, 12710 (2024).
16. Di Narzo, A. F. *et al.* High-Throughput Identification of the Plasma Proteomic Signature of Inflammatory Bowel Disease. *J. Crohns Colitis* **13**, 462–471 (2019).
17. Drobin, K. *et al.* Targeted Analysis of Serum Proteins Encoded at Known Inflammatory Bowel Disease Risk Loci. *Inflamm. Bowel Dis.* **25**, 306–316 (2019).
18. Papier, K. *et al.* Identifying proteomic risk factors for cancer using prospective and exome analyses of 1463 circulating proteins and risk of 19 cancers in the UK Biobank. *Nat. Commun.* **15**, 4010 (2024).

19. Tran, D. *et al.* Plasma Proteomic Signature Predicts Myeloid Neoplasm Risk. *Clin. Cancer Res.* OF1–OF9 (2024) doi:10.1158/1078-0432.CCR-23-3468.
20. Torres, J. *et al.* Serum Biomarkers Identify Patients Who Will Develop Inflammatory Bowel Diseases Up to 5 Years Before Diagnosis. *Gastroenterology* **159**, 96–104 (2020).
21. You, J. *et al.* Plasma proteomic profiles predict individual future health risk. *Nat. Commun.* **14**, 1–13 (2023).
22. Gadd, D. A. *et al.* Blood protein levels predict leading incident diseases and mortality in UK Biobank. 2023.05.01.23288879 Preprint at https://doi.org/10.1101/2023.05.01.23288879 (2023).
23. Ganz, P. *et al.* Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *JAMA* **315**, 2532–2541 (2016).
24. Carrasco-Zanini, J. *et al.* Proteomic signatures improve risk prediction for common and rare diseases. *Nat. Med.* 1–10 (2024) doi:10.1038/s41591-024-03142-z.
25. Møller, P. L. *et al.* Combining Polygenic and Proteomic Risk Scores With Clinical Risk Factors to Improve Performance for Diagnosing Absence of Coronary Artery Disease in Patients With de novo Chest Pain. *Circ. Genomic Precis. Med.* **16**, 442–451 (2023).
26. Møller, P. L. *et al.* Predicting the presence of coronary plaques featuring high-risk characteristics using polygenic risk scores and targeted proteomics in patients with suspected coronary artery disease. *Genome Med.* **16**, 40 (2024).
27. Yang, A. Z. & Jostins-Dean, L. Environmental variables and genome-environment interactions predicting IBD diagnosis in large UK cohort. *Sci. Rep.* **12**, 10890 (2022).
28. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
29. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825–2830 (2011).
30. Thompson, D. J. *et al.* UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. 2022.06.16.22276246 Preprint at https://doi.org/10.1101/2022.06.16.22276246 (2022).
31. Helgason, H. *et al.* Evaluation of Large-Scale Proteomics for Prediction of Cardiovascular Events. *JAMA* **330**, 725–735 (2023).
32. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).
33. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
34. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
35. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
36. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).
37. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
38. Guo, Y. *et al.* Plasma proteomic profiles predict future dementia in healthy adults. *Nat. Aging* **4**, 247–260 (2024).
39. Ngo, S. T., Steyn, F. J. & McCombe, P. A. Gender differences in autoimmune disease. *Front. Neuroendocrinol.* **35**, 347–369 (2014).
40. Borowitz, S. M. The epidemiology of inflammatory bowel disease: Clues to pathogenesis? *Front. Pediatr.* **10**, 1103713 (2023).
41. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *J. Pers. Med.* **12**, 1974 (2022).
42. Sun, B. B., Suhre, K. & Gibson, B. W. Promises and Challenges of populational Proteomics in Health and Disease. *Mol. Cell. Proteomics* **23**, (2024).
43. Stroganov, O. *et al.* Mapping of UK Biobank clinical codes: Challenges and possible solutions. *PLOS ONE* **17**, e0275816 (2022).
44. Woodfield, R., Grant, I., Group, U. B. S. O., Group, U. B. F.-U. and O. W. & Sudlow, C. L. M. Accuracy of Electronic Health Record Data for Identifying Stroke Cases in Large-Scale Epidemiological Studies: A Systematic Review from the UK Biobank Stroke Outcomes Group. *PLOS ONE* **10**, e0140533 (2015).
45. Rubbo, B. *et al.* Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. *Int. J. Cardiol.* **187**, 705–711 (2015).

46. Ungaro, R. C. *et al.* Machine learning identifies novel blood protein predictors of penetrating and stricturing complications in newly diagnosed paediatric Crohn's disease. *Aliment. Pharmacol. Ther.* **53**, 281–290 (2021).
47. Kalla, R. *et al.* Serum proteomic profiling at diagnosis predicts clinical course, and need for intensification of treatment in inflammatory bowel disease. *J. Crohns Colitis* **15**, 699–708 (2020).
48. Fabian, O. *et al.* A Current State of Proteomics in Adult and Pediatric Inflammatory Bowel Diseases: A Systematic Search and Review. *Int. J. Mol. Sci.* **24**, 9386 (2023).
49. Voskuil, M. D. *et al.* Genetic Risk Scores Identify Genetic Aetiology of Inflammatory Bowel Disease Phenotypes. *J. Crohns Colitis* **15**, 930–937 (2021).

Integrated exposomic analysis of lipid phenotypes: Leveraging GE.db in environment by environment interaction studies

ANDRE LUIS GARAO RICO and NICOLE PALMIERO
Department of Genetics, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: andreluis.rico@pennmedicine.upenn.edu, nicole.palmiero@pennmedicine.upenn.edu

MARYLYN D. RITCHIE
Department of Genetics, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: marylyn@pennmedicine.upenn.edu

MOLLY A. HALL
Department of Genetics, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: molly.hall@pennmedicine.upenn.edu

Gene-environment interaction (GxE) studies provide insights into the interplay between genetics and the environment but often overlook multiple environmental factors' synergistic effects. This study encompasses the use of environment by environment interaction (ExE) studies to explore interactions among environmental factors affecting lipid phenotypes (e.g., HDL, LDL, and total cholesterol, and triglycerides), which are crucial for disease risk assessment. We developed a novel curated knowledge base, GE.db, integrating genomic and exposomic interactions. In this study, we filtered NHANES exposure variables (available 1999-2018) to identify significant ExE using GE.db. From 101,316 participants and 77 exposures, we identified 263 statistically significant interactions (FDR $p < 0.1$) in discovery and replication datasets, with 21 interactions significant for HDL-C (Bonferroni $p < 0.05$). Notable interactions included docosapentaenoic acid (22:5n-3) (DPA) - arachidic acid (20:0), stearic acid (18:0) - arachidic acid (20:0), and blood 2,5-dimethyfuran - blood benzene associated with HDL-C levels. These findings underscore GE.db's role in enhancing -omics research efficiency and highlight the complex impact of environmental exposures on lipid metabolism, informing future health strategies.

Keywords: Knowledge-Based Filtering; Interaction Analysis; Exposome; Lipid Metabolism

1. Introduction

Understanding the intricate interplay between genetics and the environment is pivotal in unraveling the complexities of human traits and diseases. While gene-environment interaction (GxE) studies have provided valuable insights into how genetic variants interact with environmental factors, they often overlook the synergistic effects of multiple environmental variables[1,2]. This limitation

necessitates the need for utilizing environment by environment interaction (ExE) studies, which explore how different environmental factors interact with each other to influence phenotypic outcomes. The outcomes of interest used in this study are lipid traits, including high-density lipoprotein-cholesterol (HDL-C), low-density lipoprotein-cholesterol (LDL-C), total cholesterol, and triglycerides, all of which are important risk factors for a multitude of diseases[3–5]. It is well established that lipid traits are influenced by a variety of factors, including genetic inheritance, environmental and occupational exposures, medication use, ethnicity, and sex[6,7]. In this study, we define environmental exposure as any physical, chemical, or biological agent that someone is exposed to and has potential to cause a wide range of health effects. The dietary exposures in this study refer to the intake of nutrients that can either benefit, harm, or have no effect on one's health.

Due to the scale of risk variables available in contemporary cohort and biobank datasets, many researchers perform variable selection (or filtering) prior to statistical or computational modeling. The shift towards knowledge-based filtering in these studies has been shown to be an effective alternative to main effect filtering (whereby variables are filtered based on having a statistically significant independent effect), especially for variables that only exhibit an effect in the context of another variable. The incorporation of prior biological knowledge to prioritize genetic variants that are more likely to interact with one another has revealed numerous GxG for complex diseases[8-11]. However, these studies have been restricted to knowledge about genes and have not included knowledge of the biological relationship between exposures. Thus, we propose that ExE coupled with knowledge-based filtering represents a promising approach to further elucidate the complexities of ExE in health and disease. This paper introduces the Gene x Exposome database (GE.db) module of the Integrative Genome-Exposome Method (IGEM) system[12], a knowledge base of genomic and exposomic interactions derived from various public databases [see Methods]. The development of GE.db aims to leverage prior knowledge to filter high-volume research datasets, retaining only variables with known biological relationships. This approach significantly reduces the number of variables for analysis, conserves computational resources and processing time, and minimizes type I errors following multiple testing corrections.

To demonstrate the utility of GE.db, we conducted an ExE analysis with lipid traits using the National Health and Nutritional Examination Survey (NHANES)[13] data from 1999-2018. By focusing on an exposome-wide interaction approach and utilizing GE.db, this research can provide important insights for the prevention and management of lipid-based health risk factors. Additionally, this study highlights the potential of GE.db to enhance the efficiency and accuracy of -omics research by providing a knowledge base resource for filtering datasets based on known interactions, thereby facilitating more focused and reliable statistical and computational analyses.

2. Methods

2.1 NHANES Dataset

The National Health and Nutrition Examination Survey (NHANES) is an ongoing initiative conducted by the Centers for Disease Control and Prevention (CDC) aimed at evaluating the health and nutritional status of the U.S. population[14]. Its primary objectives include identifying risk factors

for prevalent diseases and informing the development of public health policies. Data collection encompasses a wide range of participant information including demographics, dietary recalls, health examinations, toxin exposures, and laboratory measurements, all obtained through structured interviews and physical examinations conducted either at participants' homes or mobile testing centers.

Datasets were extracted from the NHANES website[15], covering the cycles from 1999 to 2018. Specifically, the focus was on testing the exposomic variables only for this study. These datasets were integrated into a comprehensive table, where each row corresponds to a participant and each column represents a specific NHANES variable. This cumulative dataset consists of 101,316 participants and 11,274 variables spanning multiple domains, including demographic, dietary, health, examination, laboratory, questionnaire, socioeconomic, and occupational categories including all phenotype, exposure, and covariate information sourced from the NHANES database. From this comprehensive data, we were able to select the specified lipid phenotypes and exposures relevant to our study. It is noteworthy that NHANES fields are not consistently maintained across cycles; fields may be modified or discontinued over time, posing challenges for longitudinal analyses[16].

2.2 GE.db

The GE.db module is an integral component of the IGEM system[12], designed as a comprehensive knowledge base of genomic and exposomic interactions. This module aggregates data from various public databases, providing a curated repository of interactions that can be leveraged to filter high-volume research datasets effectively. The primary purpose of GE.db is to utilize prior knowledge of gene-exposure and exposure-exposure interactions to filter datasets, thereby retaining only variables with known biological relationships. The aim of strategic filtering is to significantly reduce the number of variables requiring analysis so as to conserve computational resources, reduce processing time, and minimize the occurrence of type I errors after multiple testing corrections.

2.2.1 Data Sources

GE.db derives its data from multiple reputable public databases that are frequently updated and maintained. As a foundational step in developing the exposure terms, IGEM incorporates an integration system of environmental and genetic data as it uses a rigorous process of standardizing and mapping terms. To facilitate this task, we use MeSH (Medical Subject Headings)[17] from the National Center for Biotechnology Information (NCBI), a widely recognized database of biomedical descriptors, as part of the word pre-processing procedure. The main function of MeSH in IGEM is to serve as a reference dictionary to standardize and consolidate different forms of terms that appear in various data sources. For instance, in the context of chemical exposures, the same chemical compound might be referred to in different ways, either by its chemical formula (e.g., "C6H12O6" for glucose), its full name (e.g., "glucose"), or a numeric code or identifier. The word pre-processing procedure in IGEM uses MeSH to identify all these variations and then assigns a unique and consolidated identifier to each term. This unified identifier ensures that IGEM recognizes all these

forms as the same concept, providing consistency across the data and facilitating the integration of external sources. Moreover, this mapping allows IGEM to link data from multiple external databases, ensuring that the same terms can be identified in different contexts, such as environmental exposures or clinical records, regardless of how they were originally represented. The final product is a standardized and unified knowledge base that simplifies the analysis of interactions of the environmental terms, improving both the efficiency and accuracy of scientific discoveries.

For this analysis, the following databases were considered as they provide relevant environmental information: Human Metabolome Database (HMDB)[18], a detailed resource containing information on small molecule metabolites found in the human body, crucial for understanding metabolic interactions and pathways; Comparative Toxicogenomics Database (CTD)[19], which integrates information on chemical-gene/protein interactions, chemical-disease, and gene-disease relationships, facilitating insights into the molecular mechanisms of environmental diseases; and Kyoto Encyclopedia of Genes and Genomes (KEGG)[20], which provides comprehensive data on gene functions, biological pathways, diseases, drugs, and chemical substances, supporting the integration of genomic and metabolic information. This methodology allowed the identification and recording of interactions where multiple exposure factors were found in the same record.

At the time of analysis, the GE.db contained 1,057,827 terms grouped into categories such as anatomy, chemicals, diseases, chromosomes, genes, metabolites, pathways, and SNPs, along with 15,667,807 interactions among these terms. The GE.db module is designed with a flexible architecture that allows for the seamless integration of new data sources. It includes several key components: Term Table, which contains key terms and concepts essential for the analysis, organized into groups and categories for efficient retrieval; Interaction Table, which stores documented interactions between various genomic and exposomic variables, providing a robust foundation for filtering datasets; and Mapping Algorithms, which utilize advanced algorithms to match external data terms to internal GE.db terms, ensuring consistency and reliability in the filtering process. To maintain the GE.db, the IGEM system employs version control routines and layers of data ingestion and data transformation to fetch data from their sources and transform them into term links (Figure 1). The GE.filter is another component of IGEM that enables various operations on the GE.db knowledge base, including term matching, interaction identification, and data reduction. The IGEM system, along with its modules GE.db and GE.filter, is deployed in a Python environment on an institutional linux computing cluster. The database utilized is SQLite, which currently has a size of 2.7 GB. For a more detailed explanation on the workflow and filtration parameters used within each command involving Ge.db and GE.filter, please refer to our user guide located on Github[21].

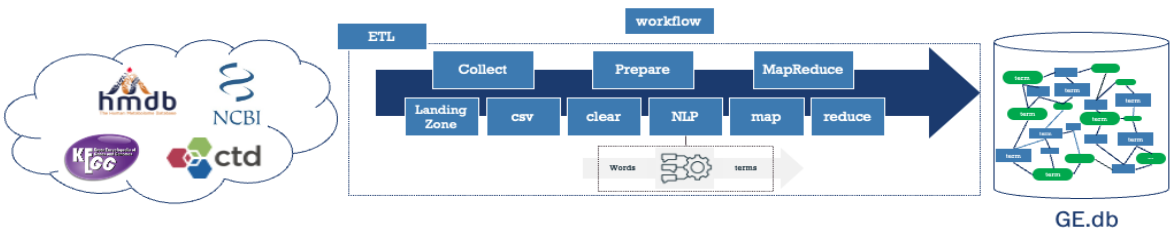


Figure 1. Visualization of GE.db workflow from database to interaction term identification.

2.3Phenotypes and Confounder Variables

Within the NHANES dataset, specific variables were identified as phenotypes and confounders for this analysis. The selected phenotypes included are listed in Table 1. For HDL-C, NHANES altered the calculation method for this indicator over different cycles. NHANES encountered method-related bias for calculating the HDL-C values for 1999-2000, 2001-2002, and 2005-2006; the bias for 2003-2004 was acceptable (<4%) and required no correction[22]. The adjustments implemented improved consistency across various years and methodologies, ensuring that the differences observed in HDL-C levels more accurately reflected true variations rather than being impacted by measurement bias. Consequently, these three fields were maintained separately, creating three distinct datasets. The selected confounders included Gender (RIAGENDR), Age (RIDAGEYR), BMI (BMXBMI), Race/Ethnicity (RIDRETH1), and Survey Cycle (SDDSRVYR).

2.4 Adjusting for Cholesterol Medications

To account for the influence of cholesterol-lowering medications on lipid measurements, we adjusted the LDL-C and Total Cholesterol (TC) values for participants who reported using statins (Figures S-1,2). This adjustment is crucial for accurately assessing lipid levels and their associations with various exposures, as statins significantly alter cholesterol levels. We utilized the NHANES dataset RXQ_RX to identify participants who reported using at least one of the following statin components: ATORVASTATIN CALCIUM, SIMVASTATIN, PRAVASTATIN SODIUM, and FLUVASTATIN SODIUM.

For these participants, we adjusted the LDL and TC values as follows: LDL-cholesterol (LBDLDL) values were divided by 0.7 to account for the reduction effect of statins, and Total Cholesterol (LBXTC) values were divided by 0.8 to adjust for statin usage[23]. By incorporating these adjustments, we enhanced the precision of our lipid measurements, ensuring that our analysis of exposure-lipid interactions was both accurate and reliable.

2.5 NHANES Exposure Filtering for the Interaction Models

To align the NHANES variables with GE.db, all NHANES variable descriptions (excluding lipid phenotypes and confounders) were processed through the GE.filter function. GE.filter utilizes an internal NLP (Natural Language Processing) engine to identify corresponding GE.db terms based on textual descriptions. This process identified 3,619 NHANES variables related to 534 GE.db terms.

A subsequent review of these related NHANES variables identified 1,136 exposure factors, corresponding to 217 unique terms. These 217 terms were then used as filter parameters for another GE.filter function run, which searched the GE.db knowledge base for all interactions among these terms, resulting in the identification of 382,613 putative Exposure x Exposure interactions. We performed this step prior to quality controlling the exposure variables, ensuring that only exposures present in the NHANES data were included for curation of the interactions to be tested.

2.6 Quality Control (QC)

Quality Control is a critical step to ensure the integrity, reliability, and validity of the dataset used in the analysis. The IGEM system includes specialized functions that accelerate and assist in the

application of QC procedures to -omics data analyses. The following procedures were applied to the NHANES dataset after filtering and modifications from previous steps.

For continuous data type QC, all variables with more than 90% missing values were removed. The distribution of phenotypes was calculated using the skewness (3(mean-median)/standard deviation.) and all phenotypes were log-transformed to normalize the distribution (Figure S-3).

Participants were then separated into discovery and replication groups for the six cohorts of phenotypes, resulting in twelve datasets. For each dataset, a minimum of 200 participants for categorical and binary exposures was maintained. Only variables present in both discovery and replication datasets for each phenotype were retained to ensure consistency and reliability (Table 1).

Table 1. Overview of lipid phenotypes sorted by survey cycle, including sample sizes, exposures, and interactions that passed quality control.

Phenotype	NHANES Cycles	NHANES ID	NHANES Description	N Discovery	N Replication	Exposures	Interactions
HDL-C	1999 – 2002	LBDHDL	HDL-cholesterol, mg/dL	4,572	4,949	96	2,073
HDL-C	2003 – 2004	LBXHDD	Direct HDL-Cholesterol, mg/dL	3,425	1,469	219	11,093
HDL-C	2005 – 2018	LBDHDD	HDL-Cholesterol, mg/dL	21,442	16,000	231	11,721
LDL-C	1999 – 2018	LBDLDL	LDL-cholesterol, mg/dL	11,453	12,695	181	6,934
Total Cholesterol	1999 – 2018	LBXTC	Total Cholesterol, mg/dL	24,836	27,023	193	7,873
Triglycerides	1999 – 2018	LBXSTR	Triglycerides, mg/dL	19,305	26,916	177	6,446

2.7 Statistical Analysis Models (Discovery and Replication)

The IGEM system, inheriting functionalities from the CLARITE system[24], performs interaction analyses by calculating the p-value of the Likelihood Ratio Test (LRT) between two models. In the full and reduced model $Y_{phenotype}$ is the outcome variable, β_0 is the intercept, $\beta_1 term1$ and $\beta_2 term2$ are the coefficients for the individual predictors, and $\beta_{n+1} cov_n$ are the coefficients for the covariates with n adding on to the number of covariates used in the model. Exclusive to the full model, $\beta_3(term1 \times term2)$ is the interaction term between term 1 and term 2.

Full Model:

$$Y_{phenotype} = \beta_0 + \beta_1 term1 + \beta_2 term2 + \beta_3(term1 \times term2) + \beta_4 cov_1 + \ldots + \beta_{n+1} cov_n \quad (1)$$

Reduced Model:

$$Y_{phenotype} = \beta_0 + \beta_1 term1 + \beta_2 term2 + \beta_3 cov_1 + \ldots + \beta_{n+1} cov_n \quad (2)$$

The LRT is utilized to compare the fit of the two models, with the full model including the interaction term ($\beta_3(term1 \times term2)$) and the reduced model excluding it. The analysis involves fitting the full model to the data to obtain the log-likelihood (L_{full}) and fitting the reduced model to obtain the log-likelihood ($L_{restricted}$). The LRT statistic represented as D with -2 used as a scaling factor that makes the likelihood ratio test statistic approximately follow a chi-squared distribution under the null hypothesis is calculated as:

$$D = -2(L_{restricted} - L_{full}) \quad (3)$$

The difference in degrees of freedom between the two models is 1, since the full model has one additional parameter ($\beta_3(term1 \times term2)$). The p-value is derived from the probability (P) that a random variable following a chi-squared distribution ($\chi 2$) with 1 degree of freedom takes a value greater than or equal to the observed test statistic (D):

$$p\text{-}value = P(\chi 2 \geq D \mid df = 1) \quad (4)$$

The LRT p-values were calculated for each interaction identified in the discovery dataset for each phenotype.

However, in some cases, the p-value of the LRT cannot be calculated. The following messages inform the user of the reasons:

- Too few complete observations (min_n filter: N < 200)
- The number of complete observations is insufficient to perform the analysis, as the minimum required is 200
- Both models are equivalent in terms of fit: the two models are equivalent in terms of fit, with no significant difference between them
- No Overlap (min_n filter: 0 < 200): there is insufficient data overlap to perform the analysis, as the minimum required is 200.

Following the interaction model analysis, the IGEM function was applied to adjust the p-values for multiple testing using both Bonferroni correction and False Discovery Rate (FDR) adjustment. From the discovery analysis, interactions with an FDR-adjusted p-value < 0.1 were filtered. These significant interactions that met the FDR adjustment threshold were then tested in the replication dataset. The same interaction analysis was conducted in the replication cohort, applying identical model specifications and LRT. The replication criteria also required that interactions exhibit consistent directional effects between the discovery and replication interaction betas, with all significant interactions retaining a Bonferroni-adjusted p-value < 0.05, across both datasets. This rigorous approach ensures that the identified interactions are robust and not due to random chance.

3. Results

In this study, we examined the interactions between various exposure variables and lipid phenotypes using the NHANES dataset. We performed a comprehensive analysis to identify significant exposure-exposure interactions (ExE) that are associated with lipid levels. Below are the key findings from our discovery and replication datasets. Of all the 26,107 interactions tested that included exposures that passed QC, a total of 263 interactions were statistically significant in the

discovery dataset with an FDR p < 0.1 (Table 2). A total of 61 interactions were found to be significant in both discovery and replication when allowing for an FDR p < 0.1 (assorted by lipid phenotype) and 21 interactions associated with the HDL-cholesterol trait was significant with a Bonferroni corrected p < 0.05 (Figure 2). Additionally, these interactions demonstrated consistent directions of effect across both discovery and replication datasets (Table S-1).

Table 2. Frequency table of all the interactions tested for every lipid phenotype.

Phenotype	Discovery Interactions	Replication Interactions	FDR p < 0.1 in both	Bonferroni p < 0.05 in both
HDL-C [1999-2002]	1,116	4	1	1
HDL-C [2003-2004]	5,459	93	2	0
HDL-C [2005-2018]	6,584	141	58	20
LDL-C	4,339	9	0	0
Total Cholesterol	4,764	10	0	0
Triglycerides	3,845	6	0	0
Total	26,107	263	61	21

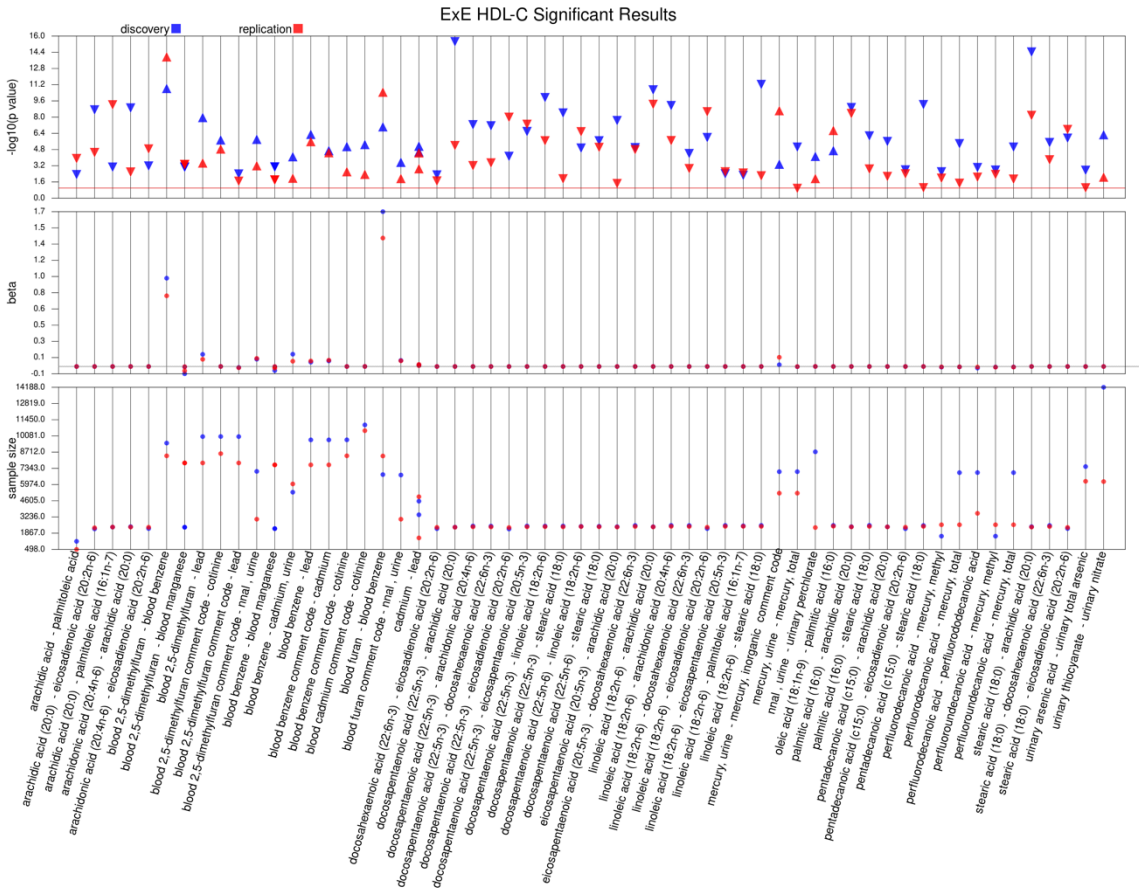
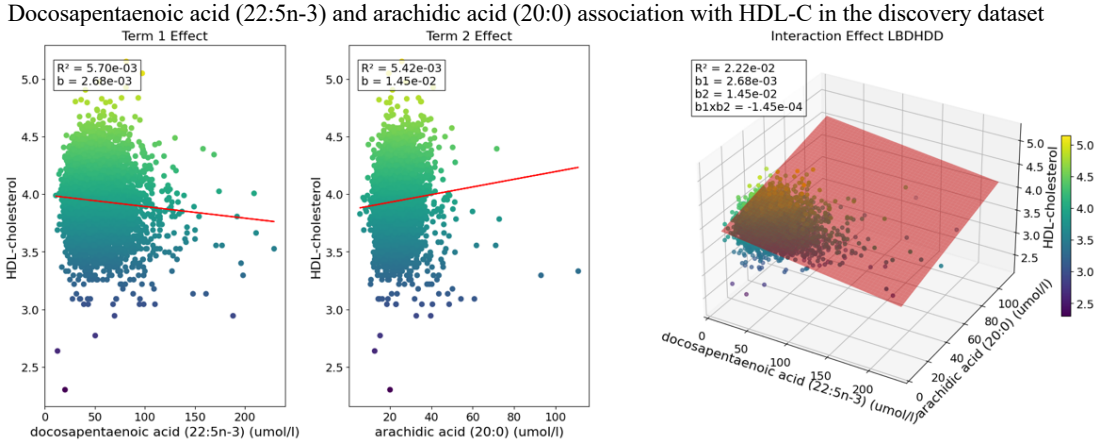


Figure 2. The sixty-one significant results starting from the top showcasing all the interactions with FDR LRT p-value < 0.1(denoted by the redline) and the twenty-one significant results with a Bonferroni adjusted LRT p-value < 0.05 (direction of effect pointing down is negative and up is positive), the interaction beta for both exposures, and the sample sizes. PheWAS-View was the software used to generate this plot[25].

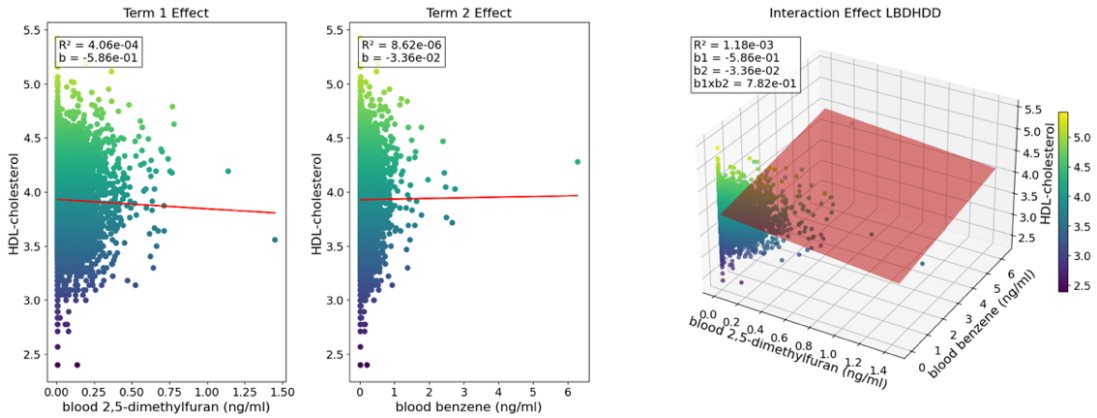
3.1 Significant Interactions

The top three results with the lowest LRT p-values associated with HDL-cholesterol include: 1) Docosapentaenoic acid (22:5n-3) (DPA) - arachidic acid (20:0) (Discovery: Bonferroni adjusted LRT p-value = 8.43×10^{-13}, $\beta = -1.4\times10^{-4}$; Replication: Bonferroni adjusted LRT p-value = 3.25×10^{-4}, $\beta = -1.2\times10^{-4}$) (Figure 3A). 2) Blood 2,5-dimethyfuran - blood benzene (Discovery: Bonferroni adjusted LRT p-value = 2.75×10^{-7}, $\beta = 0.97$; Replication: Bonferroni adjusted LRT p-value = 4.48×10^{-12}, $\beta = 0.78$) (Figure 3B). 3) Stearic acid (18:0) - arachidic acid (20:0) (Discovery: Bonferroni adjusted LRT p-value = 8.88×10^{-12}, $\beta = -7.79\times10^{-6}$; Replication: Bonferroni adjusted LRT p-value = 3.47×10^{-7}, $\beta = -1.26\times10^{-5}$) (Figure 3C).

(A) Docosapentaenoic acid (22:5n-3) and arachidic acid (20:0) association with HDL-C in the discovery dataset



(B) Blood 2,5 dimethylfuran (ng/ml) and blood benzene (ng/ml) association with HDL-C in the replicate dataset



(C) Stearic acid (18:0) (umol/l) and arachidic acid (20:0) (umol/l) association with HDL-C in the discovery dataset

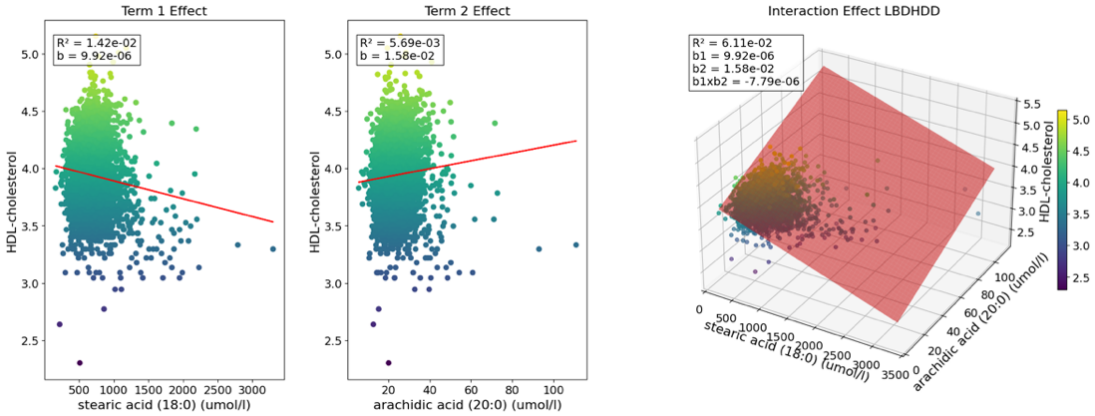


Figure 3A-C. The top three results plots observing the individual main effect correlation line, and the 3D plot showing the interaction correlation along the square plane.

4. Discussion

In this study, we leveraged the comprehensive exposomic knowledge base provided by the GE.db module of IGEM to investigate exposure-exposure interactions (ExE) associated with lipid phenotypes. By utilizing data from the NHANES dataset spanning 1999 to 2018, we identified several significant interactions between various exposures and lipid levels. The replication of these findings across independent datasets underscores the robustness of our approach and highlights the potential of GE.db in facilitating large-scale -omics research.

4.1 Clinical and Public Health Implications

Our analysis revealed several key interactions, notably DPA and stearic acid with arachidic acid associated with HDL-C. DPA is a known essential omega-3 fatty acid, and stearic and arachidic acid are saturated fatty acids[26-28]. These results suggest that specific combinations of environmental exposures may have synergistic effects on lipid metabolism, though most research only touches on their individual effects on lipid profiles. For instance, omega-3 fatty acids, such as DPA, are generally linked with increased HDL cholesterol levels[29], while high consumption of saturated fatty acids like arachidic acid may unfavorably affect lipid profiles, potentially leading to elevated LDL-C levels[27]. Our findings indicate a negative impact on HDL-C when arachidic acid interacts with fatty acids typically associated with positive HDL-C effects, suggesting that arachidic acid could potentially diminish the benefits of HDL-C promoting fatty acids. Other research suggests that stearic acid may have a neutral or even beneficial effect on cholesterol levels, possibly not adversely affecting HDL-C on its own[30]. However, as seen in our results, when combined with arachidic acid, this interaction could overall have a negative impact, counteracting any neutral or positive effects on HDL-C.

Additionally, the interaction between blood 2,5-dimethylfuran and blood benzene highlights the potential combined impact of exposure to volatile organic compounds (VOCs) on HDL-C levels. Benzene has been observed to increase LDL-C levels which would naturally displace or plateau

HDL-C levels procuring a negative effect[31,33]. Measures of 2,5-dimethyfuran though, have limited research indicating influence on lipids, but may pose health risks similar to other VOCs. These risks can include respiratory irritation, and potential systemic effects that could indirectly affect lipid metabolism and cardiovascular health[34–36]. Conversely, our results demonstrate a positive interaction effect on HDL-C with benzene and 2,5-dimethylfuran. Therefore, further study of this interaction is warranted, especially considering the known detrimental impact of VOCs on public health. In summary, all these findings have important implications for public health, as they point to the need for considering multiple concurrent exposures in dietary and environmental risk assessments. Public health strategies could be developed to mitigate the combined effects of specific dietary and environmental exposures on lipid metabolism.

4.1.1 Significant Interaction Effects Sizes

As stated previously for HDL-C, the bias adjustment was accounted for whether the survey cycle year had been corrected or not as they were all approved to use for statistical analysis. The HDL-C variable still had to be labeled and categorized differently to identify which ones were corrected vs. not corrected for transparency. Given that the LBDHDD variable spanned the largest survey cycle from 2005-2018 of the three, showed an increased sample size disparity by about 17,000 participants when comparing the other two survey cycles which had around 4,000 participants each. Thus, presuming that even if the effect size remains similar across those survey cycles, a larger sample size in one cycle can lead to a significant p-value, while a smaller sample size in another cycle could result in a non-significant p-value for the same effect size.

Regarding the effect sizes of the three significant interactions mentioned, we believe the positive beta for DPA can coexist with a slight negative trend due to the small effect size and interaction with arachidic acid (Figure 3A). The combined effect of DPA and arachidic acid as described by the interaction term, may influence the overall outcome more than the individual effect of DPA alone. In the dataset, the interaction between the two terms might reduce or counterbalance DPA's small positive effect on HDL-C. Figure 3B depicts another story where the two weaker effect sizes of blood 2,5-dimethylfuran and blood benzene alone hold less weight than compared to the larger effect size of when both blood 2,5-dimethylfuran and blood benzene increase together. Their combined effect led to an overall increase in HDL-C despite their individual negative contributions. Lastly, the negative interaction effect size for stearic acid and arachidic acid is very small in association with HDL-C, and largely driven by the positive influence of arachidic acid (Figure 3C). The overall interaction appears to slightly counteract the combined positive effects of both terms but not enough to reverse the trend significantly. Thus, a large amount of the variation is most likely not fully explained in this model and further testing is required.

4.2 Methodological Strengths, Limitations, and Future Directions

A major strength of this study is the use of the GE.db knowledge base, which allowed us to filter high-volume research datasets effectively, focusing only on variables with known biological relationships. This approach significantly reduced the computational burden and enhanced the reliability of our findings by minimizing type I errors through multiple testing corrections. By

employing multiple IGEM modules, we streamlined quality control (QC) processes, which involved variable categorization, data cleaning, and adjustment for confounders like statin use. This approach improved the integrity and accuracy of our analysis, making it user-friendly for whomever uses this tool, and ensuring alignment to bioinformatics practices. The split of data into discovery and replication datasets based on NHANES cycles further increased the validity of our results, as significant interactions identified in the discovery phase were consistently replicated. Another consideration to note is the main-effect interaction model when incorporated without the use of knowledge-driven filters, is typically performed to determine the isolated impact of each variable (in this case, each exposure factor) on phenotypes. However, the goal of this study was not to identify individual main effects but to examine how the combination of multiple exposures influences lipid phenotypes. While the standard main-effect model is valuable in other contexts like simple-trait analysis or in situations where the effects of multiple variables are purely additive, our primary focus was to highlight IGEM's strengths, particularly in capturing interactions based on the pre-existing knowledge within GE.db. GE.db was specifically designed to filter highly relevant variables based on known relationships between exposures. Using this filtering approach allows the analysis to focus on variables with biological context, avoiding the processing of many irrelevant exposures or statistical noise that could arise when including non-interactive main effects.

Despite the robustness of our findings, several limitations warrant consideration. First, since GE.db relies on public databases such as HMDB, CTD, and KEGG, the quality, completeness, and update frequency of these external databases can directly affect the accuracy and relevance of the information in GE.db. Any gaps, errors, or outdated information in these sources could introduce bias or limitations in the results. Regular updates are imperative to ensure the data remains current, but the complexity of fetching and processing new data might slow down the user's analysis pipeline. Furthermore, the interactions stored in GE.db are curated from specific sources, and their generalizability to other populations, environmental contexts, or less-studied interactions may be limited. Results may not always be applicable outside the scope of the databases from which they were derived.

In the context of the NHANES dataset, the observational nature of the data limits the ability to infer causal relationships between exposures and lipid levels. Interaction effects, as we have noted, may have opposite signs of effect when compared to the main effect betas, which complicates the interpretability of the results. Other datasets with repeated measures of QC and analysis as we have specified with the NHANES data, can help with cross checking all the betas, refining the elucidation of significant interactions. Inclusion of more datasets that host the same kinds of environmental exposures such as the UK Biobank[37] and All of Us Research Program[38], will also help address the possibility of false negatives as some interactions may not have been flagged as significant given our designated thresholds used for the NHANES dataset. Future studies could also incorporate longitudinal data and more sophisticated causal inference methods to address this limitation.

Moreover, while our analysis accounted for several covariates, there may be other unmeasured factors that could influence the observed interactions. Further research should aim to include a broader range of potential confounders and explore the underlying biological mechanisms driving these interactions. Another limitation is the reliance on self-reported data for certain exposures,

which may introduce reporting biases. The integration of more objective measures of exposure, such as well-established biomarkers, could enhance the reliability of future analyses.

4.3 Conclusion

In conclusion, this study demonstrates the utility of the GE.db module in identifying significant ExE influencing lipid traits. The consistent replication of key interactions across independent variables highlights the robustness of our approach and its potential to uncover future novel insights into the complex interplay between environmental exposures and lipid metabolism. These findings pave the way for future research aimed at understanding and mitigating the multifactorial nature of dyslipidemias, ultimately contributing to improved public health outcomes.

This project was supported by the the National Institute of Child Health and Human Development under award number U2C OD023375-06 and the National Heart Lung, and Blood Institute under awards HL169458 and HL168841. This work was additionally supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under Project #PEN04275 and Accession #1018544.

Code for GE.db, GE.db filter, and quality control steps used in this study are made available here: https://github.com/HallLab/pbs_igem/tree/main. The IGEM package and user guide are available here: https://github.com/HallLab/IGEM.
Supplemental table and figures S-1, S-2, and S-3 are available at https://ritchielab.org/publications/supplementary-data/psb-2025/igem.

References

1. Virolainen, S. J., VonHandorf, A., Viel, K. C. M. F., Weirauch, M. T. & Kottyan, L. C. Gene-environment interactions and their impact on human health. *Genes Immun.* **24**, 1–11 (2023).
2. Ottman, R. Gene-environment interaction: definitions and study designs. *Prev. Med.* **25**, 764–770 (1996).
3. Castelli, W. P. Lipids, risk factors and ischaemic heart disease. *Atherosclerosis* **124 Suppl**, S1–9 (1996).
4. Emerging Risk Factors Collaboration *et al.* Major lipids, apolipoproteins, and risk of vascular disease. *JAMA* **302**, 1993–2000 (2009).
5. Dayimu, A. *et al.* Trajectories of Lipids Profile and Incident Cardiovascular Disease Risk: A Longitudinal Cohort Study. *J. Am. Heart Assoc.* **8**, e013479 (2019).
6. Amin, K. A., Homeida, A. M., El Mazoudy, R. H., Hashim, K. S. & Garelnabi, M. Dietary Lipids in Health and Disease. *J. Lipids* **2019**, 5729498 (2019).

7. Hornburg, D. *et al.* Dynamic lipidome alterations associated with human health, disease and ageing. *Nat Metab* **5**, 1578–1594 (2023).
8. Ritchie, M. D. *et al.* Incorporation of Biological Knowledge Into the Study of Gene-Environment Interactions. *Am. J. Epidemiol.* **186**, 771–777 (2017).
9. Pendergrass, S. A. *et al.* Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData Min.* **6**, 25 (2013).
10. Kim, D. *et al.* Biofilter as a functional annotation pipeline for common and rare copy number burden. *Pac. Symp. Biocomput.* **21**, 357–368 (2016).
11. Hall, M. A. *et al.* Biology-driven gene-gene interaction analysis of age-related cataract in the eMERGE Network: Biology-driven tool to identify genetic interactions. *Genet. Epidemiol.* **39**, 376–384 (2015).
12. Term — IGEM 0.1.0 documentation. https://igem.readthedocs.io/en/latest/ge/md/term.html.
13. National health and nutrition examination survey. https://www.cdc.gov/nchs/nhanes/index.htm (2024).
14. About the national health and nutrition examination survey. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm (2024).
15. CDC. Centers for disease control and prevention. https://www.cdc.gov/ (2024).
16. Nguyen, V. K. *et al.* Harmonized US National Health and Nutrition Examination Survey 1988-2018 for high throughput exposome-health discovery. *medRxiv* (2023) doi:10.1101/2023.02.06.23284573.
17. Sievert, M., Patrick, T. & Reid, J. Need a bloody nose be a nosebleed? or, lexical variants cause surprising results. *Bull. Med. Libr. Assoc.* **89**, 68–71 (2001).
18. Wishart, D. S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **35**, D521–6 (2007).
19. Davis, A. P. *et al.* Comparative Toxicogenomics Database (CTD): update 2023. *Nucleic Acids Res.* **51**, D1257–D1262 (2023).
20. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
21. *User Guider: IGEM v.0.1.4*. (Github, 2023).

22. NHANES 2005-2006: Cholesterol - HDL Data Documentation, Codebook, and Frequencies. https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/HDL_D.htm#LBDHDD.
23. Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
24. Lucas, A. M. *et al.* CLARITE Facilitates the Quality Control and Analysis Process for EWAS of Metabolic-Related Traits. *Front. Genet.* **10**, 1240 (2019).
25. Wolfe, D., Dudek, S., Ritchie, M. D. & Pendergrass, S. A. Visualizing genomic information across chromosomes with PhenoGram. *BioData Min.* **6**, 18 (2013).
26. Human Metabolome Database: Showing metabocard for Stearic acid (HMDB0000827). https://hmdb.ca/metabolites/HMDB0000827.
27. Human Metabolome Database: Showing metabocard for Docosapentaenoic acid (22n-3) (HMDB0006528). https://hmdb.ca/metabolites/HMDB0006528.
28. Human Metabolome Database: Showing metabocard for Arachidic acid (HMDB0002212). https://hmdb.ca/metabolites/HMDB0002212.
29. Peña-de-la-Sancha, P. *et al.* Eicosapentaenoic and Docosahexaenoic Acid Supplementation Increases HDL Content in n-3 Fatty Acids and Improves Endothelial Function in Hypertriglyceridemic Patients. *Int. J. Mol. Sci.* **24**, (2023).
30. Siri-Tarino, P. W., Sun, Q., Hu, F. B. & Krauss, R. M. Saturated fatty acids and risk of coronary heart disease: modulation by replacement nutrients. *Curr. Atheroscler. Rep.* **12**, 384–390 (2010).
31. Grundy, S. M. Influence of stearic acid on cholesterol metabolism relative to other long-chain fatty acids. *Am. J. Clin. Nutr.* **60**, 986S–990S (1994).
32. Tualeka, N. A. R., Martiana, N. T., Wibrata, A. & Rahmawati, P. Effect of food consumption contain glutahione anti-oxidant towards LDL cholesterol concentrations on benzene-exposed-workers at the romokalisari shoe industry, Surabaya. *Indian J. Forensic Med. Toxicol.* (2019) doi:10.5958/0973-9130.2019.00333.5.

33. Ye, L. *et al.* Moderate body lipid accumulation in mice attenuated benzene-induced hematotoxicity via acceleration of benzene metabolism and clearance. *Environ. Int.* **178**, 108113 (2023).
34. Fu, X. *et al.* Airborne 2,5-dimethylfuran as a marker to indicate exposure to indoor tobacco and biomass burning smoke. *Atmos. Environ.* **259**, 118509 (2021).
35. Jing, L., Chen, T., Yang, Z. & Dong, W. Association of the blood levels of specific volatile organic compounds with nonfatal cardio-cerebrovascular events in US adults. *BMC Public Health* **24**, 616 (2024).
36. Chen, X. *et al.* Association of Smoking with Metabolic Volatile Organic Compounds in Exhaled Breath. *Int. J. Mol. Sci.* **18**, (2017).
37. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
38. All of Us Research Program Investigators *et al.* The 'All of Us' Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).

Frequency of adding salt is a stronger predictor of chronic kidney disease in individuals with genetic risk

Manu Shivakumar[1†] and Yanggyun Kim[1,2†], Sang-Hyuk Jung[1], Jakob Woerner[1], Dokyoon Kim[1*]

[1]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA*

[2]*Department of Internal Medicine, Kyung Hee University Hospital at Gangdong, College of Medicine, Kyung Hee University, Seoul, Korea*

[†]*Equal Contribution*

[*]*Email: Dokyoon.Kim@pennmedicine.upenn.edu*

The incidence of chronic kidney disease (CKD) is increasing worldwide, but there is no specific treatment available. Therefore, understanding and controlling the risk factors for CKD are essential for preventing disease occurrence. Salt intake raises blood pressure by increasing fluid volume and contributes to the deterioration of kidney function by enhancing the renin-angiotensin system and sympathetic tone. Thus, a low-salt diet is important to reduce blood pressure and prevent kidney diseases. With recent advancements in genetic research, our understanding of the etiology and genetic background of CKD has deepened, enabling the identification of populations with a high genetic predisposition to CKD. It is thought that the impact of lifestyle or environmental factors on disease occurrence or prevention may vary based on genetic factors. This study aims to investigate whether frequency of adding salt has different effects depending on genetic risk for CKD. CKD polygenic risk scores (PRS) were generated using CKDGen Consortium GWAS (N= 765,348) summary statics. Then we applied the CKD PRS to UK Biobank subjects. A total of 331,318 European individuals aged 40-69 without CKD were enrolled in the study between 2006-2010. The average age at enrollment of the participants in this study was 56.69, and 46% were male. Over an average follow-up period of 8 years, 12,279 CKD cases were identified. The group that developed CKD had a higher percentage of individuals who added salt (46.37% vs. 43.04%) and higher CKD high-risk PRS values compared to the group that did not develop CKD (23.53% vs. 19.86%). We classified the individuals into four groups based on PRS: low (0-19%), intermediate (20-79%), high (80-94%), very high (≥ 95%). Incidence of CKD increased incrementally according to CKD PRS even after adjusting for age, sex, race, Townsend deprivation index, body mass index, estimated glomerular filtration rate, smoking, alcohol, physical activity, diabetes mellitus, dyslipidemia, hypertension, coronary artery diseases, cerebrovascular diseases at baseline. Compared to the "never/rarely" frequency of adding salt group, "always" frequency of adding salt group had an increasing incidence of CKD proportionate to the degree of frequency of adding salt. However, the significant association of "always" group on incident CKD disappeared in the low PRS group. This study validated the signal from PRSs for CKD across a large cohort and confirmed that frequency of adding salt contributes to the occurrence of CKD. Additionally, it confirmed that the effect of frequency of "always" adding salt on CKD incidence is greater in those with more than intermediate CKD-PRS. This study suggests that increased salt intake is particularly concerning for individuals with genetic risk factors for CKD, underscoring the clinical importance of reducing salt intake for these individuals.

Keywords: chronic kidney disease; polygenic risk score; salt; lifestyle factors; UK Biobank

1. Introduction

With an aging population and changes in dietary habits, 10-16% of adults are experiencing chronic kidney disease (CKD)[1, 2]. CKD increases the incidence of cardiovascular disease and raises mortality rates, posing a threat to human health. However, aside from controlling the underlying causes, there is no definitive treatment for CKD, making it difficult for patients to be free from the disease once diagnosed. The causes of CKD are highly variable, and the heterogenous genetic backgrounds make it challenging to pinpoint the genetic predisposition for CKD[3]. While Genome-Wide Association Studies (GWAS) provide with information on thousands of candidate genetic variants associated with diseases, the individual contribution of each genetic variant to the disease is very small, making clinical application difficult. However, polygenic risk score (PRS) analysis allows us to effectively utilize GWAS results by analyzing the cumulative effect of all common variants and their correlation with disease occurrence. Several studies have used PRS to predict disease occurrence and stratify the genetic risk to enhance traditional factors in diabetes, heart disease, and obesity using PRS[4-7]. Despite high heritability, genetic factors offer relatively low predictive ability for CKD[8]. Nevertheless, Khan et al. developed a strong CKD PRS based on meta-analyzed GWAS studies using good quality medical data on large-sized populations across ancestries[9]. They demonstrated a reproducible and high-performing PRS to predict the incidence of CKD, which was consistent across various ancestries.

Lifestyle factors have also been known to play a crucial role in the development and progression of CKD. High physical activity reduces the risk of obesity, metabolic diseases, and cardiovascular diseases, and contribute to alleviating CKD[10, 11]. Excessive salt intake increases fluid volume in the body and stimulates the renin-angiotensin system (RAS) and the neurohormonal system, leading to hypertension, which increases the risk of CKD[12]. Therefore, reducing salt intake is critical not only for preventing and managing hypertension and heart disease, but also for preventing and controlling CKD. However, there are conflicting results on the relationship between salt intake and CKD[13, 14]. Accurately assessing the amount of salt consumed is challenging, so most studies indirectly estimate salt intake based on the amount of salt excreted in urine. Urine salt excretion is influenced by expression and activation of numerous renal tubular transporters and is intertwined with neurohormonal factors that control them[15]. This regulatory system can change based on age, sex, underlying diseases, and medications. Therefore, it is difficult to assert that measuring the amount of salt in urine accurately reflects the amount of salt consumed. Consequently, there have not been many studies clearly elucidating the relationship between CKD and salt intake. In the UK Biobank, participants were surveyed about adding salt to their meals, and it was found that adding salt increases the risk of cardiovascular disease, diabetes, premature mortality, and induces CKD[16-19]. However, it has not yet been clear how the salt affects the incidence of CKD in groups with high versus low genetic risk for the CKD. For CKD, which has such diverse and complex genotypes and phenotypes, personalized and customized precision medicine is crucial. This study focuses on precision medicine to examine how various dietary habits influence disease manifestation within different genetic environments. Therefore, this study aimed to develop a CKD PRS model and apply it to the UK Biobank population to identify individuals at high risk of developing CKD[20] (**Figure**

1). Additionally, we conducted a stratified analysis based on genetic risk factors to determine the extent to which adding salt contributes to CKD development in association with CKD genetic risks.

2. Methods

2.1. *Study population*

The UK Biobank is a large, prospective observational cohort study designed to improve the prevention, diagnosis, and treatment of various illnesses and to promote health[21]. Between 2006 and 2010, the study recruited over 500,000 adults aged 40 to 69 years from 22 assessment centers across England, Scotland, and Wales. Participants provided written informed consent, allowing their data and samples to be used for medical research. Initial recruitment included taking baseline measurements such as social demography, lifestyle, health information, and physical assessments through touch-screen questionnaires and direct physical measurements. Further information on UK Biobank is available in a previous study[21]. In this study, we included only 409,384 participants who had their genetic ethnicity identified as 'Caucasian'. Furthermore, we filtered out participants for whom genetic data was unavailable or did not meet quality control criteria, resulting in 377,186 participants. We further removed 39 participants who had missing salt intake information and 33,550 participants who had prevalent CKD at baseline, leading to a final sample size of 343,597.

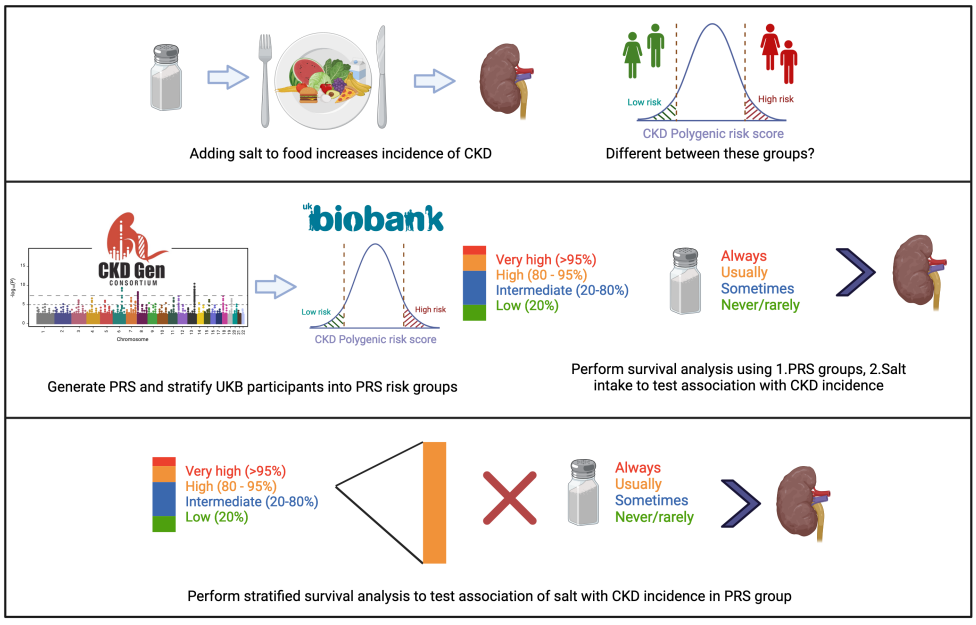


Figure 1. Study overview for assessment of the association between salt intake and CKD incidence across polygenic risk score (PRS) groups.

2.2. *Genotyping and quality control*

The genotyping process utilized by the UK Biobank has been comprehensively described in prior publications[21]. In summary, 487,409 samples were genotyped using the Affymetrix UK BiLEVE

Axiom Array and the Affymetrix UK Biobank Axiom Array (Thermo Fisher Scientific, Waltham, MA). Samples flagged for poor quality by the UK Biobank were excluded. To eliminate related samples, a greedy algorithm was employed to retain the minimal number of samples among second-degree or closer relationships. Given the predominantly European ancestry of the UK Biobank participants, we included only those with 'White British' ancestry. This classification was based on the UK Biobank showcase data field "Genetic ethnic grouping," which identifies participants who self-reported as 'White British' and exhibited very similar genetic ancestry according to principal component (PC) analysis. Samples with discrepancies between reported sex and genetically inferred sex were also excluded. For variant quality control (QC), variants were filtered out if they had an info score of <0.3 or a minor allele frequency (MAF) of <0.01. After applying these QC measures, 377,186 samples and 9,505,768 variants were included in the final dataset used in this study.

2.3. *Polygenic risk score for chronic kidney disease*

The PRS for CKD was generated using Chronic Kidney Disease Genetics (CKDGen) Consortium meta-analysis summary statistics available at https://ckdgen.imbi.uni-freiburg.de/datasets[20]. We generated PRS weights using PRS-CS[22]. PRS-CS leverages summary statistics from GWAS and linkage disequilibrium (LD) information from a reference panel, applying a Bayesian framework to estimate effect sizes. Using the weights generated by PRS-CS, we generated CKD PRS scores for 377,186 samples.

2.4. *Definitions of chronic kidney disease*

In this study, we defined both prevalent and incident CKD. Incident CKD was defined using the tenth Revision (ICD-10) codes and Office of Population Census and Surveys Classification of Interventions and Procedures, version 4 (OPCS-4) codes from primary care data, hospital inpatient data, and death register records[19]. Detailed information for CKD assessment is provided in the Supplementary Table 1. Prevalent CKD was defined as having CKD diagnosis based on the above criteria before first visit to the UK Biobank assessment center. The participants were also considered prevalent CKD if their estimated glomerular function (eGFR) was lower than 60 ml/min per 1.73m^2 or having albuminuria over 30 mg/gCr at baseline measurement. The eGFR was measured using the CKD-EPI Creatinine-Cystatin Equation 2021 with serum creatinine and cystatin C [23]. The follow-up time for the incident CKD cases was calculated from the baseline (first visit to the assessment center) to whichever date came first: diagnosis of CKD, death, date from when a person was lost to follow-up, or May 2017, last date when the “Date lost to follow-up” in UK Biobank was updated. All the events after May 2017 were ignored.

2.5. *Exposure assessment*

The frequency of adding salt phenotype in the UK Biobank was assessed through self-reported dietary questionnaires completed by participants at baseline. Specifically, participants were asked "Do you add salt to your food? (Do not include salt used in cooking)". They were asked to provide an average considering their intake over the last year. The response options included “never/rarely”, “sometimes”, “usually”, “always” and “prefer not to answer”. The “prefer not to answer” was coded as -3 and we removed all participants who selected this option.

2.6. *Covariates ascertainment*

We adjusted for a comprehensive range of covariates to ensure the robustness of our models. These covariates included age, sex, Townsend deprivation index, body mass index (BMI), eGFR, smoking status, alcohol consumption, physical activity, diabetes, dyslipidemia, hypertension, coronary heart disease, and ischemic stroke. Age was determined at the time of assessment, and sex was identified based on genetic information provided by the UK Biobank. The Townsend deprivation index and BMI data were sourced directly from the UK Biobank. Smoking and alcohol consumption statuses were reclassified to current smoker or drinker by combining the "Never" and "Previous" categories. Physical activity was quantified based on achieving either 150 minutes or more per week of moderate intensity, 75 minutes or more per week of vigorous activity, or an equivalent combination[19]. Conditions such as diabetes, dyslipidemia, hypertension, coronary heart disease, and ischemic stroke were defined based on self-reports, ICD-10, and OPCS-4 codes from primary care data, hospital inpatient data, and death register records, as detailed in Supplementary Table 1. For continuous variables, we employed mean imputation to address missing values, and a missing indicator was used for categorical covariates[19].

2.7. *Statistical analysis*

The demographic characteristics of cases and non-cases were evaluated for differences using chi-square tests for categorical variables and independent t-tests for continuous variables. All samples were divided into four groups based on the PRS scores: low (0–19th percentile), intermediate (20–79th percentile), high (80–94th percentile), and very high (>95th percentile).
We conducted three analyses:
1. The association of PRS groups with incident CKD.
2. The combined association of frequency of adding salt and PRS groups with incident CKD.
3. The association of frequency of adding salt with incident CKD, stratified by PRS groups.

For each analysis, we used three models to adjust for various confounding factors:
- Model 1: Adjusted for age and sex.
- Model 2: Adjusted for all variables in model 1, plus Townsend deprivation index, BMI, eGFR, smoking, alcohol consumption, and physical activity.
- Model 3: Adjusted for all variables in model 2, plus diabetes, dyslipidemia, hypertension, coronary artery disease, and ischemic stroke. All variables were measured at baseline.

We calculated hazard ratios (HRs), 95% confidence intervals (CIs), and p-values using univariate and multivariate Cox proportional hazards models, with follow-up time as the time scale to estimate the associations between frequency of adding salt, PRS, and CKD risk. Schoenfeld residuals were used to assess the proportional hazards assumption. Sex was found to violate the proportional hazards assumption ($P = 0.003$ in Model 3); therefore, we stratified all models by sex using the `strata` function, which allows the baseline hazard functions to differ across strata (levels of a categorical variable) while keeping the coefficients for other covariates constant across these strata. In Model 3, we also observed that age slightly violated the proportional hazards assumption ($P = 0.045$). Thus, we modeled age using a penalized smoothing spline with degree 2. After stratifying by sex and modeling age, there was no further violation of the proportional hazard assumption, as

confirmed by Schoenfeld residuals (Supplementary Table 2). The significance of trend was calculated using Jonckheere-Terpstra trend test.

3. Results

3.1. *The baseline characteristics of incident chronic kidney disease population*

The study sample included 331,318 non-CKD participants and 12,279 participants with incident CKD. **Table 1** presents the baseline characteristics of the incident CKD group compared to the non-CKD population. Individuals in the incident CKD group were more likely to fall into the higher PRS categories. Participants with incident CKD had higher prevalence rates of adding salt to their food. Among CKD cases, 53.64% reported "never/rarely" adding salt, 28.14% reported "sometimes," 12.79% reported "usually," and 5.44% reported "always." In contrast, non-CKD participants showed a distribution of 56.95%, 27.62%, 11.2%, and 4.22% for the same categories. Sex distribution showed a higher percentage of males in the incident CKD group (52.54%) compared to the non-CKD group (45.48%). The Townsend deprivation index also indicated higher levels of deprivation in the CKD group (mean = -1.36, SD = 3.07) compared to the non-CKD group (mean = -1.62, SD = 2.9). Additionally, higher prevalence rates of hypertension (74.99%), diabetes (12.54%), dyslipidemia (30.33%), and coronary heart disease (16.42%) were observed in the incident CKD group, compared to the non-CKD group with rates of 58.64%, 4.39%, 15.97%, and 6.32% respectively. The mean eGFR was significantly lower in the incident CKD group (86.11, SD = 13.92) compared to the non-CKD group (96.13, SD = 12.47). Additionally, smoking rates were slightly higher in the incident CKD group (10.51%) than in the non-CKD group (9.76%).

Table 1. Baseline Characteristics for incident CKD and Non-CKD population.

	Non-CKD	Incident CKD	P
N	331318	12279	
PRS risk			
Low	66564 (20.09)	2157 (17.57)	< 0.001
Intermediate	198928 (60.04)	7233 (58.91)	0.012
High	49444 (14.92)	2091 (17.03)	< 0.001
Very high	16382 (4.94)	798 (6.5)	< 0.001
Salt intake			
Never/rarely	188689 (56.95)	6586 (53.64)	< 0.001
Sometimes	91515 (27.62)	3455 (28.14)	0.213
Usually	37120 (11.2)	1570 (12.79)	< 0.001
Always	13994 (4.22)	668 (5.44)	< 0.001
Age (SD)	56.54 (7.94)	60.65 (6.81)	< 0.001
BMI (SD)	27.21 (4.61)	28.71 (4.98)	< 0.001
Sex			
Male	150678 (45.48)	6451 (52.54)	< 0.001

Female	180640 (54.52)	5828 (47.46)	< 0.001
Townsend deprivation index (SD)	-1.62 (2.9)	-1.36 (3.07)	< 0.001
Hypertension	194297 (58.64)	9208 (74.99)	< 0.001
eGFR (SD)	96.13 (12.47)	86.11 (13.92)	< 0.001
Smoking	32344 (9.76)	1290 (10.51)	0.007
Diabetes	14561 (4.39)	1540 (12.54)	< 0.001
Dyslipidemia	52926 (15.97)	3724 (30.33)	< 0.001
Coronary heart disease	20932 (6.32)	2016 (16.42)	< 0.001
Alcohol	311022 (93.87)	11135 (90.68)	< 0.001
Physical activity	211697 (63.9)	7097 (57.8)	< 0.001

3.2. *Chronic kidney disease occurred more in the very high-PRS group*

Categorizing PRS into risk groups revealed significant incremental trend of incident CKD across the PRS categories (P = 0.00023, **Figure 2a**). Incident CKD was significantly higher among subjects in the top 5% of PRS compared to those in other PRS groups. The hazard ratio (HR) of incident CKD for the top 5% PRS group was 1.50 (CI = 1.38 – 1.62, p-value < 2e-16) in the univariate analysis, and 1.52 (CI = 1.41 – 1.65, p-value < 2e-16) in the multivariate Cox proportional hazards model adjusting for age and sex (model 1). Even though the HR decreased when additional predictors were included in Models 2 and 3 of the Cox model, the very high PRS group still showed significantly high HR values for incident CKD (**Figure 2b,** Supplementary figure 1-3).

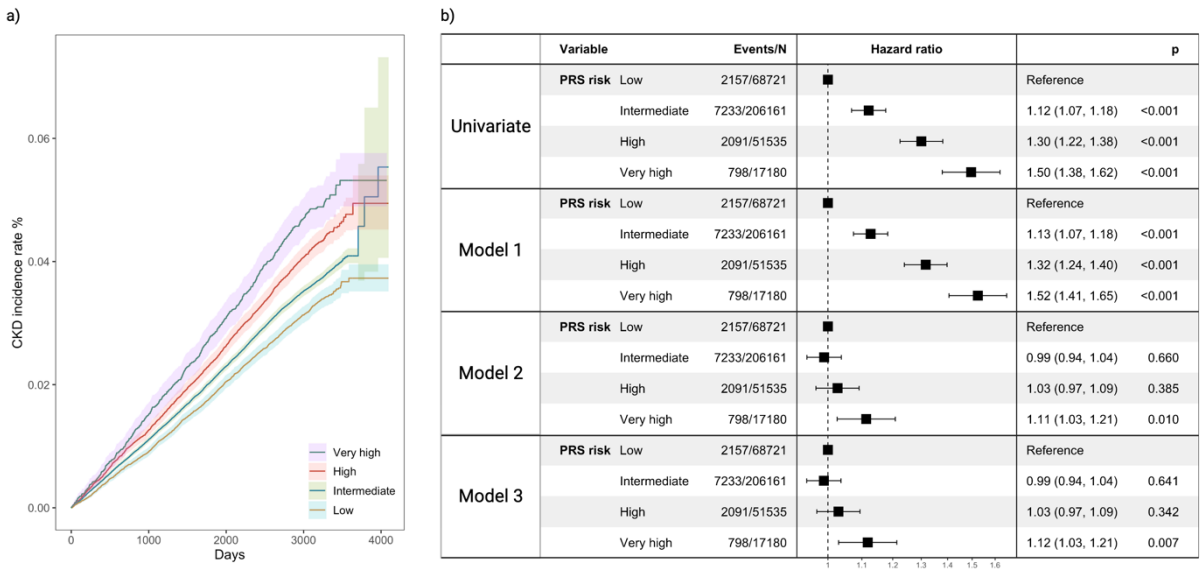


Figure 2. a) Cumulative incidence of CKD stratified by PRS groups b) Hazard ratios for PRS risk groups in different models. Model 1 includes sex and age; Model 2 adds TDI, BMI, eGFR, smoking, alcohol consumption, and physical activity; Model 3 further adds diabetes, dyslipidemia, hypertension, coronary artery disease, and ischemic stroke as covariates.

3.3. *The Frequency of adding salt contributed to incidence of chronic kidney disease*

To determine if frequency of adding salt is associated with CKD incidence, we evaluated all the 4 cox models, indicating a clear positive association between the frequency of adding salt and the incidence of CKD (Figure 3, Supplementary figure 4-6). In the univariate analysis, the HR for CKD incidence increased significantly with higher frequency of adding salt, with the "always" category showing the highest HR of 1.36. In Model 1, which adjusted for age and sex, the HRs remained

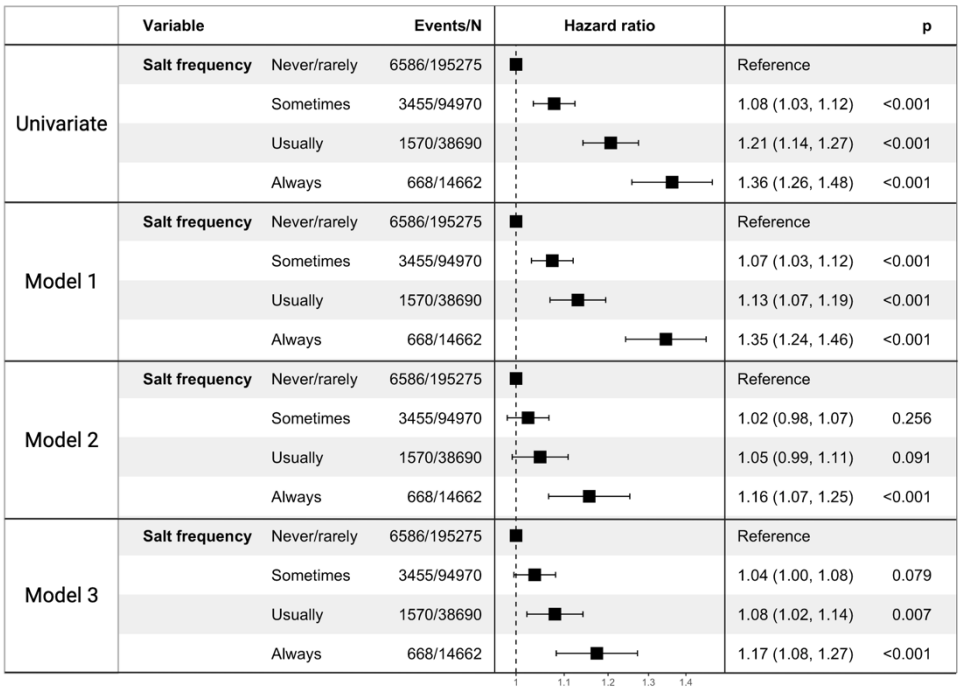


Figure 3. Hazard ratios for salt addition for all models. Model 1 includes sex and age; Model 2 adds TDI, BMI, eGFR, smoking, alcohol consumption, and physical activity; Model 3 further adds diabetes, dyslipidemia, hypertension, coronary artery disease, and ischemic stroke as covariates.

significant for all frequency of adding salt, though reduced, with the "always" category having an HR of 1.35. In Model 2, which included additional variables, the significance for the "sometimes" category diminished (P = 0.256), but the "always" categories remained significantly associated with higher CKD incidence, with HR of 1.16. Finally, in Model 3, which comprehensively adjusted for various predictors, the "sometimes" category was borderline significant (P = 0.079), while the "usually" and "always" categories continued to show significant associations with increased CKD incidence, with the highest HR of 1.17 for the "always" category. These results underscore that more frequency of adding salt is consistently linked to a higher risk of developing CKD (trend P = 0.0028). We also conducted an analysis using PRS as a continuous variable instead of categorical PRS groups as covariates and found similar results (Supplementary figure 7). No significant interaction was observed between PRS score and the frequency of adding salt.

3.4. *The effect of adding salt on incident chronic kidney disease was only significant in those with more than intermediate polygenic risk*

Finally, we stratified individuals by PRS risk groups and evaluated all the Cox models, to check how frequency of adding salt to food is associated with CKD incidence in each of the PRS groups (**Figure 5**, Supplementary figure 8-11). The stratified analysis of frequency of adding salt by PRS

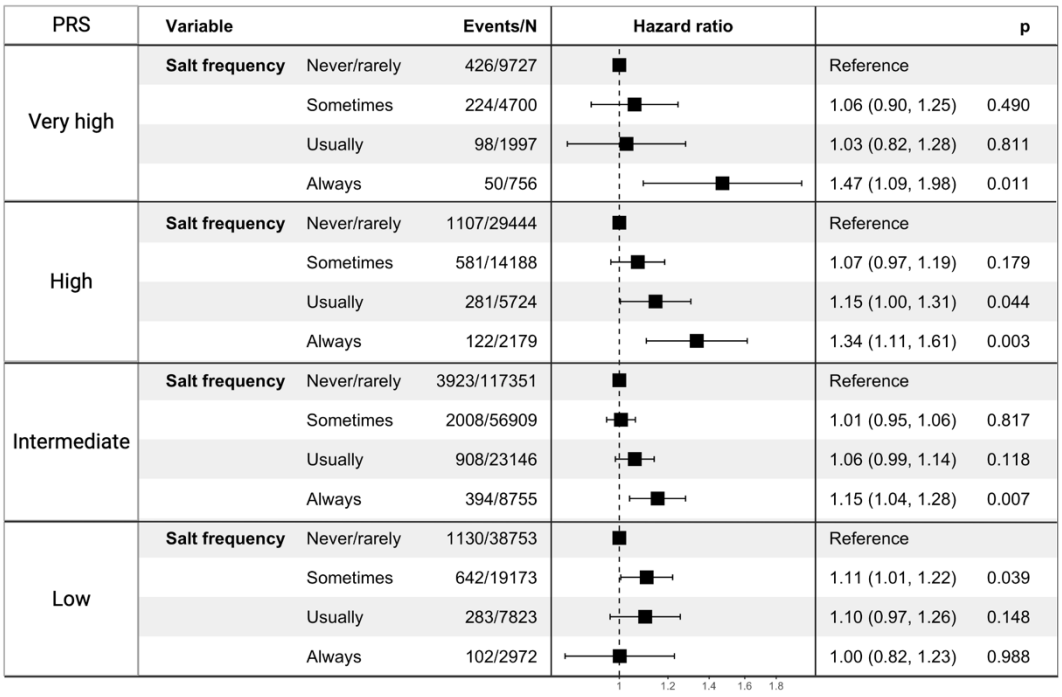


Figure 4. Hazard ratios from model 3 for Salt intake stratified by PRS risk. Model 3 includes age, sex, TDI, BMI, eGFR, smoking, alcohol consumption, physical activity, diabetes, dyslipidemia, hypertension, coronary artery disease, and ischemic stroke

categories revealed significant associations with CKD incidence in several groups. In the very high PRS category, the HR for "always" adding salt to food was significantly elevated (HR = 1.47, CI = 1.09 - 1.98, P = 0.01), indicating increased CKD risk. Similarly, in the high PRS group, both "usually" (HR = 1.15, CI = 1.00 - 1.31, P = 0.04) and "always" (HR = 1.34, CI = 1.11 - 1.61, P = 0.003) adding salt were significantly associated with higher CKD incidence. In the intermediate PRS group, "always" adding salt was also significantly associated with increased CKD risk (HR = 1.15, CI = 1.04 - 1.28, P = 0.007). The low PRS category showed a significant association for "sometimes" adding salt (HR = 1.11, CI = 1.01 - 1.22, P = 0.039), but not for "always" adding salt (HR = 1.00, CI = 0.82 - 1.23, P = 0.988). Overall, the data suggests that "always" frequency of adding salt group is associated with increased CKD risk, especially in individuals with higher genetic predisposition (intermediate, high, and very high PRS categories). There was increasing trend in high PRS (P = 0.049) and intermediate PRS group (P = 0.033), but the trend was not significant in very high PRS group (P = 0.21) and low PRS group (0.12).

4. Discussion

Individuals with a very high CKD PRS showed a significantly higher incidence of CKD, even after adjusting for other contributing factors. The addition of salt to their diet increased the incidence of CKD in proportion to the frequency of salt addition. However, in populations with a low genetic risk for CKD, the effect of adding salt on CKD generation was mitigated. In contrast, in populations with more than intermediate PRS, the incident CKD was exacerbated incrementally by the higher genetic risk.

Genetic studies have estimated the heritability of kidney diseases to between 30-75% through family studies and have identified several critical genetic loci associated with CKD, including *SHROOM3*, *UMOD,* and solute carriers[24-28]. However, most kidney diseases are etiologically complex and heterogenous, making it difficult to identify clear causal pathways and common susceptible genes. The most common causes of CKD are diabetes and hypertension, so CKD often arises as a secondary complication due to these other diseases rather than from primary kidney issues[29]. Both diabetes and hypertension have diverse genetic backgrounds, and the genetic background of kidney damage resulting from these conditions can vary depending on the underlying cause and the stage of the disease. For complex diseases that cannot be explained by candidate genes, PRS aggregates the associations of numerous single nucleotide polymorphisms (SNPs) associated with the disease or trait for a large population. Therefore, a well-validated PRS is a valuable tool for understanding the genetic background of CKD and stratifying risk factors. This study utilized meta-analysis GWAS data generated from SNPs associated with eGFR levels below 60 ml/min per $1.73m^2$ [20]. One of the challenges of optimizing PRS is its application to diverse ancestries with significantly different genetic backgrounds. African ancestry individuals have a higher risk for developing CKD than the other population because they tend to have high-risk alleles in the *APOL1* gene[30]. For this reason, we did not include the small portion of African ancestry in UK Biobank. Finally, when we applied the PRS to the enrolled UK Biobank population in this study, a significant association with actual CKD incidence was found. CKD occurrence proportionally increased with higher PRS, but the predictive power was markedly augmented in the population of top 5% PRS, even after adjusting for many critical CKD risk factors. In line with our study, the extreme tail with the top 1-5% CKD PRS showed about a threefold increase in incident CKD[9].

Dietary salt is known to elevate blood pressure, particularly in individuals with hypertension, those over the age of 55, and those consuming more than 4g of sodium daily[31]. Salt sensitivity refers to the physiological response to blood pressure with sodium intake. In salt sensitive individuals, kidneys retain more sodium by up-regulating the sodium transporters, increasing sympathetic nervous tone, and activating RAS, which leads to higher blood pressure and increased risk of cardiovascular diseases[32]. Elderly individuals, African Americans, and those with CKD are more likely to be salt-sensitive[33]. Our study found that individuals who developed CKD had a significantly higher frequency of adding salt to their food compared to those who did not develop CKD. The "always" frequency of adding salt was significantly associated with the occurrence of CKD. However, this association was observed in the high and intermediate CKD-PRS groups but not in

the low PRS group, where the "always" frequency of adding salt did not show a significant association with CKD incidence. The kidneys play a significant role in blood pressure regulation, and conversely, hypertension can worsen kidney disease. Additionally, both hypertension and CKD share common genetic factors to a considerable extent[34, 35]. Kidney aging also contributes to salt sensitivity by increasing the activation of sodium channels in renal tubules[36]. Therefore, it is hypothesized that populations with genetic variants related to salt sensitivity or renal aging may experience an increased incidence of CKD due to up-regulation of salt sensitivity. Experimental models revealed several candidate genes that increased salt-sensitivity and induce kidney damage, suggesting renal tubular sodium transporters could be involved in the pathogenesis[37, 38]. For those with CKD, reducing salt intake not only helps lower blood pressure but also reduces proteinuria and improves composite kidney outcomes[39, 40]. This study suggests that frequency of adding salt is particularly concerning for individuals with genetic risk factors for CKD, highlighting the clinical importance of reducing salt intake for these individuals. This study has some limitations. The precise amount of salt intake was not available, as we only had information on the frequency of adding salt to meals, which could introduce bias. However, the frequency of adding salt has been shown to be correlated with the 24-hour urinary sodium excretion[17, 41]. n Models 2 and 3, significant hazard ratios were only observed in the 'Very high' PRS-CKD group (Figure 3) and for the 'Usually' or 'Always' frequency of adding salt (Figure 4). This may be due to the inclusion of covariates that are strong predictors of the outcome, which absorb part of the risk previously attributed to PRS or salt frequency alone, reducing the significance of their associations. Additionally, some of these covariates may be masking the true effects of PRS or salt frequency, as they could act as mediators in the causal pathway. In addition, this study did not collect national health insurance data, so the incidence of CKD was only identified through self-report, ICD codes and follow-up eGFR values. This limitation may have led to the underestimation of the actual number of CKD cases. Although we adjusted for socioeconomic status and some lifestyle factors, unmeasured factors may still confound the association. We were also unable to replicate our findings in other datasets due to the lack of comparable definitions of salt intake.

This study developed and validated a PRS for predicting CKD and analyzed how the frequency of adding salt, a crucial trigger, impacts individuals based on their genetic risk factors. While salt restriction has long been considered a vital lifestyle factor in CKD management, this study demonstrated that the influence of frequency of adding salt is more pronounced in individuals with higher genetic risk. Looking ahead, it is anticipated that personalized salt intake recommendations based on genetic risk will become available, allowing for more tailored and effective lifestyle interventions for individuals.

Acknowledgement

This work was partially supported by NIGMS R01 GM138597 and NHLBI R01 HL169458.

Supplementary Material https://biomedinfolab.s3.amazonaws.com/supp/CKD_salt_supp.pdf

References

1. Collaborators GBDCoD. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet.* 2017;390(10100): 1151-1210.

2. Coresh J, Selvin E, Stevens LA, et al. Prevalence of chronic kidney disease in the United States. *JAMA.* 2007;298(17): 2038-2047.

3. Gillion V, Devresse A, Olinger E, et al. Monogenic Kidney Diseases in Kidney Transplantation. *Kidney Int Rep.* 2024;9(3): 549-568.

4. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* 2018;50(9): 1219-1224.

5. Lall K, Magi R, Morris A, Metspalu A, Fischer K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet Med.* 2017;19(3): 322-329.

6. Gladding PA, Legget M, Fatkin D, Larsen P, Doughty R. Polygenic Risk Scores in Coronary Artery Disease and Atrial Fibrillation. *Heart Lung Circ.* 2020;29(4): 634-640.

7. Khera AV, Chaffin M, Wade KH, et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell.* 2019;177(3): 587-596 e589.

8. Sinnott-Armstrong N, Tanigawa Y, Amar D, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet.* 2021;53(2): 185-194.

9. Khan A, Turchin MC, Patki A, et al. Genome-wide polygenic score to predict chronic kidney disease across ancestries. *Nat Med.* 2022;28(7): 1412-1420.

10. Lavie CJ, Ross R, Neeland IJ. Physical activity and fitness vs adiposity and weight loss for the prevention of cardiovascular disease and cancer mortality. *Int J Obes (Lond).* 2022;46(12): 2065-2067.

11. Oh W, Cho M, Jung SW, et al. High physical activity alleviates the adverse effect of higher sedentary time on the incidence of chronic kidney disease. *J Cachexia Sarcopenia Muscle.* 2023;14(1): 622-631.

12. Kalantar-Zadeh K, Fouque D. Nutritional Management of Chronic Kidney Disease. *N Engl J Med.* 2017;377(18): 1765-1776.

13. Yoon CY, Noh J, Lee J, et al. High and low sodium intakes are associated with incident chronic kidney disease in patients with normal renal function and hypertension. *Kidney Int.* 2018;93(4): 921-931.

14. Kieneker LM, Bakker SJ, de Boer RA, Navis GJ, Gansevoort RT, Joosten MM. Low potassium excretion but not high sodium excretion is associated with increased risk of developing chronic kidney disease. *Kidney Int.* 2016;90(4): 888-896.

15. Minegishi S, Luft FC, Titze J, Kitada K. Sodium Handling and Interaction in Numerous Organs. *Am J Hypertens.* 2020;33(8): 687-694.

16. Ma H, Wang X, Li X, Heianza Y, Qi L. Adding Salt to Foods and Risk of Cardiovascular Disease. *J Am Coll Cardiol.* 2022;80(23): 2157-2167.

17. Ma H, Xue Q, Wang X, et al. Adding salt to foods and hazard of premature mortality. *Eur Heart J.* 2022;43(30): 2878-2888.

18. Wang X, Ma H, Kou M, et al. Dietary Sodium Intake and Risk of Incident Type 2 Diabetes. *Mayo Clin Proc.* 2023.

19. Tang R, Kou M, Wang X, et al. Self-Reported Frequency of Adding Salt to Food and Risk of Incident Chronic Kidney Disease. *JAMA Netw Open.* 2023;6(12): e2349930.

20. Wuttke M, Li Y, Li M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet.* 2019;51(6): 957-972.

21. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726): 203-209.

22. Clark K, Leung YY, Lee WP, Voight B, Wang LS. Polygenic Risk Scores in Alzheimer's Disease Genetics: Methodology, Applications, Inclusion, and Diversity. *J Alzheimers Dis.* 2022;89(1): 1-12.

23. Inker LA, Eneanya ND, Coresh J, et al. New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race. *N Engl J Med.* 2021;385(19): 1737-1749.

24. Satko SG, Freedman BI. The familial clustering of renal disease and related phenotypes. *Med Clin North Am.* 2005;89(3): 447-456.

25. O'Seaghdha CM, Fox CS. Genome-wide association studies of chronic kidney disease: what have we learned? *Nat Rev Nephrol.* 2011;8(2): 89-99.

26. Wu HH, Kuo CF, Li IJ, et al. Family Aggregation and Heritability of ESRD in Taiwan: A Population-Based Study. *Am J Kidney Dis.* 2017;70(5): 619-626.

27. Akrawi DS, PirouziFard M, Fjellstedt E, Sundquist J, Sundquist K, Zoller B. Heritability of End-Stage Renal Disease: A Swedish Adoption Study. *Nephron.* 2018;138(2): 157-165.

28. Canadas-Garre M, Anderson K, Cappa R, et al. Genetic Susceptibility to Chronic Kidney Disease - Some More Pieces for the Heritability Puzzle. *Front Genet.* 2019;10: 453.

29. Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. *Kidney Int Suppl (2011).* 2022;12(1): 7-11.

30. Parsa A, Kao WH, Xie D, et al. APOL1 risk variants, race, and progression of chronic kidney disease. *N Engl J Med.* 2013;369(23): 2183-2196.

31. Mente A, O'Donnell MJ, Rangarajan S, et al. Association of urinary sodium and potassium excretion with blood pressure. *N Engl J Med.* 2014;371(7): 601-611.

32. Ellison DH, Welling P. Insights into Salt Handling and Blood Pressure. *N Engl J Med.* 2021;385(21): 1981-1993.

33. Balafa O, Kalaitzidis RG. Salt sensitivity and hypertension. *J Hum Hypertens.* 2021;35(3): 184-192.

34. Corredor Z, Filho M, Rodriguez-Ribera L, et al. Genetic Variants Associated with Chronic Kidney Disease in a Spanish Population. *Sci Rep.* 2020;10(1): 144.

35. Zhang C, Fang X, Zhang H, et al. Genetic susceptibility of hypertension-induced kidney disease. *Physiol Rep.* 2021;9(1): e14688.

36. Frame AA, Wainford RD. Mechanisms of altered renal sodium handling in age-related hypertension. *Am J Physiol Renal Physiol.* 2018;315(1): F1-F6.

37. Trudu M, Janas S, Lanzani C, et al. Common noncoding UMOD gene variants induce salt-sensitive hypertension and kidney damage by increasing uromodulin expression. *Nat Med.* 2013;19(12): 1655-1660.

38. Bovee DM, Cuevas CA, Zietse R, Danser AHJ, Mirabito Colafella KM, Hoorn EJ. Salt-sensitive hypertension in chronic kidney disease: distal tubular mechanisms. *Am J Physiol Renal Physiol.* 2020;319(5): F729-F745.

39. Shi H, Su X, Li C, Guo W, Wang L. Effect of a low-salt diet on chronic kidney disease outcomes: a systematic review and meta-analysis. *BMJ Open.* 2022;12(1): e050843.
40. McMahon EJ, Bauer JD, Hawley CM, et al. A randomized trial of dietary sodium restriction in CKD. *J Am Soc Nephrol.* 2013;24(12): 2096-2103.
41. Quader ZS, Zhao L, Harnack LJ, et al. Self-Reported Measures of Discretionary Salt Use Accurately Estimated Sodium Intake Overall but not in Certain Subgroups of US Adults from 3 Geographic Regions in the Salt Sources Study. *J Nutr.* 2019;149(9): 1623-1632.

Enhancing Privacy-Preserving Cancer Classification with Convolutional Neural Networks

Aurora A. F. Colombo†, Luca Colombo, Alessandro Falcetta, and Manuel Roveri

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

†*E-mail: auroraanna.colombo@mail.polimi.it*

Precision medicine significantly enhances patients prognosis, offering personalized treatments. Particularly for metastatic cancer, incorporating primary tumor location into the diagnostic process greatly improves survival rates. However, traditional methods rely on human expertise, requiring substantial time and financial resources. To address this challenge, Machine Learning (ML) and Deep Learning (DL) have proven particularly effective. Yet, their application to medical data, especially genomic data, must consider and encompass privacy due to the highly sensitive nature of data. In this paper, we propose OGHE, a convolutional neural network-based approach for privacy-preserving cancer classification designed to exploit spatial patterns in genomic data, while maintaining confidentiality by means of Homomorphic Encryption (HE). This encryption scheme allows the processing directly on encrypted data, guaranteeing its confidentiality during the entire computation. The design of OGHE is specific for privacy-preserving applications, taking into account HE limitations from the outset, and introducing an efficient packing mechanism to minimize the computational overhead introduced by HE. Additionally, OGHE relies on a novel feature selection method, VarScout, designed to extract the most significant features through clustering and occurrence analysis, while preserving inherent spatial patterns. Coupled with VarScout, OGHE has been compared with existing privacy-preserving solutions for encrypted cancer classification on the iDash 2020 dataset, demonstrating their effectiveness in providing accurate privacy-preserving cancer classification, and reducing latency thanks to our packing mechanism. The code is released to the scientific community.

Keywords: Computational genomics; Deep Learning; Homomorphic encryption; Privacy.

1. Introduction

Precision medicine is fundamentally changing the landscape of cancer treatment by tailoring medical care to individual genetic profiles, enhancing the efficacy of therapies.[1] This personalized approach not only targets treatments more effectively but also significantly improves patient outcomes and survival rates.[2] Nowadays, however, precision medicine mainly relies on human-performed processes, which require high expertise, lots of time, and finances.[3] From this perspective, the advancement of Machine Learning (ML) and Deep Learning (DL) techniques offers researchers the potential to improve cancer classification accuracy, particularly in identifying primary tumor sites from patients' genomic data,[4] which can lead to more precise and effective treatment strategies.[5] Medical clinics and hospitals often lack expertise in ML and DL and may struggle to afford the necessary computing infrastructure. To address this issue, third-party *as-a-service* solutions have emerged as a promising alternative.[6] However, exposing medical and personal data to third-party providers raises significant privacy concerns, especially when dealing with sensitive genomic information.[7] This vulnerability is a major obstacle to the widespread adoption of ML and DL-*as-a-service* (DLaaS) in healthcare.

In recent years, the application of Homomorphic Encryption (HE) within the DLaaS framework has gained considerable momentum in addressing privacy concerns. HE is an encryption method that encrypts data using a public key, making it unreadable to unauthorized entities. Only the holder of the corresponding private key can decrypt and access the original information. A key advantage of HE is its ability to perform computations on encrypted data without requiring decryption.[8] This enables the encrypted processing of patient genomic data by third-party ML and DL algorithms while maintaining data confidentiality, as the raw genomic data remains encrypted and inaccessible during analysis.[9]

With this method, healthcare institutions encrypt genomic data before transmitting it to a third-party ML and DL service, ensuring that the service provider remains unaware of the underlying data during processing. The service provider receives the keys needed to perform computations on encrypted data, and returns the encrypted results to the client for decryption. This privacy-preserving computation *as-a-service* not only addresses the shortage of ML and DL expertise while reducing costs, but also offers scalability and flexibility to meet the growing computational needs of medical research, data analysis, and clinical decision-making. In recent years, numerous privacy-preserving solutions have been developed for various healthcare applications, leveraging HE to protect sensitive data during analysis. For instance, studies have demonstrated the use of HE in securely processing medical images,[10] and conducting genome-wide association studies.[11] These advancements highlight the potential of HE to maintain data confidentiality while enabling valuable insights in the healthcare domain.

Nonetheless, the task of cancer classification on encrypted genomic data is quite new. Existing solutions have explored ML techniques such as Logistic Regressions (LR)[12] and Shallow Neural Networks (SNN).[13] Interestingly, despite their effectiveness in genomics,[14] Convolutional Neural Networks (CNNs) have received little attention due to their developmental complexity in the HE framework. Indeed, HE poses considerable limitations on the type and number of operations that can be performed on encrypted data. Since only addition and multiplication are supported by HE, several layers and activation functions commonly used in DL

models cannot be directly computed on encrypted inputs. Additionally, HE constraints the number of consecutive encrypted multiplications, thereby limiting the depth of DL models.[15]

In this perspective, our work introduces *Oncological Genomic analysis over HE and CNN* (OGHE), a CNN-based approach for cancer classification designed to operate on encrypted genomic data. Featuring parallel convolutional layers, OGHE separately analyzes Single Nucleotide Variants (SNVs) and Copy Number Variations (CNVs) to enhance accuracy and effectiveness. Additionally, OGHE employs a novel feature selection method, *Variant Scout* (VarScout), to extract the most significant features while preserving the inherent spatial patterns in genomic data. This approach effectively complements the characteristics of OGHE convolutional layers, while maintaining compatibility with HE limitations.

Overall, this work introduces the following innovations: (1) OGHE, a privacy-preserving CNN that incorporates parallel one-dimensional (1D) convolutional layers to independently capture SNVs and CNVs spatial patterns, as they provide distinct and uncorrelated information; (2) a novel feature selection technique, VarScout, which uses clustering and mutation frequency to identify key SNVs and CNVs, thereby reducing computational complexity; and (3) a novel packing mechanism to efficiently encrypt data, weights, and biases into ciphertexts, resulting in high computational performance and reduced latency. The efficacy and efficiency of OGHE and VarScout have been evaluated on the iDASH2020 competition dataset.[16] Compared to State-of-The-Art (SoTA) privacy-preserving cancer classification solutions, our approach achieves an accuracy improvement of 0.8% while reducing the inference time per sample to less than 30 seconds. The code has been made available to the scientific community.[a]

The paper is organized as follows. Sec. 2 presents the related literature. The background is given in Sec. 3. OGHE and VarScout are described in Sec. 4, whereas the experimental results are presented in Sec. 5. Conclusions are finally drawn in Sec. 6.

2. Related Works

In this section, we review the literature on cancer classification task on genomic data. We first discuss solutions for processing plain data and then those for encrypted data.

Over the past few years, both supervised and unsupervised learning techniques have been extensively explored for cancer classification based on genomic data.[17,18] However, the preference has leaned towards supervised classifiers as they result more reliable, interpretable, and precise. CNNs have largely conquered the genomic scenario thanks to their ability to extract spatial patterns.[14] For example, AlShibli et al.[19] proposed ResCNN6, a 6-layers Residual-CNN, to perform CNV-based cancer classification over six tumor types.[20] The architecture encompassed four 2D convolutional layers coupled with MaxPooling to extract relevant features, while two fully connected layers are exploited for classification. ResCNN6 presented shortcut connections to ensure the lowest training error possible by avoiding one or more convolutional layers. On the same task, Chen et al.[21] explored a simpler CNN architecture composed of two 1D convolutional layers. Each convolution is followed by MaxPooling and batch normalization, while a fully connected layer ended the processing pipeline. Despite their effectiveness, how-

[a]Code is available at `https://github.com/AI-Tech-Research-Lab/OGHE.git`.

ever, the reported solutions are not feasible for privacy-preserving computation based on HE due to the inability to compute several layers and activation functions on encrypted inputs. HE imposes stringent limitations, permitting only linear functions and operations. Additionally, the depth of these solutions would surpass the number of consecutive multiplications allowed by HE constraints, potentially leading to data corruption and unreliability.[22]

Due to the aforementioned HE limitations, ML solutions are still preferred over DL ones in privacy-preserving computation. In 2020, iDASH[16] competition challenged its competitors with the development of a cloud-based solution for privacy-preserving classification of eleven cancer locations exploiting genetic mutations and HE. Among the presented solutions, Sarkar et al.[12] developed a logistic regression approach, incorporating a feature engineering strategy to encode somatic mutations based on biological intuition and statistical tests. They advanced a technique to reduce the feature space from over $50,000$ features to $43,000$, implementing a HE-based model through an optimized matrix multiplication algorithm. Differently, Mağara et al.[23] investigated two ML algorithms, i.e., Support Vector Machine (SVM) and XGboost. Given that XGBoost internally utilizes comparisons not supported by the HE scheme, an efficient encoding method for encrypted comparison operations was devised for inference. Moreover, Hong et al.[13] proposed a Shallow Neural Network (SNN) consisting of one hidden layer with 64 nodes and a linear activation function. In the preprocessing of the input genomic data, the feature selection step incorporated both clustering and data filtering methods. Lastly, in 2024, Song et al.[24] introduced ReActHE, a family of CNNs characterized by a novel type of activation layer, i.e., the *Residue activation layer*, and a scaled power activation function. In particular, by selecting the $1,000$ most significant features by means of a L1 normalized logistic regression, they outperformed alternative privacy-preserving ML solutions, achieving low approximation errors in the cancer classification task.

Differently from the existing literature, our solution proposes two key aspects which are fundamental for privacy-preserving cancer classification. First, VarScout selects the most representative features in an effective way, helping in reducing input and model dimensions. Second, OGHE exploits spatial information from genomic data by employing only HE-compliant operations to allow encrypted computation.

3. Background

This section will present the basics needed to understand OGHE and VarScout implementation. Sec. 3.1 will provide a brief overview of the HE scheme employed, while Sec. 3.2 will present the characteristics of the genetic mutations analyzed.

3.1. *Homomorphic Encryption*

HE is a family of encryption schemes that enables a set of operations to be performed directly on encrypted data.[8] Mathematically, two functions $E(k_p, \cdot), D(k_s, \cdot)$ are said to be homomorphic with respect to a set of functions $\mathcal{F}$ if, for any $f \in \mathcal{F}$, a function g can be found that:

$$f(m) = D(k_s, g(E(k_p, m))) \tag{1}$$

for any set of input m.[25] In particular, $E(k_p, \cdot)$ and $D(k_s, \cdot)$ represent the encryption and

decryption functions, respectively, whereas k_p denotes the public key and k_s the secret key. The ability of HE to provide encrypted operations relies on the maintenance of the datum algebraic structure during the processing pipeline.[26] In this way, the result obtained from ciphertext computation is guaranteed to match the one from the same operation in plaintext. In this study, we adopted the Cheon–Kim–Kim–Song (CKKS) scheme,[27] which is based on the Ring Learning With Errors (RLWE) problem,[28] a computational problem commonly used in quantum-resistant cryptography.[29] The CKKS scheme supports encrypted additions and multiplications between real values. More in detail, it belongs to the family of leveled HE schemes, i.e., schemes that allow only a finite number of consecutive encrypted operations to be performed before the information is lost. This limit is called scheme level, denoted by l, and it is due to noise injection performed by the scheme itself in order to guarantee the probabilistic encryption properties.[30] In CKKS scheme, the algebraic structure of plaintexts and ciphertexts is defined through a set of encryption parameters $\Theta = \{N, q, \Delta\}$, where N is the polynomial modulus, q is the list of $l + 2$ coefficient modulus, and Δ is the scaling factor. More in detail, plaintexts are in the polynomial ring $\mathcal{R} = \mathbb{Z}[X]/(X^N + 1)$, while ciphertexts are in the polynomial ring $\mathcal{R}_q = \mathbb{Z}_{q_0}[X]/(X^N + 1)$.

When dealing with the CKKS scheme, two key factors must be considered. The former deals with the choice of the encryption parameters Θ, defining the security level, which in this work is set to 128 bit, the polynomial order, and the encoding precision. They represent a trade-off between the scheme level l and the overhead added with respect to plain computation. The latter is called *batching* technique: it enables parallel processing through *Single Instruction, Multiple Data (SIMD)* operations.[27] By using *batching*, a single ciphertext can store up to $N/2$ values, reducing the computational overhead both in terms of time and memory requirements.

The adopted CKKS scheme supports two main operations. Let $\underline{a} = [\underline{a}_0, \underline{a}_1, \ldots, \underline{a}_n]$ and $\underline{b} = [\underline{b}_0, \underline{b}_1, \ldots, \underline{b}_n]$ be two encrypted CKKS vectors. Then, the encrypted element-wise addition can be defined as follows:

$$\underline{A} + \underline{B} = [\underline{a}_0 + \underline{b}_0, \underline{a}_1 + \underline{b}_1, \ldots, \underline{a}_n + \underline{b}_n]. \tag{2}$$

Conversely, the encrypted element-wise product is defined as:

$$\underline{A} * \underline{B} = [\underline{a}_0 * \underline{b}_0, \underline{a}_1 * \underline{b}_1, \ldots, \underline{a}_n * \underline{b}_n]. \tag{3}$$

Additionally, matrices can be represented in ciphertext as their flattened forms. Aggregate operations that perform homomorphic sums across specific dimensions of encrypted data can be defined. Let $\underline{C}$ be an encrypted and flattened matrix with dimensions $M \times N$. The sum over columns is then expressed as:

$$\underline{S} = \left[\sum_{j=1}^{N} \underline{C}_{i,j} \right]_{i=1}^{M} \tag{4}$$

where $\underline{S}$ is an encrypted matrix containing the row sums of $\underline{C}$, repeated to match the dimensions of the input matrix. Similarly, this operation can be applied to the other dimension.

3.2. *Single Nucleotide Variants and Copy Number Variations*

Information from SNVs and CNVs is vital in the diagnostic process of metastatic cancer, as it helps in identifying the origin of the primary tumor mass. Being common for a certain population, SNVs, which involve the alteration of a single nucleotide in DNA strands, serve as biomarkers for specific diseases. When occurring in protein-coding regions, SNVs can lead to *missense variations*, i.e., the substitution of an amino acid altering protein structure and function, and *nonsense mutations*, i.e., the premature truncation of the protein-coding process.[31] SNVs are categorized as [LOW, MODERATE, MODIFIER, HIGH] based on their impact on disease onset, as determined by the Variant Effect Predictor software.

Conversely, CNVs are structural variations involving the rearrangement of more than 50 base pairs in the genome. Entire genes can be altered in the number of copies, compromising normal gene expression levels and affecting critical cellular processes like cell cycle regulation, apoptosis, and cell signaling. Thus, CNVs are strongly associated with genetic disorders and complex diseases such as cancer.[32] In this work, CNVs are represented by five mutation levels, i.e., $\{0.0, \pm 1.0, \pm 2.0\}$, where the absolute value indicates the number of strands involved, while the sign denotes either a positive duplication or a negative deletion. This information is directly inferred from the Copy Number Segmentations generated by the ASCAT software.

4. Proposed Solution

This section details the proposed solution, composed of VarScout and OGHE, designed to address the considered primary tumor location problem, formalized as follows. Let X_{CNV} and X_{SNV} be two vectors of size L_{CNV} and L_{SNV}, respectively. We define the primary tumor location as $\hat{y} = \arg\max_i \ y_i$, where

$$y = \varphi(X_{CNV}, X_{SNV}) \in \mathbb{R}^C \tag{5}$$

is the output vector, C is the number of classes, and $\varphi(\cdot)$ is the model. In the rest of the section, we will consider the encrypted version of this problem. In particular, Sec. 4.1 introduces VarScout, our proposed feature selection method designed to reduce the CNV and SNV feature space dimension, while Sec. 4.2 details OGHE, the model architecture specifically designed to provide encrypted primary tumor classification on encrypted SNV and CNV inputs.

4.1. *VarScout method*

VarScout has been designed to reduce CNV and SNV feature space dimension while keeping the highest data representation, which is crucial to design OGHE being HE compliant. Inspired by Hong et al.,[13] VarScout aims at enhancing OGHE accuracy by prioritizing the most impactful mutations while maintaining typical spatial patterns.

After organizing CNV and SNV mutations in chromosomal order, agglomerative cluster analysis is employed for CNV filtering to eliminate redundant information. Specifically, our method comprises three main steps: (1) similarity computation between adjacent genes (g_i, g_{i+1}) through the Hamming distance d, i.e., $d(g_i, g_{i+1})$, for each gene g_i in the original dataset; (2) formation of clusters O_l such that $O_l = \{g_i \,|\, min(d(g_i, g_{i+1}))\}$, i.e., genes character-

ized by the least distance are chosen to form a cluster; and (3) selection of the first in order gene g_i to represent the cluster l. These steps are repeated until l reaches L_{CNV}.

Conversely, SNVs are numerically encoded within the range $\{0.0, 0.20, 0.50, 0.90, 1.0\}$, as proposed by Hong et al.,[13] to denote the impact of the mutation on the disease insurgence. In particular, 0.0 represents the absence of genetic alteration, whereas 1.0 denotes the highest level of influence. To reduce SNV feature space, our feature selection method is based on mutation occurrences across different cancer types. Frequencies are calculated based on the impact of genetic alterations, defined as $F_j = \sum_i f_{ij}$, for $i = [1, ..., |Z_c|]$. More specifically, F_j denotes the weighted frequency of occurrence of a gene j within a sub-population $Z_c = \{x \,|\, y = c\} \, \forall c = [1, \ldots, C]$ characterized by a specific cancer class y, and f_{ij} represents the impact of the j-th gene for each individual sample i. The process ranks mutations by sequentially adding the most recurrent gene for each tumor type to the feature space until the desired feature dimension L_{SNV} is reached.

4.2. *OGHE Architecture*

To exploit spatial patterns in genomic data while addressing HE constraints, we propose OGHE. OGHE takes as input VarScout-selected CNV and SNV features encoded as two separate ciphertexts, namely $\underline{\widetilde{X}_{CNV}}$ and $\underline{\widetilde{X}_{SNV}}$, and outputs an encrypted vector $\underline{\widetilde{Y}_{pred}}$ of length C. Once decrypted, the output $Y_{pred} = D(k_s, \underline{\widetilde{Y}_{pred}})$, where $D(k_s, \cdot)$ is the decryption function described in Sec. 3, reveals the predicted cancer class, identified by the index of the highest value in Y_{pred}. OGHE architecture is designed to work within HE constraints while maintaining high accuracy and computational performance. The training of OGHE was performed on plain data, although it is specifically designed for encrypted inference.

As shown in Fig. 1, OGHE is a shallow CNN composed of two parallel 1D convolutional layers and a fully connected layer. Parallel convolutions are chosen to separate CNV and SNV information, ensuring independent processing of uncorrelated data until the fully connected layer. In its training configuration, a square activation function was chosen as commonly used in the privacy-preserving DL[15] framework, and a Spatial-Dropout layer is incorporated to mitigate the risk of overfitting. The feature maps resulting from the convolutional layers, i.e., $Y_{conv_{h,CNV}}$ and $Y_{conv_{h,SNV}}$, for each kernel $h = [1, \ldots, H]$, are then concatenated and passed to the flatten function. A fully connected layer ends the processing pipeline to provide the output vector Y_{pred}, where the index of the highest value indicates the predicted primary tumor mass location. This strategic design ensures compatibility with the limitations posed by HE while exploiting the available genetic information for accurate cancer classification.

Conversely, OGHE encrypted processing is designed to provide optimal computational performance by efficiently managing the ciphertext space through a well-defined packing mechanism. By strategically organizing data within ciphertexts, our approach enables efficient encrypted computations, and significantly enhances the performance of the network. OGHE includes *Encrypted Convolutional Blocks* (Sec. 4.2.1) and an *Encrypted Fully Connected Block* (Sec. 4.2.2). To streamline the notation we will consider the case where both the input data and OGHE model are encrypted. However, the following formulations can easily be extended to the scenario where the model is kept unencrypted by the service provider.

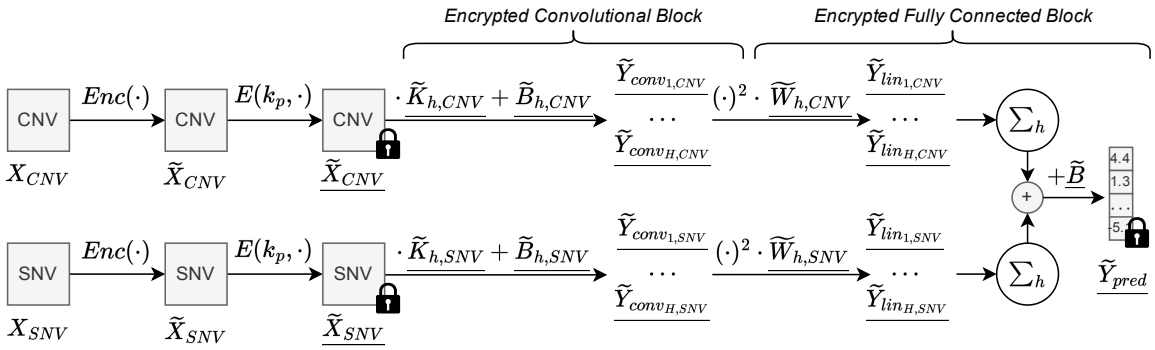


Fig. 1. OGHE encrypted pipeline. Each sample, composed of X_{CNV} and X_{SNV}, is encoded and encrypted into $\underline{\widetilde{X}_{SNV}}$ and $\underline{\widetilde{X}_{CNV}}$, respectively, before being processed by the *Encrypted Convolutional Block* and the *Encrypted Fully Connected Block*.

4.2.1. *Encrypted Convolutional Block*

This block proposes a 1D re-elaboration of the *im2col* method[33] to facilitate the computation of CNV and SNV convolutional layers. In our approach, data, weights, and feature maps are efficiently packed to be encrypted into single ciphertexts, to maximize computational efficiency.

Let $K = [k_1, \ldots, k_D]$ denote a 1D convolutional kernel of dimension D and stride S, and let $X = [x_1, \ldots, x_{L_x}]$ represent the 1D input vector. Our method encodes the input X into a matrix $\widetilde{X}$ of size $L_y \times V$, built as follows:

$$\widetilde{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{D-1} & x_D & 0 & \cdots & 0 \\ x_{(S+1)} & x_{(S+1)+1} & \cdots & x_{(S+1)+D-1} & x_{(S+1)+D} & 0 & \cdots & 0 \\ x_{(2S+1)} & x_{(2S+1)+1} & \cdots & x_{(2S+1)+D-1} & x_{(2S+1)+D} & 0 & \cdots & 0 \\ \vdots & & & & & & & \\ x_{(iS+1)} & x_{(iS+1)+1} & \cdots & x_{(iS+1)+D-1} & x_{(iS+1)+D} & 0 & \cdots & 0 \end{bmatrix}. \tag{6}$$

Similarly, the convolutional kernel K is encoded into $\widetilde{K}$, a $L_y \times V$ matrix where each row contains a copy of K:

$$\widetilde{K} = \begin{bmatrix} k_1\, k_2 \cdots k_{D-1}\, k_D\, 0 \cdots 0 \\ k_1\, k_2 \cdots k_{D-1}\, k_D\, 0 \cdots 0 \\ k_1\, k_2 \cdots k_{D-1}\, k_D\, 0 \cdots 0 \\ \vdots \\ k_1\, k_2 \cdots k_{D-1}\, k_D\, 0 \cdots 0 \end{bmatrix}. \tag{7}$$

For computational reasons, both $\widetilde{X}$ and $\widetilde{K}$ are padded with zeros at the end of each row to maintain a power of 2 number of columns V, which will be set to $V = 2^{\lceil log_2(max(D,C)) \rceil}$. Thus, $\widetilde{X}$ and $\widetilde{K}$ share the same shape $L_y \times V$, where

$$L_y = \left\lfloor \frac{L_x - D}{S} \right\rfloor + 1, \tag{8}$$

being L_x the input length, D the kernel dimension, S the stride, and C the number of classes.

The matrices $\widetilde{X}$, $\widetilde{K}$ are then flattened and encrypted into the ciphertexts $\underline{\widetilde{X}}$ and $\underline{\widetilde{K}}$, respectively. This encoding ensures that both the input vector X and the convolutional kernel K are appropriately formatted for efficient encrypted computation.

Computing the h-th convolution $\underline{\widetilde{Y}_{conv_h}}$ is reduced to a single Hadamard multiplication between ciphertexts, followed by a sum over the columns, as described in Eq. (4):

$$\underline{\widetilde{Y}_{conv_h}} = \left[\sum_{j=1}^{V} (\underline{\widetilde{X}_h} \cdot \underline{\widetilde{K}_h})_{i,j} \right]_{i=1}^{L_y} \tag{9}$$

where i denotes the i-th row, j the j-th column, and L_y and V the dimensions of the output $\underline{\widetilde{Y}_{conv_h}}$. This operation is repeated for each of the $h = [1, \ldots, H]$ kernels of the convolutional block. The resulting ciphertext $\underline{\widetilde{Y}_{conv_h}}$ will be encoded as:

$$\underline{\widetilde{Y}_{conv_h}} = \begin{bmatrix} \underline{y_1} & \underline{y_1} & \cdots & \underline{y_1} & \underline{y_1} & \underline{y_1} & \cdots & \underline{y_1} \\ \underline{y_2} & \underline{y_2} & \cdots & \underline{y_2} & \underline{y_2} & \underline{y_2} & \cdots & \underline{y_2} \\ \underline{y_3} & \underline{y_3} & \cdots & \underline{y_3} & \underline{y_3} & \underline{y_3} & \cdots & \underline{y_3} \\ \vdots & & & & & & & \\ \underline{y_{L_y}} & \underline{y_{L_y}} & \cdots & \underline{y_{L_y}} & \underline{y_{L_y}} & \underline{y_{L_y}} & \cdots & \underline{y_{L_y}} \end{bmatrix} \tag{10}$$

in its flattened form, where $\underline{y_i}$ is the encrypted result of a single window convolution. Lastly, the bias is encoded to match $\underline{\widetilde{Y}_{conv_h}}$ packing, replicated $L_y * V$ times, and added to it.

It is worth noting that in OGHE, CNV and SNV inputs are processed in separate, parallel 1D convolutional layers. Eq. (9) is effectively used to compute *Encrypted Convolutions* for each kernel and parallel branch, after which the square activation is applied.

4.2.2. *Encrypted Fully Connected Block*

The output of each parallel convolutional branch is subsequently forwarded through the final fully connected layer. However, since CKKS ciphertexts cannot be concatenated, the operation has to be decomposed. The weight matrix W associated to the layer is split into $2H$ submatrices W_h, which are then flattened. Specifically, each W_h represents the portion of weights W that has to be multiplied by the h-th output channel per each parallel branch. This way, $2H$ reduced fully connected layers can be performed to compute the output:

$$\underline{\widetilde{Y}_{lin_{h,CNV}}} = \left[\sum_{i=1}^{L_y} (\underline{\widetilde{Y}_{conv_{h,CNV}}} \cdot \underline{\widetilde{W}_{h,CNV}})_{i,j} \right]_{j=1}^{V}, \underline{\widetilde{Y}_{lin_{h,SNV}}} = \left[\sum_{i=1}^{L_y} (\underline{\widetilde{Y}_{conv_{h,SNV}}} \cdot \underline{\widetilde{W}_{h,SNV}})_{i,j} \right]_{j=1}^{V}. \tag{11}$$

A Hadamard multiplication followed by a summation over the rows, described in Eq. (4), effectively emulates a vector-matrix multiplication. All the $2H$ values are then summed together along with the bias vector to provide the final prediction $\underline{\widetilde{Y}_{pred}}$, which will be encoded having the resulting vector repeated along the ciphertext.

Lastly, by multiplying the output $\underline{\widetilde{Y}_{pred}}$ by a binary mask, the output vector $\underline{Y_{pred}}$ will result in a single prediction vector of size C, where each element corresponds to a class. After decrypting with the secret key, i.e., $D(k_s, \cdot)$, the index of the element with the highest value will correspond to OGHE prediction.

5. Experimental Results

The experimental campaign is organized into two parts. First, Sec. 5.2 compares OGHE to SoTA solutions in terms of accuracy, micro Area Under the Curve (mAUC), and computational performance for privacy-preserving cancer classification tasks. Then, Sec. 5.3 shows the effectiveness of OGHE and VarScout when compared to Hong et al.[13] SNN and a baseline model, i.e., a single fully connected layer network, referred to as FCM. Our solution, implemented using OpenFHE-python[34] library, has been tested on a workstation equipped with 2 Intel Xeon Gold 5318 S CPUs and 384GBs of RAM.

5.1. *Procedure*

To evaluate stability and consistency, OGHE and VarScout were tested alongside the literature on the iDASH2020 dataset,[16] sourced from The Cancer Genome Atlas (TCGA). This dataset includes $3,622$ samples with CNV and SNV information for $25,128$ genes, and eleven cancer classes representing the primary tumor mass location.

Moreover, for the in-depth comparison of Sec. 5.3, we employed a 5-fold cross-validation technique to evaluate all the considered models. For each fold, we allocated data in a 7:1:2 ratio for training, validation, and testing, respectively. We also employed a hyperparameter selection based on the validation loss for all the considered models. In particular, we optimized the hyperparameters for OGHE considering the following ranges: kernel sizes of $\{16, 32, 64\}$, strides of $\{4, 8, 16\}$, and number of kernels $\{4, 8, 16\}$, along with activation functions either linear or square. The spatial dropout rate has been fixed to 0.5. We fixed an Adam optimizer with a weight decay of 0.0001, learning rate of 0.001 and cosine annealing, and batch size of 16. Instead, for the SNN[13] and FCM, the learning rate was evaluated in $\{0.001, 0.0001\}$, the batch size in $\{4, 8, 16, 32\}$, and the weight decay of the Adam optimizer in $\{0, 0.0001, 0.0005\}$. All the solutions were trained for 200 epochs using a weighted cross-entropy loss function, whose weights are inversely proportional to class frequencies.

For the encrypted computations, we employed the following CKKS encryption parameters: $\Theta = \{N = 32,768, q = [60, 50, 50, 50, 50, 50, 60], \Delta = 2^{50}\}$, yielding results that are consistent with those obtained from processing in plaintext, and ensuring a 128-bits security level.[35]

5.2. *Comparison with SoTA Solutions*

As a first analysis, OGHE accuracy was compared to privacy-preserving cancer classification SoTA solutions. Note that we compared OGHE only to models specifically designed for HE applications, as they share the same characteristics and limitations.

As demonstrated in Table 1, our solution outscores all other models in the literature in terms of accuracy, highlighting the exceptional capabilities of OGHE and VarScout. This accuracy improvement is attributed to the simultaneous learning from multiple sources, namely CNV and SNV, which enhances the model's robustness to variations and noise, thereby increasing its reliability. Nonetheless, OGHE shows a slight decrease in mAUC, which can be attributed to a more distributed error among the classes.

Furthermore, the computational performance of OGHE has been assessed in comparison to DL models in literature, as shown in Table 2. Since ReActHE[24] was evaluated in its orig-

Table 1. Accuracy and mAUC of our proposed solution compared to the existing literature.

	Model name	Model class	Accuracy	mAUC
Mağara et al.[23]	XGBoost	XGBoost	–	93.80%
Sarkar et al.[12]	LR	LR	83.61%	98.00%
Song et al.[24]	ReActHE	CNN	83.82%	–
Hong et al.[13]	SNN	NN	85.15%	98.82%
Ours	OGHE	CNN	85.94%	98.44%

Table 2. Comparison of computational performance with respect to FCM and SNN[13] in terms of encryption, computation, and decryption time, and in terms of latency per sample (L_1), and in the encrypted inference of 100 samples (L_{100}). All values are in seconds.

	Model name	Enc[s]	Comp[s]	Dec[s]	L_1[s]	L_{100}[s]
Hong et al.[13]	SNN	13.50	227.20	0.10	240.80	–
Song et al.[24]	ReActHE	–	–	–	–	685.35
Ours	OGHE	2.97	23.17	0.013	27.59	190.02

inal work by encrypting the model weights, the weights and biases of OGHE have also been encrypted to ensure a fair comparison. Additionally, single-sample inference utilized only 4 threads, whereas for the inference of 100 samples we limited our machine to use 40 threads, to align with the ReActHE[24] experimental setting. Table 2 proves the efficiency of our method both in single-sample and high-throughput inference. Notably, the computational times are the same for all the feature sizes up to [1024, 2286], given that the inputs, weights, and feature maps fit into a single ciphertext. For larger models, two ciphertexts must be used, leading to a slight increase in latency. However, the performance remains highly competitive, outperforming current state-of-the-art solutions. Additionally, the potential for further optimization through parallelization ensures scalability and efficiency in future implementations. Moreover, OGHE encryption time encompasses both model and data encryption. However, model encryption takes 2.88 seconds, making it the most time-consuming aspect of the encryption process. This evaluation considers the worst-case scenario where both the model and data are encrypted. If the model were in plaintext, a significant amount of time (around 15%) would be saved, highlighting the efficiency potential in less stringent encryption scenarios.

5.3. *VarScout and OGHE evaluation*

The aim of this part is to rigorously evaluate the effectiveness of VarScout and OGHE. To do so, we used the preprocessing method of Hong et al.[13] as feature extractor for OGHE, FCM, and the SoTA SNN model.[13] Subsequently, we applied our VarScout preprocessing method to our solution to determine if VarScout effectively enhances OGHE's performance. Note that Song et al.[24] is not included in this comparison, as they did not release the implementation.

Initially, we applied Hong et al.[13] feature selection method to all the models under consideration. The models were then tested with different input sizes, reflecting the number of CNV and SNV features after preprocessing. Along with the size [709, 1198], identified by Hong et al.[13]

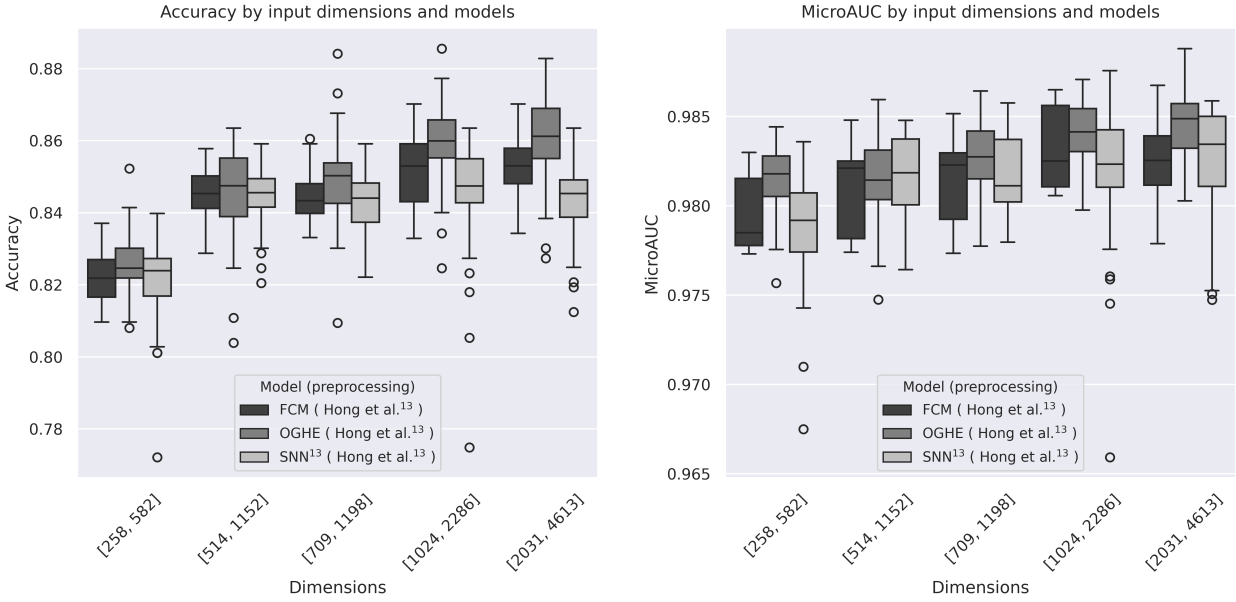


Fig. 2. Accuracy and MicroAUC boxplots of the considered models, for different input dimensions, i.e., the number of CNV and SNV features, respectively. The model names are followed by the preprocessing feature selection procedure in parenthesis. They show the metrics over 10 runs of the 5-fold cross validation.

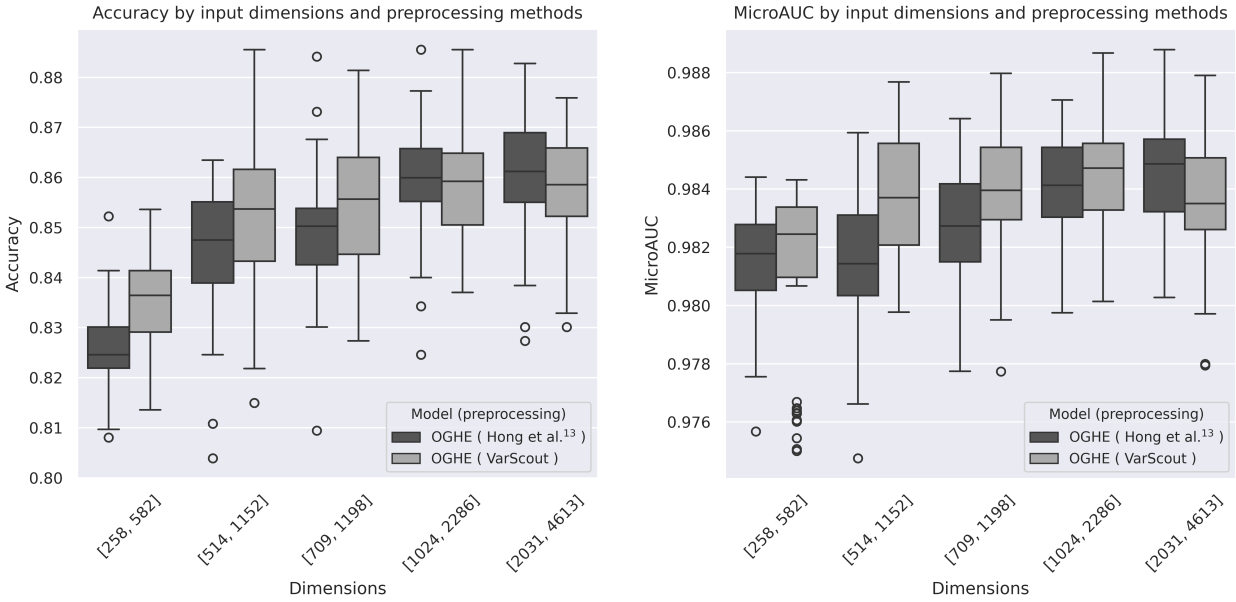


Fig. 3. Accuracy and MicroAUC boxplots of OGHE with different feature selection methods, for different input dimensions, i.e., the number of CNV and SNV features, respectively. The model names are followed by the preprocessing feature selection procedure in parenthesis. They show the metrics over 10 runs of the 5-fold cross validation.

as optimal, we also evaluated the input sizes $[258, 582]$, $[514, 1152]$, $[1024, 2286]$, and $[2031, 4613]$.

Fig. 2 demonstrates the effectiveness of OGHE, showing that it outperforms both FCM and SNN.[13] OGHE shows greater improvement over the other models as input feature size increases. Achieving a higher median accuracy with a narrower interquartile range, OGHE confirms that genomic data contains useful spatial and hierarchical information, effectively captured by the convolutional layers. Moreover, OGHE shows reduced variance and a fewer outliers across all input sizes when compared to the SNN,[13] indicating its robustness to variations in the input data and parameter initialization when dealing with complex tasks.

Furthermore, the statistical difference between SNN[13] and OGHE was evaluated using the McNemar-Bowker[36] test when both models were provided with inputs of size $[1024, 2286]$ as it ensures optimal performance for both models. The comparison was based on the run providing the highest test accuracy for each fold of the cross validation. The test indicated a statistically significant difference at a 5% confidence level between OGHE and the SNN[13] in four out of five folds, confirming the improvement our approach provides over existing literature.

Further improvements rise from the introduction of VarScout as feature selection method. To demonstrate its effectiveness, OGHE integrated with VarScout was compared to OGHE using the feature selection method proposed by Hong et al.[13] Fig. 3 shows that, for smaller input sizes, the model trained on VarScout-extracted features outperforms the one trained with Hong et al.[13] method, demonstrating that our feature selection method is superior in capturing spatial patterns and extracting the most important features first. This characteristic helps in maintaining a smaller network without sacrificing performance. Specifically, reducing the feature space to $[514, 1152]$, which is half the size of the configuration providing the best performance, results in only a 0.7% loss in accuracy. This is a key aspect when dealing with HE computations as it allows the use of ciphertexts characterized by smaller polynomial rings, resulting in a significant reduction in memory footprint and computation time.

6. Conclusion

This work proposes OGHE, a HE-friendly CNN for privacy-preserving cancer classification, and VarScout, a preprocessing method designed to maximize OGHE performance. OGHE architecture exploits spatial correlations in genomic data, separately processing the most relevant SNVs and CNVs extracted by VarScout, while preserving their spatial patterns. Together, these techniques achieve SoTA performance in encrypted cancer classification.

Despite advancements in privacy-preserving computing, HE introduces significant limitations in Artificial Intelligence applications, particularly regarding reduced computational efficiency. Future research will focus on minimizing the computational overhead and developing encrypted training, enabling researchers to analyze genomic data securely while preserving privacy, unlocking new possibilities for medical research and discovery.

Additionally, leveraging Neural Architecture Search (NAS) to optimize OGHE's architecture under HE constraints could further enhance its performance by automating the search for optimal architectures. Lastly, the release of new datasets will enable further validation and refinement of OGHE, expanding its potential applications.

References

1. M. R. Kosorok and E. B. Laber, Precision medicine, *Annual review of statistics and its application* **6**, 263 (2019).
2. M. Schwaederle, M. Zhao, J. J. Lee, A. M. Eggermont, R. L. Schilsky, J. Mendelsohn, V. Lazar and R. Kurzrock, Impact of precision medicine in diverse cancers: a meta-analysis of phase ii clinical trials, *Journal of clinical oncology* **33**, p. 3817 (2015).
3. Y.-F. Sun, X.-R. Yang, J. Zhou, S.-J. Qiu, J. Fan and Y. Xu, Circulating tumor cells: advances in detection methods, biological issues, and clinical relevance, *Journal of cancer research and clinical oncology* **137**, 1151 (2011).
4. A. Bhola and A. K. Tiwari, Machine learning based approaches for cancer classification using gene expression data, *Machine Learning and Applications: An International Journal* **2**, 01 (2015).
5. D. van Uden, M. van Maaren, L. Strobbe, P. Bult, M. Stam, J. van der Hoeven, S. Siesling, J. de Wilt and C. Blanken-Peeters, Better survival after surgery of the primary tumor in stage iv inflammatory breast cancer, *Surgical Oncology* **33**, 43 (2020).
6. S. Ghosh and R. Dasgupta, Cloud computing infrastructure in healthcare industry, in *Machine Learning in Biological Sciences: Updates and Future Prospects*, (Springer, 2022) pp. 169–176.
7. J. Santaló and M. Berdasco, Ethical implications of epigenetics in the era of personalized medicine, *Clinical epigenetics* **14**, p. 44 (2022).
8. A. Acar, H. Aksu, A. S. Uluagac and M. Conti, A survey on homomorphic encryption schemes: Theory and implementation, *ACM Computing Surveys (Csur)* **51**, 1 (2018).
9. A. Wood, K. Najarian and D. Kahrobaei, Homomorphic encryption for machine learning in medicine and bioinformatics, *ACM Computing Surveys (CSUR)* **53**, 1 (2020).
10. A. Vizitiu, C. I. Niţă, A. Puiu, C. Suciu and L. M. Itu, Towards privacy-preserving deep learning based medical imaging applications, in *2019 IEEE international symposium on medical measurements and applications (MeMeA)*, (IEEE, 2019).
11. M. Blatt, A. Gusev, Y. Polyakov and S. Goldwasser, Secure large-scale genome-wide association studies using homomorphic encryption, *Proceedings of the National Academy of Sciences* **117**, 11608 (2020).
12. E. Sarkar, E. Chielle, G. Gursoy, L. Chen, M. Gerstein and M. Maniatakos, Privacy-preserving cancer type prediction with homomorphic encryption, *Scientific reports* **13**, p. 1661 (2023).
13. S. Hong, J. H. Park, W. Cho, H. Choe and J. H. Cheon, Secure tumor classification by shallow neural network using homomorphic encryption, *BMC genomics* **23**, 1 (2022).
14. C. Gunavathi, K. Sivasubramanian, P. Keerthika and C. Paramasivam, A review on convolutional neural network based deep learning methods in gene expression data for disease diagnosis, *Materials Today: Proceedings* **45**, 2282 (2021).
15. A. Falcetta and M. Roveri, Privacy-preserving deep learning with homomorphic encryption: An introduction, *IEEE Computational Intelligence Magazine* **17**, 14 (2022).
16. X. Jiang, A. O. Harmanci, M. Kim, H. Tang, X. Wang, T.-T. Kuo and L. Ohno-Machado, Idash privacy & security workshop 2020 - secure genome analysis competition (2020).
17. D. Wu, D. Wang, M. Q. Zhang and J. Gu, Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification, *BMC genomics* **16**, 1 (2015).
18. Y. Chen, J. Sun, L.-C. Huang, H. Xu, Z. Zhao *et al.*, Classification of cancer primary sites using machine learning and somatic mutations, *BioMed research international* **2015** (2015).
19. A. AlShibli and H. Mathkour, A shallow convolutional learning network for classification of cancers based on copy number variations, *Sensors* **19**, p. 4207 (2019).
20. J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson *et al.*, Integrative analysis of complex cancer genomics and clinical profiles

using the cbioportal, *Science signaling* **6**, pl1 (2013).
21. H. Attique, S. Shah, S. Jabeen, F. G. Khan, A. Khan and M. ELAffendi, Multiclass cancer prediction based on copy number variation using deep learning, *Computational Intelligence and Neuroscience* **2022** (2022).
22. T. Lepoint and M. Naehrig, A comparison of the homomorphic encryption schemes fv and yashe, in *International Conference on Cryptology in Africa*, (Springer, 2014).
23. Ş. S. Mağara, C. Yıldırım, F. Yaman, B. Dilekoğlu, F. R. Tutaş, E. Öztürk, K. Kaya, Ö. Taştan and E. Savaş, Ml with he: Privacy preserving machine learning inferences for genome studies, *arXiv preprint arXiv:2110.11446* (2021).
24. C. Song and X. Shi, Reacthe: A homomorphic encryption friendly deep neural network for privacy-preserving biomedical prediction, *Smart Health* **32**, p. 100469 (2024).
25. F. Boemer, A. Costache, R. Cammarota and C. Wierzynski, ngraph-he2: A high-throughput framework for neural network inference on encrypted data, in *Proceedings of the 7th ACM workshop on encrypted computing & applied homomorphic cryptography*, (ACM, 2019).
26. M. Ogburn, C. Turner and P. Dahal, Homomorphic encryption, *Procedia Computer Science* **20**, 502 (2013).
27. J. H. Cheon, A. Kim, M. Kim and Y. Song, Homomorphic encryption for arithmetic of approximate numbers, in *Advances in Cryptology–ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part I 23*, (Springer, 2017).
28. V. Lyubashevsky, C. Peikert and O. Regev, On ideal lattices and learning with errors over rings, in *Advances in Cryptology–EUROCRYPT 2010: 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, French Riviera, May 30–June 3, 2010. Proceedings 29*, (Springer, 2010).
29. J.-P. Bossuat, R. Cammarota, J. H. Cheon, I. Chillotti, B. R. Curtis, W. Dai, H. Gong, E. Hales, D. Kim, B. Kumara *et al.*, Security guidelines for implementing homomorphic encryption, *Cryptology ePrint Archive* (2024).
30. G. J. Fuchsbauer, An introduction to probabilistic encryption, *Osječki matematički list* **6**, 37 (2006).
31. P. Katsonis, A. Koire, S. J. Wilson, T.-K. Hsu, R. C. Lua, A. D. Wilkins and O. Lichtarge, Single nucleotide variations: biological impact and theoretical interpretation, *Protein Science* **23**, 1650 (2014).
32. C. N. Henrichsen, E. Chaignat and A. Reymond, Copy number variants, diseases and gene expression, *Human molecular genetics* **18**, R1 (2009).
33. K. Chellapilla, S. Puri and P. Simard, High performance convolutional neural networks for document processing, in *Tenth international workshop on frontiers in handwriting recognition*, (Suvisoft, 2006).
34. A. Al Badawi, J. Bates, F. Bergamaschi, D. B. Cousins, S. Erabelli, N. Genise, S. Halevi, H. Hunt, A. Kim, Y. Lee *et al.*, Openfhe: Open-source fully homomorphic encryption library, in *Proceedings of the 10th workshop on encrypted computing & applied homomorphic cryptography*, (ACM, 2022).
35. D. Joseph, R. Misoczki, M. Manzano, J. Tricot, F. D. Pinuaga, O. Lacombe, S. Leichenauer, J. Hidary, P. Venables and R. Hansen, Transitioning organizations to post-quantum cryptography, *Nature* **605**, 237 (2022).
36. A. Krampe and S. Kuhnt, Bowker's test for symmetry and modifications within the algebraic framework, *Computational statistics & data analysis* **51**, 4124 (2007).

One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data

Michal Golovanevsky[§], Eva Schiller[§], Akira Nair[§], Eric Han[§], Ritambhara Singh[§‡ †*]
[§]*Department of Computer Science, Brown University,*
[‡] *and Center for Computational Molecular Biology, Brown University*
Providence, RI 02912, USA
[†] *E-mail: ritambhara@brown.edu*

Carsten Eickhoff*
School of Medicine,
Institute for Bioinformatics and Medical Informatics,
University of Tübingen,
Tübingen, 72074, Germany
E-mail: c.eickhoff@acm.org
**Co-corresponding authors*

Multimodal models have become increasingly important as they surpass single-modality approaches on diverse tasks ranging from question-answering to disease diagnosis. Despite the importance of multimodal learning, existing efforts focus on vision-language applications, where the number of modalities rarely exceeds four (images, text, audio, video). However, data in healthcare domain, may include many more modalities like X-rays, PET scans, MRIs, genetic screening, genomic data, and clinical notes, creating a need for both efficient and accurate data integration. Many state-of-the-art multimodal models rely on cross-attention or self-attention for effective data integration, which do not scale well for applications with more than two modalities. The complexity per layer of computing attention in either paradigm is, at best, quadratic with respect to the number of modalities, posing a computational bottleneck that impedes broad adoption. To address this, we propose a new attention mechanism, One-Versus-Others (OvO) attention, that scales *linearly* with the number of modalities, thus offering a significant reduction in computational complexity compared to existing multimodal attention methods. Using three clinical datasets with multiple diverse modalities, we show that our method decreases computation costs while maintaining or increasing performance compared to popular integration techniques. Across all clinical datasets, OvO reduced the number of required floating point operations (FLOPs) by at least 91.98%, demonstrating its significant impact on efficiency and enabling multi-modal predictions in healthcare.*

Keywords: Multimodal learning; deep learning; attention mechanism; clinical decision support.

*Code and Appendix are available at `https://github.com/rsinghlab/OvO`

1. Introduction

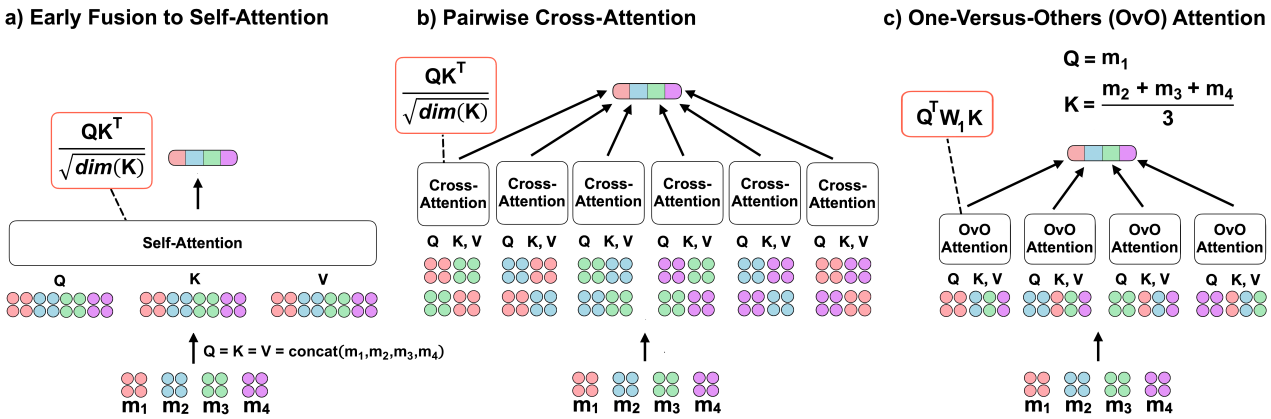


Fig. 1. **Integration scheme comparison**. (a) Early fusion to self-attention with scaled dot product attention,[1] and (b) Pairwise cross-attention integration with scaled dot product attention.[1] (c) Our proposed method, One-Versus-Others (OvO), does not rely on pairwise interactions or long concatenated sequences but rather captures all modalities in a single attention score. A modality embedding is represented by m_i and W is a learnable parameter (see Section 3.1).

Multimodal learning has emerged as a promising approach, which enables joint learning from multiple modalities of data (e.g., text and images). Combining different modalities allows for a more comprehensive and accurate understanding of tasks such as clinical decision support,[2–4] image and video captioning,[5,6] audio-visual speech recognition,[7] and sentiment analysis.[8] Multimodal learning has been explored through various methods in machine learning and deep learning. While feature-level integration was mostly used in more traditional machine learning algorithms, Neural Networks have allowed for the intermediate fusion of modalities at any layer and late fusion at the decision-making stage. However, these fusion paradigms lack a key component - capturing explicit interaction between modalities. For example, in detecting Alzheimer's Disease, genetic features help reinforce and ground the clinical information and thus lead to more robust decision-making.[3] Such relevant interactions can be captured through the attention mechanism. Popular multimodal models, such as LXMERT[9] and BLIP,[10] use a fusion method that captures interactions between modalities using *cross-attention*. On the other hand, models such as VisualBERT[11] and LLaVA[12] use early fusion, where vision and language inputs are concatenated early to learn multimodal through *self-attention*. The clinical domain embraced these approaches, with multimodal models like MedFuseNet[13] and ARMOUR[14] using cross-attention for medical vision question answering and mortality prediction. In parallel, models such as BioViL-T[15] and MMBERT[16] employ early fusion through self-attention for disease prediction and report generation.

However, both self-attention and cross-attention grow quadratically in computational burden with the number of modalities, posing a scalability challenge. While in popular vision-language integration tasks, the number of modalities rarely exceeds four (images, text, audio, video), a significant bottleneck can arise in other domains. The healthcare domain exempli-

fies this issue, as a single task may involve integrating data from complex and rich sources spanning multiple modalities from radiology, pathology, genomics, genetics, and clinical data. Therefore, with the influx of many modalities, the use of cross-attention or self-attention will remain limited in the clinical domain as their computational demands escalate even further. To address this gap, we propose a new attention mechanism, One-Versus-Others (OvO) attention. OvO attention is calculated by comparing *one* modality against a combined representation of all *other* modalities (hence the name, One-Versus-Others). Our approach significantly reduces computational complexity as it grows linearly with the number of modalities (see Section 3.3). Figure 1 sketches a four-modality example to demonstrate the difference between our approach (scales linearly) and self-attention/cross-attention (scales quadratically). OvO is a general attention scheme that can be integrated into existing clinical multimodal architectures instead of cross-attention or self-attention.

We first, present a complexity analysis and validate it through a simulated dataset. Our simulation results show scalability gains in an extreme multimodal setting (20 modalities). Next, we use three diverse clinical datasets that vary in modalities, encoder types (pre-trained and not), number of samples, and tasks to show our model's improved scalability in different clinical multimodal settings. Our results demonstrate that our method dramatically decreases computation costs (offering at minimum a 91.98% reduction in computations), compared to self-attention and cross-attention while maintaining or even exceeding performance.

Overall, OvO is a novel attention scheme for multimodal integration that scales linearly to the number of modalities, enabling the practical application of deep learning models in healthcare, where computational efficiency and accuracy are vital for deployment.

2. Related work

Multimodal attention-based models are increasingly pivotal in clinical decision support systems, paralleling their widespread use in vision-language applications. In the medical domain, these models have shown remarkable utility in diverse scenarios, such as cancer classification,[17] biomarker discovery,[18,19] prognosis prediction,[20,21] and more. These applications highlight the versatility and potential of multimodal learning in handling complex and rich medical data. The attention mechanism serves as a core component in these models. Attention measures the similarity among individual representations, like word embedding vectors or, in the multimodal scenario, modality-specific embeddings. Each input embedding can assume one of three roles: (1) Query (Q), representing the current focus of attention when compared against other input embeddings; (2) Key (K), signifying an input embedding being compared to the Query; and (3) Value (V), which contributes to computing the output for the Query.

Commonly, the representations from each modality in the multimodal models are passed through one of two paradigms - early fusion followed by self-attention or fusion through cross-attention. The early fusion group (e.g., Transformer-based models like UNITER,[22] VisualBERT,[11] LLaVA,[12] BioViL-T,[15] MedViLL,[23] etc.) concatenates the visual embeddings and the textual embeddings as a single input, before passing the inputs through attention (see Figure 1 (a)). Given modalities m_1 and m_2, queries (Q), keys (K), and values (V) are computed from their concatenated sequence (e.g., $Q_{1,2} = concat(m_1, m_2)$). The final output from a

standard Transformer block is denoted by Z, Equation 1 shows the early fusion paradigm.

$$\begin{cases} \mathrm{Z}_{1,2} = \textit{Multiheaded Attention}\,(\mathrm{Q}_{1,2}, \mathrm{K}_{1,2}, \mathrm{V}_{1,2}) \\ \mathrm{Z} = \textit{Transformer}(\mathrm{Z}_{1,2}) \end{cases} \tag{1}$$

The cross-attention scheme (used in Transformer-based models like ViLBERT,[24] LXMERT,[9] MedFuseNet,[13] MADDi,[3] etc.) inputs each modality into its own Transformer, the outputs of which are fed to a cross-modal Transformer (see Figure 1 (b)). For such models, the cross-modal interactions are captured in a pairwise manner through cross-attention, where queries (Q), keys (K), and values (V) are computed from the modality inputs (m_1 and m_2), and then the keys and values from each modality are fed to the multi-headed attention block of the other modality. The output, Z, is shown in Equation 2.

$$\begin{cases} \mathrm{Z}_1 = \textit{Multiheaded Attention}\,(\mathrm{Q}_2, \mathrm{K}_1, \mathrm{V}_1) \\ \mathrm{Z}_2 = \textit{Multiheaded Attention}\,(\mathrm{Q}_1, \mathrm{K}_2, \mathrm{V}_2) \\ \mathrm{Z} = \textit{Transformer}\,(\textit{concat}\,(\mathrm{Z}_1, \mathrm{Z}_2)) \end{cases} \tag{2}$$

While the early fusion and cross-attention paradigms could be extended to three modalities, seen in TriBERT[25] and VATT,[26] these models face scalability challenges for more than three modalities. Cross-attention methods can leverage joint representations formed from cross-attention but do not scale well to larger numbers of modalities as they are computed in a pairwise fashion. Thus, if there are k modalities, computing pairwise fusion between each pair will result in $\binom{k}{2}$ matrix computations. Moreover, attention is not a symmetric calculation, which means that most commonly, it is computed bi-directionally (e.g., image to text and text to image), leading to an even greater computational burden. Early fusion involves the concatenation of modalities before the Transformer layer, which similarly does not scale well with the number of modalities. Self-attention is quadratic with respect to sequence length,[1] and since early fusion methods concatenate inputs before attention, the computational complexity will increase quadratically as the number of modalities increases (see Section 3.3). Furthermore, concatenation is not order invariant, making the ordering of modalities an important consideration, potentially requiring similar bi-directional computations as cross-attention. Our integration method, OvO, addresses the limitations mentioned above in a scalable and domain-agnostic manner.

3. Methods

3.1. *One-Versus-Others (OvO) attention*

We propose a new attention mechanism, One-Versus-Others (OvO) Attention, which grows linearly with the number of modalities rather than quadratically, as would be the case for cross-attention or self-attention (see Section 3.3). OvO computes attention between one modality at a time with respect to all other modalities. Given modality m_i, which is an embedding obtained from a dedicated encoder (e.g., CNN, ClinicalBERT, etc.) and $i \in 1, 2, \ldots, k$ where k is the number of modalities, OvO takes in one modality and computes the dot product against all the other modalities with a weight matrix W_i. W_i is a learnable parameter that can help scale the importance of each attention calculation (see Figure 1 (c)) and can learn

interactions throughout training. The modality embeddings m_i in OvO attention function like queries, while the weighted sum of the other modalities behaves like keys and values, akin to the dot-product attention mechanisms, with the naming choice of m_i adapted for multimodal applications. This dot product The similarity score function, representing the degree of alignment between the chosen modality and others, calculated for modality m_i with respect to a set of other modalities $(m_j : j \neq i)$ is shown in Equation 3. This produces a vector of scores which measure the relevance of m_i with respect to the other modalities. The context vector in OvO for modality m_i, which is a combined representation of information from the other modalities, is shown in Equation 4:

$$score\left(m_i, \{m_j : \ j \ \neq \ i\}\right) = m_i^T \ W_i \ \frac{\sum_{j \ \neq \ i}^{k} \ m_j}{k-1} \tag{3}$$

$$OvO\left(m_i, \{m_j : \ j \ \neq \ i\}\right) = softmax(score\left(m_i, \{m_j : \ j \ \neq \ i\}\right)) \cdot m_i \tag{4}$$

In Equation 4, the softmax is applied across the input dimension of the embeddings, normalizing across the attention scores. The result is then multiplied by the original modality embeddings to compute the final output. We chose to sum over the "other" modalities instead of concatenation because: (1) the concatenation vector will continue to increase in length with the number of modalities, which will result in a less scalable framework; (2) concatenation is not invariant to the order of modalities, which could affect the model prediction, whereas a sum provides position invariance.

Unlike cross-attention and self-attention, OvO provides a more interpretable mechanism for analyzing interactions between multiple modalities. In cross-attention, interactions are captured in isolated pairs (e.g., m_1 compared to m_2 or m_1 compared to m_3), limiting the ability to see how one modality integrates information from all others. Similarly, self-attention condenses modality interactions into a single operation, which can obscure explicit cross-modal interactions and make it difficult to disentangle their contributions. OvO, however, creates one attention matrix per modality, with each modality interacting with all others through the learnable weight matrix W_i, which ensures a flexible and adaptive scaling mechanism. In future work, W_i will be studied to better understand the relative importance of each modality's contributions to the final prediction.

3.2. *Multi-headed OvO Attention*

We extend OvO attention to the multi-headed attention framework to directly compare with early fusion through self-attention and pairwise cross-attention. Multi-headed attention allows the model to attend to the input embeddings in different ways simultaneously. This is achieved by splitting the input embeddings into multiple linear projections, each processed independently through a self-attention mechanism. The outputs of each attention head are then combined to obtain the final output of the multi-headed attention layer. Formally, taking the input modality m_i with respect to a set of other modalities $(m_j : j \neq i)$, the multi-headed

attention layer for OvO attention is defined as follows:

$$\begin{cases} MultiheadedOvO(m_i, \{m_j : j \neq i\}) \\ = concat(head_1, \ldots, head_h)W^O \\ head_k = OvO(m_i W_k^{m_i}, \{m_j W_k^{m_j} : j \neq i\}) \end{cases} \tag{5}$$

Here, h is the number of attention heads, W_k is a learnable weight matrix for the k-th attention head, W^O is a learnable weight matrix that projects the concatenated outputs of the attention heads back to the original dimension, and Equation 5 defines OvO Attention.

3.3. *Model Complexity*

This section highlights the complexities of the two commonly used paradigms: early fusion followed by self-attention and pairwise cross-attention, as well as our One-Versus-Others (OvO) attention. Table 1 summarizes the complexity per layer. Let k represent the number of modalities, n be the feature-length of each modality (assuming equal), and d be the representation dimension of the respective weight matrices. As established in,[1] self-attention has complexity of $\mathcal{O}(n^2 \cdot d)$. In the multimodal case, self-attention concatenates modalities before attention, leading to a sequence length of $k \cdot n$, influencing the quadratic term. Thus, the complexity of self-attention is $\mathcal{O}((k \cdot n)^2 \cdot d) = \mathcal{O}(k^2 \cdot n^2 \cdot d)$. Cross-attention computes attention over all pairwise permutations of modalities: ${}_kP_2 = \frac{k!}{(k-2)!} = k(k-1)$. Thus, the number of operations required by cross-attention is $\mathcal{O}(k \cdot (k-1) \cdot n^2 \cdot d) = \mathcal{O}((k^2 - k) \cdot n^2 \cdot d)$. When focusing on the fastest-growing terms in big O notation, the final complexity per layer simplifies to $\mathcal{O}(k^2 \cdot n^2 \cdot d)$. One-Versus-Others (OvO) Attention requires one attention calculation per modality, making it linear with respect to k. Thus, the complexity per layer for OvO is $\mathcal{O}(k \cdot n^2 \cdot d)$. Appendix Section 1 provides step-by-step details for the complexity calculations.

Table 1. **Per-Layer complexities of model paradigm.** Let k be the number of modalities, n the feature-length of a modality, and d the representation dimension.

Model	Complexity Per Layer
Self-Attention	$\mathcal{O}(k^2 \cdot n^2 \cdot d)$
Cross-Attention	$\mathcal{O}(k^2 \cdot n^2 \cdot d)$
One-Versus-Others (OvO) Attention	$\mathcal{O}(\mathbf{k} \cdot n^2 \cdot d)$

3.4. *Illustration through simulation*

To illustrate the linearity of OvO compared to the other integration paradigms, we simulated 20 artificial modalities. We consider two classes: (1) 20 random feature values that sum up to 1.0, and, (2) 20 random feature values that are each less than 0.15. These classes were created such that the correct label can only be inferred after inspecting all features. For example, 0.14

is less than 0.15, but it could also be a value that adds to 1. For more details on the simulation dataset and how the threshold was chosen, see Appendix Section 2.

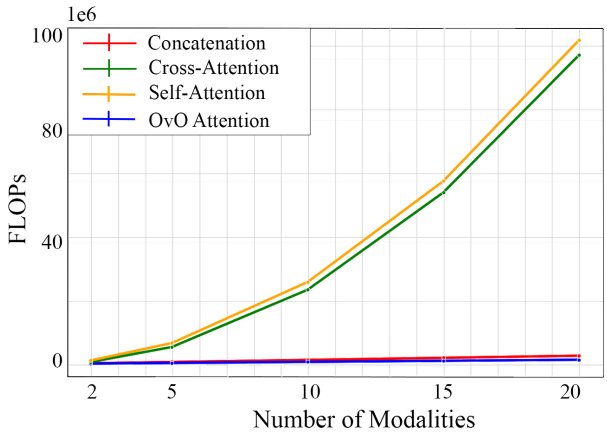


Fig. 2. **The impact of using OvO attention to fuse simulated data.** Using FLOPs as a measure of compute, we demonstrate that OvO grows linearly with respect to the number of modalities, while self and cross-attention grow quadratically.

Each value was then vectorized by sampling randomly around the selected number, such that each modality is a vector of size 20 rather than a single number, leading to a combined total of 400 features. Overall, the dataset contains 2,000 samples (1,000 for each class). Our constructed simulation dataset tests the scaling capabilities of our method to an extent that real-world datasets do not usually reach.

We examine the computation cost across the three integration methods using 2, 5, 10, 15, and 20 simulated modalities. Most notably, while self-attention and cross-attention grow quadratically with respect to the number of modalities, k, ($\mathcal{O}(k^2 \cdot n^2 \cdot d)$), our method scales linearly ($\mathcal{O}(k \cdot n^2 \cdot d)$), as shown in Figure 2.

4. Experiments

We used three diverse clinical datasets to examine our method against three standard integration techniques: concatenation with no attention (baseline), early fusion with self-attention, and pairwise cross-attention. These clinical tasks feature a range of rich modalities that, despite their high integration costs, remain essential to solve.

4.1. *Dataset descriptions*

4.1.1. *MIMIC-IV and CXR data*

MIMIC-IV[27] covers 431K visits for 180K patients admitted to the ICU in the Beth Israel Deaconess Medical Center. MIMIC Chest X-ray (MIMIC-CXR)[28] contains chest radiographs in DICOM format with free-text radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at Beth Israel Medical Center. We follow the pre-processing of MedFuse[29] to extract the clinical time-series data from MIMIC-IV along

with the associated chest X-ray images in MIMIC-CXR. We further expand the number of modalities by adding a demographics table and discharge notes, resulting in four modalities. We also follow MedFuse in the construction of the phenotyping task. The goal of this multi-label classification task is to predict whether a set of 25 chronic, mixed, and acute care conditions are assigned to a patient in a given ICU stay. This is a 25-class multi-label task with four modalities.

4.1.2. *The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) data*

The Alzheimer's Disease Neuroimaging Initiative (ADNI)[30] database provides neuroimaging data, cognitive test scores, biomarker profiles, and genetic information for Alzheimer's disease (AD), mild cognitive impairment (MCI), and normal patients. We use the processed data from the Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge.[31] We focus on a one-time diagnosis prediction task, using the most recent available data point for each patient across all modalities. This ensures that each data sample includes information from the same time point, which aligns with our goal of evaluating cross-modal integration rather than longitudinal progression. We utilize six modalities that have the least missing information per patient: cognitive tests - neuropsychological tests administered by a clinical expert; MRI ROIs (generated from Freesurfer) - measures of brain structural integrity; FDG PET ROI averages - measure cell metabolism, where cells affected by AD show reduced metabolism; AV45 PET ROI averages - measures amyloid-beta load in the brain; demographic information (e.g., age, gender, education); and CSF biomarkers - amyloid and tau levels in the cerebrospinal fluid. The preprocessing provided by TADPOLE turned every modality into a tabular form (including imaging). After removing patients with missing modalities, we had 767 MCI patients, 493 normal patients, and 143 AD patients. Thus, this is a three-class classification task with six modalities.

4.1.3. *eICU data*

The eICU collaborative database includes data from ICUs across the continental United States between 2014 and 2015.[32] It consists of tables linked through a patient unit stay ID. For our mortality prediction task, we focus on six tabular modalities: patient, diagnosis, treatment, medication, lab, and apacheApsVar tables. The apacheApsVar table contains numerical variables used to calculate the Acute Physiology Score (APS), an established method within the Acute Physiology Age Chronic Health Evaluation (APACHE) system for summarizing patient's severity of illness on ICU admission and predicting outcomes. The patient table includes demographic, admission, and discharge details, and is used to determine mortality status. The diagnosis table lists active diagnoses for each patient, the treatment table includes active treatments, and the medication table contains active medication orders. We extract the features from these tables by one-hot encoding the relevant conditions, treatment types, and drug names, respectively. The lab table includes lab results, with features extracted by summing commonly recorded lab types. Our dataset includes 75,845 unique patients with 93,784 ICU stays, 86,012 recorded as alive and 7,772 as dead. This is a 2-class classification task with six modalities.

4.2. *Baselines*

Our multimodal baselines include a conventional concatenation fusion with no attention, early fusion followed by self-attention, and pairwise cross-attention fusion. The architectures of all models are identical except for their integration stage. For example, since modality-specific encoders can produce different dimension sizes, we add a linear layer before integration to create the same input dimensions. Although this step is not strictly necessary for concatenation, we still add the layer there so that no additional factors influence computation costs and performance. While there are many multimodal Transformers available for the vision-language domain, our focus is on examining the underlying fusion mechanism and creating a general integration paradigm for any application, especially ones outside of vision-language. In Appendix Section 5, we touch on the limitations of our experiments and future work that we did not cover.

4.3. *Implementation details*

For the MIMIC dataset, we follow the established train, validation, and test split in Hayat *et al.*[4] Similarly, for the TADPOLE task, we use the provided data splits but add a constraint that repeating patients cannot appear across data splits to avoid information leakage. In the other datasets, for consistency, we randomly sampled 80% of the data for the training set and 10% each for test and validation sets, as there was not an established split. To evaluate our model against other integration techniques, we use the domain-accepted metrics for each task: For MIMIC and eICU, we use area under the receiver operating characteristic (AUROC) and area under Precision-Recall (PR) curve (AUPRC) as established in past works;[4,33] For TADPOLE we use the multi-class area under the receiver operating curve (mAUC) and the overall balanced classification accuracy (BCA), as established by the competition creators.[31] For all datasets, we used the number of floating-point operations (FLOPs) as the measure of runtime complexity. FLOPs were measured per sample and reported as the difference between concatenation, the simplest integration setting, and multimodal attention (ΔFLOPs).

4.4. *Hyperparameter Tuning*

Our hyperparameter tuning scheme was consistent for each dataset and each model. For each experiment, we used the evaluation metrics on the validation set to determine the best hyperparameters. We tuned the learning rate (0.01 - 1 x 10^{-8}, dividing by 10 for each interval), batch size (16, 32, 64, 128), epochs (200 epochs with early stopping if validation performance did not increase for 5 epochs), and number of attention heads for OvO, self-attention, and cross-attention models (1, 2, 4, 8, 16). For the neural network encoders, we tuned the number of linear layers ranging from 1 to 4. Similarly, for the convolutional neural network, we tuned the number of convolution layers ranging from 1 to 4. For compute times and GPU details used for hyperparameter tuning, see Appendix Section 2. Lastly, we randomly picked 10 random seeds for every experiment - once the best hyperparameters were picked, ten models initialized with those seeds and parameters were run. Then, using the trained ten models, we evaluated on the test set and took the average of the 10 runs along with the standard deviation, which is reported in Section 5.

5. Results

Using three real-world clinical datasets, diverse in terms of the number of modalities, feature space, and classification tasks, we demonstrate that our method consistently and drastically reduces computational costs compared to early fusion and pairwise fusion while simultaneously maintaining or enhancing performance. This is demonstrated on one four-modality dataset and two six-modality datasets.

For the four-modality MIMIC task, we used pre-trained ClinicalBERT model for the text modality and fine-tuned it for the unimodal baseline and the multimodal task, separately, ensuring adaptation in each setting. For all other modalities, we used the appropriate neural network architecture (i.e., CNN for images, LSTM for time series, and a multi-layer perception for all tabular data). We perform significance testing between OvO attention and the next best-performing model, detailed in Appendix Section 3.

Table 2. **MIMIC IV+CXR results**.(*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 93.73% compared to self and cross-attention.

Model	Modalities	↓ Δ FLOPs	↑ AUROC	↑ AUPRC
LSTM	Time Series	-	58.8 ±0.6	28.5 ±0.4
CNN	Images	-	56.9 ±0.3	26.7 ±0.2
Neural Net	Demographics	-	64.1 ±0.4	32.4 ±0.3
ClinicalBERT	Text	-	79.3 ±0.4	58.7 ±0.3
Concatenation	All	*	82.7 ±0.6	65.1 ±1.8
Cross-Attention	All	52,723,712	78.2 ±2.1	54.1 ±2.7
Self-Attention	All	67,633,152	78.5 ±2.0	55.7 ±3.1
OvO Attention	**All**	**4,227,072**	**83.6 ±1.1**	**66.2 ±2.6**

The results on MIMIC are presented in Table 2, clearly demonstrating the scalability and performance advantages of OvO attention. OvO's 4,227,072 FLOPs notably reduce computational costs compared to cross-attention (52,723,712 FLOPs) and self-attention (67,633,152 FLOPs), achieving reductions by **91.98%** and **93.75%**, respectively, thus highlighting OvO's superior efficiency. The unimodal results show that the textual modality is most valuable in phenotype prediction, and ClinicalBERT alone performs better than self-attention and cross-attention. This indicates that the added complexity and forced interactions are not necessarily conducive to result quality. However, OvO attention can extract information from the other modalities for a significant performance increase rather than a decrease (p-value <0.01, see Appendix Section 3).

For the six-modality Alzheimer's detection task from TADPOLE, we show our results in Table 3. OvO's 405,504 FLOPs significantly undercut cross-attention (8,921,088 FLOPs) and

Table 3. **TADPOLE results**. (*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 95.45% compared to self and cross-attention.

Model	Modalities	↓ Δ FLOPs	↑ MAUC	↑ BCA
Neural Net	AV45 PET ROI	-	63.5 ±3.1	56.4 ±3.8
Neural Net	CSF Biomarkers	-	64.4 ±1.1	53.6 ±2.7
Neural Net	MRI ROIs	-	67.0 ±1.3	57.2 ±1.0
Neural Net	FDG PET ROI	-	66.6 ±0.3	60.8 ±0.7
Neural Net	Demographics	-	74.6 ±0.9	62.0 ±0.6
Neural Net	Cognitive Tests	-	97.8 ±0.2	88.6 ±0.7
Concatenation	All	*	97.7 ±0.8	91.9 ±1.9
Cross-Attention	All	8,921,088	97.1 ±0.6	90.7 ±1.7
Self-Attention	All	9,633,792	94.8 ±1.1	86.6 ±2.6
OvO Attention	**All**	**405,504**	**98.3 ±0.4**	**93.0 ±1.4**

self-attention (9,633,792 FLOPs), achieving reductions of **95.45%** and **95.79%**, respectively, highlighting OvO's remarkable efficiency. Similarly to the MIMIC results, the unimodal results show that the cognitive tests modality is most valuable in disease prediction, and performs on its own better than self-attention and cross-attention. However, OvO attention can extract information from the other modalities for a significant performance increase rather than a decrease (p-value <0.01).

Lastly, the results on the six-modality eICU mortality prediction task are shown in Table 4, demonstrating the scalability and performance advantages of OvO attention.

OvO's 6,340,608 FLOPs significantly undercut those of cross-attention (129,957,888 FLOPs) and self-attention (151,781,376 FLOPs), achieving reductions of approximately **95.12%** and **95.82%**, respectively, thereby highlighting OvO's efficiency. Mirroring the trends observed in the MIMIC and TADPOLE datasets, we note a dominant unimodal modality, specifically Lab modality, in this experiment as well. While concatenating modalities does enhance performance, this improvement is not seen in self and cross-attention models. In contrast, OvO attention not only reflects these performance gains but does so significantly (p-value <0.01). We hypothesize that this is due to the overfitting of more complex integration frameworks of self and cross-attention on relatively smaller clinical datasets. OvO, in its simplicity akin to concatenation, manages to strike a balance by maintaining flexibility and capturing inter-modal interactions through its attention mechanism, thus offering an edge in performance without excessive complexity.

In summary, across diverse clinical datasets and modalities, OvO attention consistently outperforms traditional fusion techniques in both predictive performance and computational efficiency, underlining its robustness in handling complex multimodal healthcare data.

Table 4. **eICU results**. We report the average of 10 random seeds for AUROC and AUPRC, along with standard deviations. (*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 95.12% compared to self and cross-attention.

Model	Modalities	↓ Δ FLOPs	↑ AUROC	↑ AUPRC
Neural Net	Demographics	-	50.2 ±0.6	91.8 ±0.2
Neural Net	Medication	-	56.3 ±1.3	93.1 ±0.3
Neural Net	Diagnosis	-	58.2 ±2.1	93.3 ±0.4
Neural Net	Treatment	-	66.1 ±0.5	94.8 ±0.1
Neural Net	APACHE APS	-	77.6 ±0.2	97.0 ±0.1
Neural Net	Laboratory	-	81.5 ±0.4	97.0 ±0.1
Concatenation	All	*	81.7 ±1.6	97.5 ±0.3
Cross-Attention	All	129,957,888	77.6 ±1.6	95.4 ±0.3
Self-Attention	All	151,781,376	80.2 ±2.0	96.8 ±0.4
OvO Attention	**All**	**6,340,608**	**82.5 ±0.9**	**97.8 ±0.2**

6. Conclusion

We present One-Versus-Others (OvO), a new scalable multimodal attention mechanism. The proposed formulation significantly reduces the computational complexity compared to the widely used early fusion through self-attention and cross-attention methods. Notably, OvO achieves, at minimum, a reduction of 91.98% in FLOPs when benchmarked against self and cross-attention methods across a range of clinical datasets containing up to six modalities. We provide both a detailed theoretical complexity analysis and empirical evidence from a simulated experiment, illustrating that OvO's computational demand scales linearly with the number of modalities, in contrast to the quadratic scaling observed in other methods. Our proposed method provides a way to overcome one of the major challenges associated with multimodal datasets - computational resource demand and cost, thus enabling adoption in resource-constrained domains, such as clinical decision support. Overall, the results unequivocally establish that OvO not only significantly reduces computational expenses but also exceeds the performance of existing state-of-the-art fusion methodologies.

References

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* **30** (2017).
2. Y. Ming, X. Dong, J. Zhao, Z. Chen, H. Wang and N. Wu, Deep learning-based multimodal image analysis for cervical cancer detection, *Methods* **205**, 46 (2022).
3. M. Golovanevsky, C. Eickhoff and R. Singh, Multimodal attention-based deep learning for alzheimer's disease diagnosis, *Journal of the American Medical Informatics Association* **29**, 2014 (2022).

4. N. Hayat, K. J. Geras and F. E. Shamout, Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images, in *Machine Learning for Healthcare Conference*, 2022.
5. J. Yu, J. Li, Z. Yu and Q. Huang, Multimodal transformer with multi-view visual representation for image captioning, *IEEE transactions on circuits and systems for video technology* **30**, 4467 (2019).
6. P. H. Seo, A. Nagrani, A. Arnab and C. Schmid, End-to-end generative pretraining for multimodal video captioning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
7. G. Sterpu, C. Saam and N. Harte, Attention-based audio-visual fusion for robust automatic speech recognition, in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018.
8. S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh and A. Hussain, Multimodal sentiment analysis: Addressing key issues and setting up the baselines, *IEEE Intelligent Systems* **33**, 17 (2018).
9. H. Tan and M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, *arXiv preprint arXiv:1908.07490* (2019).
10. J. Li, D. Li, C. Xiong and S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in *International conference on machine learning*, 2022.
11. L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh and K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, *arXiv preprint arXiv:1908.03557* (2019).
12. H. Liu, C. Li, Q. Wu and Y. J. Lee, Visual instruction tuning, *Advances in neural information processing systems* **36** (2024).
13. D. Sharma, S. Purushotham and C. K. Reddy, Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain, *Scientific Reports* **11**, p. 19826 (2021).
14. J. Liu, D. Capurro, A. Nguyen and K. Verspoor, Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities, *Journal of Biomedical Informatics* **145**, p. 104466 (2023).
15. S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme *et al.*, Learning to exploit temporal structure for biomedical vision-language processing, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
16. Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar and C. Jawahar, Mmbert: Multimodal bert pretraining for improved medical vqa, in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021.
17. B. LI, T. Nakaguchi, Y. Yoshimura and P. Xuan, Robust multi-modal prostate cancer classification via feature disentanglement and dual attention, *Transactions of Japanese Society for Medical and Biological Engineering* , 308 (2021).
18. N. Braman, J. W. Gordon, E. T. Goossens, C. Willis, M. C. Stumpe and J. Venkataraman, Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, 2021.
19. S. E. Ilyin, S. M. Belkowski and C. R. Plata-Salamán, Biomarker discovery and validation: technologies and integrative approaches, *Trends in biotechnology* **22**, 411 (2004).
20. S. Schulz, A.-C. Woerl, F. Jungmann, C. Glasner, P. Stenzel, S. Strobl, A. Fernandez, D.-C. Wagner, A. Haferkamp, P. Mildenberger *et al.*, Multimodal deep learning for prognosis prediction in renal cancer, *Frontiers in oncology* **11**, p. 788740 (2021).

21. L. A. V. Silva and K. Rohr, Pan-cancer prognosis prediction using multimodal deep learning, in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020.
22. Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng and J. Liu, Uniter: Universal image-text representation learning, in *European conference on computer vision*, 2020.
23. J. H. Moon, H. Lee, W. Shin, Y.-H. Kim and E. Choi, Multi-modal understanding and generation for medical images and text via vision-language pre-training, *IEEE Journal of Biomedical and Health Informatics* **26**, 6070 (2022).
24. J. Lu, D. Batra, D. Parikh and S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, *Advances in neural information processing systems* **32** (2019).
25. T. Rahman, M. Yang and L. Sigal, Tribert: Full-body human-centric audio-visual representation learning for visual sound separation, *arXiv preprint arXiv:2110.13412* (2021).
26. H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui and B. Gong, Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, *Advances in Neural Information Processing Systems* **34**, 24206 (2021).
27. A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, Mimic-iv, a freely accessible electronic health record dataset, *Scientific data* **10**, p. 1 (2023).
28. A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz and S. Horng, Mimic-cxr-jpg-chest radiographs with structured labels, *PhysioNet* (2019).
29. M. Mohammed, H. Mwambi, I. B. Mboya, M. K. Elbashir and B. Omolo, A stacking ensemble deep learning approach to cancer type classification based on tcga data, *Scientific reports* **11**, 1 (2021).
30. R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga *et al.*, Alzheimer's disease neuroimaging initiative (adni): clinical characterization, *Neurology* **74**, 201 (2010).
31. R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, P. Golland, S. Klein *et al.*, Tadpole challenge: Accurate alzheimer's disease prediction through crowdsourced forecasting of future data, in *Predictive Intelligence in Medicine: Second International Workshop, PRIME 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 2*, 2019.
32. T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark and O. Badawi, The eicu collaborative research database, a freely available multi-center database for critical care research, *Scientific data* **5**, 1 (2018).
33. S. Sheikhalishahi, V. Balaraman and V. Osmani, Benchmarking machine learning models on multi-centre eicu critical care dataset, *Plos one* **15**, p. e0235424 (2020).

Translating Big Data Imaging Genomics Findings to the Individual: Prediction of Risks and Outcomes in Neuropsychiatric Illnesses

Peter Kochunov
Department of Psychiatry and Behavioral Sciences, University of Texas Health Science Center at Houston and UT Health Houston School of Behavioral Health Sciences, Houston, TX, USA

Li Shen
Department of Biostatistics, Epidemiology and Informatics
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
Email: li.shen@pennmedicine.upenn.edu

Zhongming Zhao
Center for Precision Health, McWilliams School of Biomedical Informatics
University of Texas Health Science, Center at Houston, Houston, TX, USA

Paul M. Thompson
USC Mark and Mary Stevens Neuroimaging and Informatics Institute
Keck School of Medicine, University of South California, Los Angeles, CA, USA
Email: pthomp@usc.edu

This PSB 2025 session is focused on opportunities, challenges and solutions for translating Big Data Imaging Genomic findings toward powering decision making in personalized medicine and guiding individual clinical decisions. It combines many of the scientific directions that are of interest to PSB members including Big Data analyses, pattern recognition, machine learning and AI, electronic health records and others.

1. Introduction

National and international scientific efforts are expanding toward collection, sharing and analyses of large and inclusive epidemiological and illness-focused datasets that combine genetic, imaging, metabolic and electronic health records (EHRs) data to enable examination of the contribution of genetic, environmental and interventional factors to human illness and health. High-resolution neuroimaging (~10^{4-6} voxels), genetic (10^{6-8} single nucleotide polymorphic variants (SNPs)) and EHRs (~10^{2-5} structured features + clinical notes) per individual are available in statistically powerful (N=10^{3-5}) epidemiological and disorder-focused samples. This also leads to major challenges on collection, sharing and homogenization of data, including how to identify reproducible signatures of complex polygenic illnesses. Research findings in such illnesses, e.g., neuropsychiatric, neurodegenerative, metabolic and other complex disorders, have historically suffered from a substantial variability and heterogeneity both within and across disorders - including genetics, environmental risk factors, mean age of onset, symptom presentations, treatment response, and long-term prognosis. Sources of heterogeneity have long remained a challenge to clinicians and scientists and have contributed to a surprisingly poor reproducibility

and difficulty in translating research findings to personalized risk assessments that can guide clinical decisions.

Presentations in this session demonstrate how Big Data collaborations such as IBM Watson Health, UK Biobank (UKBB), Enhancing Neuro Imaging Genetics through Meta Analyses (ENIGMA), the Human Connectome Project (HCP), Alzheimer's Diseases Neuroimaging Initiative (ADNI), Psychiatric Genetics Consortium (PGC), Penn Medicine EHRs and others have enabled novel principled approaches to reduce false positive findings and improve sensitivity, specificity and reproducibility of true findings. This session is focused on the methodological breakthrough that used multi-cohort/national Big Data collaborations to derive imaging and genetics signatures of complex illnesses from depression to cancer and translate them to guide personalized clinical decisions. The objective of our session is to encourage and disseminate novel analytical concepts, approaches, and applications to speed up the development of innovative technologies for hypothesis testing and data-driven discovery and translation to personalized medicine. Here we summarize the six submissions accepted for the session, with an emphasis on the diversity and coverage of the novel approaches. The accepted submissions were selected to cover novel analytic developments and applications with a focus on deriving novel risk measures for neuropsychiatric illnesses. The computational methods range from linear algebra to Artificial Intelligence and Machine Learning with imaging and omics data. *The first two contributions* focus on methodological developments intended to answer such fundamental questions as causality of identified genetic variants, preserving individual privacy in the Big Data genetic studies, and testing novel approaches for deriving genomic-trait association. *The second two contributions* report novel findings, including linking hypotheses generation and analyses across multiple Big Data samples. *The final two contributions* report on novel approaches for translating Big Data findings to the level of the individual in mental health and oncology.

2. Overview of Contributions

Childhood-to-adolescence is a critical period for brain development that corresponds to maturation of cerebral grey matter, that peaks at puberty, and maturation of cerebral white matter that peaks in late adolescence [1]. This supports the development and maturation of structural and functional networks that support higher cognitive skills [2-5]. It is also the period associated with development of lifelong, severe neuropsychiatric illnesses including autism spectrum disorder, schizophrenia, bipolar disorder, major depressive disorder and others [1, 6-11]. These illnesses are characterized by deviations from the normal brain maturation trajectory caused by action of risk factors that include genetic predisposition, pre/perinatal complications, childhood adversity and others. Early-life malnutrition has among the largest effect sizes and also is a key target for intervention and prevention. The manuscript by Gurkas and Karakurt describes a study where lifelong impact of early life malnutrition was quantified via EHRs data collected in adulthood. The greatest effects of childhood malnutrition in adults included problems with pregnancy/fetal abnormalities (20%), development of psychological/psychiatric illness (up to 16%), development of speech disorder (11%), followed by higher rates of various infection. Thus, childhood malnutrition can have lasting impact on both those who experienced it and their offspring.

The paper by Jacokes and colleagues considers advanced neuroimaging measures and blood-derived measures of gene expression to improve our understanding of autism spectrum disorder (ASD). Specifically, this paper uses logistic regression based on imaging and gene expression measures to predict ASD diagnosis, in a classification task, by using two different PCA-based approaches for feature reduction. The authors' integration of multiple methods is important for the field to advance. The lack of significant gene expression predictors suggests that brain microstructure anomalies may more tightly associated with ASD; even so, there may be a partial dissociation between blood-based and brain-based gene expression.

The paper by Noshin et al. explores the use of Electronic Health Records (EHR) to identify important diagnostic features for three types of Neuro-Degenerative Disorders (NDD), including Alzheimer's Disease (AD), Parkinson's Disease (PD), and other dementias (OD). By analyzing the EHR data from a cohort of 70,420 Alzheimer's Disease and Related Dementisa (ADRD) patients treated at Penn Medicine, the research aims to uncover key risk factors for these neurodegenerative disorders. The study employed both univariate and multivariate machine learning (ML) approaches and compared their performance in identifying risk features. A key finding is that the univariate approach was effective in uncovering rare but clinically important features specific to each disorder, while the features common across all methods represent the most robust indicators. The study also highlights the advantages and limitations of each ML method in the context of EHR data. This work is significant for researchers interested in using real-world clinical data to study neurodegenerative diseases, offering insights into the strengths and weaknesses of various ML approaches for ADRD and NDD research.

The effects of neuro-psychiatric illnesses on the brain are not regionally uniform. Neuropsychiatric disorders exert large pathological effects on some areas and circuits of the brain, while sparing others. Presently, Big Data meta-analytic studies of mental and neurological illnesses tabulate regional effect sizes using structural and/or functional brain atlases that are based on the anatomical boundaries, landmarks and connectivity patterns in healthy brains. Researchers have translated these findings to individual level predictors using approaches such as the Regional Vulnerability Index (RVI). RVI and other similar approaches quantify the agreement between individual brain patterns and the expected illness patterns identified by Big Data case control studies. Standard anatomical or connectomics-based atlases that were derived from healthy subjects are typically used to tabulate these effect sizes. However, these atlases are unlikely to capture the regional deficit pattern expressed in specific disorders, whereby the regions affected by illness may be averaged with regions that are spared, reducing the specificity and sensitivity of individual-level predictions. The study by Huang, Labate and colleagues posited that disorder-specific atlases derived using the Kullback-Leibler (KL) distance may offer a solution. KL-distance is a statistical measure of the dissimilarity between two arbitrary distributions. This offers a more stable approach to identifying areas of contrast between cases and controls than for example effect size-based measurements, because it is more stable in the presence of the non-Gaussian effects such as kurtosis, skewness and outliers. This study applied this approach to pilot a novel cortical template for Regional Homogeneity (ReHo) measurements in the subjects with the

Major Depressive Disorder (MDD). ReHo is measurement of homogeneity in the time course of blood oxygenation level dependence signal in functional MRI that was hypothesized to capture regional hypoperfusion deficits in this disorder. The MDD specific template the cerebral cortex was created by subdividing cortical landscape into contiguous region with 10 level. Each level constituted the compromise between the effects of MDD and size of the parcel to maximize contrast to noise ratio. They showed that the RVI metric--calculated using an MDD-specific parcellation--showed numerically higher effect sizes for separating patients and controls vs. those calculated using the standard Desikan-Killiany Atlas.

The contribution by He and colleagues addressed an important topic in the study of Alzheimer's disease (AD), which is to quantify Alzheimer's progression through multi-modal imaging-based pseudotime approaches. AD is a neurodegenerative disorder with no cures, and early detection is critical for successful intervention. This study explored pseudotime methods, which convert cross-sectional brain imaging data into 'faux' longitudinal data, to model the progression of AD and better understand how this complex process unfolds over time. Using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, the study evaluated pseudotime scores derived from individual imaging modalities and multi-modal data. The study found that most pseudotime analysis tools did not perform well on brain imaging data, with issues like reversed progression scores or poor distinction between diagnosis groups, likely due to assumptions designed for single-cell data. However, one tool showed promising results, where pseudotime from both single imaging modalities and multi-modal data captured the progression of diagnosis groups. Multi-modal pseudotime confirmed the hypothetical order of imaging phenotypes, and was primarily driven by amyloid and tau imaging, indicating their continuous changes across the full spectrum of Alzheimer's disease progression.

The manuscript by Ozdemir et al. tackles the long-standing question of predicting the future development of Alzheimer's disease (AD) in people who have mild cognitive impairment (MCI) – a condition that increases risk for AD, where people tend to develop AD at a rate of around 15% per year. The authors introduce a novel dynamic deep learning model for early prediction of AD (DyEPAD) to predict pro-gression from MCI to AD using EHR data. In the first step of DyEPAD, embeddings for each timestep or visit are captured through Graph Convolutional Networks (GCN) and aggregation functions. In the final step, DyEPAD employs tensor algebraic operations for frequency domain analysis of these embeddings, capturing the full scope of evolutionary patterns across all time steps. Their experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) and National Alzheimer's Coordinating Center (NACC) datasets show that their proposed model outperforms or is on a par with other state-of-the-art methods.

References

1. Kochunov, P. and L.E. Hong, *Neurodevelopmental and neurodegenerative models of schizophrenia: white matter at the center stage.* Schizophr Bull, 2014. **40**(4): p. 721-8.
2. Adibpour, P., et al., *Anatomo-functional correlates of auditory development in infancy.* Dev Cogn Neurosci, 2020. **42**: p. 100752.
3. Caffarra, S., et al., *Development of the visual white matter pathways mediates development of electrophysiological responses in visual cortex.* Hum Brain Mapp, 2021. **42**(17): p. 5785-5797.
4. Flechsig, P., *Developmental (myelogenetic) localisation of the cerebral cortex in the human.* Lancet, 1901. **158**: p. 1027-30.
5. Natu, V.S., et al., *Apparent thinning of human visual cortex during childhood is associated with myelination.* Proc Natl Acad Sci U S A, 2019. **116**(41): p. 20750-20759.
6. Rapoport, J.L., A. Addington, and S. Frangou, *The neurodevelopmental model of schizophrenia: what can very early onset cases tell us?* Curr Psychiatry Rep, 2005. **7**(2): p. 81-2.
7. Casey, B.J., J.T. Nigg, and S. Durston, *New potential leads in the biology and treatment of attention deficit-hyperactivity disorder.* Curr Opin Neurol, 2007. **20**(2): p. 119-24.
8. Kalia, M., *Brain development: anatomy, connectivity, adaptive plasticity, and toxicity.* Metabolism, 2008. **57 Suppl 2**: p. S2-5.
9. Feinberg, I., *Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence?* J Psychiatr Res, 1982. **17**(4): p. 319-34.
10. Kochunov, P., et al., *Translating ENIGMA schizophrenia findings using the regional vulnerability index: Association with cognition, symptoms, and disease trajectory.* Hum Brain Mapp, 2020.
11. Kochunov, P., et al., *Ancestral, Pregnancy, and Negative Early Life Risks Shape Children's Brain Dis/Similarity to Schizophrenia.* Biol Psychiatry, 2023. **94**(4): p. 332-340.

Electronic Health Record Analysis for Personalized Medicine: Predicting Malnutrition-Related Health Outcomes and Secondary Neuropsychiatric Health Concerns

Pinar Gurkas
Clayton State University, Department of Psychology
Morrow, GA 30260, USA
Email: PinarGurkas@clayton.edu

Gunnur Karakurt
Case Western Reserve University, Department of Psychiatry
Cleveland, OH 44109, USA
Email: gunnur.karakurt@case.edu

Malnutrition poses risks regarding cognitive, behavioral, and physical well-being. The aim of this study was to investigate the prevalent health issues associated with malnutrition by utilizing electronic health records (EHR) data. The IBM Watson Health, Explorys platform was used to access the EHR data. Two cohorts were created by two queries; patients with a history of malnutrition (n=5180) and patients without a history of malnutrition diagnosis (n= 413890). The log odds ratio and $\chi 2$ statistic were used to identify the statistically significant differences between these two cohorts. We found that there were 35 terms that were more common among the cohort with the malnutrition diagnosis. These terms were categorized under developmental anomalies, infectious agents, respiratory system issues, digestive system issues, pregnancy/prenatal problems, mental, behavioral, or neurodevelopmental disorders, diseases of the ear or mastoid process, diseases of the visual system, and chromosomal anomalies. The management of malnutrition in children is a complex problem that can be addressed with a multifactorial approach. Based on the key themes emerging from among the commonly prevalent terms identified in our study, infection prevention, education in appropriate nutritional solutions for digestive health issues, supportive services to address neurodevelopmental needs, and quality prenatal healthcare would constitute beneficial prevention efforts. Improving our understanding of malnutrition is necessary to develop new interventions for prevention and treatment.

Keywords: Malnutrition, mental health, developmental anomalies, Electronic Health Records

1. Background

1.1. *Malnutrition as a Public Health Issue*

Malnutrition is a global health concern with long-lasting negative health implications regarding children's mental and physical well-being. The World Health Organization [1, 2] defines malnutrition as "deficiencies, excesses, or imbalances in a person's intake of energy and/or nutrients

“with pediatric undernutrition accounting for about 45% of all child deaths globally. Severe malnutrition increases the risk of serious illness and death as well as acute infectious diseases for children with malnutrition [3]. Malnutrition is a critical public health problem in many parts of the world. Every form of malnutrition poses a significant threat to health and well-being and consequently improving nutrition is an important global health priority [1, 2].

Protein-energy malnutrition (PEM) is a global problem present in both developing and industrialized countries. In developing countries, malnutrition is often associated with several risk factors, including social and environmental factors such as poverty, poor education, limited access to health care, and a polluted environment. Living in areas with high rates of infectious diseases such as respiratory infections, diarrheal diseases, human immunodeficiency virus (HIV), tuberculosis, and nutritional factors such as acute and chronic food shortages and suboptimal feeding practices are also associated with malnutrition [3, 4]. In industrialized countries, malnutrition frequently observed in the form of micronutrient deficiencies in childhood can have long-term effects on health and productivity in adulthood. Not only detection and prevention of severe malnutrition but also subclinical deficiencies can be important to reverse the adverse effects of these deficiencies (especially iron and B vitamins) on the social, cognitive, and physical well-being of children [5].

Although malnutrition is a major global health problem that threatens human well-being, until recently there was no consensus on diagnostic criteria [6]. Many researchers used international growth standards to diagnose and treat severe malnutrition. The 2021 edition of the Joint Child Malnutrition Estimates (JCME) published by the UNICEF/ World Health Organization / World Bank Group provided country-level assessments regarding types of malnutrition including stunting and wasting among children under the age of five [7]. Children who are affected by stunting are too short for their age. Stunting is associated with irreversible physical and cognitive deficits. It is expected that the coronavirus disease 2019 (COVID-19) pandemic will also impact rates of malnutrition negatively. Based on the JCME, the global number of children affected by undernutrition is expected to have increased by up to 15% due to the negative impact of COVID-19 on household income, access to nutritious food, and essential nutrition services [7].

Children need good nutrition to grow, learn, play, and participate. A review study conducted in 1995 found that children with a history of early childhood malnutrition were likely to have lower cognitive functioning, school performance, and more behavioral problems as compared to matched controls and their siblings [8]. A recent systematic review found that children with a history of early childhood malnutrition were likely to have impaired cognition and higher levels of behavioral problems during childhood and adolescence [9]. However, the causal pathways linking malnutrition to neurodevelopment, cognition, behavior, and mental health are not clearly identified by the previous research due to shared environmental complexity with a multitude of risk factors including poverty, socioeconomic adversity, risk of infectious disease, lack of parental engagement, and school truancy [9]. The same systematic review concluded that studies examining the effects of malnutrition on children’s mental health are inconclusive. It was indicated that only a few studies

have examined specific areas of mental health, such as depression, and that more research is needed to investigate the long-term care needs of children with a malnutrition history and to investigate the prevalent health concerns co-occurring with malnutrition to improve patient outcomes.

2. Methods

2.1. *Motivation*

Utilizing big data in the form of Electronic Health Records (EHR) is a relatively recent approach to conducting research regarding mental and physical health with various significant knowledge discovery implications. However, big data has been rarely used to study the implications of malnutrition on human development particularly in industrialized countries such as the U.S. and Canada. Given the lack of current evidence on this increasingly important topic, in this study, we aim to identify the health correlates that are highly prevalent among patients with a diagnosis of malnutrition using large-scale EHR data. The objective of this study is to investigate EHR data to explore the associations between malnutrition and negative health outcomes. We hypothesized that malnutrition diagnosis will be adversely linked to mental and physical health outcomes among participants in the EHR data from the US and Canada.

2.1.1. *Data-Source*

Data for this study was pulled through IBM Watson Health, Explorys EHR platform. Explorys is a commercial platform that provides access to EHR data on clinical health information. It provides data from various hospital-affiliated providers from 40 acute care facilities, from more than 400,000 providers and Physicians, from 2.0 billion patient visits, and 4.4 million emergency care visits (IBM Explorys, 2020). It utilizes the Systematized Nomenclature of Medicine (SNOMED) via the International Classification of Diseases (ICD), one of the designated standards for use in EHRs, to systematically structure medical terms. The data is accessed through queries. Quarries produce data for the frequencies of selected medical diagnostic terms. It is also possible to select a subset of patients from the system. For example, the Explorys platform has been successfully used to investigate associations between multiple sclerosis and COVID-19 [10], colorectal cancer [11], and substance use and mental health concerns [12].

2.2. *Sample*

In this study, cohorts were created by two queries in the late fall of 2021, i) all patient records containing malnutrition diagnosis and ii) all patient records not containing malnutrition diagnosis. In the present study, malnutrition is a clinician-based diagnosis that is coded through ICD10 into the patient chart. The IBM Explorys data provides these patient cohorts based on the diagnosis. Although there is no consensus on the diagnostic criteria, many clinicians in the US and Canada use the Global Leadership Initiative on Malnutrition's (GLIM) criteria for the diagnosis of malnutrition [6]. Specifically, a two-step methodology is frequently used in the diagnosis of malnutrition, the initial stage involves screening individuals to identify those who are at risk by utilizing validated

screening tools followed by the subsequent stage, where the assessment is conducted to diagnose the condition and determine its severity. The diagnosis of malnutrition can often be made through a comprehensive assessment of physical observations and a thorough history of the patient's dietary and health status by taking into account the notably non-volitional weight loss, low body mass index, and diminished muscle mass (i.e., phenotypic criteria), in addition to reduced food intake and occurrences of inflammation or disease burden (i.e, etiologic criteria).To diagnose malnutrition at least one phenotypic criterion and one etiologic criterion should be present. Healthcare practitioners additionally may conduct a body mass index (BMI) or a child's arm circumference assessment. The following terms were used in diagnostic decision-making by the clinicians: Deficiency of macronutrients, Disease Malnutrition (calorie), Nutritional deficiency disorder, Nutritional disorder, and Undernutrition. Clinicians also included their observations such as body measurements, pediatric percentile measurement, body mass index, chemistry hematology, cell fractions/differential myeloid cells, and pediatric weight percentile. If feasible, a blood sample is procured for the purpose of examining the potential problems in specific micronutrients. In the present study, the malnutrition cohort had 5180 patient records and the cohort with no malnutrition had 413890 patient records. The majority of the patient records in the malnutrition cohort were children; 40% were younger than 4 years of age, 35% were between 5 to 9 years of age, 25% were between 10 to 14 years of age, and 5% were older than 18 years of age. Fifty-six percent of the malnutrition cohort were males.

2.3. *Analysis*

We conducted the analysis at the level of frequencies. Our analytical strategy is addressing this difference based on unequal group sizes. Specifically, we use odds ratio statistics to investigate the proportional prevalence rate of the health concerns. This statistical framework has been used in other research publications including traumatic brain injury [TBI, 13] intimate partner violence [14], and mental health predictions [15]. The odds ratio takes into account the expected observation in larger cohorts as compared to observation in smaller cohort sizes.

Two statistical analyses, namely the log odds ratio and the $\chi 2$ statistic, were utilized in this study. Specifically, the log odds ratio was utilized to compute the logarithm of the ratio between the frequencies of Malnutrition and No Malnutrition. The $\chi 2$ statistic was utilized to determine if there is a significant association between observed and expected frequencies of diagnostic terms in the two cohorts. We ranked the terms based on both the log odds ratio and $\chi 2$ statistic. Two rankings based on frequencies were produced, and the highest rank allocated to every term was established. Subsequently, diagnostic terms were ranked based on the premise that a term could only attain a high ranking if it satisfied the criteria in both the log odds ratio and $\chi 2$ statistic rankings. Furthermore, one-tailed z-tests ($p<0.05$) were conducted as a conservative approach to reduce the likelihood of Type 1 errors to evaluate the extent of the evidence against the null hypothesis. Using the extent of the standard deviations and the sample mean's distance from the population mean, we concentrated the inferences concerning the diagnostic terms grounded in sample data with higher confidence.

To conduct a one-tailed z-test for an odds ratio, we first calculated the log odds ratio (logOR), the standard error of the logOR, and then we compared the calculated z-score to a critical value-based $p<.05$ significance level and the direction of our hypothesis (one-tailed test) to determine if the observed odds ratio is statistically significant. To make this calculation we used a commercially available statistical package (IBM SPSS Statistics).

Null Hypothesis (H0): The odds ratio is equal to 1 (no association between malnutrition and non-malnutrition cohorts).
Alternative Hypothesis (Ha): The odds ratio is greater than 1 (testing for a positive association between malnutrition diagnosis and health concerns).

In order to calculate the z-score, SPSS uses Log OR to SE ratio: z = (logOR) / (SE).

2.4. *Tables*

For more information about the sample characteristics see Table 1.

Table 1. Characteristics of Study Population

Variables	**Malnutrition**	**Non- Malnutrition**
Race	**n (%)**	**n (%)**
Caucasian	1865 (36%)	289723 (70%)
African American	829 (16%)	74500 (18%)
Asian	104 (2%)	4139 (1%)
Hispanic/Latino	52 (1%)	4139 (1%)
Native American/Alaskan Native	0 (0%)	0 (0%)
Multi-racial (race)	207 (4%)	12417 (3%)
Other (race)	207 (4%)	8278 (2%)
Gender		
Female	2279 (44%)	215223 (52%)
Male	2901 (56%)	198667 (48%)
Vitals		
Severely underweight body mass index <16.49	2849 (55%)	53806 (13%)
Underweight body mass index 16.5< x <18.49	1968 (38%)	95195 (23%)
Prehypertension sys. 120-139 dias. 80-89	1295 (25%)	322834 (78%)

3. Results

The age of a patient was utilized to categorize them into one of the 3 age groups; juniors (i.e., younger than 18), adults (i.e., between 18 and 65), and seniors (i.e., older than 65). Ninety-five percent of the records were those from juniors majority of whom were younger than 4 years of age. The majority of the records with no malnutrition diagnosis were from patients who were adults older than 50 years of age. Investigation of the most frequently observed diagnostic terms among those diagnosed with malnutrition indicated that; attention deficit hyperactivity disorder, deficiency of macronutrients, developmental disorder, a developmental disorder of motor function, disorder by body site, disorder of body system, disorders of attention and motor control, malnutrition (calorie), the mental disorder usually first evident in infancy, childhood AND/OR adolescence were among the most frequently observed terms for those in that cohort under the age of 4.

Our comparison of the malnutrition cohort to no malnutrition background cohort indicated that 35 terms were significantly more prevalent among the malnutrition cohort (See Table 2). Identified terms were classified into broader categories based on ICD 11 classification system, and SNOMED. We present the distribution of the 35 terms into these categories in Table 2. These broader parent codes include developmental anomalies, infectious agents, respiratory system, digestive system, pregnancy/prenatal problems, mental, behavioral, or neurodevelopmental disorders, diseases of the ear or mastoid process, diseases of the visual system, and chromosomal anomalies (See Figure 1).

Table 2. Comparison of the malnutrition cohort to no malnutrition background

Diagnostic Terms	**Malnutrition Frequency n=5180 n (%)**	**Non-Malnutrition Frequency n= 413890 n (%)**	**Log Odds Ratio**	**p**	**Z score**
Infective laryngitis	370 (7%)	380 (0.09%)	3.50	<0.001	47
Coxsackie virus infection of oral cavity	250 (5%)	260 (0.06%)	3.32	<0.001	37
Enlargement of tonsil or adenoid	200 (4%)	210 (0.05%)	3.22	<0.001	32
Tonsil and/or adenoid hypertrophy	200 (4%)	210 (0.05%)	3.22	<0.001	32
Infection of larynx	370 (7%)	390 (0.09%)	3.20	<0.001	43
Arterial malformation	250 (5%)	270 (0.07%)	3.02	<0.001	34
Developmental speech disorder	580 (11%)	630 (0.15%)	3.01	<0.001	51
Congenital anomaly of pulmonary artery	110 (2%)	120 (0.03%)	2.95	<0.001	22
Congenital anomaly of tricuspid valve	110 (2%)	120 (0.03%)	2.95	<0.001	22
Deletion of part of autosome	80 (2%)	90 (0.02)	2.81	<0.001	18
Tracheitis	80 (2%)	90 (0.02%)	2.81	<0.001	18

Condition in fetus originating in the perinatal period	460 (9%)	530 (0.13%)	2.76	<0.001	42
Disorder of psychological development	840 (16%)	980 (0.24%)	2.75	<0.001	56
Cow's milk protein sensitivity	70 (1%)	80 (0.02%)	2.75	<0.001	16
Fetal disorder	310 (6%)	360 (0.09%)	2.72	<0.001	34
Disorder of fetus or newborn	1020 (20%)	1220 (0.29%)	2.70	<0.001	60
Acute suppurative otitis media without spontaneous rupture of ear drum	450 (9%)	530 (0.13%)	2.69	<0.001	41
Abnormal ventriculoarterial connection	120 (2%)	140 (0.03%)	2.69	<0.001	22
Congenital biliary atresia	60 (1%)	70 (0.02%)	2.68	<0.001	15
Disease caused by Adenovirus	60 (1%)	70 (0.02%)	2.68	<0.001	15
Disease caused by Poxviridae	60 (1%)	70 (0.02%)	2.68	<0.001	15
Disease caused by unassigned Poxviridae	60 (1%)	70 (0.02%)	2.68	<0.001	15
Molluscum contagiosum infection	60 (1%)	70 (0.02%)	2.68	<0.001	15
Coxsackie virus disease	280 (5%)	330 (0.08%)	2.67	<0.001	32
Alternating esotropia	50 (1%)	60 (0.01%)	2.60	<0.001	13
Hypoplasia of the optic nerve	50 (1%)	60 (0.01%)	2.60	<0.001	13
Inflammation of bronchiole (Human metapneumovirus)	50 (1%)	60 (0.01%)	2.60	<0.001	13
Neuromuscular scoliosis	50 (1%)	60 (0.01%)	2.60	<0.001	13
Anomaly of jaw size	90 (2%)	110 (0.03%)	2.56	<0.001	13
Disease caused by Enterovirus	90 (2%)	110 (0.03%)	2.56	<0.001	13
Overriding aorta	90 (2%)	110 (0.03%)	2.56	<0.001	18
Right ventricular hypertrophy	90 (2%)	110 (0.03%)	2.56	<0.001	18
Tetralogy of Fallot	90 (2%)	110 (0.03%)	2.56	<0.001	18
Congenital anomaly of muscle AND/OR tendon	130 (3%)	160 (0.04%)	2.54	<0.001	18
Congenital anomaly of skeletal muscle	130 (3%)	160 (0.04%)	2.54	<0.001	18

4. Discussion

The aim of the study was to examine prevalent health issues related to malnutrition diagnosis and associated neuropsychiatric health issues utilizing EHR data available for the U.S. and Canada. The EHR platform utilizes the Systematized Nomenclature of Medicine (SNOMED) via the International Classification of Diseases (ICD), one of the designated standards for use in EHRs, to systematically structure medical terms [16]. The study underscores the importance of bridging the gap between imaging for malnutrition and the early detection of malnutrition signs, particularly in children. By supporting clinical insights that focus on predicting risks and outcomes, this approach aims to prevent secondary neuropsychiatric illnesses that can arise from nutrition-related problems. This paper explores the challenges and solutions involved in translating findings from electronic health records (EHRs) into actionable insights for personalized medicine, ultimately facilitating informed clinical decisions at the individual level.

Building on the consensus criteria established by the European Society of Clinical Nutrition and Metabolism (ESPEN)—which include weight loss, reduced BMI, and reduced fat-free mass index (FFMI)—this research emphasizes the limitations and regional variations in the prevalence of malnutrition [6]. In North America, where rates of moderate and severe stunting and wasting are classified as low and very low, respectively, the study suggests that these disparities in malnutrition outcomes may be even more pronounced in regions characterized by higher food insecurity and political instability [17]. This highlights the need for context-specific analyses and the integration of diagnostic imaging and biomarker criteria in addressing malnutrition across diverse populations.

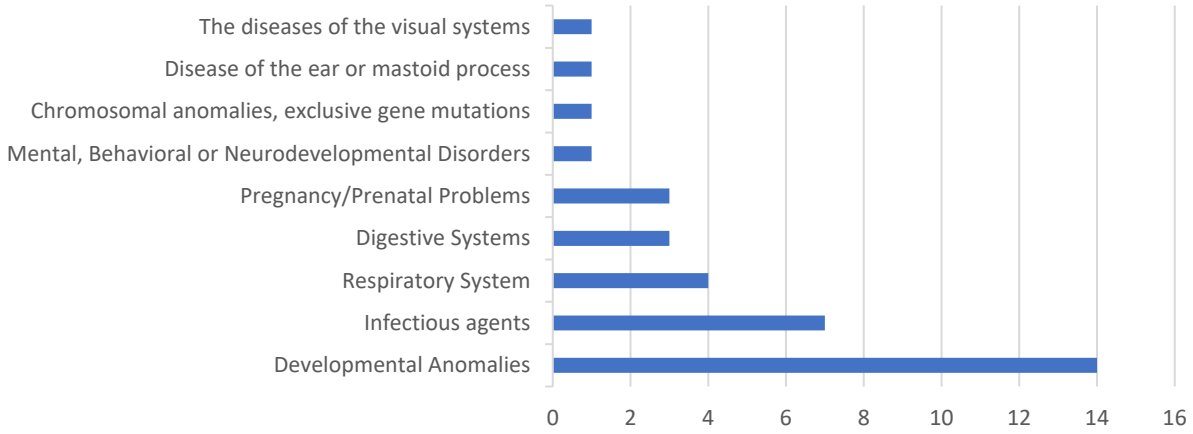


Figure 1. Distribution of Significant Terms into Broader Categories
This figure provides information regarding the 35 terms that were significantly more prevalent in the malnutrition cohort classified into broader categories based on ICD 10 classification system, and SNOMED. These broader categories were developmental anomalies, infectious agents, respiratory system, digestive system, pregnancy/prenatal problems, mental, behavioral, or neurodevelopmental disorders, diseases of the ear or mastoid process, diseases of the visual system, and chromosomal anomalies.

Our results showed that 35 health correlates were significantly more common among the undernourished cohort. Key themes emerging from among the commonly prevalent terms included developmental anomalies, infectious agents, respiratory system problems, digestive system problems, pregnancy/prenatal problems, mental, behavioral, or neurodevelopmental disorders, diseases of the ear or mastoid process, diseases of the visual system, and chromosomal anomalies.

Prevalence of these diagnostic themes could be due to various reasons, including (i) terms such as mental, behavioral, or neurodevelopmental disorders are a direct consequence of current malnutrition (ii) terms such as chromosomal abnormalities and pregnancy/prenatal problems are risk factors that increase the vulnerability to malnutrition (iii) terms such as frequent infections, respiratory system issues, gastrointestinal system issues might be more prevalent in the demographics such as regions with climate inequality, among malnourished children (iv) terms such as ear or mastoid disorders and visual system disorders are associated with the occurrence of malnutrition and linked to a potentially long-term consequence.

Examination of the most frequently observed diagnostic terms in a cohort aged 0-4 years diagnosed with malnutrition revealed that this cohort was also frequently diagnosed with attention deficit hyperactivity disorder and macronutrient deficiencies. Past research children with ADHD were also found to have lower levels of protein intake and lower levels of vitamins B1, B2, and C and lower levels of zinc, iron, and calcium in comparison to typically developing children in the control group [17]. Significant terms might also have potentially indirect mental and social developmental effects such as musculoskeletal disorders. Neuromuscular scoliosis is one of the terms that was observed as significantly more prevalent among malnourished children. Multiple surgeries (traditional growing rods) during the treatment process in combination with the condition itself might have a negative impact on the mental health of children resulting in lower self-esteem, depression, and anxiety [18].

Previous studies have also emphasized potential ethnic and racial differences in malnutrition and their impact on biological reference intervals. For instance, Colantonio et al. (2012) investigated the impact of ethnicity on biochemical markers of health and disease in a healthy cohort of 179 multi-ethnic and racially diverse Canadian children and adolescents [19]. The study found variations in levels of ALT, iron, total cholesterol, triglycerides (TG), and SHBG. Additionally, Vitamin D levels were higher in Caucasian pediatric participants compared to Black, Asian, and Hispanic participants. These biochemical markers play a critical role in diagnosing and understanding the manifestations of malnutrition. Therefore, establishing accurate reference intervals is essential for detecting the true extent of these health concerns in developing children, which is vital for improving health equity [20].

Social, political, and economic factors can be linked to chronic infections play an important role in the emergence of malnutrition through systemic and individual interactions [21]. In our data from the US and Canadian hospital systems, we observed that boys are slightly more likely to have a history of malnutrition [22]. A recent systematic review found that there are sex differences regarding undernutrition with boys more likely than girls to be wasted, underweight and stunted [22]. The occurrence of secondary malnutrition in developed countries is linked to abnormal nutrient loss, increased energy expenditure, or decreased food intake, often associated with chronic diseases such as cystic fibrosis, chronic renal failure, childhood malignancies, congenital heart disease, and neuromuscular disease.

The extent to which early impairments in neurodevelopment due to malnutrition impact future cognition and function, and the extent to which other relevant environmental factors such as prenatal nutrition, family characteristics, and infectious diseases influence these findings are unclear. Therefore, future research comparing the health effects of malnutrition while controlling for these interactions can help us better understand the impact of malnutrition [22, 23].

Imaging Genetics and Neuropsychiatric Illnesses and Future Research. Nutritional biomarkers serve as objective indicators of normal biological or pathogenic processes and are categorized into three main types: biomarkers of exposure, biomarkers of status, and biomarkers of function [24]. These biomarkers are assessed through a variety of methods, ranging from traditional dietary exposure assessments, such as self-reports, to the evaluation of biological fluids, tissues, or urine, which reflect the body's total nutrient reserves or the degree of nutrient depletion [24]. Functional biomarkers provide insights into physiological and behavioral variations and nutrient imbalances, such as enzyme activity or the presence of abnormal metabolic by-products, signaling early signs of subclinical deficiencies [24]. As an emerging field, functional biochemical markers underscore the significance of studying alterations in DNA, gene expression, and immune function caused by malnutrition [24]. Key laboratory biomarkers frequently discussed in the literature for diagnosing malnutrition, assessing nutritional risk, and monitoring the efficacy of nutritional interventions include albumin, pre-albumin, transferrin, C-reactive protein, α1-acid glycoprotein, hemoglobin, cholesterol, and lymphocyte count [20]. These biomarkers are often used in combination, such as albumin, pre-albumin, cholesterol, and lymphocyte count, or albumin, hemoglobin, and total lymphocyte count, to provide a comprehensive assessment of nutritional status [20].

There is limited research utilizing neuroimaging to document the effects of malnutrition on the brain. Among the most commonly utilized neuroimaging methods examining this impact are Electroencephalogram (EEG), Functional Near-Infrared Spectroscopy (fNIRS), and Nuclear Magnetic Resonance (NMR). Research utilizing the EEG method to study malnutrition reveals alpha wave disturbances that are linked to neurodevelopmental delays in children [25]. The fNIRS is another neuroimaging technique that can be particularly valuable for assessing brain functionality among young children [26, 27, 28]. Studies using this technology found that cerebral blood flow in malnourished children is a biological marker of cognitive functional difficulties [29]. Due to the challenges regarding the affordability of NMR technology, this method has not been utilized often. The majority of the limited number of studies utilizing this methodology are case studies with small samples and no controls. Nonetheless, available evidence so far demonstrates the significant potential of NMR technology in early diagnosis and prevention of long-lasting impacts of malnutrition on cognitive function [30].

In a recent review of neuroimaging studies, Ayaz et al. (2023) documented cerebral atrophy of malnourished children, with or without ventricular dilation [31]. The extent of atrophy or ventricular dilation were not assessed in any of the studies reviewed. Consequently, the authors emphasized the need for a universal scoring system to quantify the extent of brain atrophy and correlate it with the severity of malnutrition that would enable healthcare providers to better assess the impact of nutritional interventions [31].

Whole brain mapping, an emerging technology for investigating the functional impact of neurocircuitry in specific regions and neural networks, holds significant promise for understanding the effects of malnutrition on neuropsychiatric concerns [32, 33]. This approach is particularly valuable for studying mechanistic pathways such as the blood-brain barrier, glial cell proliferation, and brain-body interactions [32, 33]. Additionally, gene ontology analyses offer promising mechanistic insights, including pathways related to oxidative stress, myeloid interactions with the immune system, and stress-related responses [32, 33, 34].

Limitations and Future Research. There are limitations due to the nature of this type of data. We utilized an established statistical framework to analyze the EHR data to address these limitations. Although malnutrition is a significant global health concern posing a threat to well-being, until recently there was no consensus on the diagnostic criteria. The lack of such diagnostic criteria hinders progress regarding the design and implementation of successful interventions. Children who are affected by wasting are too thin for their height either due to recent rapid weight loss or failure to gain weight are often diagnosed with malnutrition. Another limitation of the current study is that, due to the nature of the available data, it was not feasible to calculate the effects across all BMI ranges. The extent to which early impairments in neurodevelopment due to malnutrition impact future cognition and function, and the extent to which other relevant environmental factors such as prenatal nutrition, family characteristics, and infectious diseases influence these findings are unclear. Therefore, future research comparing the health effects of malnutrition while controlling for these interactions can help us better understand the impact of malnutrition [22, 35, 36].

Malnutrition is a global health concern with long-lasting negative health implications regarding children's mental and physical well-being and children living in high-income countries are not immune to this impact. Utilizing big data in the form of Electronic Health Records (EHR) is a relatively recent approach to conducting research regarding mental and physical health with various significant knowledge discovery implications. We found that for the malnourished cohort, the terms that were more significantly common were grouped under the following categories: developmental anomalies, infectious agents, respiratory system issues, digestive system issues, pregnancy/prenatal problems, mental, behavioral, or neurodevelopmental disorders, diseases of the ear or mastoid process, diseases of the visual system and chromosomal anomalies. Our findings underscore the significance of prevention strategies such as providing support during pregnancy, educating primary caregivers and family members on general child nutritional needs, basic health, and hygiene; as well as a community-based approach including affordable access to basic health services [7, 19].

In conclusion, the management of malnutrition in children is a complex problem that can be addressed with a multifactorial approach. Health policies on prevention of malnutrition such as educating primary caregivers and family members on general child nutritional needs, basic health, and hygiene might be helpful in improving children's health [21, 22]. A community-based approach including access to affordable healthy and nutritious food, basic health, water, hygiene, and sanitation services, and opportunities for safe physical activity can be beneficial to address the issue of malnutrition among children under 5 years of age [35, 36].

Furthermore, our findings underscore the importance of ensuring adequate nutrition before and during pregnancy as one of the vital pathways to improving the health and well-being of children [7]. Infection prevention, education in appropriate nutritional solutions, supportive services and quality prenatal healthcare would constitute beneficial prevention efforts.

4.1. *Footnotes*

The IBM Explorys Therapeutic Dataset is used in this manuscript in the form of aggregate statistics (number of records) in a specified population. The data is unidentified and aggregated, and no individual records were used due to privacy concerns. The study is considered an exempt study.

5. Acknowledgments

This publication was made possible by R01-LM012518 from the National Library of Medicine. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

1. World Health Organization. Nutrition. Accessed January 30, 2023. https://www.who.int/health-topics/nutrition#tab=tab_3 .
2. World Health Organization. Malnutrition. Published June 9, 2021. Accessed December 16, 2022. https://www.who.int/news-room/fact-sheets/detail/malnutrition .
3. Bhutta ZA, Berkley JA, Bandsma RHJ, Kerac M, Trehan I, Briend A. Severe childhood malnutrition. Nat Rev Dis Primers. 2017 Sep 21;3:17067. doi: 10.1038/nrdp.2017.67. PMID: 28933421; PMCID: PMC7004825.
4. Katoch OR. Determinants of malnutrition among children: A systematic review. Nutrition. 2022 Apr;96:111565. doi: 10.1016/j.nut.2021.111565. Epub 2021 Dec 11. PMID: 35066367.
5. Swaminathan S, Edward BS, Kurpad AV. Micronutrient deficiency and cognitive and physical performance in Indian children. Eur J Clin Nutr. 2013;67(5):467-474. doi:10.1038/ejcn.2013.14.
6. Cederholm T, Bosaeus I, Barazzoni R, Bauer J, Van Gossum A, Klek S, Muscaritoli M, Nyulasi I, Ockenga J, Schneider SM, de van der Schueren MA, Singer P. Diagnostic criteria for malnutrition - An ESPEN Consensus Statement. Clin Nutr. 2015 Jun; 34(3):335-40. doi: 10.1016/j.clnu.2015.03.001. Epub 2015 Mar 9. PMID: 25799486.
7. Levels and trends in child malnutrition: UNICEF/ World Health Organization / World Bank Group joint child malnutrition estimates: key findings of the 2021 edition. Published May 5, 2021. Accessed December 16, 2022. https://www.who.int/publications/i/item/9789240025257

8. Grantham-McGregor S. A review of studies of the effect of severe malnutrition on mental development. J Nutr. 1995 Aug;125(8 Suppl):2233S-2238S. doi: 10.1093/jn/125.suppl_8.2233S. PMID: 7542705.
9. Kirolos A, Goyheneix M, Kalmus Eliasz M, Chisala M, Lissauer S, Gladstone M, Kerac M. Neurodevelopmental, cognitive, behavioural and mental health impairments following childhood malnutrition: a systematic review. BMJ Glob Health. 2022 Jul;7(7):e009330. doi: 10.1136/bmjgh-2022-009330. PMID: 35793839; PMCID: PMC9260807.
10. Reder AT, Centonze D, Naylor ML, Nagpal A, Rajbhandari R, Altincatal A, Kim M, Berdofe A, Radhakrishnan M, Jung E, Sandrock AW, Smirnakis K, Popescu C, de Moor C. COVID-19 in Patients with Multiple Sclerosis: Associations with Disease-Modifying Therapies. CNS Drugs. 2021 Mar;35(3):317-330. doi: 10.1007/s40263-021-00804-1. Epub 2021 Mar 20. PMID: 33743151; PMCID: PMC7980129.
11. Elangovan A, Skeans J, Landsman M, Ali SMJ, Elangovan AG, Kaelber DC, Sandhu DS, Cooper GS. Colorectal Cancer, Age, and Obesity-Related Comorbidities: A Large Database Study. Dig Dis Sci. 2021 Sep;66(9):3156-3163. doi: 10.1007/s10620-020-06602-x. Epub 2020 Sep 21. PMID: 32954457.
12. Sprunger JG, Johnson K, Lewis D, Kaelber DC, Winhusen TJ. Five-year incidence of substance use and mental health diagnoses following exposure to opioids or opioids with benzodiazepines during an emergency department encounter for traumatic injury. Drug Alcohol Depend. 2022 Sep 1;238:109584. doi: 10.1016/j.drugalcdep.2022.109584. Epub 2022 Jul 22. PMID: 35933891; PMCID: PMC9680036.
13. Liu, Larry Y., et al. "Interplay between traumatic brain injury and intimate partner violence: data driven analysis utilizing electronic health records." *BMC women's health* 20 (2020): 1-16.
14. Karakurt G, Patel V, Whiting K, Koyutürk M. Mining Electronic Health Records Data: Domestic Violence and Adverse Health Effects. J Fam Violence. 2017 Jan;32(1):79-87. doi: 10.1007/s10896-016-9872-5. Epub 2016 Sep 23. PMID: 28435184; PMCID: PMC5397110.
15. Karakurt, Günnur, et al. "Comprehensive analysis of electronic health records to characterize the association between intimate partner violence and mental health." *AMIA Summits on Translational Science Proceedings* 2023 (2023): 310.
16. Cornet, Ronald, and Nicolette de Keizer. "Forty years of SNOMED: a literature review." *BMC medical informatics and decision making* 8 (2008): 1-6.
17. Salvat H, Mohammadi MN, Molavi P, Mostafavi SA, Rostami R, Salehinejad MA. Nutrient intake, dietary patterns, and anthropometric variables of children with ADHD in comparison to healthy controls: a case-control study. *BMC Pediatr* 22, 70 (2022).
18. Yildiz Mİ, Goker B, Demirsöz T, et al. A Comprehensive Assessment of Psychosocial Well-being Among Growing Rod Graduates: A Preliminary Investigation. J Pediatr Orthop. 2023;43(2):76-82. doi:10.1097/BPO.0000000000002298.
19. Colantonio DA, Kyriakopoulou L, Chan MK, Daly CH, Brinc D, Venner AA, et al. Closing the gaps in pediatric laboratory reference intervals: a CALIPER database of 40 biochemical markers in a healthy and multiethnic population of children. Clin Chem. 2012;58:854–68.

20. Keller, Ulrich. "Nutritional laboratory markers in malnutrition." *Journal of clinical medicine* 8.6 (2019): 775. 10.3390/jcm8060775. PMID: 31159248; PMCID: PMC6616535.
21. Zhang X, Zhang L, Pu Y, Sun M, Zhao Y, Zhang D, Wang X, Li Y, Guo D, He S. Global, Regional, and National Burden of Protein-Energy Malnutrition: A Systematic Analysis for the Global Burden of Disease Study. Nutrients. 2022 Jun 22;14(13):2592. doi: 10.3390/nu14132592. PMID: 35807772; PMCID: PMC9268010.
22. Thurstans S, Opondo C, Seal A, et al.. Boys are more likely to be undernourished than girls: a systematic review and meta-analysis of sex differences in undernutrition. BMJ Glob Health 2020;5:e004030. 10.1136/bmjgh-2020-004030.
23. Govender I, Rangiah S, Kaswa R, Nzaumvila D. Malnutrition in children under the age of 5 years in a primary health care setting. S Afr Fam Pract (2004). 2021 Sep 7;63(1):e1-e6. doi: 10.4102/safp.v63i1.5337. Erratum in: S Afr Fam Pract (2004). 2021 Dec 14;63(1):5416. PMID: 34677078; PMCID: PMC8517826.
24. Gibson RS. Principles of Nutritional Assessment. Biomarkers (2024).https://nutritionalassessment.org/biomarkers/
25. Bringas Vega, Maria L., et al. "An age-adjusted EEG source classifier accurately detects school-aged barbadian children that had protein energy malnutrition in the first year of life." *Frontiers in neuroscience* 13 (2019): 1222.
26. Gallagher, Anne, Renée Béland, and Maryse Lassonde. "The contribution of functional near-infrared spectroscopy (fNIRS) to the presurgical assessment of language function in children." *Brain and language* 121.2 (2012): 124-129.
27. Lloyd-Fox, S., Blasi, A., Elwell, C.E., 2010. Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy. Neurosci. Biobehav. Rev. 34, 269–284. doi: 10.1016/j.neubiorev.2009.07.008 .
28. Pinti, Paola, et al. "A review on the use of wearable functional near-infrared spectroscopy in naturalistic environments." *Japanese Psychological Research* 60.4 (2018): 347-373.
29. Roberts, S.B., Franceschini, M.A., Krauss, A., Lin, P.-Y., Braima de Sa, A., Có, R., Taylor, S., Brown, C., Chen, O., Johnson, E.J., Pruzensky, W., Schlossman, N., Balé, C., Wu, K.-C.(Tony), Hagan, K., Saltzman, E., Muentener, P., 2017. A pilot randomized controlled trial of a new supplementary food designed to enhance cognitive performance during prevention and treatment of malnutrition in childhood. Curr. Dev. Nutr. 1, e000885. doi: 10.3945/cdn.117.000885 .
30. Galler, Janina R., et al. "Neurodevelopmental effects of childhood malnutrition: A neuroimaging perspective." *Neuroimage* 231 (2021): 117828.
31. Ayaz, Ahmed, et al. "Structural changes in the brain on magnetic resonance imaging in malnourished children: a scoping review of the literature." *Pediatric Neurology* (2023);149:151-158. doi: 10.1016/j.pediatrneurol.2023.08.020. Epub 2023 Aug 29. PMID: 37890309.
32. Cathomas, Flurin, et al. "Circulating myeloid-derived MMP8 in stress susceptibility and depression." *Nature* 626.8001 (2024): 1108-1115
33. Lutter, Michael, et al. "Orexin signaling mediates the antidepressant-like effect of calorie restriction." *Journal of Neuroscience* 28.12 (2008): 3071-3075.
34. Bergamini, Giorgio, et al. "Mouse psychosocial stress reduces motivation and cognitive function in operant reward tests: a model for reward pathology with effects of agomelatine." *European neuropsychopharmacology* 26.9 (2016): 1448-1464.

35. Charle-Cuéllar P, Lopez-Ejeda N, Toukou Souleymane H, Yacouba D, Diagana M, Dougnon AO, Vargas A, Briend A. Effectiveness and Coverage of Treatment for Severe Acute Malnutrition Delivered by Community Health Workers in the Guidimakha Region, Mauritania. Children (Basel). 2021 Dec 4;8(12):1132. doi: 10.3390/children8121132. PMID: 34943328; PMCID: PMC8700149.
36. Tahmasebi, Houman, et al. "Influence of ethnicity on biochemical markers of health and disease in the CALIPER cohort of healthy children and adolescents." *Clinical Chemistry and Laboratory Medicine (CCLM)* 58.4 (2020): 605-617.

Unsupervised Dimensionality Reduction Techniques for the Assessment of ASD Biomarkers

Zachary Jacokes[1], Ian Adoremos[2,3], Arham Rameez Hussain[4], Benjamin T. Newman[4], Kevin A. Pelphrey[5], and John Darrell Van Horn[1,4] for the ACE GENDAAR Consortium

[1]*School of Data Science, University of Virginia*
[2]*College of Computer, Mathematical, and Natural Sciences, University of Maryland*
[3]*The Human Genetics Branch, National Institute of Mental Health*
[4]*Department of Psychology, University of Virginia*
[5]*Department of Neurology, University of Virginia*
Charlottesville, VA 22903, United States of America
Contact Email: jdv7g@virginia.edu

Autism Spectrum Disorder (ASD) encompasses a range of developmental disabilities marked by differences in social functioning, cognition, and behavior. Both genetic and environmental factors are known to contribute to ASD, yet the exact etiological factors remain unclear. Developing integrative models to explore the effects of gene expression on behavioral and cognitive traits attributed to ASD can uncover environmental and genetic interactions. A notable aspect of ASD research is the sex-wise diagnostic disparity: males are diagnosed more frequently than females, which suggests potential sex-specific biological influences. Investigating neuronal microstructure, particularly axonal conduction velocity offers insights into the neural basis of ASD. Developing robust models that evaluate the vast multidimensional datasets generated from genetic and microstructural processing poses significant challenges. Traditional feature selection techniques have limitations; thus, this research aims to integrate principal component analysis (PCA) with supervised machine learning algorithms to navigate the complex data space. By leveraging various neuroimaging techniques and transcriptomics data analysis methods, this methodology builds on traditional implementations of PCA to better contextualize the complex genetic and phenotypic heterogeneity linked to sex differences in ASD and pave the way for tailored interventions.

Keywords: Autism; Neuroimaging; Copy number variation; Gene expression; Conduction velocity

1. Introduction

Autism Spectrum Disorder (ASD) encompasses a broad range of developmental conditions characterized by persistent deficits in social functioning, cognition, and restricted, repetitive behavior[1]. Individuals with ASD often experience challenges in communication, social interactions, and engage in repetitive behaviors or have narrowly focused interests[2]. The prevalence of ASD has been steadily increasing worldwide, affecting between 1 in 36 children and 1 in 45 children according to recent meta-analyses and research by the Centers for Disease Control and Prevention (CDC)[3,4]. This diagnostic increase has brought significant attention to the urgent need for a deeper understanding of the underlying mechanisms of ASD.

Research indicates that ASD is a heterogeneous condition, meaning that it can present very differently from one person to another, complicating efforts to pinpoint its causes[5]. Although it is widely accepted that both genetic and environmental factors contribute to the development of ASD, the exact etiological factors and their interactions remain unclear. Genetic studies have identified numerous genes associated with ASD, suggesting a strong hereditary component[6–8]. However, environmental factors such as prenatal exposure to certain drugs, complications during birth, and advanced parental age have also been identified as potential risk factors for developing ASD[9,10].

Moreover, neuroimaging studies have revealed differences in brain structure and function in individuals with ASD[11–13]. These studies have shown abnormalities in areas of the brain responsible for social behavior, communication, and sensory processing. Despite these advances, there is still much to learn about how these genetic and environmental factors interact to influence brain development and lead to the diverse array of symptoms observed in ASD.

Neuroimaging and genomics exploration is essential for understanding ASD because these approaches provide complementary insights into the biological underpinnings of the condition. Neuroimaging techniques, such as MRI and fMRI, allow researchers to observe structural and functional differences in the brains of individuals with ASD; this imaging data helps to identify patterns and variations in brain development and connectivity that may contribute to ASD symptoms. Concurrently, genomics offers a window into the genetic factors influencing ASD risk, uncovering specific genes and genetic variants associated with the disorder. By integrating genomic information with neuroimaging data, research efforts can better explore how genetic predispositions affect brain structure and function, and vice versa. This combined approach is crucial for elucidating the complex interplay between genetic and neural mechanisms, ultimately enhancing our understanding of ASD and guiding the development of more targeted interventions.

1.1. *Sex-wise disparity in ASD*

A significant aspect of ASD research is the observed sex-wise disparity in its prevalence. Males are diagnosed with ASD more frequently than females, with a ratio of approximately four-to-one[3]. This disparity suggests potential sex-specific biological factors that may influence the development of ASD. Several hypotheses have been proposed to explain this difference, including genetic differences in sex chromosomes, hormonal influences, and differences in brain structure and function between males and females[12,14]. Understanding these sex-specific factors is crucial for developing tailored diagnostic and therapeutic approaches for ASD.

1.2. *Neuronal microstructure analysis in ASD*

Neuroscientific research has increasingly focused on the neuronal microstructure to uncover the subtle differences in brain form and function associated with ASD. Using diffusion MRI, microstructural analysis allows for the examination of small-scale variations in the brain's cellular architecture and can provide insights into the neural underpinnings of ASD. A recently developed microstructural analysis measures axonal conduction velocity, which is derived from parameters such as the g-ratio (the ratio of the inner to the outer diameter of the myelin sheath) and axon diameter[15,16]. Conduction velocity approximates the speed at which action potentials travel along axons, and deviations from the optimal speed can result in impaired neuronal communication.

1.3. *Genetic factors and the pseudo-autosomal region*
Genetic research has identified several candidate genes associated with ASD, many of which are in the pseudo-autosomal regions of the sex chromosomes[17–19]. These regions are of particular interest because they escape the usual X-inactivation process in females, resulting in a unique expression pattern that may contribute to the sex-wise disparity observed in ASD. Exploring these genetic factors, combined with microstructural data, can provide a more comprehensive understanding of the biological basis of ASD.

1.4. *The ACE Network and NDA*
The Autism Centers of Excellence (ACE) program is an initiative funded by the National Institute of Mental Health (NIMH) aimed at advancing the understanding, diagnosis, and treatments of ASD. Established to support large-scale multidisciplinary research projects, the ACE program brings together leading experts from various fields like genetics, neuroimaging, and phenotypic science to foster collaboration. Its structure allows for the integration of novel methodologies and state-of-the-art technologies to ensure that research efforts are at the forefront of scientific discovery. Complementary to the ACE program is the NIMH Data Archive (NDA), a comprehensive database managed by the NIMH that serves to centralize and disseminate the vast array of data collected on mental health research. Together, the ACE program and the NDA create a synergistic environment to nurture and advance the field of ASD research. The ACE program generates rich multimodal datasets that feed into the NDA. By leveraging the comprehensive data available through the NDA, researchers can explore new hypotheses, validate findings, and translate discoveries into clinical applications more effectively.

1.5. *Dimensionality in microstructural analysis*
A significant challenge in the analysis of neuronal microstructure data is the so-called “curse of dimensionality”. Microstructural processing pipelines typically generate data from over 200 distinct brain regions for each individual participant, which when performed on a voxel-wise level results in millions of datapoints for each individual. In our study, which includes 213 participants, this results in a vast multidimensional dataset. Although an N=213 might be considered respectable in human neuroimaging research, the sheer number of predictors poses a challenge for attaining sufficient statistical power, reproducibility, and interpretation. As an addendum to the concept of “big data,” we suggest that researchers consider highly dimensional datasets such as this one as “wide data” that is subject to a different set of equally important challenges.

Traditional approaches to address this issue involve feature selection to reduce the analytic search space. However, such techniques have inherent limitations. Firstly, they rely heavily on domain expertise, which may not always be available or infallible. Secondly, feature selection excludes certain predictors from the analysis before any machine learning algorithms can utilize them, thereby potentially limiting the scope of the analysis. While this approach can be beneficial when domain expertise is available, it can hinder exploratory analyses of new datasets.

1.6. *Multimodal data fusion in health sciences*
The integration of multimodal neuroimaging and genetic data presents a significant opportunity to improve model performance resulting from the synergy of shared and complementary information across modalities. For ASD research, the known genetic and neurological bases provide a strong foundation for exploring the rich multimodal data space afforded by large-scale data repositories like NDA provides for the ACE program. However, emphasizing interpretable methods is of paramount importance if research findings are to be translated into clinical application. It is through this framework this study has sought to provide insights into the multimodal data space generated by combining neuroimaging and genetic features.

1.7. *Novel approach: PCA and machine learning integration*
Our analysis aims to navigate the complex multidimensional space created by combining genetic and microstructural data modalities. To achieve this, we employ a novel implementation of principal component analysis (PCA) to identify unique characteristics of the dataset in an

unsupervised manner. PCA allows us to reduce the dimensionality of the dataset while retaining the within-class variation, thereby addressing the curse of dimensionality without relying on traditional feature selection methods, as well as retaining generalizability to unseen data.

Following the unsupervised feature selection through PCA, we integrate the results into a traditional classification machine learning framework. This approach enables us to leverage the strengths of both unsupervised and supervised learning techniques, providing a more robust analysis of the data. By doing so, we aim to uncover novel insights into the relationship between genetic factors, neuronal microstructure, and ASD.

The integration of advanced neuroimaging techniques and genetic data analysis holds great promise for unraveling the complex etiology of ASD. By addressing the challenges posed by the curse of dimensionality and leveraging advanced analytical methods, we can enhance our understanding of how the neuronal microstructure and genetic factors combine to form the autistic phenotype. This comprehensive approach not only advances our knowledge of ASD but also paves the way for the development of more effective diagnostic and therapeutic strategies tailored to the unique needs of individuals with ASD.

2. Methods

2.1. *Participants*

Participants included 213 (mean age=153.20 [in months], standard dev.=±35.22; age range=96–215; 99 female [46.48%]) volunteers from Wave 1 of an NIH-sponsored Autism Centers for Excellence network. The study sample included 113 autistic individuals (mean age=150.19, standard dev.=±34.56; age range=96–215; 51 female [45.13%]) and 100 non-autistic individuals (mean age=156.60, standard dev.=±35.81; age range=97–215; 48 female [48.00%]). The diagnostic and sex ratios were intended to be balanced. All ACE GENDAAR Wave 1 (9/04/2012-7/31/2022) neuroimaging, phenotypic, and genetic data were collected, processed, and archived on secure local compute servers under the following Internal Review Board (IRB) approvals: USC Approval #HS-13-00668; USC Approval #HS-18-00467; UVA Approval #22078; UVA IRB HSR #21361; GMU #00000169; and UVA #HSR-22-0423. As per the requirements of the US NIMH, de-identified and de-linked copies of all data were regularly submitted to the NDA as part of Collection #2021, where they are freely available for access to approved investigators. Data obtained by subsequent ACE GENDAAR Waves 2 and 3 (ongoing data collection) were not considered in this analysis. Informed consent was obtained from all participants and their legally authorized representatives.

2.2. *Genetic data preparation*

2.2.1. *Analysis of copy number variant densities*

Using Bioconductor R, a karyotype map was created to visualize mutation densities[20]. Statistical differences were assessed between groups to determine mutation loci present in exclusively in ASD females, and vice versa. Loci were systematically compared to the locations of known genes using the UCSC genome browser, along with their exonic sections and prior association with ASD[21]. Copy number variants (CNVs) were identified from a set (N=196) of Manta-annotated variant-call format (VCF) files. The New York Genome Institute preprocessed and designed these files. Manta is a structural variant (SV) calling tool from Chen et al. that utilizes discordant read-pair and split-read evidence to identify various CNVs, including insertions, deletions, translocations, inversions, and tandem duplications[22]. Manta-annotated VCF files for each subject were compared against a Homo sapiens (assembly GRCh38.p14) reference genome, which contains base-pair positions for transcripts, genes, exons, and introns for all 24 chromosomes, including sex-linked chromosomes X and Y.

2.2.2. *Analysis of differential expression and functional enrichment analytics*

Whole blood transcriptome sequencing was performed on 370 individuals. Transcript-level abundances were quantified using Kallisto[23]. Tximport was employed to aggregate these transcript-level abundances into gene-level counts[24]. Differential expression analysis was conducted using the R package DESeq2, facilitating the identification of statistically significant

changes in gene expression across ASD-diagnosed individuals, and were compared across neurotypical cohorts with sex and diagnosis were examined for interaction effects[25].

2.3. *Conduction velocity data preparation*

2.3.1. *Image acquisition*

Diffusion, T1-weighted, and T2-weighted images were acquired from each participant. Diffusion images were acquired with an isotropic voxel size of 2x2x2mm^3, 64 non-colinear gradient directions at b=1000 s/mm^2, and 1 b=0, TR=7300ms, TE=74ms. T1-weighted MPRAGE images with a FOV of 176x256x256 and an isotropic voxel size of 1x1x1mm^3, TE=3.3; T2-weighted images were acquired with a FOV of 128x128x34 with a voxel size of 1.5x1.5x4mm^3, TE=35. All images were preprocessed to correct for common sources of error and bias in accordance with prior published work[11,26]. T1w/T2w ratio was calculated by performing N4-bias correction, rescaling image intensity, then dividing on a voxel-wise basis[27,28]. Diffusion images were analyzed using a single-shell constrained spherical deconvolution (CSD) to obtain 3 tissue CSD (3T-CSD) microstructure compartments (intra- and extra-cellular isotropic signal, and intra-cellular anisotropic signal) and a fixel-based analysis was used to measure axonal fiber density and cross-section on a voxel-wise basis[11,26,29,30]. Despite obtaining multiple microstructure metrics using this methodology, only conduction velocity was examined here.

2.3.2. *Conduction velocity determination*

The aggregate g-ratio was calculated on a voxel-wise basis and was used as Mohammadi & Callaghan suggest; this is displayed in Equation 1[16,31–33]. As a measure of intra-axonal volume, the fiber density cross section was used as the intra-axonal volume fraction (AVF), and as a metric of myelin density, the T1w/T2w ratio was used as the myelin volume fraction (MVF)[34]. Both metrics represent the total sums of each respective compartment across the volume of the voxel and are a volume-based equivalent to the original formulation of g as the ratio of axon diameter (d) to fiber diameter (D).

$$g = \frac{d}{D} = \sqrt{1 - \frac{\mathrm{MVF}}{\mathrm{MVF+AVF}}} \tag{1}$$

Aggregate conduction velocity was calculated based on the calculations of Rushton and Berman et al.; reiterating Rushton's calculation that conduction velocity (θ) is proportional to the length of each fiber segment (l), and that this is roughly proportional to D, which in turn can be defined as the ratio between d and the g-ratio[15,35]. A value proportional to conduction velocity can be calculated using axon diameter and the g-ratio as in equation 2[35]:

$$\theta \propto l \propto Dg\sqrt{-\ln(g)} \propto d\sqrt{-\ln(g)} \tag{2}$$

All imaging metrics, 3T-CSD compartments, T1w/T2w ratio, aggregate g-ratio, and aggregate conduction velocity were averaged across each of 214 ROIs taken from the JHU-ICBM WM atlas (48 ROIs) and the Destrieux Cortical Atlas (164 ROIs)[27,28]. Additionally, two composite ROIs were included, one of all 48 JHU ROIs and one of 150 neocortical regions from the Destrieux Atlas.

2.4. *Initial analysis*

2.4.1. *Data preprocessing*

All conduction velocity and gene expression predictors were included in an initial traditional model for a total of 245 predictors. Participants were removed from the sample if missing either modality. The data was randomly split into training and testing sets, stratified by diagnostic cohort, at a 75-25 ratio. For feature preprocessing, all numeric predictors were normalized; the two

modalities do not occur on the same scale, so in this way we ensured equitable contributions from each in the analysis.

2.4.2. *Principal component analysis*

PCA was performed to reduce the dimensionality of the data. 40 principal components (PCs) were determined to be the maximum number of PCs examined: this number is equal to approximately 25% of the training data points (n=159), and 40 PCs account for approximately 85% of the cumulative explained variance.

2.4.3. *Logistic regression*

Logistic regression modeling for classification was employed to determine how well the PCs separate the two classes. Model complexity was managed by tuning the number of PCs. 10-fold cross-validation was employed to further validate the modeling procedure. The workflow examined a range between one and 40 PCs to identify the number of PCs that maximized the area under the receiver operating characteristic (AUROC) curve, a metric that balances true positive rate against false positive rate. The final model configuration was applied to the entire training data set with the optimal hyperparameters determined by the tuning process. The final model was deployed on the unseen testing dataset, evaluated using both AUROC and accuracy. The results of the training and testing sets for this analysis are displayed in Table 1.

2.5. *Experimental analysis*

2.5.1. *Data preprocessing*

For the second comparative analysis, the existing training data set was split by participant cohort such that all autistic participants comprised one data frame and all non-autistic participants comprised another data frame. All conduction velocity and gene expression predictors were included in each of these data frames (again a total of 245 predictors). All numeric predictors were normalized again for the same reasons outlined above.

2.5.2. *Principal component analyse*s

Separate PCAs were performed on each of the cohort data frames to reduce the dimensionality of the cohort-specific data by exploring the underlying structures. The number of PCs retained were determined independently for each group. First, the number of PCs that account for 70% of the cumulative variance was identified. Then, the number of PCs with a corresponding eigenvalue greater than or equal to one was identified. If these numbers were not equal, the number of retained PCs was decided to be the midpoint between them (rounded down). The results of this process are displayed in scree plots in Figure 1. Consequently, 17 PCs were retained for the autistic cohort and 14 PCs were retained for the non-autistic cohort.

2.5.3. *Feature selection*

Salient features for each group were extracted from the selected PCs systematically using the following procedure. First, the top 25% (75th percentile) of variable loadings (in terms of absolute value) were identified per selected component to focus on those that contributed most to the within-class variance. Then, instances of each of the predictors present in the top 25% were aggregated to identify the unique predictors among and across these PCs, defined as those only appearing once across all selected PCs. This resulted in seven predictors for the autistic group and 29 for the non-autistic group. Finally, four common predictors between the two classes were removed; the remaining 32 predictors were selected for modeling. A full accounting of these predictors is reported in Tables 2 and 3.

2.5.4. *Logistic regression*

Logistic regression modeling for classification was again employed to determine the effectiveness of this dimensionality reduction technique as compared to the traditional method. Predictors for this model included the 36 predictors selected from the procedure above. Model complexity was managed by tuning the number of PCs. 10-fold cross-validation was employed to further validate

the modeling procedure. The workflow examined a range between one and 36 PCs to identify the number of PCs that maximized the AUROC curve. The final model configuration was applied to the entire training data set with the optimal hyperparameters determined by the tuning process. The final model was deployed on the unseen testing dataset, evaluated using both AUROC and accuracy. The results of the training and testing sets for this analysis are displayed in Table 1. All machine learning analyses and plot visualizations were created using the R package TidyModels[36].

3. Results

3.1. *Genetic analysis*

3.1.1. *Sex-wise analysis of CNV densities in autistic participants*

CNVs were detected in 196 Manta-annotated VCF files from the New York Genome Institute. VCF files benchmarked against the reference genome were assessed for sex-wise differences in the pseudo-autosomal region using pairwise *t*-tests; the results were statistically significant (*T*-statistic = -7.21; $p < 0.001$).

3.1.2. *Differential expression analysis*

Differential expression analysis in DESeq2 showed that 3,707 genes exhibited significant differences when sex and diagnosis are considered as interacting factors. Differentially expressed genes showed statistical significance ($p < 0.01$ after false discovery correction) within or near the pseudo-autosomal boundary and the heterochromatic regions of the Y chromosome. Among these, the homologously encoded zinc finger transcription factors ZFX and ZFY emerged as highly significant genes. After adjustment, ZFX and ZFY showed exceptionally low *p*-values.

3.2. *Traditional analysis*

The results of the traditional modeling procedure were as follows. The 10-fold cross validation procedure for tuning the number of principal components showed that the best training AUROC was 0.693 at 12 PCs. The associated training accuracy was 61.267%. For the unseen testing dataset, the AUROC was 0.618, and the accuracy was 57.407%. These values are reported in Table 1; ROC curves are displayed in Figure 2.

3.3. *Experimental analysis*

3.3.1. *Scree plot description*

Scree plots were generated for each of the autistic and non-autistic cohort PCAs. Thresholds were determined based on the intersection of cumulative percent variance explained (greater than 70% were considered) as well as the principal components with eigenvalues greater than or equal to one; the average PC of these two metrics was used as the final threshold. These thresholds are shown in Figure 1; the former is indicated in smaller dashed red lines, the latter is indicated by longer dashed red lines, and the average of these two is also displayed as a solid red line. These

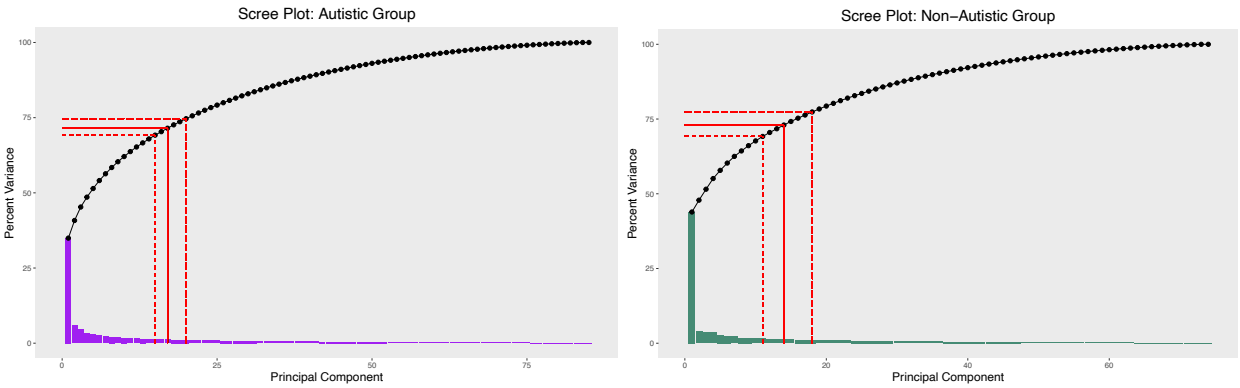


Fig. 1. Scree plots of the autistic and non-autistic cohort PCAs. Short-dashed lines indicate the number of PCs that account for 70% of the cumulative variance, long-dashed lines indicate the number of PCs with eigenvalues greater than or equal to one; the solid red lines indicate the average of these two values, rounded down.

values were as follows: greater than 70% cumulative variance was explained by 15 PCs in the autistic group and 11 PCs in the non-autistic group, eigenvalues greater than or equal to one included 20 PCs in the autistic group and 18 PCs in the non-autistic group, and the final threshold for the autistic group was 17 PCs and 14 PCs for the non-autistic group.

3.3.2. *Model evaluation*

Table 1 contains the logistic regression performance results from the two approaches. The training AUROC and accuracy values were comparable across both approaches, while the testing AUROC of 0.668 was greatly improved in the experimental approach, indicating more robust generalizability. Overall, the accuracy metrics were poor for both models, but an accuracy value of 59.259% for the experimental approach showed improvement over the traditional approach. Visualizations of the AUROC curves are available in Figure 2.

Table 1. Area under ROC curve and accuracy for the traditional model and experimental model.

	Training		Testing	
	AUROC	Accuracy	AUROC	Accuracy
Traditional	0.6931	61.2672%	0.6181	57.4074%
Experimental	0.6935	60.2892%	0.6676	59.2593%

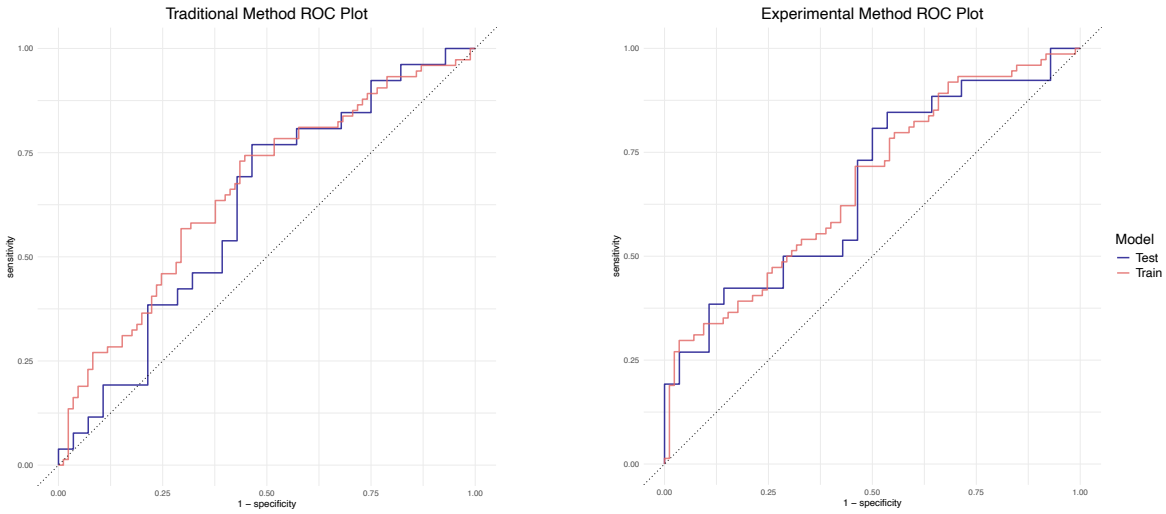


Fig. 2. ROC curves for the traditional logistic regression results (left) and experimental results (right).

3.3.3. *Feature selection*

Tables 2 and 3 display the features selected by the experimental procedure, ordered by PC number and then loading value. After removing the predictors that appeared in PCA procedure for both the autistic and non-autistic group, the experimental analysis contained 36 predictors. These features were mostly loaded onto the first principal component for each group (25/40; 62.5%). The value reported in the final column of these tables represents the loading value of a given predictor on the PC where higher absolute values represent a stronger relationship between predictor and PC. Directionality is also relevant here: positive values indicate a positive relationship between predictor and PC, whereas negative values indicate the opposite. These values only apply within the context of a given PC and should not be compared across PCs. The relevant cortical, subcortical, and white matter regions can be found highlighted in Figure 3.

Table 2. Top predictors from the autistic cohort PCA, sorted by component number, then loading value within each component.

Predictor	Region type	Hemisphere	Component	Value
Frontal superior gyrus	Gray matter	Right	PC1	0.0990
Lateral fissure (posterior part)	Gray matter	Right	PC1	0.0890
Lateral superior temporal gyrus	Gray matter	Right	PC2	0.1511
Frontal inferior sulcus	Gray matter	Right	PC9	-0.1263
Dorsal posterior cingulate gyrus	Gray matter	Left	PC9	-0.1068

Table 3. Top predictors from the non-autistic cohort PCA, sorted by component number, then loading value with each component.

Predictor	Region Type	Hemisphere	Component	Value
Superior corona radiata	White matter	Right	PC1	0.0898
Body of corpus callosum	White matter	-	PC1	0.0893
Posterior corona radiata	White matter	Right	PC1	0.0890
Anterior corona radiata	White matter	Left	PC1	0.0888
Posterior limb of internal capsule	White matter	Left	PC1	0.0886
Posterior thalamic radiation	White matter	Right	PC1	0.0881
Superior circular sulcus of the insula	Gray matter	Left	PC1	0.0877
Posterior corona radiata	White matter	Left	PC1	0.0875
Mid./posterior cingulate gyrus/sulcus	Gray matter	Right	PC1	0.0864
External capsule	White matter	Right	PC1	0.0859
Genu of corpus callosum	White matter	-	PC1	0.0850
Posterior thalamic radiation	White matter	Left	PC1	0.0847
Caudate	Subcortical	Left	PC1	0.0845
Sub-parietal sulcus	Gray matter	Left	PC1	0.0833
Precuneus gyrus	Gray matter	Left	PC1	0.0814
Superior temporal sulcus	Gray matter	Left	PC2	-0.0744
Superior temporal gyrus (transverse)	Gray matter	Left	PC5	-0.0755
Anterior circular sulcus of the insula	Gray matter	Right	PC5	0.0676
Inferior frontal sulcus	Gray matter	Left	PC7	0.1026
H-shaped orbital sulcus	Gray matter	Left	PC7	0.0806
Superior occipital gyrus	Gray matter	Right	PC8	0.0814
Hippocampus	Subcortical	Left	PC9	-0.1093
Superior temporal gyrus (transverse)	Gray matter	Right	PC10	0.1104
Tapetum	White matter	Left	PC11	-0.1104
Inferior parietal gyrus (supramarginal)	Gray matter	Left	PC12	-0.1337
Paracentral lobule gyrus and sulcus	Gray matter	Right	PC13	0.1560
Transverse temporal sulcus	Gray matter	Left	PC14	-0.1019

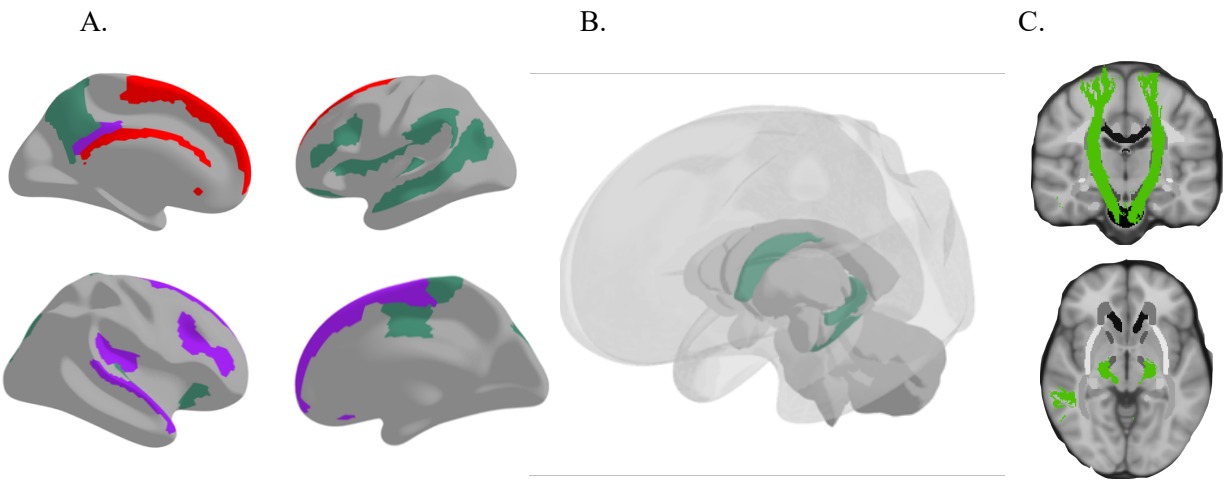


Fig. 3. (A) Cortical regions extracted from the PCA procedure. Top left: medial view of the left hemisphere; top right: lateral view of the left hemisphere; bottom left: lateral view of the right hemisphere; bottom right: medial view of right hemisphere. (B) Subcortical regions extracted from the PCA procedure. (C) White matter tracts extracted from the PCA procedure. Purple regions were found to be characteristic of the autistic group; green regions were found to be characteristic of the non-autistic group; red regions represent the overlapping regions between both the autistic and non-autistic group.

4. Discussion

The results of the experimental dimensionality reduction procedure are promising. In the context of classification for autistic vs. non-autistic individuals using neuroimaging and genetic features, the AUROC performance in this study is acceptable, especially using traditional machine learning frameworks (and not deep learning, which brings its own set of challenges)[37–40]. PCA is effective in this context because it better addresses the issue of overfitting, as evidenced by the improved testing AUROC metric. By capturing within-class variability, the modeling effort performs better on unseen testing data and generalizes more readily to other datasets. Despite failing to achieve performance that could provide actionable clinical insights and true inference of the underlying mechanisms, the feature selection methodology succeeded for multiple reasons.

First, the marked improvement in testing AUROC performance (over the traditional approach) demonstrates that the extracted features capture many of the relevant aspects that differentiate the classes. AUROC is better suited to many classification tasks, including this one, since it provides a balance between true positive rate and false positive rate, whereas accuracy is a simpler metric that measures the ratio of correct predictions to total predictions. AUROC is also the preferred metric for datasets with imbalanced classes; while the classes in this study are not exceptionally imbalanced, AUROC is equipped to handle even slight imbalances and, as such, is the preferred metric here. AUROC improvements in the experimental analysis demonstrate this methodology's internal validity and robustness to variations in unseen testing data.

Additionally, many of the extracted features represent notable regions of cortical, subcortical, and white matter connectivity in ASD research. ASD is characterized by abnormalities in brain structure, function, and connectivity, and many of the established areas of study are present in the extracted features[12,41,42]. The ability of the proposed procedure to pinpoint differences in relevant brain regions validates the methodology and necessitates further exploration both within and without the context of ASD research.

This analysis does not provide much evidence for the role of the pseudo-autosomal region on autism development, as none of the examined genetic predictors outperformed the microstructural predictors in terms of principal component loading. The low N of the sample is

likely a contributing factor to this phenomenon, though it is also possible that the pseudo-autosomal region is not nearly as contributory to the etiology of ASD as microstructural metrics. Indeed, when the two modalities were examined separately, the genetic data performed poorly as predictors for classification within the same framework.

4.1. *Feature selection*

4.1.1. *Cortical features*

Of the many cortical gray matter regions extracted by this methodology, two have been implicated in ASD research previously: the superior occipital gyrus and the frontal superior gyrus[43,44]. The frontal superior gyrus in particular is known to play a role in executive functioning, a domain previously identified as having deficits for autistic individuals relative to non-autistic individuals[45,46]. Other extracted cortical regions not directly implicated in ASD research do pertain to neurological processes relevant to areas previously identified as lacking in ASD individuals, including social cognition (anterior circular sulcus of the insula, inferior parietal gyrus), language processing (inferior frontal sulcus, superior temporal sulcus) and executive function (inferior frontal sulcus)[47–50]. It should be noted that certain cortical regions previously identified as differentially active in autistic and non-autistic individuals were not highlighted by this method, including the dorsal medial frontal cortex, anterior cingulate cortex, and orbitofrontal cortex[51–53].

4.1.2. *Subcortical features*

Subcortical features extracted using this method included the hippocampus and caudate nucleus. The hippocampus is known to be heavily involved in memory-related functions, and specific to ASD, both encoding and retrieval processes of episodic memory have been implicated as altered in ASD[54]. The caudate nucleus has been shown to have decreased connectivity in autistic individuals and is implicated in restricted and repetitive behavior development and increased autistic symptom severity as well[55–57].

4.1.3. *White matter features*

Many of the white matter features extracted in this study are also characteristic of the differences observed between autistic and non-autistic individuals. Corpus callosum tracts are most relevant here (body and genu of corpus callosum, superior/anterior/posterior corona radiata), but the tapetum has also been found to be under-connected in ASD relative to non-autistic individuals[58,59].

4.2. *Alternative approaches*

4.2.1. *PCA procedure on different data frames*

This experimental technique was deployed on this dataset in other ways to assess its effectiveness in different contexts. PCA was performed on each modality without first separating classes to attempt to capture modality-specific variability. Many of the extracted microstructure predictors remained the same as the focus of this study; however, this method also incorporated several genetic predictors as well. The resulting logistic regression yielded poor classification performance, likely due to an inability to extract the most salient features for each class.

Further, separate PCAs were performed on the four groups defined by the two different modalities and the two classes (autistic genetic, autistic microstructure, non-autistic genetic, non-autistic microstructure). Again, the microstructure metrics were comparable to those extracted in the main analysis of the study, and again this method allowed for more genetic predictors to contribute to the machine learning framework. This methodology performed even worse than before, however. The results of both attempts further cements the conclusion that the pseudo-autosomal region does not contribute to differences between autistic and non-autistic participants in this study and it is possible the genetic basis of ASD may lie elsewhere on the genome.

4.2.2. *Other machine learning models*

Two other types of machine learning models were employed for the classification part of this analysis: random forest (RF) and quadratic discriminant analysis (QDA). These models are appropriate for data that is not expected to display a linear decision boundary and as such are more

flexible. Logistic regression does expect the data to be linearly separable, and while that may appear to be a significant limitation of the modeling efforts of this study, RF and QDA performed far worse than logistic regression in both the traditional and experimental dimensionality reduction frameworks. One explanation for this could be that the data is not complete enough to allow for flexible models to generalize well. Microstructure and genetics are only two pieces of a larger puzzle that can include many other modalities like functional imaging, EEG, and behavioral data. Relatedly, while the extracted features comprise the major group differences in this dataset, they only capture part of the global within-group variability and therefore further limit the generalizability of the results; a phenomenon exacerbated by flexible machine learning methods.

4.3. *Future directions*

In the pursuit of assessing putative neurogenetic markers of ASD through the integration of neuroimaging, genomic, and phenotypic data, built upon the approach described here, several critical future directions emerge. One primary consideration is the utilization of data imputation to increase the sample size. While genetic data imputation may not be valid due to the potential introduction of biases and inaccuracies, it can be more appropriately applied to other metrics such as conduction velocity, pending further exploration and validation of the technique in this context.

In terms of machine learning applications, while classification remains a viable approach, regression-based predictive modeling presents an avenue with the potential for more nuanced and informative results. Incorporating behavioral phenotyping outcome surveys, including measures of language, executive function, and social interaction, could provide rich data for these models, enhancing their predictive power and relevance.

An interesting observation from the experimental model is the failure of the gene expression features to contribute significantly following the selection procedure. When modalities were analyzed independently absent the experimental procedure, the resulting classification performance was suboptimal compared to traditional methods. This issue was further compounded when PCA was applied separately to four classes based on diagnostic groups and modalities (e.g., gene expression-autistic, gene expression-non-autistic, etc.). This suggests that the variance captured through the main PCA feature selection approach is sufficient for robust case classification, outperforming more granular feature selection strategies. Some recent studies have attempted to balance modality-specific contributions; these procedures tend to utilize regularization and differential weighting to achieve modality balance and could provide a more nuanced representation of the influence of each modality[60,61].

The feature selection approach could be applicable in individual nuances in autistic individuals; the initial provenance of salient features provides a starting point from which individual similarities and differences can be assessed. Additionally, sex-specific disparities in ASD are another critical area that warrants further examination and could be addressed by an exacting feature selection approach. Conducting separate PCAs for different sexes within the autistic group may reveal unique and actionable insights, potentially improving the performance of downstream machine learning models.

Moreover, several advanced analytical methods offer promising future directions, in particular deep learning**.** Employing deep learning techniques for data fusion to integrate multimodal data could capture complex relationships between neuroimaging, genomic, and phenotypic data. This is an emerging area with promising results but no unified optimal strategy as of yet[62,63].

In summary, future research in the integration of neuroimaging, genomic, and phenotypic data in ASD will need to explore advanced data imputation techniques, leverage regression-based predictive modeling, and consider sex-specific analyses. Employing deep learning, sophisticated weighting and thresholding strategies, and advanced dimensionality reduction methods could significantly enhance the understanding and predictive power of these complex datasets.

4.4. *Conclusions*

The results of the experimental dimensionality reduction procedure for classifying autistic versus non-autistic individuals using neuroimaging and genetic features are promising. The AUROC performance achieved in this study is acceptable, especially within traditional machine learning

frameworks. PCA effectively addresses overfitting, as indicated by the improved testing AUROC metric. By capturing within-class variability, the model performs better on unseen testing data and generalizes more readily to other datasets.

Firstly, the marked improvement in testing AUROC performance over traditional approaches indicates that the extracted features capture many relevant aspects differentiating the classes. AUROC is a balanced metric that accounts for both true positive and false positive rates, making it particularly suitable for datasets with even slight class imbalances. The improvements in AUROC demonstrate the methodology's internal validity and robustness to variations in unseen testing data.

Many of the extracted features represent notable regions of cortical, subcortical, and white matter connectivity, which are well-documented in ASD research. Interestingly, the analysis did not provide substantial evidence for the role of the pseudo-autosomal region in autism development. None of the examined genetic predictors outperformed the microstructural predictors in terms of principal component loading. This may be due to the low sample size, but it also raises the possibility that the pseudo-autosomal region is not as contributory to the etiology of ASD as microstructural metrics. When examined separately, genetic data performed poorly as predictors for classification within the same framework, further supporting this conclusion.

Cortical features extracted from the analysis highlight critical regions involved in ASD, such as areas related to social cognition, language processing, and executive function. These regions are consistent with the existing literature on ASD, reinforcing their importance in understanding the disorder's neurobiological underpinnings. Likewise, subcortical features identified include regions involved in emotion regulation, reward processing, and motor functions. Abnormalities in these areas are frequently reported in ASD studies, underscoring their relevance to the disorder's phenotype and supporting the validity of the feature selection process. Finally, white matter features point to connectivity issues between different brain regions, which are a hallmark of ASD. Disruptions in white matter integrity can affect communication between cortical and subcortical regions, contributing to the diverse symptomatology of ASD.

Applying PCA to each modality without separating classes aimed to capture modality-specific variability. While some microstructure predictors remained consistent, this approach also included several genetic predictors. However, the resulting logistic regression yielded poorer than expected classification performance, likely due to an inability to extract the most salient features for each class. Separate PCAs for the four groups (autistic genetic, autistic microstructure, non-autistic genetic, non-autistic microstructure) also performed poorly, reaffirming that the pseudo-autosomal region may not significantly contribute to ASD classification.

Exploring more flexible machine learning methods, such as quadratic discriminant analysis and tree-based models, did not improve performance over logistic regression. This suggests that the proposed feature selection method is most effective with less flexible machine learning models, highlighting the need for careful selection of analytical techniques based on the data and research goals. The identification of critical cortical, subcortical, and white matter features aligns with existing ASD research, reinforcing their relevance in understanding the disorder's neurobiological underpinnings. While the role of genetic predictors remains less clear, these findings highlight the need for meticulous selection of analytical techniques tailored to the specific characteristics of the data. Such comprehensive and data-driven strategies are vital for understanding the nuances of ASD and advancing for the field toward more effective and personalized diagnostics and interventions.

References

1. Werling DM, Geschwind DH. Sex differences in autism spectrum disorders: *Curr Opin Neurol*. 2013;26(2):146-153. doi:10.1097/WCO.0b013e32835ee548

2. Elsabbagh M, Johnson MH. Infancy and autism: progress, prospects, and challenges. *Prog Brain Res*. 2007;164:355-383. doi:10.1016/S0079-6123(07)64020-5

3. Zeidan J, Fombonne E, Scorah J, et al. Global prevalence of autism: A systematic review update. *Autism Res*. 2022;15(5):778-790. doi:10.1002/aur.2696

4. Maenner MJ, Warren Z, Williams AR, et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *MMWR Surveill Summ*. 2023;72(2):1-14. doi:10.15585/mmwr.ss7202a1

5. Masi A, DeMayo MM, Glozier N, Guastella AJ. An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options. *Neurosci Bull*. 2017;33(2):183-193. doi:10.1007/s12264-017-0100-y

6. Ramaswami G, Geschwind DH. Genetics of autism spectrum disorder. In: *Handbook of Clinical Neurology*. Vol 147. Elsevier; 2018:321-329. doi:10.1016/B978-0-444-63233-3.00021-X

7. Choi L, An JY. Genetic architecture of autism spectrum disorder: Lessons from large-scale genomic studies. *Neurosci Biobehav Rev*. 2021;128:244-257. doi:10.1016/j.neubiorev.2021.06.028

8. Woodbury-Smith M, Scherer SW. Progress in the genetics of autism spectrum disorder. *Dev Med Child Neurol*. 2018;60(5):445-451. doi:10.1111/dmcn.13717

9. Bagasra O, Heggen C, Hossain MI. *Autism and Environmental Factors*. 1st ed. Wiley; 2018. doi:10.1002/9781119042280

10. Landrigan PJ. What causes autism? Exploring the environmental contribution. *Curr Opin Pediatr*. 2010;22(2):219-225. doi:10.1097/MOP.0b013e328336eb9a

11. Newman BT, Jacokes Z, Venkadesh S, et al. Conduction velocity, G-ratio, and extracellular water as microstructural characteristics of autism spectrum disorder. Bray S, ed. *PLOS ONE*. 2024;19(4):e0301964. doi:10.1371/journal.pone.0301964

12. Ilioska I, Oldehinkel M, Llera A, et al. Connectome-wide Mega-analysis Reveals Robust Patterns of Atypical Functional Connectivity in Autism. *Biol Psychiatry*. 2023;94(1):29-39. doi:10.1016/j.biopsych.2022.12.018

13. Yoon N, Huh Y, Lee H, et al. Alterations in Social Brain Network Topology at Rest in Children With Autism Spectrum Disorder. *Psychiatry Investig*. 2022;19(12):1055-1068. doi:10.30773/pi.2022.0174

14. Gata-Garcia A, Porat A, Brimberg L, Volpe BT, Huerta PT, Diamond B. Contributions of Sex Chromosomes and Gonadal Hormones to the Male Bias in a Maternal Antibody-Induced Model of Autism Spectrum Disorder. *Front Neurol*. 2021;12:721108. doi:10.3389/fneur.2021.721108

15. Rushton W a. H. A theory of the effects of fibre size in medullated nerve. *J Physiol*. 1951;115(1):101-122. doi:10.1113/jphysiol.1951.sp004655

16. Mohammadi S, Callaghan MF. Towards in vivo g-ratio mapping using MRI: Unifying myelin and diffusion imaging. *J Neurosci Methods*. 2021;348:108990. doi:10.1016/j.jneumeth.2020.108990

17. Maxeiner S, Benseler F, Krasteva-Christ G, Brose N, Südhof TC. Evolution of the Autism-Associated Neuroligin-4 Gene Reveals Broad Erosion of Pseudoautosomal Regions in Rodents. Nowick K, ed. *Mol Biol Evol*. 2020;37(5):1243-1258. doi:10.1093/molbev/msaa014

18. Mclellan A, Wynne F, Ball M, Moore T. *Sexual Antagonism and Autism Susceptibility in the Xq/Yq Pseudoautosomal Region (PAR2).*; 2007.

19. Wang S, Wang B, Drury V, et al. Rare X-linked variants carry predominantly male risk in autism, Tourette syndrome, and ADHD. *Nat Commun*. 2023;14(1):8077. doi:10.1038/s41467-023-43776-0

20. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80. doi:10.1186/gb-2004-5-10-r80

21. Kent WJ, Sugnet CW, Furey TS, et al. The Human Genome Browser at UCSC. *Genome Res*. 2002;12(6):996-1006. doi:10.1101/gr.229102

22. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-1222. doi:10.1093/bioinformatics/btv710

23. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519

24. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*. 2016;4:1521. doi:10.12688/f1000research.7563.2

25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

26. Newman BT, Dhollander T, Reynier KA, Panzer MB, Druzgal TJ. Test-retest reliability and long-term stability of three-tissue constrained spherical deconvolution methods for analyzing diffusion MRI data. *Magn Reson Med*. 2020;84(4):2161-2173. doi:10.1002/mrm.28242

27. Mori S, Crain BJ. *MRI Atlas of Human White Matter*. Elsevier; 2006.

28. Destrieux C, Fischl B, Dale A, Halgren E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*. 2010;53(1):1-15. doi:10.1016/j.neuroimage.2010.06.010

29. Raffelt DA, Tournier JD, Smith RE, et al. Investigating white matter fibre density and morphology using fixel-based analysis. *NeuroImage*. 2017;144(Pt A):58-73. doi:10.1016/j.neuroimage.2016.09.029

30. Newman BT, Patrie JT, Druzgal TJ. An intracellular isotropic diffusion signal is positively associated with pubertal development in white matter. *Dev Cogn Neurosci*. 2023;63:101301. doi:10.1016/j.dcn.2023.101301

31. Campbell JSW, Leppert IR, Narayanan S, et al. Promise and pitfalls of g-ratio estimation with MRI. *NeuroImage*. 2018;182:80-96. doi:10.1016/j.neuroimage.2017.08.038

32. Stikov N, Perry LM, Mezer A, et al. Bound pool fractions complement diffusion measures to describe white matter micro and macrostructure. *NeuroImage*. 2011;54(2):1112-1121. doi:10.1016/j.neuroimage.2010.08.068

33. Stikov N, Campbell JSW, Stroh T, et al. In vivo histology of the myelin g-ratio with magnetic resonance imaging. *NeuroImage*. 2015;118:397-405. doi:10.1016/j.neuroimage.2015.05.023

34. Raffelt D, Tournier JD, Rose S, et al. Apparent Fibre Density: a novel measure for the analysis of diffusion-weighted magnetic resonance images. *NeuroImage*. 2012;59(4):3976-3994. doi:10.1016/j.neuroimage.2011.10.045

35. Berman S, Filo S, Mezer AA. Modeling conduction delays in the corpus callosum using MRI-measured g-ratio. *NeuroImage*. 2019;195:128-139. doi:10.1016/j.neuroimage.2019.03.025

36. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. Published online 2020. https://www.tidymodels.org

37. Reiter MA, Jahedi A, Fredo ARJ, Fishman I, Bailey B, Müller RA. Performance of machine learning classification models of autism using resting-state fMRI is contingent on sample heterogeneity. *Neural Comput Appl*. 2021;33(8):3299-3310. doi:10.1007/s00521-020-05193-y

38. Mellema CJ, Nguyen KP, Treacher A, Montillo A. Reproducible neuroimaging features for diagnosis of autism spectrum disorder with machine learning. *Sci Rep*. 2022;12(1):3057. doi:10.1038/s41598-022-06459-2

39. Santana CP, De Carvalho EA, Rodrigues ID, Bastos GS, De Souza AD, De Brito LL. rs-fMRI and machine learning for ASD diagnosis: a systematic review and meta-analysis. *Sci Rep*. 2022;12(1):6030. doi:10.1038/s41598-022-09821-6

40. Yassin W, Nakatani H, Zhu Y, et al. Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis. *Transl Psychiatry*. 2020;10(1):278. doi:10.1038/s41398-020-00965-5

41. Walsh MJM, Wallace GL, Gallegos SM, Braden BB. Brain-based sex differences in autism spectrum disorder across the lifespan: A systematic review of structural MRI, fMRI, and DTI findings. *NeuroImage Clin*. 2021;31:102719. doi:10.1016/j.nicl.2021.102719

42. Khundrakpam BS, Lewis JD, Kostopoulos P, Carbonell F, Evans AC. Cortical Thickness Abnormalities in Autism Spectrum Disorders Through Late Childhood, Adolescence, and Adulthood: A Large-Scale MRI Study. *Cereb Cortex N Y N 1991*. 2017;27(3):1721-1731. doi:10.1093/cercor/bhx038

43. Arunachalam Chandran V, Pliatsikas C, Neufeld J, et al. Brain structural correlates of autistic traits across the diagnostic divide: A grey matter and white matter microstructure study. *NeuroImage Clin*. 2021;32:102897. doi:10.1016/j.nicl.2021.102897

44. Zhao X, Zhu S, Cao Y, et al. Abnormalities of Gray Matter Volume and Its Correlation with Clinical Symptoms in Adolescents with High-Functioning Autism Spectrum Disorder. *Neuropsychiatr Dis Treat*. 2022;Volume 18:717-730. doi:10.2147/NDT.S349247

45. Ball G, Stokes PR, Rhodes RA, et al. Executive Functions and Prefrontal Cortex: A Matter of Persistence? *Front Syst Neurosci*. 2011;5. doi:10.3389/fnsys.2011.00003

46. Jacokes Z, Jack A, Sullivan CAW, et al. Linear discriminant analysis of phenotypic data for classifying autism spectrum disorder by diagnosis and sex. *Front Neurosci*. 2022;16:1040085. doi:10.3389/fnins.2022.1040085

47. Ruland SH, Palomero-Gallagher N, Hoffstaedter F, Eickhoff SB, Mohlberg H, Amunts K. The inferior frontal sulcus: Cortical segregation, molecular architecture and function. *Cortex*. 2022;153:235-256. doi:10.1016/j.cortex.2022.03.019

48. Wymbs NF, Nebel MB, Ewen JB, Mostofsky SH. Altered Inferior Parietal Functional Connectivity is Correlated with Praxis and Social Skill Performance in Children with Autism Spectrum Disorder. *Cereb Cortex*. 2021;31(5):2639-2652. doi:10.1093/cercor/bhaa380

49. Kortz M, Lillehei K. Insular Cortex. In: *StatPearls [Internet]*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK570606/

50. Beauchamp MS. The social mysteries of the superior temporal sulcus. *Trends Cogn Sci*. 2015;19(9):489-490. doi:10.1016/j.tics.2015.07.002

51. Zoltowski AR, Lyu I, Failla M, et al. Cortical Morphology in Autism: Findings from a Cortical Shape-Adaptive Approach to Local Gyrification Indexing. *Cereb Cortex*. 2021;31(11):5188-5205. doi:10.1093/cercor/bhab151

52. Moradi E, Khundrakpam B, Lewis JD, Evans AC, Tohka J. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. *NeuroImage*. 2017;144:128-141. doi:10.1016/j.neuroimage.2016.09.049

53. Di Martino A, Ross K, Uddin LQ, Sklar AB, Castellanos FX, Milham MP. Functional Brain Correlates of Social and Nonsocial Processes in Autism Spectrum Disorders: An Activation Likelihood Estimation Meta-Analysis. *Biol Psychiatry*. 2009;65(1):63-74. doi:10.1016/j.biopsych.2008.09.022

54. Banker SM, Gu X, Schiller D, Foss-Feig JH. Hippocampal contributions to social and cognitive deficits in autism spectrum disorder. *Trends Neurosci*. 2021;44(10):793-807. doi:10.1016/j.tins.2021.08.005

55. Turner KC, Frost L, Linsenbardt D, McIlroy JR, Müller RA. Atypically diffuse functional connectivity between caudate nuclei and cerebral cortex in autism. *Behav Brain Funct*. 2006;2(1):34. doi:10.1186/1744-9081-2-34

56. Qiu T, Chang C, Li Y, et al. Two years changes in the development of caudate nucleus are involved in restricted repetitive behaviors in 2–5-year-old children with autism spectrum disorder. *Dev Cogn Neurosci*. 2016;19:137-143. doi:10.1016/j.dcn.2016.02.010

57. O'Dwyer L, Tanner C, Van Dongen EV, et al. Decreased Left Caudate Volume Is Associated with Increased Severity of Autistic-Like Symptoms in a Cohort of ADHD Patients and Their Unaffected Siblings. Hu VW, ed. *PLOS ONE*. 2016;11(11):e0165620. doi:10.1371/journal.pone.0165620

58. Payabvash S, Palacios EM, Owen JP, et al. White Matter Connectome Edge Density in Children with Autism Spectrum Disorders: Potential Imaging Biomarkers Using Machine-Learning Models. *Brain Connect*. 2019;9(2):209-220. doi:10.1089/brain.2018.0658

59. Gibbard CR, Ren J, Seunarine KK, Clayden JD, Skuse DH, Clark CA. White matter microstructure correlates with autism trait severity in a combined clinical–control sample of high-functioning adults. *NeuroImage Clin*. 2013;3:106-114. doi:10.1016/j.nicl.2013.07.007

60. Sheng J, Xin Y, Zhang Q, Wang L, Yang Z, Yin J. Predictive classification of Alzheimer's disease using brain imaging and genetic data. *Sci Rep*. 2022;12(1):2405. doi:10.1038/s41598-022-06444-9

61. Bi X an, Hu X, Wu H, Wang Y. Multimodal Data Analysis of Alzheimer's Disease Based on Clustering Evolutionary Random Forest. *IEEE J Biomed Health Inform*. 2020;24(10):2973-2983. doi:10.1109/JBHI.2020.2973324

62. Calhoun VD, Sui J. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2016;1(3):230-244. doi:10.1016/j.bpsc.2015.12.005

63. Kalamkar S, A. GM. Multimodal image fusion: A systematic review. *Decis Anal J*. 2023;9:100327. doi:10.1016/j.dajour.2023.100327

Uncovering Important Diagnostic Features for Alzheimer's, Parkinson's and Other Dementias Using Interpretable Association Mining Methods

Kazi Noshin[1*], Mary Regina Boland[3*], Bojian Hou[4], Victoria Lu[1],
Carol Manning[2], Li Shen[4†], Aidong Zhang[1†]

[1]*Department of Computer Science,* [2]*Department of Neurology*
University of Virginia, VA 22903, USA
[3]*Data Science Program, Department of Mathematics,*
Saint Vincent College, Latrobe, PA 15650, USA
[4]*Department of Biostatistics, Epidemiology and Informatics,*
University of Pennsylvania, Philadelphia, PA 19104, USA
E-mail: epw9kz@virginia.edu, mary.boland@stvincent.edu
Bojian.Hou@Pennmedicine.upenn.edu, gbp7sb@virginia.edu
CM4R@uvahealth.org, li.shen@pennmedicine.upenn.edu, aidong@virginia.edu

Alzheimer's Disease and Related Dementias (ADRD) afflict almost 7 million people in the USA alone. The majority of research in ADRD is conducted using post-mortem samples of brain tissue or carefully recruited clinical trial patients. While these resources are excellent, they suffer from lack of sex/gender, and racial/ethnic inclusiveness. Electronic Health Records (EHR) data has the potential to bridge this gap by including real-world ADRD patients treated during routine clinical care. In this study, we utilize EHR data from a cohort of 70,420 ADRD patients diagnosed and treated at Penn Medicine. Our goal is to uncover important risk features leading to three types of Neuro-Degenerative Disorders (NDD), including Alzheimer's Disease (AD), Parkinson's Disease (PD) and Other Dementias (OD). We employ a variety of Machine Learning (ML) Methods, including uni-variate and multivariate ML approaches and compare accuracies across the ML methods. We also investigate the types of features identified by each method, the overlapping features and the unique features to highlight important advantages and disadvantages of each approach specific for certain NDD types. Our study is important for those interested in studying ADRD and NDD in EHRs as it highlights the strengths and limitations of popular approaches employed in the ML community. We found that the uni-variate approach was able to uncover features that were important and rare for specific types of NDD (AD, PD, OD), which is important from a clinical perspective. Features that were found across all methods represent features that are the most robust.

Keywords: Electronic Health Records; Machine Learning; Alzheimers Disease and Related Dementias; Data Mining

*Equal-Contribution First-Authors.
†Equal-Contribution Senior-Authors.

1. Introduction

1.1 Alzheimer's Disease and Related Dementias

ADRD afflicts an estimated 6.9 million people in the United States of America (USA), using current July 2024 statistics.[1] ADRD and dementia collectively kill more patients per year than breast and prostate cancers combined.[1] However, despite its frequency of incidence, not much is known about ADRD patients in community-based settings. This is because the majority of Alzheimer's Disease (AD) research focuses on post-mortem (after the patient has died) samples or patients recruited through expensive clinical trials (that often lack racial/ethnic diversity). In addition, there remains a paucity of research among diverse populations, including investigating sex-disparities[2] and racial disparities[3] in outcomes. Additionally, many state-of-the-art studies on ADRD have limited generalizability because of the almost exclusive use of trials that lack race/ethnicity/socioeconomic inclusiveness,[4] leading to a diversity dearth.[5]

1.2 Electronic Health Records (EHRs)

The recent development and implementation of EHRs now provide a tremendous opportunity to evaluate ADRD patients from community-based settings that includes in-patient and outpatient medical records data obtained through routine clinical care. EHR data contain information on millions of patients from both in-patient and out-patient settings. They often contain more representative patient populations (in terms of race, ethnicity, and socioeconomic inclusiveness) than clinical trials due to their community-based settings. Several studies have used EHR data for AD research. Xu et al.[6] developed a data-driven method to uncover four subphenotypes of AD from EHRs. Their subphenotypes were correlated with common comorbidities of ADRD, including mental health diseases and cardiovascular disease.[6] None of these prior studies (as far as we are able to glean from the reported papers) have incorporated socioeconomic or racial/ethnic disparities into their algorithm development. This is important as not properly capturing these features can lead to biased research results.[7,8]

1.3 Uni-variate Association Mining

While Xu et al.[6] utilized unsupervised Machine Learning (ML) methods to learn types of ADRD (a form of neurodegenerative disorder (NDD)) from the data itself, another common method for uncovering important features or characteristics of a dataset is to utilize association mining. Association mining is used extensively in EHR research through a process called Phenotype-Wide Association Studies (PheWAS) first introduced in 2010 by Denny et al.[9] In their study they held the genetic variant constant while looping over a wide range of clinical EHR-derived phenotypes.[9] This process was then employed by BioBanks throughout the USA and abroad, but also applied to EHR datasets not linked to BioBank data.[9–23] Others used EHR data without genetic information to perform association mining or PheWAS style studies.[18] Boland et al. employed a similar algorithmic approach when exploring the relationship between birth seasonality and later risk of disease through a method first published in 2015[19] and later replicated in several studies.[20–23] The essence of association mining is to test for an association (using some statistical method, e.g., chi-square test, fisher's exact test, or regression) between each phenotype (typically represented as columns in a matrix) and the outcome of interest. In this study, our outcomes are three different NDD types. Therefore, the outcome is set (in this work either AD, PD or OD) *a priori* and then each phenotype

(i.e., covariate/feature/column) is tested for association with that outcome of interest. If one wants to investigate more than one outcome (in this case our different NDD types) then one simply repeats the entire process over again with each outcome. We construct our algorithm such that outcome Y is always the same (in this case a binary indicator variable for whether or not the patient has/had a particular NDD type, e.g., PD). We then have our intercept term (β_0) and our term related to that particular feature that is being tested (or iterated over) is X and the coefficient term related to the feature is represented as β_x. We will loop over all potential features and therefore with each iteration the actual feature in X and the corresponding coefficient β_x will change. A sample regression equation for a binary outcome of interest (NDD type: Parkinson's Disease) is as follows:

$$Y_{(NDD\ type:\ Parkinson'sDisease)} = \beta_0 + \beta_x * X,$$

with β_x indicating the term for each phenotype (or feature) that will be iterated over. Therefore, in our example the first feature would be some clinical or demographic feature, followed by the second feature until all features have been iterated over. Typically, there are a large number of associations explored (often into the thousands) requiring multiple hypothesis correction methods to adjust for multiple comparisons.

1.4 Multivariate Association Mining: SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) is a method to explain individual predictions based on Shapley values from cooperative game theory.[24] It assigns each feature an importance value for a particular prediction,[25] aiming to fairly distribute the 'payout' (prediction) among features. SHAP provides both local explanations for individual predictions and global interpretation methods, linking optimal credit allocation with local explanations using Shapley values.[26]

In the context of EHR data, SHAP can be a powerful tool for interpreting the predictions made by the models. This study uses PD, AD, and OD as separate outcomes. Each patient in the dataset can be considered as an instance for which a prediction is made. The features are the 'players' in the game. The 'payout' is the prediction of whether a patient has PD, AD, or OD. For example, if a model predicts a certain patient has a high risk of developing PD, SHAP can help us understand how each feature contributes to this prediction. This can provide valuable insights into which factors are most influential in predicting PD, AD, or OD.

The Shapley value is the average of all the marginal contributions to all possible coalitions.[24] For a set N of n features, the Shapley value $\phi_i(v)$ of feature i is:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] \tag{1}$$

where S is a subset of features not including feature i, v is a value function that represents the model's prediction for a subset of features, $v(S)$ is the prediction for subset S, and $|S|$ is the number of features in S.

The SHAP explanation method computes Shapley values. Let g be is the explanation model, $z' \in \{0,1\}^M$ the coalition vector, M the maximum coalition size, and $\phi_j \in \mathbb{R}$ the feature

attribution, i.e., the Shapley values for feature j. SHAP is defined mathematically as follows:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2}$$

In Equation 2, an entry of 1 in the coalition vector indicates that the corresponding feature value is 'present', whereas an entry of 0 signifies that it is 'absent'. Within the framework of SHAP, the Shapley values help us understand each feature's contribution to the prediction.

2. Dataset

2.1 Dataset Description

We obtained de-identified EHR data from Penn Medicine for patients with ADRD using a set of diagnosis codes. The age range of our medical records indicate that the majority of the EHR data was collected between 2002 and 2022 with some diagnosis dates occurring earlier (all the way back to the 1920s indicating manually entered diagnosis information that was pertinent for specific patients). The internal Clinical Data Warehouse at Penn Medicine converted the International Classification of Diseases (ICD) version 9 (ICD-9) codes to version 10 (ICD-10). We have cross-mapped our list of ADRD diagnostic codes using existing resources[27] to provide researchers with our full list of ICD-9 and ICD-10 diagnosis codes for ADRD identification.[28] The EHR data comes in the Observational Health Data Sciences and Informatics (OHDSI) Common Data Model (CDM) format with relevant data broken down into several files corresponding to tables in a SQL database. The dataset contains information on patients' encounters, diagnoses, medications, procedures, vitals, laboratory findings, chemotherapy, and laboratory values. This study was approved by the University of Pennsylvania's Institutional Review Board (IRB) with approval id: 851588. We mapped our entire dataset consisting of 70,420 ADRD patients to their corresponding PheCodes. This allowed us to identify 14,911 patients with AD diagnoses specifically (PheCode:290.11), 16,216 patients with PD diagnoses specifically (PheCode:332) and 14,911 patients with 'Dementias' (PheCode:290.10) called in this paper Other Dementias (OD), which is an unspecified generic dementia category. Demographics are provided in Table 1 and visualized in Figure 2. The Venn diagram in Figure 1 represents the overlap of patients diagnosed with PD, AD, and OD.

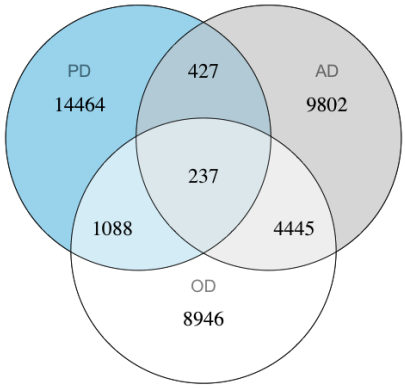


Fig. 1. Venn Diagram of Patients Diagnosed with PD, AD and OD.

Demographic factors differs among the three NDD subtypes. Figure 2 shows the Racial and Sex distributions across the NDD types. The bars represent percentages of different racial groups for four categories: AD, PD, OD, and Overall. White individuals have a higher percentage of PD, whereas Black or African Americans have a higher percentage of OD diagnoses (Figure 2). Figure 2 shows that females have a higher percentage of NDD types that include AD and OD compared to males. On the other hand, males have a higher percentage of PD

compared to females. Overall, across NDD types, there was a higher proportion of females with ADRD diagnoses than males (54.43% vs. 45.56%).

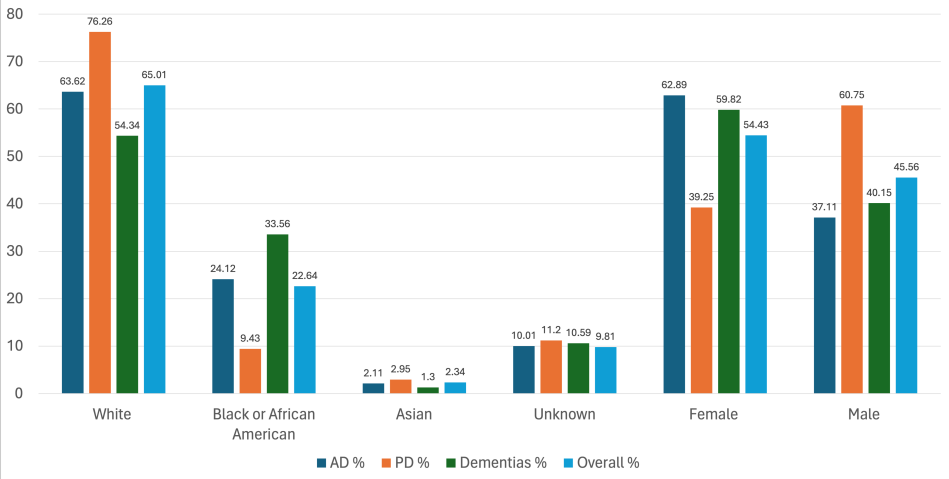


Fig. 2. Racial and Sex Distribution by NDD type.

Table 1. Demographics of ADRD Patients by NDD type.

Attribute	Value	AD(%) (N = 14911)	PD(%) (N = 16216)	OD(%) (N = 14911)	Overall(%) (N = 70420)
Race	White	63.62	76.26	54.34	65.01
	BAA	24.12	9.43	33.56	22.64
	Asian	2.11	2.95	1.3	2.34
	NHOPI	0.1	0.09	0.16	0.12
	AIAN	0.05	0.07	0.04	0.08
	Unknown	10.01	11.2	10.59	9.81
Gender	Female	62.89	39.25	59.82	54.43
	Male	37.11	60.75	40.15	45.56
	Missing	0.01	0	0.03	0.01

AD: Alzheimer's Disease, PD: Parkinson's Disease, OD: Other Dementias, BAA: Black or African American, NHOPI: Native Hawaiian or Other Pacific Islande, AIAN: American Indian or Alaska Native.

2.2 Dataset Preprocessing

Our raw data consisted of diagnosis code information in both ICD version 9 (ICD-9) and version 10 (ICD-10). We mapped these codes to their respective PheCodes.[29] These PheCodes were used for each terminology system (ICD-9 and ICD-10), aligning on the 'code_system' and 'code' fields. This also allowed us to collapse results to the PheCode level rather than using individual ICD-9 and ICD-10 codes. To enable quantitative analysis, we used one-hot encoding of these categorical data to transform those data into binary format with one column per unique PheCode. For each unique phenotype (PheCode) identified, we created a new column in the diagnosis data and assigned binary values indicating the presence (1) or absence (0) of the phenotype (PheCode) for each patient. The final dataset comprised of patient identifiers,

demographic information, and binary-encoded phenotypes, providing a structured and analyzable representation of the patient diagnosis data. We also transformed the Race variable using one-hot encoding with Race_White, Race_Black, Race_Asian and Race_Other with each corresponding to a binary relationship with the race variable. We also transformed the Hispanic and Sex_Male columns to binary variables. We also included features pertaining to the type of hospital-visit, including: chemotherapy, emergency_visit, inpatient_visit, ambulatory_visit, and other_unknown_visit. Like the demographic features, each of these was binary indicating that a patient had at least one occurrence of that particular type of visit or chemotherapy.

For the multi-variate analysis no missing data was allowed, and therefore the missing data for Hispanic and Sex_Male were coded with -1 to indicate that those values were missing. We decided not to use imputation methods because that could result in other biases. For the univariate analysis, this was not needed as each feature was assessed one at a time and therefore, if there was missing data for a feature then those rows would be dropped automatically from the analysis via the glm() function in R.

3. Methodology

3.1 Uni-variate Logistic Regression Association Mining

We utilize traditional EHR association mining methods.[18] To do this, we evaluate each NDD type as an outcome separately to compare the features association with that particular type of NDD. This allows us to identify features that are strongly associated with a particular NDD type, and also features that are only associated with one NDD type. For each outcome (AD, PD, OD), we test each feature for its association with the outcome. Each feature (N=1796) was tested for association with each outcome (hence uni-variate association mining). The majority of features were conditions/diseases represented by PheCodes. The non-PheCode features included demographic features: Race_White, Race_Black, Race_Asian, Race_Other, Hispanic and Sex_Male. Hospital-visit characteristic features were also explored including: chemotherapy, emergency visit, ambulatory visit, inpatient visit, and other unknown visit. Each of these was binary indicating that a patient had at least one occurrence of that particular type of visit or chemotherapy. Once all features were tested for association with each NDD type, we then removed the intercept terms from our model results and corrected for multiple hypothesis testing using the Bonferroni adjustment method, defined as:

$$\text{corrected p-value} = \alpha/N = 0.05/1796$$

where alpha represents our significance cutoff (0.05 in this case) and N represents the number of tests (1796 in this case).

We used Logistic Regression (LR) to test for the association between each feature and the NDD type, given that the outcome variables are binary. This analysis was performed in the statistical programming language R using the glm() function with the statistical family set to binomial (i.e., to perform LR). Importantly, while we tested 1796 features for association with each NDD type, in the Venn Diagrams we only show 1794 features because we removed the features that consist of the NDD types themselves (AD, PD, and OD).

3.2 Machine Learning (ML) Methods

We employed three distinct models: LR, Ridge Regression (RR), and a Residual Network

(ResNet) based Neural Network, to predict the occurrence of PD, AD and OD separately. The dataset, after preprocessing, consisted of a feature set of 1796 features per NDD type model. The data was split into training and testing sets in an 80:20 ratio. We used 'LogisticRegression' and 'Ridge' from Python package 'sklearn'.[30] We implemented a ResNet model using Keras, starting with an input layer for feature vectors, followed by a dense layer with 64 units, batch normalization, ReLU activation, and a dropout layer (rate 0.5) to prevent over-fitting. The model's core has five ResNet blocks, each comprising two dense layers with batch normalization, ReLU activation, and dropout (rate 0.5). The output of the second dense layer was added to the block's input tensor, followed by ReLU activation. The final output was generated by a dense layer with a single unit and sigmoid activation. We used the Adam optimizer (learning rate 0.001), binary cross-entropy loss, and accuracy as the metric. Early stopping with a patience of 5 epochs was employed to mitigate over-fitting. The model was trained for up to 50 epochs with a batch size of 32, using 20% of the training data for validation.

We performed 5-fold cross-validation on the training set for all the above-mentioned models. The models were then trained on the entire training dataset. A bootstrapping procedure generated multiple bootstrap samples from the test data, evaluated the model's accuracy on each sample, and used those accuracies to compute the 95% confidence intervals.

3.3 Analysis with SHAP

We aim to identify factors contributing to the progression of AD, PD, and OD using the SHAP method. To do this, we construct separate models for each NDD type, using patient attributes as predictors. The target variable was defined as the presence or absence of AD. Similarly, separate models were constructed for PD and OD. The value of the target variable is 1 if the targeted event happened to the subject during the whole project and 0 otherwise. For each outcome (AD, PD, OD), we train the models mentioned in subsection 3.2 to determine the presence or absence of the disease. The SHAP method from the 'shap'[25] Python package was used to identify significant features using LR and RR. We used LinearExplainer for both LR and RR models. [a]

3.4 Feature Selection and Top 5% Subset

For methods that used multi-variate approaches, we selected features as being important if the mean shapley value for that feature was greater than or equal to the overall mean shapley value for that NDD type.[31] For the uni-variate approach, we selected features as important if their Bonferroni adjusted p-value was statistically significant. For the 5% subset, we selected the top 5% of features from each method and each NDD type. The top 5% of features amounts to 90 features from our overall feature set. The features are ranked based on their mean shapley value if a multi-variate method, or their p-value if a uni-variate method.

4. Results

4.1 Uni-Variate Assoication Mining Results

We found 340 significant associations with AD, 590 significant associations with PD and

[a] For the ResNet-based Neural Network, we encountered significant computational constraints using KernelExplainer due to its inherent complexity and the large size of the dataset. Consequently, to maintain consistency in our analysis of feature importance, we proceeded without considering the contributions derived from the neural network model.

583 significant associations with OD using uni-variate LR. Table 2 reports findings based on nominal significance, Bonferroni adjusted significance and a combination of Bonferroni significance and Odds Ratio ≥ 2. We visualized the uni-variate LR results using Manhattan plots for each NDD type: AD (in Figure 3A), PD (in Figure 3B) and OD (in Figure 3C). One can see that there are many Bonferroni significant results spread across the NDD types, but that AD has only a few significant results (see Figure 3A).

Table 2. Number of Association Mining Results

Results	AD	PD	OD
Number of Nominal Significant Results	723	1017	971
Number of Bonferroni Significant Results	340	590	583
Number of Bonferroni Significant Results and OR ≥ 2	3	16	278

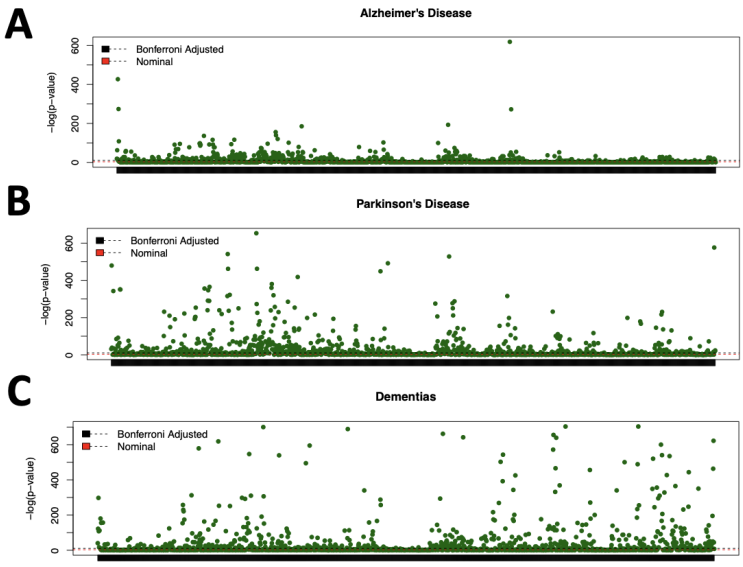


Fig. 3. Manhattan Plot for NDD type: AD, PD, OD

4.2 Performance of Multi-variate Methods for: AD, PD, OD

Table 3. Accuracy Performance of ML Methods by NDD type (All Features, N=1796).

Out-come	Logistic Regression (LR)				Ridge Regression (RR)				Neural Net (ResNet)			
	cv (mean)	Train	Test	CI(95%) [lower, upper]	cv (mean)	Train	Test	CI(95%) [lower, upper]	cv (mean)	Train	Test	CI(95%) [lower, upper]
AD	78.48	80.1	79.44	[78.74, 80.08]	78.8	79.85	79.43	[78.73, 80.07]	79.03	79.87	79.26	[78.59, 79.89]
PD	85.07	86.51	85.17	[84.58, 85.77]	83.34	84.29	83.21	[82.58, 83.83]	84.83	88.49	84.89	[84.32, 85.5]
OD	84.91	86.37	84.5	[83.91, 85.11]	84.37	85.5	84.15	[83.56, 84.75]	86.49	88.51	86.34	[85.76, 86.92]

Table 4. Accuracy Performance of ML Methods by NDD type (Intersection Features).

Outcome	Logistic Regression (LR)				Ridge Regression (RR)				Neural Net (ResNet)			
	cv (mean)	Train	Test	CI(95%) [lower, upper]	cv (mean)	Train	Test	CI(95%) [lower, upper]	cv (mean)	Train	Test	CI (95%) [lower, upper]
AD (N=180)	78.99	79.2	79.54	[78.88, 80.18]	78.99	79.03	79.47	[78.76, 80.13]	79.95	80.25	80.32	[79.67, 80.94]
PD (N=225)	85.54	85.76	85.05	[84.46, 85.63]	83.28	83.47	83.2	[82.56, 83.84]	86.15	87.87	86.16	[85.57, 86.71]
OD (N=205)	85.06	85.29	84.8	[84.24, 85.41]	84.45	84.56	84.29	[83.73, 84.88]	88.22	89.95	87.98	[87.44, 88.51]

Tables 3 and 4 present the performance metrics of patients with NDD types of PD, AD, and OD, assessed using different ML models: LR, RR, and ResNet. For each NDD type, both of the tables display cross-validation mean accuracy (cv mean), the training and testing accuracy percentages alongside the 95% confidence intervals (CIs) for both the lower and upper bounds with respect to testing accuracy. The test accuracies are obtained using a held-out independent test set. The main difference between the two tables is the number of features used for training the models. In Table 3, all features were used, while in Table 4, only selected intersectional subsets of features mentioned in Section 5 were used. In both tables, we additionally included PD and AD as features with OD as the outcome, PD and OD as features with AD as the outcome, and AD and OD as features with PD as the outcome while evaluating the models' performances.

Results for all 1796 features shown in Table 3. LR, RR and ResNet models show slight variations in the mean cv, training, testing accuracies and CIs. In Table 4, the models show comparable results, highlighting the contribution of the reduced feature sets of 180 features for AD, 225 for PD, and 205 for OD. Tables 3 and 4 demonstrate that the use of selected features, as opposed to all features, does not significantly degrade model performance. Specifically, the slight differences in test accuracy, e.g., 79.44% vs. 79.54% for AD classification using LR, indicate that the models maintain robust performance even with reduced feature sets.

4.3 Overlap of Features Across Methods per NDD type

Characteristics of important features are given in Table 5. Non-overlapping features represent those that are unique to one method. We show the results for the 5% subset and the entire set of important features. AD had the lowest amount of non-overlapping features at 30.7% indicating that many of the features found by methods when applied to AD were similar Table 5. However, both PD and OD had higher amounts of non-overlapping features (i.e., unique) with 48.3% and 51.3% respectively in Table 5. Depending on the particular use case, some researchers may want to use only the top important features, which is our rationale for the top 5% feature subset from each method. This results in the same number of features being selected per method (i.e., 90 features). We found that for the 2 NDD types with less overlap (i.e., PD and OD) there were fewer non-overlapping features in the top 5% subset with 35.8% vs. 48.3% for PD, and 41.8% vs. 51.3% for OD. However, for AD, which already had a high agreement across methods, the top 5% of features actually had more non-overlapping features with 38.6% vs. 30.7% in the 5% subset in Table 5.

Table 5. Characteristics of Important Features

Results	AD	PD	OD
All Important Features			
Total Number of Important Features Across Methods	567	681	708
Number of Non-Overlapping Features	174	329	363
Percentage of Non-Overlapping Important Features	30.7%	48.3%	51.3%
Top 5% Feature Subset			
Total Number of Important Features Across Methods	140	137	146
Number of Non-Overlapping Features	54	49	61
Percentage of Non-Overlapping Important Features	38.6%	35.8%	41.8%

Feature penetrance indicates how often a feature was determined to be important by one of the three methods used for each NDD type. We also calculated penetrance across all NDD types, therefore a feature could have a maximum of 9 to indicate that it was found across all 3 NDD types and methods.[28] However, in some situations differences across methods maybe important. The Venn diagrams show the results for all important features and for only the top 5% of features for AD Figure 4A, for PD Figure 4B and for OD Figure 4C. Results for the intersections appear similar across the NDD types with OD having a larger number of features overall, mainly resulting from the large number of OD results generated by the uni-variate LR approach. Figure 4C.

Many interesting unique features were identified using the uni-variate LR method, including the association between Creutzfeldt-Jakob Disease or (CJD) and OD with a large reported Odds Ratio (OR=51.87, 95% CI: 18.88, 214.32). Note that CJD is listed in the PheCodes as Jakob Creutzfeldt Disease (PheCode:324.1). The percentage of individuals with CJD and OD was 93.18% versus 4.55% with AD and 2.27% with PD (Figure 5).

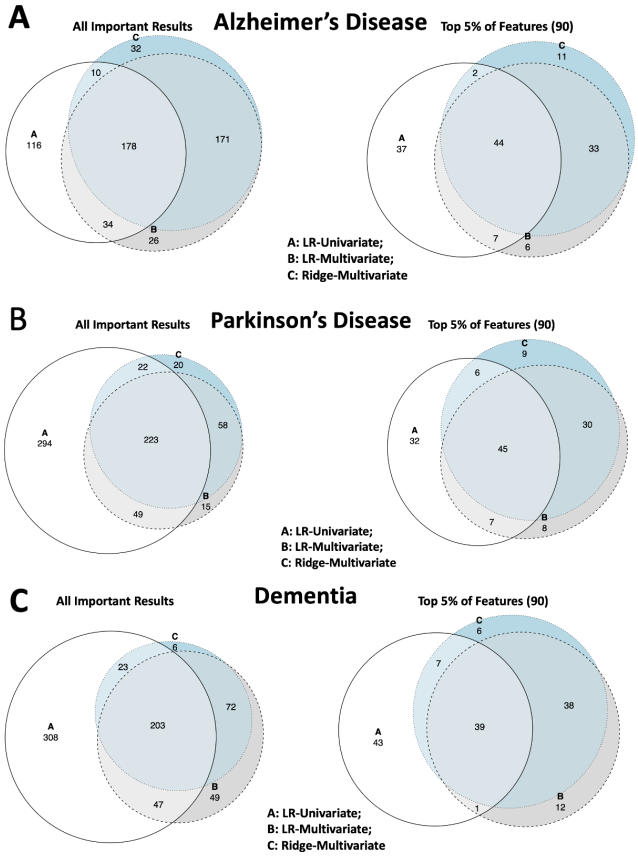


Fig. 4. Venn Diagram of Important Features for NDD type: AD, PD, and OD.

5. Discussion

5.1 Overview of Study

Overall, our study found that it is possible to identify important features for different NDD types, specifically AD, PD and OD. We found that the performance obtained using the specific method (LR, RR or ResNet) in terms of accuracy varied somewhat by NDD type, with all achieving similar performance. We also found that while methods achieved similar performance overall, there were substantial differences in 'important' features revealed by each

method. We identified features that were common (i.e., found by each method) and also features that were unique to one particular method. Therefore, our findings are important for others using EHR data for ADRD analyses because the 'important' features identified not only varies by statistical method used, but also by NDD type. Because of the heterogeneity of EHR data, the exact prevalence of each NDD type may vary by site to site, making this finding of importance for those utilizing EHR data for ADRD analyses.

The features found in the intersection of all 3 methods, namely uni-variate LR, multivariate LR (identified using shapley values), and multi-variate RR (identified via shapley values) may represent the most significant and robust features. These features are of particular interest because they are important across multiple methods, suggesting they are less likely to be influenced by confounding factors. In contrast, features identified by only one method may be less reliable and could be artifacts of the specific analytical approach used. Therefore, focusing on the intersecting features provides a more comprehensive and reliable understanding of the key predictors in the dataset. However, we will describe below circumstances that illustrate the advantages and disadvantages of various methods and features identified by the methods indicating that the intersection features may include only a subset of the truly 'important' features.

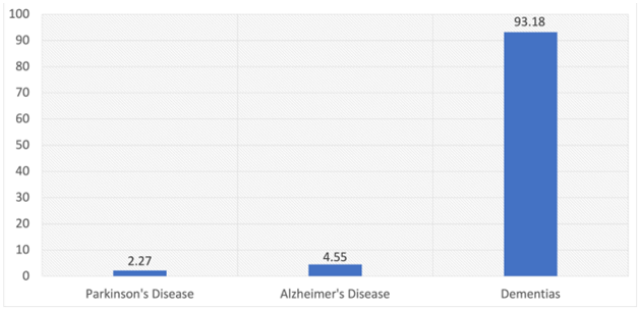


Fig. 5. Distribution of Creutzfeldt-Jakob disease (CJD) by NDD types.

5.2 Uni-variate versus Multivariate Models

CJD Disease Identified via Uni-variate Method Alone. There are some findings that were only uncovered via the uni-variate LR approach. It was the only method that revealed that CJD was significant in OD (one of the NDD types), and clearly there is a dramatic difference in our dataset for the prevalence observed among those with CJD with the majority of individuals having OD (see Figure 5). CJD is established as a rare cause of dementia[32] and therefore, this finding is of clinical significance and would be missed in multivariate approaches due to the overall rarity of this disease. However, there have been studies that found that CJD could be mistaken for AD,[33] indicating that clinically distinguishing these various diseases can be challenging in different circumstances. CJD is an example of one of the 308 features identified for OD that were only identified using the uni-variate LR approach (see Figure 4C). There were 6 OD features uniquely identified via the Multi-variate RR approach, but these features were odd, and included 'late pregnancy and failed induction' along with 'genital prolapse', which indicates that perhaps these findings were associated with a lower chance of OD. However, our population only includes those who are 65 and older and therefore, these features existing in our cohort remains somewhat odd. Features unique to the multi-variate LR approach also appeared somewhat unusual, including 'elevated Prostate Specific Antigen'(PSA test). This is an unusual finding given that our OD patients were predominantly female.

Sleep Apnea. Interestingly, the PheCode for Sleep Apnea (PheCode:327.30) was found

to be significant for AD across all 3 methods, including both multi-variate and uni-variate approaches. However, the uni-variate LR approach also identified another related PheCode for Obstructive Sleep Apnea (PheCode:327.32) as being significantly important for the AD type. The 'Obstructive Sleep Apnea' or OSA PheCode was **not** identified as being important by the other multi-variate approaches, and indicates a finding unique to the uni-variate method for the AD type, one of the 116 unique features in Figure 4A. This also represents a clinical relevant finding as OSA has been linked with AD specifically in a number of studies[34,35] indicating its importance in the AD type.

Overall Advantages of Uni-variate Alone. Overall, these findings highlight a main advantage of uni-variate LR (sometimes referred to as a 'traditional approach' for ML) in that it enables one to calculate Odds Ratios (OR) and to determine whether a finding increases or decreases the risk of diagnosis for each NDD type. Shapley values on the other hand provide the importance of the feature without the directionality of the finding, which in some cases makes them more difficult to interpret, and might be the reason for some of these results identified as unique to the multi-variate approaches. Overall, our findings suggest that uni-variate LR may be better at detecting NDD-type-specific differences, even with smaller sample sizes like we observed with CJD. Features supported across the methods appear to be more robust than features identified by just one method - unless that method was a uni-variate approach (again due to the advantages of ORs and directionality of the result).

5.3 Performance Varies by NDD type: AD, PD, and OD

Based on the models' performance presented in Tables 3 and 4, it is evident that the performance of different ML methods varies depending on both the NDD type and the method used. When using LR with all features, the highest accuracy was achieved when detecting PD with a test accuracy of 85.17%, while the lowest accuracy was for AD with a test accuracy of 79.44%. In contrast, when using RR and ResNet, the highest accuracy was for OD with a test accuracy of 84.15% and 86.34%, but the lowest accuracy was again for AD with a test accuracy of 79.43% and 79.26%. The performances using the subset of the features also demonstrate similar pattern. RR had the lowest test accuracy for all 3 NDD types across all features (in table 3) and the intersectional subsets of features (in table 4). While ResNet demonstrated the highest test accuracy using all features in classifying OD, LR had the highest test accuracy for AD and PD. On the other hand, in Table 4 the ResNet model significantly outperformed the regression models' test accuracies for all NDD types. The ResNet's superior performance implies the possibility that this increased performance is due to the neural network's capacity to learn complex, non-linear relationships, which might be more important for certain NDD types.

5.4 Comparison of Feature Results Across Methods

Interestingly, when applying Neural Net with selected features for prediction purposes, there was an increase in accuracy across all NDD types compared to using all features as presented in Tables 3 and 4. Both AD and OD predictions yield higher test accuracies with intersection features rather than the full feature set with all the 3 models. Although PD results demonstrate a slight decrease in test accuracies for both LR and RR while using instersectional features, the train accuracies also decrease. These observations suggest that

our feature selection enhances model performance by reducing noise and focusing on the most relevant information. The use of intersection features, which encapsulate the most critical and discriminative attributes, facilitates better generalization across models, reducing overfitting and improving robustness.

5.5 Implications of Our Findings on Other ADRD ML Studies

Spectrum bias[36,37] occurs when a test is studied among a population that is not representative of the intended target population. For example, if a study is conducted on an ADRD population using EHR data in Florida with a large population of ADRD patients having OD and then that method was applied in a population from Delaware where the majority of ADRD patients have PD, that could result in spectrum bias. Therefore, it is important to understand the important disease features that are unique to each NDD type: AD, PD and OD because the case-mix distribution of patients among ADRD patients may vary across the USA. Therefore, to develop robust ML models, we must understand the relationships between these features and each NDD type to understand if models (ours and others) will validate adequately at other locations in the USA treating ADRD patients.

6. Conclusion

In conclusion, we utilized a large (70,420 patients) ADRD cohort derived from EHR data collected during routine clinical care. Our cohort is an order of magnitude larger in size (70k versus 7k) than another recent ML ADRD study using EHR data.[6] Using this large and comprehensive dataset, we aimed to identify important diagnostic features for the NDD types using a variety of ML methods. Our study demonstrates the strengths and weakness of univariate and multivariate ML methods in detecting features specific to certain NDD types, namely, AD, PD and OD. We report accuracies of these methods and report what NDD types where each method worked best. We also identified features that were found across all methods, and features that were unique to a particular method. We share these findings with the research community with the goal of mitigating spectrum bias in ADRD studies as the NDD types vary from site to site across the USA and could therefore introduce biases if not accounted for.

7. Acknowledgments and Appendices

Research reported in this publication/ presentation was supported by the National Institute On Aging of the National Institutes of Health under Award Number P30AG073105. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. ALZ.org, Alzheimer's disease facts and figures, *Accessed in July 2024* **https://www.alz.org/alzheimers-dementia/facts-figures** (2024).
2. A. S. Tang, T. Oskotsky, S. Havaldar, W. G. Mantyh, M. Bicak, C. W. Solsberg, S. Woldemariam, B. Zeng, Z. Hu, B. Oskotsky *et al.*, Deep phenotyping of alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations, *Nature communications* **13** (2022).
3. G. M. Babulal, Y. T. Quiroz, B. C. Albensi, E. Arenaza-Urquijo, A. J. Astell, C. Babiloni, A. Bahar-Fuchs, J. Bell, G. L. Bowman, A. M. Brickman *et al.*, Perspectives on ethnic and racial disparities in alzheimer's disease and related dementias: update and areas of immediate need, *Alzheimer's & Dementia* **15**, 292 (2019).
4. A. K. Manrai, B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron, P. Szolovits, D. M. Margulies, J. Loscalzo and I. S. Kohane, Genetic misdiagnoses and the potential for health disparities, *New England Journal of Medicine* **375**, 655 (2016).
5. A. L. Chin, S. Negash and R. Hamilton, Diversity and disparity in dementia: the impact of ethnoracial differences in alzheimer disease, *Alzheimer Disease & Associated Disorders* **25** (2011).
6. J. Xu, F. Wang, Z. Xu, P. Adekkanattu, P. Brandt, G. Jiang, R. C. Kiefer, Y. Luo, C. Mao, J. A. Pacheco *et al.*, Data-driven discovery of probable alzheimer's disease and related dementia subphenotypes using electronic health records, *Learning Health Systems* **4**, p. e10246 (2020).
7. D. S. Char, N. H. Shah and D. Magnus, Implementing machine learning in health care—addressing ethical challenges, *New England Journal of Medicine* **378**, 981 (2018).
8. I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman and M. Ghassemi, Ethical machine learning in healthcare, *Annual review of biomedical data science* **4**, 123 (2021).
9. J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden and D. C. Crawford, Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations, *Bioinformatics* **26**, 1205 (2010).
10. M. Saad, A. El-Menyar, K. Kunji, E. Ullah, J. Al Suwaidi and I. J. Kullo, Validation of polygenic risk scores for coronary heart disease in a middle eastern cohort using whole genome sequencing, *Circulation: Genomic and Precision Medicine* **15**, p. e003712 (2022).
11. A. Verma, A. Lucas, S. S. Verma, Y. Zhang, N. Josyula, A. Khan, D. N. Hartzel, D. R. Lavage, J. Leader, M. D. Ritchie *et al.*, Phewas and beyond: the landscape of associations with medical diagnoses and clinical measures across 38,662 individuals from geisinger, *The American Journal of Human Genetics* **102**, 592 (2018).
12. X. Li, X. Meng, A. Spiliopoulou, M. Timofeeva, W.-Q. Wei, A. Gifford, X. Shen, Y. He, T. Varley, P. McKeigue *et al.*, Mr-phewas: exploring the causal effect of sua level on multiple disease outcomes by using genetic instruments in uk biobank, *Annals of the rheumatic diseases* **77** (2018).
13. J. Pathak, R. C. Kiefer, S. J. Bielinski and C. G. Chute, Applying semantic web technologies for phenome-wide scan using an electronic health record linked biobank, *Journal of biomedical semantics* **3**, 1 (2012).
14. R. W. Read, K. A. Schlauch, G. Elhanan, W. J. Metcalf, A. D. Slonim, R. Aweti, R. Borkowski and J. J. Grzymski, Gwas and phewas of red blood cell components in a northern nevadan cohort, *PLoS One* **14**, p. e0218078 (2019).
15. B. Namjou, K. Marsolo, R. J. Caroll, J. C. Denny, M. D. Ritchie, S. S. Verma, T. Lingren, A. Porollo, B. L. Cobb, C. Perry *et al.*, Phenome-wide association study (phewas) in emr-linked pediatric cohorts, genetically links plcl1 to speech language development and il5-il13 to eosinophilic esophagitis, *Frontiers in genetics* **5**, p. 401 (2014).
16. X. Chang, M. March, F. Mentch, H. Qu, Y. Liu, J. Glessner, P. Sleiman and H. Hakonarson, Genetic architecture of asthma in african american patients, *Journal of Allergy and Clinical*

Immunology **151**, 1132 (2023).
17. Y.-C. A. Feng, I. B. Stanaway, J. J. Connolly, J. C. Denny, Y. Luo, C. Weng, W.-Q. Wei, S. T. Weiss, E. W. Karlson and J. W. Smoller, Psychiatric manifestations of rare variation in medically actionable genes: a phewas approach, *BMC genomics* **23**, p. 385 (2022).
18. M. R. Boland, S. Alur-Gupta, L. Levine, P. Gabriel and G. Gonzalez-Hernandez, Disease associations depend on visit type: results from a visit-wide association study, *BioData Mining* **12**, 1 (2019).
19. M. R. Boland, Z. Shahn, D. Madigan, G. Hripcsak and N. P. Tatonetti, Birth month affects lifetime disease risk: a phenome-wide method, *Journal of the American Medical Informatics Association* **22**, 1042 (2015).
20. L. Li, M. R. Boland, R. Miotto, N. P. Tatonetti and J. T. Dudley, Replicating cardiovascular condition-birth month associations, *Scientific reports* **6**, p. 33166 (2016).
21. M. R. Boland, M. Fieder, L. H. John, P. R. Rijnbeek and S. Huber, Female reproductive performance and maternal birth month: a comprehensive meta-analysis exploring multiple seasonal mechanisms, *Scientific Reports* **10**, p. 555 (2020).
22. M. R. Boland, P. Parhi, L. Li, R. Miotto, R. Carroll, U. Iqbal, P.-A. Nguyen, M. Schuemie, S. C. You, D. Smith *et al.*, Uncovering exposures responsible for birth season–disease effects: a global study, *Journal of the American Medical Informatics Association* **25**, 275 (2018).
23. M. R. Boland, M. S. Kraus, E. Dziuk and A. R. Gelzer, Cardiovascular disease risk varies by birth month in canines, *Scientific Reports* **8**, 1 (2018).
24. C. Molnar, *Interpretable machine learning : a guide for making Black Box Models interpretable* (Lulu, 2019).
25. S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, *Advances in neural information processing systems* **30** (2017).
26. L. S. Shapley and A. E. Roth, *The Shapley value : essays in honor of Lloyd S. Shapley* (Cambridge University Press, 1988).
27. ICDCodes, Icd10 to icd9 code coverter, *Accessed in July 2024* **https://icd.codes/convert/icd10-to-icd9-cm** (2024).
28. M. R. Boland, Boland lab github: Alzheimer's disease and related dementias (adrd) project, *Accessed in July 2024* **https://github.com/bolandlab/AlzheimersDiseaseandRelatedDementias** (2024).
29. PheWAS, Phewas - phenome wide association studies.
30. L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt and G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013.
31. L. Bloch, C. M. Friedrich and A. D. N. Initiative, Data analysis with shapley values for automatic subject selection in alzheimer's disease data sets using interpretable machine learning, *Alzheimer's Research & Therapy* **13**, 1 (2021).
32. R. Knight, Creutzfeldt-jakob disease: a rare cause of dementia in elderly persons, *Clinical infectious diseases* **43**, 340 (2006).
33. H. J. Tschampa, M. Neumann, I. Zerr, K. Henkel, A. Schröter, W. J. Schulz-Schaeffer, B. Steinhoff, H. A. Kretzschmar and S. Poser, Patients with alzheimer's disease and dementia with lewy bodies mistaken for creutzfeldt-jakob disease, *Journal of Neurology, Neurosurgery & Psychiatry* **71**, 33 (2001).
34. F. Emamian, H. Khazaie, M. Tahmasian, G. D. Leschziner, M. J. Morrell, G.-Y. R. Hsiung, I. Rosenzweig and A. A. Sepehry, The association between obstructive sleep apnea and alzheimer's disease: a meta-analysis perspective, *Frontiers in aging neuroscience* **8**, p. 78 (2016).
35. A. G. Andrade, O. M. Bubu, A. W. Varga and R. S. Osorio, The relationship between obstructive

sleep apnea and alzheimer's disease, *Journal of Alzheimer's Disease* **64**, S255 (2018).
36. C. Goehring, A. Perrier and A. Morabia, Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance, *Statistics in medicine* **23**, 125 (2004).
37. S. A. Mulherin and W. C. Miller, Spectrum bias or spectrum effect? subgroup variation in diagnostic test evaluation, *Annals of internal medicine* **137**, 598 (2002).

Exploring the Granularity of the Illnesses-Related Changes in Regional Homogeneity in Major Depressive Disorder using the UKBB Data

Yewen Huang, Syed Ibrar Hussain, Demetrio Labate, Robert Azencott

Department of Mathematics, University of Houston, Houston, TX 77203, USA
Email: yhuang55@central.uh.edu, shussa44@central.uh.edu, dlabate@uh.edu, razencot@central.uh.edu

Paul Thompson

Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA 90007, USA
Email: pthomp@usc.edu

Bhim Adhikari, Peter Kochunov

Louis A. Faillace, MD, Department of Psychiatry and Behavioral Sciences, University of Texas Health Science Center at Houston, Houston, TX 77054, USA
Email: Bhim.M.Adhikari@uth.tmc.edu, Peter.Kochunov@uth.tmc.edu

Illness related brain effects of neuropsychiatric disorders are not regionally uniform, with some regions showing large pathological effects while others are relatively spared. Presently, Big Data meta-analytic studies tabulate these effects using structural and/or functional brain atlases that are based on the anatomical boundaries, landmarks and connectivity patterns in healthy brains. These patterns are then translated to individual level predictors using approaches such as Regional Vulnerability Index (RVI), which quantifies the agreement between individual brain patterns and the canonical pattern found in the illness. However, the atlases from healthy brains are unlikely to align with deficit pattern expressed in specific disorders such as Major Depressive Disorder (MDD), thus reducing the statistical power for individualized predictions. Here, we evaluated a novel approach, where disorder specific templates are constructed using the Kullback-Leibler (KL) distance to balance granularity, signal-to-noise ratio and the contrast between regional effect sizes to maximize translatability of the population-wide illness pattern at the level of the individual. We used regional homogeneity (ReHo) maps extracted from resting state functional MRI for $N = 2,289$ MDD sample (mean age $\pm$ s.d.: 63.2 ± 7.2 years) and $N = 6104$ control subjects (mean age $\pm$ s.d.: 62.9 ± 7.2 years) who were free of MDD and any other mental condition. The cortical effects of MDD were analyzed on the 3D spherical surfaces representing cerebral hemispheres. KL-distance was used to organize the cortical surface into 28 regions of interest based on effect sizes, connectivity and signal-to-noise ratio. The RVI values calculated using this novel approach showed significantly higher effect size of the illness than these calculated using standard Desikan brain atlas.

Keywords: Kullback-Leibler Distance; Major Depressive Disorder; Neuroimaging

1. Introduction

The effects of neuropsychiatric illnesses on brain structure and function are not regionally uniform; pathological processes impact some areas while sparing others, leading to formation of illness-specific deficit patterns.[1,2] Neuroimaging can capture these deficit patterns as case-control differences in functional and structural biomarkers including cortical grey matter thickness, white matter integrity, hypoperfusion, etc. The summary of the findings is tabulated as regional effect sizes for brain areas derived from atlases that parcellate the cerebral landscape using structural landmarks, cellular organization or functional connectivity patterns. The underlying premise stems from basic neuroscience, lesion studies and functional brain mapping that shows that cortical landscape can be represented as parcels of functionally specific and interconnected areas. Large and inclusive meta-analytic studies conducted by big data consortia, such as the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) Consortium[3,4] use these maps as a principle way to report disease-related brain findings. These studies perform "Big Data" level analyses that use these summary regional deficit data aggregative findings from thousands of subjects from multiple studies/cohorts/geographic location to further refine these patterns by eliminating regionally/ethnic or site-specific heterogeneity leading to illness patterns that are reproducible across diverse cohorts.[5]

The patterns published by ENIGMA across neuropsychiatric disorders served as the basis for translating Big Data to the individual level by measuring the agreements between individual brain and those expressed in a disorder.[5] A shortcoming of this approach is that the standard atlases used by ENIGMA to summarize illness-related regional effect sizes may or may not capture the regional pattern and the granularity of the illness effects on the brain. ENIGMA uses atlases that are designed to maximize the regional contrast of the effects of neuropsychiatric illnesses. An alternative strategy is to report the effects of the illness at the spatial resolution of the neuroimaging data. However, this approach also has shortcomings: a) the voxel-wise data has much lower signal-to-noise ratio (SNR) because averaging across regions with uniform effects of the illness minimizes the noise, while maintaining the signal and b) spatial resolution, brain coverage and other data collection parameters vary from cohort to cohort, making reporting and meta-analytical aggregation a challenge. Here, we propose an alternative approach aiming to derive a disorder specific brain parcellation method that maximizes the regional ability to separate cases from controls while maximizing the SNR (through the size of the regions) and contrast between affected/unaffected areas of the brain.

We developed this approach to study the effect of major depressive disorder (MDD) on the regional homogeneity (ReHo) index to pilot this approach. MDD is the most common severe mental illness affecting up to 30% in the lifetime of the population.[6] Despite its high prevalence, the neuroimaging findings in MDD have been affected by poor reproducibility.[7] MDD does not exert a strong neurodegenerative effect on brain structure and findings of meta-analytical studies suggest that MDD is associated with only small (Cohen's $d = 0.01 - 0.1$) structural effect sizes.[8–11] Instead, the MDD-related effects on the brain are likely manifested as a reduction in regional cerebral blood flow (rCBF)[12] especially in cingulate, prefrontal and temporal areas.[13,14] ReHo is a coefficient measuring the temporal coherence of the resting state BOLD functional MRI time-series in neighboring voxels.[15,16] Lower regional ReHo in patients

with MDD versus controls are commonly reported and is often interpreted as evidence for less synchronized local neural activity/connectivity.[17–19] Interpreting the temporal correlations of BOLD signals between neighboring voxels as deficits in neural connectivity is speculative,[20,21] nonetheless, ReHo is a robust and replicable measure that has been validated in human and animal research.[22–24] We and others have shown that ReHo is physiologically linked to rCBF and about 40-60% of the variance in underlying rCBF variations.[25] We hypothesize that reduced ReHo reported in MDD captures the hypoperfusion in affected individuals.

Specifically, we show that the use of the ad-hoc MDD-specific atlas based on the maximizing the separation of regional illness effects can also maximize the translatability of the Big Data findings to the individual level by deriving the contrast between affected and unaffected areas. We chose MDD because it is associated with regionally specific reductions in cerebral blood flow (rCBF)[12] including cingulate, prefrontal and temporal areas, while other parts of the brain show no or even elevated rCBF.[13,14] Creating an ad-hoc MDD-specific brain parcellation that averages the signal across regions with consistent effect of the illnesses, based on the ability to separate cases and controls, can increase SNR and provide disorder and functionally specific pattern of illness related changes. The overall intention is to develop this parcellation approach that is based on capturing the contrast between areas that show deficits and unaffected areas for future meta-analytical studies of MDD where participating sites will use the map for reporting the effect sizes and eventually will use the combined meta-analytical effect size pattern to perform individual prediction of similarity to the illness. Specifically, we propose to use the maps to power the Regional Vulnerability Index (RVI) that measures the similarity between an individual brain and the expected patterns derived from large scale meta-analyses using a representative psychiatric illness, such as MDD.

The RVI approach assumes that the meta-analytic effect-sizes derived from such large meta-analyses can serve as the 'ground truth' for expected disorder-specific deficit patterns and that the similarity between individual and disorder pattern may serve as a biomarker. The utility of this approach has been demonstrated by showing that the white matter RVI for schizophrenia predicted treatment resistance in schizophrenia better than any individual imaging measure.[1] We later demonstrated the similarity in white matter deficit patterns across psychiatric illnesses, suggesting that RVI serves as an important index for cross-disorder research.[26,27] Here, we present an RVI that is based on the MDD-specific atlas that was built based on the regions that show effect size of MDD versus unaffected regions. Specifically, we show that optimizing the granularity of the underlying brain parcellation schema based on the balance between regional specificity and SNR can improve the power of RVI when translating these at level of the individual. Another novelty of our approach is to use the Kullback-Leibler (KL) distance – a rigorously defined distance between probability distributions – to optimize the underlying disorder-specific atlas by balancing granularity, SNR and effects of the illness.

2. Methods

Out of 22,000 available datasets from the first release, 1,780 ($\sim 8\%$) datasets were unusable, and 1,322 ($\sim 6\%$) datasets failed to pass the AFNI processing steps due to poor image quality. The usable data sample consisted of 18,898 participants (8,833 males, 10,065 females; mean age $\pm$ s.d.: 63.2 $\pm$ 7.5 years) with resting state functional MRI (rsfMRI). We used the

UKBB parser software (https://github.com/USC-IGC/ukbb_parser) to identify participants with MDD and non-psychiatric controls based on ICD codes, medication information, symptom severity, hospital records and self-reported diagnoses and other variables using previously published schema. Recurrent MDD was defined as experiencing at least two major depressive episodes in lifetime that required medication or hospitalization. Recurrent MDD subjects experienced on average 3.5 major depressive episodes in their lifetime.[28] Regional homogeneity (ReHo) maps were extracted from rsfMRI for the MDD sample (N=2,289, mean age= 63.2 ± 7.2 years) (ReHo) and control subjects (N=6,104, mean age= 62.9 ± 7.2 years) who were free of MDD and any other mental condition. Other participants were left unclassified because the definitive conclusion on certain criteria could not be made or neurological and psychiatric conditions (including stroke, cerebral ischemia or other disorders)[28] were present.

2.1. *Resting state functional MRI data acquisition, processing, analysis*

UKBB rsfMRI data were acquired on 3 T Siemens Skyra scanners with the standard Siemens 32-channel receive head coil using the following parameters: TR = 735 ms, TE = 39 ms, spatial resolution of 2.4-mm isotropic voxels, matrix size = 88 × 88 with 64 axial slices, number of volumes = 490, flip angle = 52° and multi-band acceleration factor = 8. A separate single-band reference image was acquired and used as the reference scan for head motion correction and alignment to other modalities.[29] The resting state analysis workflow developed by the ENIGMA consortium was used to process the rsfMRI data; processing steps have been described in full detail in prior publications.[30,31] The analysis workflow uses Marchenko-Pastur principal component analysis denoising[32] to improve SNR/temporal SNR of the time series data. In this workflow, a transformation is computed registering the base volume to the ENIGMA EPI template, which is used as a common anatomical spatial reference frame for registration purposes. This step was followed by 3D deconvolution of methodological covariates, and regression of the global signal.[33] Each functional volume was registered to the volume with the minimum outlier fraction for head motion correction, where each transformation was concatenated with the transformation to standard space, to avoid unnecessary interpolation. We removed the effects of the following nuisance variables by using them as covariates using multiple linear regression analysis: the six motion parameters and their temporal derivatives, and time courses from the local white matter and cerebrospinal fluid from lateral ventricles. Motion was estimated as the magnitude of displacement from one time point to the next including neighboring time points and outlier voxels fraction (> 0.1). Time points with excessive motion ($> 0.2mm$) were excluded from further statistical analysis. Images were spatially normalized to the ENIGMA EPI template in MNI standard space for group analysis. The preprocessed data was then used for ReHo calculations. In the whole sample, the average motion, average outlier voxels fraction and average time points censored fraction were 0.12 mm, 0.004 and ~ 0.13 respectively. The preprocessed data was then used for ReHo calculations.

2.2. *Regional Homogeneity (ReHo) analysis*

ReHo was designed to investigate changes in local spontaneous brain activity by performing a nearest neighbor analysis of similarity of the BOLD time-series and assigning a score, called Kendall's coefficient of concordance (KCC)[16] per voxel. The KCC score is calculated per voxel

based on signals from neighboring voxels as: $W = (\sum_i R_i^2 - nR^2)/(\frac{1}{12}K^2(n^3 - n))$. Here, W is the KCC among given voxels, ranging from 0 to 1; R_i is the sum rank of the ith time point; $R = ((n+1)K)/2$ is the mean of the R_i's; K is the number of time series within a measured cluster (K is set to be 7, 19, or 27), and n is the number of ranks (= number of volumes).[16] K was set to be 27, which is appropriate for covering all directions in 3D space and to optimize the trade-off between mitigation of partial volume effects and generation of Gaussian random fields.[17] For each subject, the ReHo map was computed in 3D volumetric space using the AFNI-command '3dReHo'. These maps were used to extract regional ReHo values for the regions of interest for post-processing.

2.3. *Computation of reference brain image and hemispheric mesh*

Clustering analyses were performed using the 2D spherical manifold defined by the hemispheric surfaces from the **average ReHo brain image** (**avBrain**), that was obtained as the arithmetic average over the entire dataset of N=8,393 (registered and normalized) ReHo images. After computing avBrain, we manually edited it to separate it into cerebrum and two hemispheres. The marching cube algorithm was used to extract uniform triangulated mesh of $K = 5068 vertices$ $\{\text{VER}_1, \text{VER}_2, \ldots, \text{VER}_K\}$ at uniform 1 mm spacing. We refer to this set of vertices as the **average boundary mesh** (**avMesh**). Each vertex VER_i in avMesh is identified by its 3 coordinates $[x_i, y_i, z_i]$ in 3D space. This procedure is illustrated in Fig. 1.

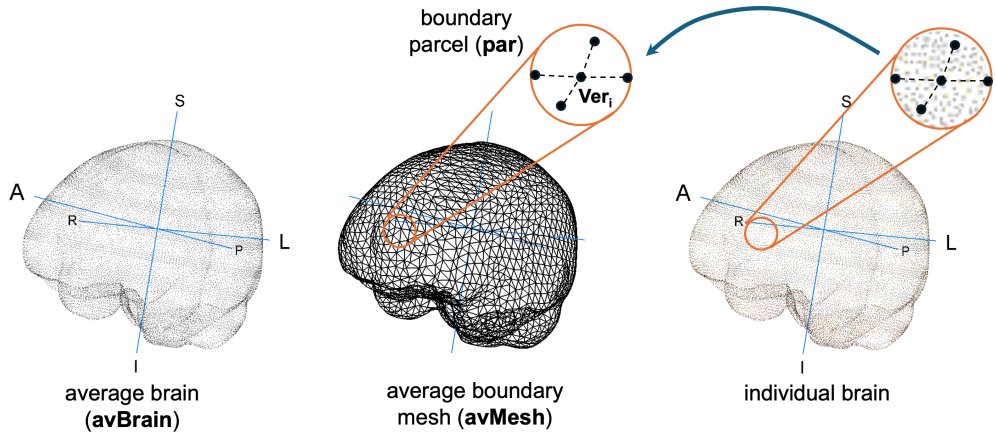


Fig. 1. Reference brain image avBrain and a hemispheric mesh computation to generate avMesh.

2.3.1. *Registration of the average cortical mesh onto individual ReHo images*

Matching between the discretized boundary avMesh and each ReHo image $I(n)$, for a given subject n, cannot be performed using affine transformations and required a nonlinear registration where each cortical vertex VER_i of avMesh was matched to a single voxel in $I(n)$, namely the voxel of $I(n)$ that was geometrically closest to VER_i. Other common approaches to archive this registration includes averaging over a spherical kernel and averaging over the normal projection. The **intensity** $J_i(n)$ of a cortical vertex VER_i is the intensity of its matching voxel in the image $I(n)$. After registration was completed for all vertices of the average boundary

mesh avMesh, then each subject could be characterized by a boundary pattern $J(n)$ where each cortical vertex VER_i of avMesh has the image intensity $J_i(n)$ computed via registration. That is, the $n-$th subject is described by a list of K numerical features, i.e., the K intensities $\{J_1(n), J_2(n), \ldots, J_K(n)\}$ indexed by the K cortical vertices $\{\text{VER}_1, \text{VER}_2, \ldots, \text{VER}_K\}$ on the average boundary mesh avMesh.

2.4. *Parcellation of the average boudary mesh: SNR vs. granularity*

For each cortical vertex VER_i on avMesh, we computed three **boundary parcels** $parD(i)$ centered at VER_i and having three different sizes $D = 1, 3, 5$, to capture local information in the ReHo image at different **granularity levels**.

The parcel $parD(i)$ includes the D closest neighbors of the vertex VER_i in avMesh. Hence, parcel $par1(i)$ only contains the boundary vertex VER_i, parcel $par3(i)$ contains VER_i and its 2 closest neighbors in avMesh and $par5(i)$ contains VER_i and its 4 closest neighbors in avMesh. The radius of $par3(i)$ roughly ranges between 1 and 3.1 voxels with a mean radius of 1.9 voxels; for $par5(i)$, the parcel radius ranges between 1.4 and 3.6 voxels, with a mean radius of 2.5 voxels.

2.5. *Kullback-Leibler distance between two probability distributions*

Let us consider two multivariate probability distributions P and Q on $\mathbb{R}^D$; for instance, P and Q can be determined by two probability density functions $f_P(x)$ and $f_Q(x)$ defined for $x \in \mathbb{R}^D$. The **Kullback-Leibler (KL) divergence**[34,35] between P and Q, denoted as $KL(P, Q)$, classically measures how much Q differs from P. When the density functions f_P and f_Q are known, the KL divergence $KL(P, Q)$ is given by the explicit formula

$$KL(P, Q) = \int f_P(x) \log \frac{f_P(x)}{f_Q(x)} dx \geq 0.$$

Note that $KL(P, Q) \neq KL(Q, P)$ in general. One defines a symmetric **KL distance**[36] between P and Q by the formula:

$$KLdis(P, Q) = KLdis(Q, P) = KL(P, Q) + KL(Q, P) \geq 0.$$

We have that $KLdis(P, Q) = 0$ if and only if $P = Q$

The KL distance has played a key part in theoretical and applied statistics for more than 40 years.[37] It quantifies the discriminating power of the well-known log-likelihood statistical test to discriminate between two probability models P and Q based on N random observed vectors $x_1, \ldots, x_N$ of dimension D generated by an unknown probability distribution. This log-likelihood test between P and Q is an *optimal test* that is based on comparing the log-likelihood of observations $x_1, \ldots, x_N$ under P with their log-likelihood under Q. As such it provides a most sensitive measure to compare signal intensities over multiple locations.

When the probability distributions P and Q have multivariate normal density functions $f_P(x)$ and $f_Q(x)$, respectively, defined for all $x \in \mathbb{R}^D$, then f_P and f_Q are determined by their respective mean vectors m_P, m_Q and their $D \times D$ covariance matrices S_P, S_Q. In this case, the KL distance between P and Q is computed using the explicit formula:

$$KLdis(P, Q) = -D + \frac{1}{2} trace(S_P^{-1} S_Q + S_Q^{-1} S_P) + \frac{1}{2} (m_P - m_Q)^T (S_Q^{-1} S_P^{-1})(m_P - m_Q)$$

where A^T denotes the transpose of matrix A.

2.6. *Discriminating score of boundary parcels*

Within the average boundary mesh avMesh, we intend to identify which small cortical parcels $parD(i)$, $D = 1, 3, 5$, have high discriminating power between CTL and MDD subjects. To this end, for each parcel, we introduce a notion of discriminating score as follows.

For any vertex VER_i and boundary parcel $parD(i)$ of size D centered around VER_i, the boundary vertices belonging to $parD(i)$ are indexed by their D indices $i1, i2, \ldots, iD$, with $i1 = i$. For the n-th subject, these D vertices define a vector $W_i(n)$ of D intensities pre-computed above by registration of avMesh to the subject's ReHo image $I(n)$. Namely we set $W_i(n) = [J_{i1}(n), J_{i2}(n), \ldots, J_{iD}(n)]$.

Hence, we define two subsets of vectors H_1 and H_0 in $\mathbb{R}^D$ of respective sizes N_1 and N_0 ($N_1 = 2,289, N_0 = 6,104$, here) by:

- H_1 = set of all vectors $W_i(n)$ such that subject n belongs to the MDD class
- H_0 = set of all vectors $W_i(n)$ such that subject n belongs to the CTL class

The $N_1 = 2,289$ observed vectors $W_i(n)$ in H_1 will be viewed as a sample of N_1 random vectors generated by a multivariate normal P_1 with mean vector m_1 and $D \times D$ covariance matrix S_1. Similarly, the $N_0 = 6,104$ vectors $W_i(n)$ in H_0 provide a sample of size N_0 generated by a multivariate normal P_0 with mean vector m_0 and $D \times D$ covariance matrix S_0. The vectors m_0, m_1 and the matrices S_0, S_1 are unknown but can be estimated by the following sample means and sample covariances:

$$m_\ell = \frac{1}{N_\ell} \sum_{W_i(n) \in H_\ell} W_i(n), \; cov_\ell = \frac{1}{N_\ell} \sum_{W_i(n) \in H_\ell} (W_i(n) - m_\ell)(W_i(n) - m_\ell)^T, \quad \text{for } \ell = 0, 1.$$

When $KLdis(P_1, P_0)$ is large, the two probabilities P_1 and P_0 are very different, meaning that the boundary parcel $parD(i)$ has potentially high discriminating power between the CTL and MDD groups. Hence, we define the **discriminating score** $scoreD(i)$ of boundary parcel $parD(i)$ as the KL distance $KLdis(P_1, P_0)$ computed above. Clearly, the computation of $scoreD(i)$ for parcel $parD(i)$ need to be repeated separately for each boundary vertex VER_i, where $i = 1, \ldots, K$. This yielded $K = 5,020$ discriminating scores $scoreD(i)$ at granularity sizes $D = 1, 3, 5$, with one score per boundary vertex VER_i. We can then re-order the vertices VER_i in decreasing order of their discriminating scores and display them in 3D space on the average boundary mesh avMesh. We applied the method described for granularity sizes $D = 1, 3, 5$.

2.7. *Implementation of discriminating score analysis*

The methods outlined above required about 20 hours of computing time on a standard laptop to compute the three discriminating scores $scoreD(i)$, $D = 1, 3, 5$, for each one of the $K = 5,020$ boundary vertices VER_i generated at scale 5x5x5. This led to a positive first assessment of our methodology at a reasonable computing cost, and a substantial analysis of the spatial continuity of our 3 discriminating scores.

Since the brain ReHo images of our UKBB-MDD dataset and the corresponding average brain images were actually discretized at a finer scale 2x2x2 (mm), we have then used the freely

available Mango software[38] to generate, at scale 2x2x2, a new triangulated mesh avMesh2 of $L = 20,025$ boundary vertices Z_s, $s = 1, 2, \ldots, L$, densely located on the boundary of the average brain. Of course, avMesh2 contained our initial coarser avMesh of $K = 5,020$ boundary vertices VER_i, $i = 1, 2, \ldots, K$, discretizing the brain boundary at scale 5x5x5.

Our three discriminating $scoreD(i)$, $D = 1, 3, 5$, initially computed for each VER_i of avMesh were then smoothly extended to scores $scoreD(i)$, $D = 1, 3, 5$, for each Z_s of avMesh2 using the following fast spatial propagation algorithm. Specifically, for each vertex Z_s of avMesh2, we identified the list $G(s)$ of all vertices VER_i in avMesh which are at distance less than 5 mm from Z_s, and computed the discriminating score $scoreD(s)$ of Z_s as the average of $scoreD(i)$ over all vertices VER_i belonging to $G(s)$; this procedure was carried out for $D = 1, 3, 5$. This method for spatial extension was useful to generate better 3D visualizations of the most significant discriminating brain boundary vertices on avBrain.

2.8. *Regional Vulnerability Index (RVI) calculations*

RVI scores were calculated using the 'RVIpkg' in the R software based on our previous publication[39] with some revisions. The original RVI calculated the correlational agreement between an individual's regional brain measures and the pattern of regional MDD-related brain calculated using a standard atlas.[5] Here, we used the regions identified by the cluster analysis to compute the effect sizes for MDD cases vs. controls. Next, we used these effect sizes to calculate RVI for each subject as the dot product between vectors $Z = (Z_i)$ and $E = (E_i)$, normalized by the dimensions of the vector using the equation $RVI = \sum_{i=1}^{N} Z_i E_i$, where Z is the vector of deviation from the mean and E is the vector of meta-analytical effect size (*Cohen's d coefficients*) for the i-th regional measure for MDD. N is the dimension of the vector, i.e., the total number of imaging phenotypes for that modality. The modified RVI was calculated for the whole-brain phenotype vector and for cortical, subcortical and white matter modalities. Positive RVI values indicate that the regional pattern of an individual coincides with the expected pattern of MDD based on the overall effect sizes. We compared effect sizes for the RVI-MDD vs. the effect sizes obtained for individual regions identified by cluster analysis.

3. Results

3.1. *Calculation of discriminatory boundary vertices*

According to the procedure described in Sec. 2.3, for each one of the $K = 5,020$ vertices VER_i from the brain boundary mesh, we computed three discriminating scores, namely, $score1(i)$, $score3(i)$, $score5(i)$, corresponding to the boundary parcels $par1(i), par2(i), par5(i)$ of sizes $D = 1, 3$ and 5, respectively, centered at each boundary vertex VER_i. As explained above, these scores *quantify the discriminating power of statistical tests based on log-likelihood values*, using boundary parcels of different sizes. We implemented numerical simulations to compute the respective statistical significance thresholds $thr1, thr3, thr5$ for $score1(i), score3(i), score5(i)$. For instance, whenever $score3(i) < thr3$, then vertex VER_i has high probability of *being not discriminatory* between the CTL and MDD groups. Table 1 reports the minimum (min), median or 50%-quantile, 80%-quantile, significant thresholds (threshold) of the computed discriminating scores and finally the percentages of significant mesh boundary vertices for

$score1, score3, score5$. We note that the highest percentage (30%) of significant vertices is reached with $score3$.

Table 1. Discriminating scores for granularity sizes 1, 3 and 5.

	min	median	80% quant	threshold	significant vertices (%)
score1	0	.009	.020	.015	27%
score3	0	.030	.051	.045	30%
score5	0	.040	.070	.070	20%

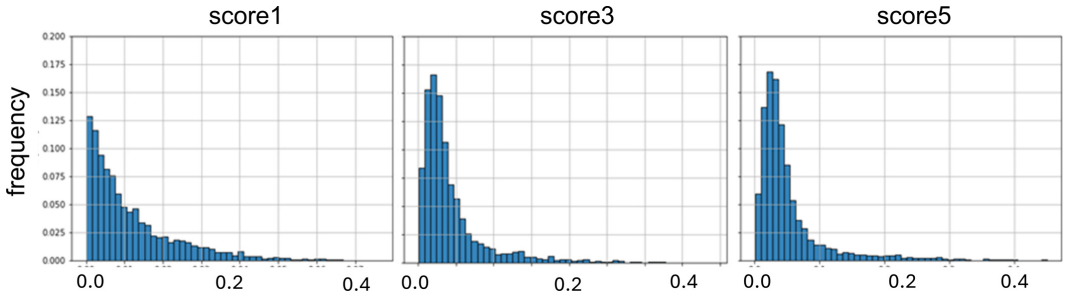


Fig. 2. Histograms of the discriminating scores $score1, score3, score5$ computed from the 5,020 brain boundary vertices VER_i of avMesh. Each plot displays horizontally the values of the discriminating score and vertically the associated frequencies.

Fig. 2 displays the histograms of the discriminating scores $score1, score3, score5$. The figure shows that the distribution of $score1$, which has a peak at $score1 = 0$, is very different from the distributions of $score3$ and $score5$. In fact, after horizontal rescaling of scores values, the distributions of $score3$ and $score5$ become similar. This remark is confirmed by the calculation of the correlation matrix of $score1, score3, score5$, showing that $score3$ and $score5$ are highly correlated (corr = 0.89), while $score1$ is weakly correlated to either $score3$ or $score5$.

3.1.1. *Spatial stability of discriminating scores*

Clearly, the value of the discriminating score $scoreD(i)$ computed at a vertex VER_i may be sensitive to the granularity size D. To address this problem, we proceeded as follows.

For each brain boundary vertex VER_i in avMesh, we denote as VER_{i*} its closest neighbor within avMesh and let $d(i)$ be the Euclidean distance between vertices VER_i and VER_{i*}.

For each vertex VER_i and any dimension D, the relative change $relchD(i)$ of the discriminating score $scoreD(i)$ when one replaces VER_i by its closest neighbor VER_{i*} is defined by $relchD(i) = \frac{|scoreD(i) - scoreD(i*)|}{scoreD(i)}$. To quantify the *spatial stability* of $scoreD(i)$ around a boundary vertex VER_i, we compute its Lipschitz coefficient $LipD(i) = \frac{relchD(i)}{d(i)}$, measuring the sensitivity of the discriminating score to small spatial changes. Hence, for each discriminating score, *lower values of its Lipschitz coefficients indicate higher spatial stability of the score.*

Fig. 3(a) displays the histograms of $Lip1(i), Lip2(i), Lip3(i)$ and shows that $Lip1(i)$ is stochastically larger than $Lip3(i)$ or $Lip5(i)$. This implies that $score1$ is significantly less spatially stable than $score3$ and $score5$. This observation is consistent with the plots of the three

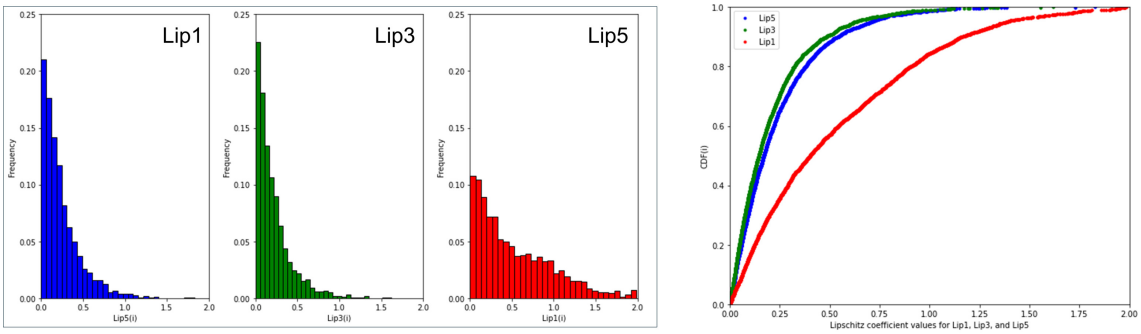


Fig. 3. (a) Histograms of the Lipschitz coefficients $Lip1$, $Lip3$ and $Lip5$ of the discriminating scores $score1$, $score3$, and $score5$, respectively; (b) corresponding cumulative distribution functions of $Lip1$, $Lip3$ and $Lip5$. Plots show that $score1$ is significantly less spatially stable than $score3$ or $score5$.

Cumulative Distribution Functions of LipD, for D=1,2,3, denoted as CDFD and shown in Fig. 3(b). The figure shows that, for all $x = LipD$, one has that $CDF1(x) \leq CDF3(x)$ and $CDF1(x) \leq CDF5(x)$, again confirming that $score1$ is significantly less spatially stable than $score3$ and $score5$. In fact, the plot in Fig. 3(b) confirms that $score3$ is stochastically the most spatially stable of the 3 scores computed.

3.2. *3D visualization of discriminatory boundary vertices*

Based on the analysis presented above, we concluded that $score3$, at granularity size $D = 3$, is the most spatially stable, in a precise mathematical sense, hence providing the most valuable information about which boundary vertices have most discriminatory power. This granularity size was shown perform better than granularity sizes $D = 1$ and $D = 5$. We focus henceforth on this granularity size to report our graphic illustration of the most discriminatory boundary vertices. Accordingly, Fig. 4 displays the boundary vertices having highest discriminatory score using $score3$, the discriminating score computed at granularity size $D = 3$, overlaid on rendered brain.

3.3. *Score based parcellation of brain boundary and RVI results*

Using the Mango software on avBrain at scale 2x2x2, we generated a triangulated mesh avMesh2 of 20,025 vertices and, for each such vertex, we computed a discriminating score, namely score3, as seen above.

The triangulated mesh avMesh2 contained p triangles $\{TR_1, \ldots, TR_p\}$, with roughly $p \approx 40,000$. We extended our discriminating score3 to each triangle TR_j, by computing the average of the scores of the 3 vertices of TR_j. After reordering the list of our p triangles in decreasing order of their discriminating scores, we partitioned this ordered list of triangles into 6 successive sub-lists $S_1, S_2, \ldots, S_6$ of similar sizes, ranging from 6,000 to 7,000 triangles. Next, for each $i = 1, \ldots, 6$ we computed disjoint "watertight" connected components of the sub-list S_i. Recall that a set W of triangles is called watertight connected if any two triangles T and T' in W can be linked by some chain of triangles $T_1 = T, T_2, \ldots, T_r = T'$ where any pair of successive triangles have one edge in common, and where r is any integer.

After completion of this procedure, for each sub-list S_i, we obtained a partition of our

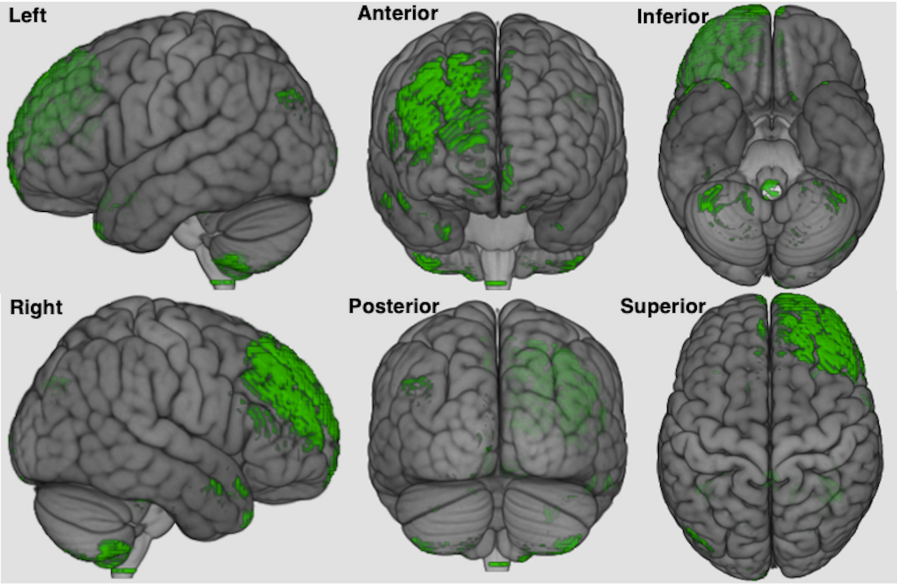


Fig. 4. Boundary vertices with highest discriminating scores (using score3 above 80% quartile), overlaid on rendered brain with different views.

initial set of p triangles into q disjoint watertight connected components $C_1, \ldots, C_q$. Each C_s is a finite set of small triangles on the surface of the average brain, defining a connected sub-region REG_s of the brain surface. We kept only the regions REG_s which have a large enough area (i.e., more than 250 vertices) and implemented a simple procedure for pragmatic regrouping of the very small sub-regions. At this point we obtained a score-based parcellation of the brain surface into sub-regions denoted REG_s, which we used to compute RVI values.

3.4. *Effect sizes for ReHo values in clusters vs. RVI*

The Cohen's d effect sizes were calculated for the 28 spatial distinct components that were identified based on the ability to discriminated between MDD cases and controls and were larger than 200 connected vertices. The effect sizes varied from $d = 0.25, p = 10^{-24}$ for the ReHo values calculated for the highly discriminating component in Fig. 4 to $d = 0.00, p = 1.00$ for the component that showed no effects of MDD. We next calculated the individual RVI using the effect sizes for the 28 components. Subjects with MDD showed highly elevated RVI-MDD versus controls (0.16 ± 0.01 vs -0.01 ± 0.01), Cohen's $d = 0.36, p = 10^{-55}$. We observed that the effect size for RVI were numerically stronger than effect sizes for any of the individual components, as shown in Fig. 5.

4. Discussion

Regional cerebral hypoperfusion in individuals with MDD was first described over three decades ago.[14,40,41] Specifically, with the hypoperfusion of limbic-frontal-temporal circuitry,[42] this approach was deemed replicable, informative of the clinical state and predictive of treatment outcome.[14,43] Our study proposes and evaluates a novel approach to categorize and parcellate the cortical landscape based on the balance between the size of the continuous region and the effect size to discriminate between cases and controls. This method achieves the

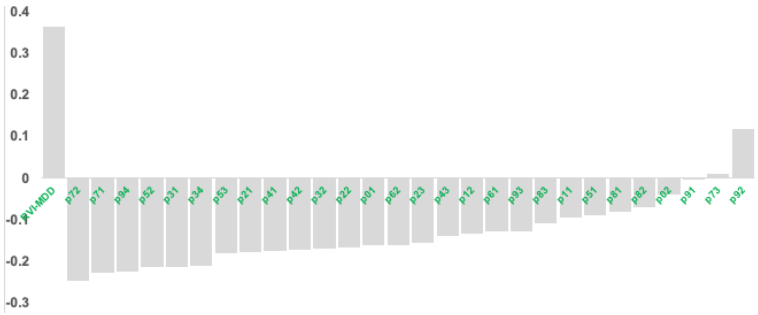


Fig. 5. Effect-size of RVI compared to effect-size for individual components. For instance, the component p_{11} represents the first parcellated region and the first cluster and and p_{01} represents the tenth parcellated region and the first cluster respectively.

optimal granularity to quantify the effects of the illness on the brain that maintains high SNR while preserving functional specificity to the areas that show effects of the illness. The proposed methodologies can be viewed as development of a disorder specific atlas that is focused on subdivision of the overall cerebral cortex into a set of 10 levels based on their ability to discriminate cases versus controls. This regional variability in effect sizes is then used to build a novel biomarker - RVI - that summarizes the agreement between individual brain patterns and the expected pattern of the illness. We found that the RVI built on these regions showed numerically higher effect size of the illness than the best effect sizes (level 1) for regional measures (Cohen's d=0.36 versus 0.25). This suggested that individual agreement with the pattern served as a better phenotype than hypoperfusion in any single area of the brain, even those specifically chosen to provide the best description.

We showed that effects of the ReHo-based measurements can be used to summarize hypoperfusion patterns in MDD. The granularity analyses demonstrated that areas with best discrimination between MDD cases and controls involve middle and inferior frontal and parietal regions that were previously identified by PET and SPECT studies.[44,45] However, the effect sizes for these regions were modest (∼0.22-0.25) suggesting that the lower ReHo values in these areas in an individual are not specific to MDD and are unlikely to have clinical relevance; a further detailed study on the clinical significance of these findings will be needed. In contrast, the ReHo-based RVI for MDD built on the overall contrast across all ten levels showed significant elevation in individuals with MDD and overall had numerically larger effect sizes than these for regional ReHo. The RVI approach is not new or specific to ReHo. It was first applied in schizophrenia, where higher RVI values for schizophrenia have been linked to treatment resistance, cognitive deficits, to family risks for the disease[2,5,5,39] and was shown to be applicable to other illnesses or conditions.[46,47] The RVI approach has been proposed as a prospective tool for early detection of brain patterns shifting towards a particular condition, and may provide an early predictive signal for other neurological and psychiatric illnesses.[39,48] However, all previous work was focused on structural brain deficits. The novelty of this work is to show the applicability of RVI for functional fMRI measures, here using ReHo, which we posited as a proxy for rCBF signal and can potentially be altered by therapies such as medication. It is also novel because the RVI was built on the regional measurements that were

specifically identified for this purpose and balanced the size of the regions versus the disorder related contrast. Still, how RVI measures derived using ReHo correlate with clinical measures of MDD severity and whether there exist identifiable brain regions or patterns need to be explored.

There are limitations to this study. This analysis was focused on evaluation of the MDD-specific versus cortical-area parsed parcellation to tabulate effect sizes of MDD for further RVI analyses. Future studies need to evaluate MDD-specific versus other connectivity-atlases. It is likewise important to evaluate the ability of RVI calculated using MDD-specific versus standard atlas approaches to predict clinical and cognitive variance among subjects. We calculated rsfMRI ReHo signal that included global signal regression (GSR). The global signal in rsfMRI data is associated with head motion, respiration and cardiac rhythms.[49,50] GSR is a necessary pre-processing step for ReHo analysis because these non-physiological factors can artifactually increase ReHo through global autocorrelation[51,52] and exerts complex effects on ReHo measures.[19,53] However, other studies demonstrated that GSR also carries diagnosis related variance.[54–56] No analyses were performed to examine whether our study findings would differ without GSR, a shortcoming of the rsfMRI data processing. In addition, testing of the outcomes in this study were limited to subjects in the same cohort. The narrow aim of this study was to evaluate the novel parcellation approach and show that RVI derived from these regions carried higher effect size than the regions specifically selected for their high discrimination of illness effect. Follow up studies will evaluate if the pattern of MDD ReHo contrasts can be replicated in other cohorts and further this pattern by performing meta-aggregation to improve this disorder specific parcellation schema. In addition, our analysis did not focus on MDD subtypes and this need to be considered in future studies.

The method presented in the manuscript can be extended to other types of neuropsychiatric disorders where similar datasets are available with minimal changes. Future studies need to examine the granularity size more extensively to allow for a wider range of scales. From the viewpoint of computational cost, the method presented is scalable as the most computationally expensive steps of the algorithm, namely vertex registration and score computation, are highly parallelizable.

Acknowledgments

DL acknowledges support from Simons Foundation grant MP-TSM-00002738 and NIH grant 1RF1MH123163-01A1 (subaward); PK acknowledges support from NIH grants R01MH133812, U01MH108148, R01NS114628, DP1DA048968, R01EB015611, and R01MH121246.

References

1. P. Kochunov, P. M. Thompson and L. E. Hong, Toward high reproducibility and accountable heterogeneity in schizophrenia research, *JAMA Psychiatry* **76**, 680 (2019).
2. P. Kochunov, F. Fan, M. C. Ryan, K. S. Hatch, S. Tan, N. Jahanshad, P. M. Thompson, T. G. van Erp, J. A. Turner, S. Chen *et al.*, Translating ENIGMA schizophrenia findings using the regional vulnerability index: association with cognition, symptoms, and disease trajectory, *Human Brain Mapping* **43**, 566 (2022).
3. P. M. Thompson, O. A. Andreassen, A. Arias-Vasquez, C. E. Bearden, P. S. Boedhoe, R. M.

Brouwer, R. L. Buckner, J. K. Buitelaar, K. B. Bulayeva, D. M. Cannon *et al.*, ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide, *Neuroimage* **145**, 389 (2017).
4. P. M. Thompson, N. Jahanshad, C. R. Ching, L. E. Salminen, S. I. Thomopoulos, J. Bright, B. T. Baune, S. Bertolín, J. Bralten, W. B. Bruin *et al.*, ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries, *Translational Psychiatry* **10**, p. 100 (2020).
5. P. Kochunov, J. Huang, S. Chen, Y. Li, S. Tan, F. Fan, W. Feng, Y. Wang, L. M. Rowland, A. Savransky *et al.*, White matter in schizophrenia treatment resistance, *American Journal of Psychiatry* **176**, 829 (2019).
6. P. Waraich, E. M. Goldner, J. M. Somers and L. Hsu, Prevalence and incidence studies of mood disorders: a systematic review of the literature, *The Canadian Journal of Psychiatry* **49**, 124 (2004).
7. J. P. Ioannidis, How to make more published research true, *Revista Cubana de Información en Ciencias de la Salud (ACIMED)* **26**, 187 (2015).
8. M. Farrell, T. Werge, P. Sklar, M. J. Owen, R. Ophoff, M. C. O'Donovan, A. Corvin, S. Cichon and P. F. Sullivan, Evaluating historical candidate genes for schizophrenia, *Molecular psychiatry* **20**, 555 (2015).
9. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder Working Group.
10. L. Schmaal, D. Hibar, P. G. Sämann, G. Hall, B. Baune, N. Jahanshad, J. Cheung, T. G. van Erp, D. Bos, M. A. Ikram *et al.*, Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group, *Molecular psychiatry* **22**, 900 (2017).
11. S. Kelly, N. Jahanshad, A. Zalesky, P. Kochunov, I. Agartz, C. Alloza, O. Andreassen, C. Arango, N. Banaj, S. Bouix *et al.*, Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA schizophrenia DTI working group, *Molecular psychiatry* **23**, 1261 (2018).
12. J. P. Gray, V. I. Müller, S. B. Eickhoff and P. T. Fox, Multimodal abnormalities of brain structure and function in major depressive disorder: a meta-analysis of neuroimaging studies, *American Journal of Psychiatry* **177**, 422 (2020).
13. K. Matsuo, D. Glahn, M. Peluso, J. Hatch, E. Monkul, P. Najt, M. Sanches, F. Zamarripa, J. Li, J. Lancaster *et al.*, Prefrontal hyperactivation during working memory task in untreated individuals with major depressive disorder, *Molecular Psychiatry* **12**, 158 (2007).
14. H. S. Mayberg, S. K. Brannan, R. K. Mahurin, P. A. Jerabek, J. S. Brickman, J. L. Tekell, J. A. Silva, S. McGinnis, T. G. Glass, C. C. Martin *et al.*, Cingulate function in depression: a potential predictor of treatment response, *Neuroreport* **8**, 1057 (1997).
15. L. Wang, M. Song, T. Jiang, Y. Zhang and C. Yu, Regional homogeneity of the resting-state brain activity correlates with individual intelligence, *Neuroscience letters* **488**, 275 (2011).
16. Y. Zang, T. Jiang, Y. Lu, Y. He and L. Tian, Regional homogeneity approach to fMRI data analysis, *Neuroimage* **22**, 394 (2004).
17. L. Jiang and X.-N. Zuo, Regional homogeneity: a multimodal, multiscale neuroimaging marker of the human connectome, *The Neuroscientist* **22**, 486 (2016).
18. C. Qiu, W. Liao, J. Ding, Y. Feng, C. Zhu, X. Nie, W. Zhang, H. Chen and Q. Gong, Regional homogeneity changes in social anxiety disorder: a resting-state fMRI study, *Psychiatry Research: Neuroimaging* **194**, 47 (2011).
19. X.-N. Zuo, T. Xu, L. Jiang, Z. Yang, X.-Y. Cao, Y. He, Y.-F. Zang, F. X. Castellanos and M. P. Milham, Toward reliable characterization of functional homogeneity in the human brain: preprocessing, scan duration, imaging resolution and computational space, *Neuroimage* **65**, 374

(2013).
20. H. Li, Y. Lin, J. Chen, X. Wang, Q. Wu, Q. Li and Z. Chen, Abnormal regional homogeneity and functional connectivity in adjustment disorder of new recruits: a resting-state fMRI study, *Japanese journal of radiology* **35**, 151 (2017).
21. J. Chen, Y. Xu, K. Zhang, Z. Liu, C. Xu, Y. Shen and Q. Xu, Comparative study of regional homogeneity in schizophrenia and major depressive disorder, *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **162**, 36 (2013).
22. Y. Han, J. Wang, Z. Zhao, B. Min, J. Lu, K. Li, Y. He and J. Jia, Frequency-dependent changes in the amplitude of low-frequency fluctuations in amnestic mild cognitive impairment: a resting-state fMRI study, *Neuroimage* **55**, 287 (2011).
23. Z. Zhang, Y. Liu, T. Jiang, B. Zhou, N. An, H. Dai, P. Wang, Y. Niu, L. Wang and X. Zhang, Altered spontaneous activity in alzheimer's disease and mild cognitive impairment revealed by Regional Homogeneity, *Neuroimage* **59**, 1429 (2012).
24. L. Jiang, Y. Xu, X. Zhu, Z. Yang, H. Li and X. Zuo, Local-to-remote cortical connectivity in early-and adulthood-onset schizophrenia, *Translational Psychiatry* **5**, e566 (2015).
25. J. Zhu, Y. Jin, K. Wang, Y. Zhou, Y. Feng, M. Yu and X. Jin, Frequency-dependent changes in the regional amplitude and synchronization of resting-state functional MRI in stroke, *PLoS One* **10**, p. e0123850 (2015).
26. P. Kochunov, L. E. Hong, E. L. Dennis, R. A. Morey, D. F. Tate, E. A. Wilde, M. Logue, S. Kelly, G. Donohoe, P. Favre *et al.*, ENIGMA-DTI: Translating reproducible white matter deficits into personalized vulnerability metrics in cross-diagnostic psychiatric research, *Human Brain Mapping* **43**, 194 (2022).
27. P. Kochunov, A. Zavaliangos-Petropulu, N. Jahanshad, P. M. Thompson, M. C. Ryan, J. Chiappelli, S. Chen, X. Du, K. Hatch, B. Adhikari *et al.*, A white matter connection of schizophrenia and Alzheimer's disease, *Schizophrenia Bulletin* **47**, 197 (2021).
28. D. J. Smith, B. I. Nicholl, B. Cullen, D. Martin, Z. Ul-Haq, J. Evans, J. M. Gill, B. Roberts, J. Gallacher, D. Mackay *et al.*, Prevalence and characteristics of probable major depression and bipolar disorder within UK biobank: cross-sectional study of 172,751 participants, *PloS One* **8**, p. e75362 (2013).
29. F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee *et al.*, Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank, *Neuroimage* **166**, 400 (2018).
30. B. M. Adhikari, N. Jahanshad, D. Shukla, D. C. Glahn, J. Blangero, P. T. Fox, R. C. Reynolds, R. W. Cox, E. Fieremans, J. Veraart *et al.*, Comparison of heritability estimates on resting state fMRI connectivity phenotypes using the ENIGMA analysis pipeline, *Human Brain Mapping* **39**, 4893 (2018).
31. B. M. Adhikari, N. Jahanshad, D. Shukla, D. C. Glahn, J. Blangero, R. C. Reynolds, R. W. Cox, E. Fieremans, J. Veraart, D. S. Novikov *et al.*, Heritability estimates on resting state fmri data using ENIGMA analysis pipeline, in *Pacific Symposium on Biocomputintg 2018: Proceedings of the Pacific Symposium*, 2018.
32. J. Veraart, D. S. Novikov, D. Christiaens, B. Ades-Aron, J. Sijbers and E. Fieremans, Denoising of diffusion MRI using random matrix theory, *Neuroimage* **142**, 394 (2016).
33. B. M. Adhikari, N. Jahanshad, D. Shukla, J. Turner, D. Grotegerd, U. Dannlowski, H. Kugel, J. Engelen, B. Dietsche, A. Krug *et al.*, A resting state fmri analysis pipeline for pooling inference across diverse cohorts: an ENIGMA rs-fMRI protocol, *Brain Imaging and Behavior* **13**, 1453 (2019).
34. M. Tumminello, F. Lillo and R. N. Mantegna, Kullback-Leibler distance as a measure of the information filtered from multivariate data, *Physical Review E—Statistical, Nonlinear, and Soft*

Matter Physics **76**, p. 031123 (2007).
35. F. Pérez-Cruz, Kullback-Leibler divergence estimation of continuous distributions, in *2008 IEEE international symposium on information theory*, 2008.
36. D. H. Johnson, S. Sinanovic *et al.*, Symmetrizing the Kullback-Leibler distance, *IEEE Transactions on Information Theory* **1**, 1 (2001).
37. J. Shlens, Notes on Kullback-Leibler divergence and likelihood theory, *Systems Neurobiology Laboratory* **92037**, 1 (2007).
38. Mango (Multi-image Analysis GUI) `https://mangoviewer.com/`, [Accessed 10-08-2024].
39. P. Kochunov, Y. Ma, K. S. Hatch, S. Gao, A. Acheson, N. Jahanshad, P. M. Thompson, B. M. Adhikari, H. Bruce, J. Chiappelli *et al.*, Ancestral, pregnancy, and negative early-life risks shape children's brain (dis) similarity to schizophrenia, *Biological Psychiatry* **94**, 332 (2023).
40. H. S. Mayberg, P. J. Lewis, W. Regenold and H. N. Wagner, Paralimbic hypoperfusion in unipolar depression, *Journal of Nuclear Medicine* **35**, 929 (1994).
41. H. Mayberg, Depression and frontal-subcortical circuits: Focus on prefrontal-limbic interactions. (2001).
42. A. Silva-dos Santos, M. Sales, A. Sebastião and R. Gusmão, A new viewpoint on the etiopathogenesis of depression: insights from the neurophysiology of deep brain stimulation in Parkinson's disease and treatment-resistant depression, *Frontiers in Psychiatry* **12**, p. 607339 (2021).
43. D. G. Amen and M. Easton, A new way forward: how brain SPECT imaging can improve outcomes and transform mental health care into brain health care, *Frontiers in Psychiatry* **12**, p. 715315 (2021).
44. K. Helm, K. Viol, T. M. Weiger, P. A. Tass, C. Grefkes, D. Del Monte and G. Schiepek, Neuronal connectivity in major depressive disorder: a systematic review, *Neuropsychiatric disease and treatment* , 2715 (2018).
45. S.-G. Kang and S.-E. Cho, Neuroimaging biomarkers for predicting treatment response and recurrence of major depressive disorder, *International journal of molecular sciences* **21**, p. 2148 (2020).
46. P. Kochunov, Y. Ma, K. S. Hatch, L. Schmaal, N. Jahanshad, P. M. Thompson, B. M. Adhikari, H. Bruce, J. Chiappelli, E. L. Goldwaser *et al.*, Separating clinical and subclinical depression by big data informed structural vulnerability index and its impact on cognition: ENIGMA Dot Product, in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2022.
47. K. S. Hatch, S. Gao, Y. Ma, A. Russo, N. Jahanshad, P. M. Thompson, B. M. Adhikari, H. Bruce, A. Van der Vaart, A. Sotiras *et al.*, Brain deficit patterns of metabolic illnesses overlap with those for major depressive disorder: A new metric of brain metabolic disease, *Human Brain Mapping* **44**, 2636 (2023).
48. N. R. Karcher, H. Modi, P. Kochunov, S. Gao and D. M. Barch, Regional vulnerability indices in youth with persistent and distressing psychoticlike experiences, *JAMA Network Open* **6**, e2343081 (2023).
49. R. M. Birn, J. B. Diamond, M. A. Smith and P. A. Bandettini, Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI, *Neuroimage* **31**, 1536 (2006).
50. J. D. Power, A. Mitra, T. O. Laumann, A. Z. Snyder, B. L. Schlaggar and S. E. Petersen, Methods to detect, characterize, and remove motion artifact in resting state fMRI, *Neuroimage* **84**, 320 (2014).
51. J. D. Kruschwitz, A. Meyer-Lindenberg, I. M. Veer, C. Wackerhagen, S. Erk, S. Mohnke, L. Pöhland, L. Haddad, O. Grimm, H. Tost *et al.*, Segregation of face sensitive areas within the fusiform gyrus using global signal regression? A study on amygdala resting-state functional connectivity, *Human Brain Mapping* **36**, 4089 (2015).
52. B. T. Yeo, J. Tandi and M. W. Chee, Functional connectivity during rested wakefulness predicts

vulnerability to sleep deprivation, *Neuroimage* **111**, 147 (2015).
53. Z. Qing, Z. Dong, S. Li, Y. Zang and D. Liu, Global signal regression has complex effects on regional homogeneity of resting state fMRI signal, *Magnetic Resonance Imaging* **33**, 1306 (2015).
54. A. Anticevic, M. W. Cole, G. Repovs, A. Savic, N. R. Driesen, G. Yang, Y. T. Cho, J. D. Murray, D. C. Glahn, X.-J. Wang *et al.*, Connectivity, pharmacology, and computation: toward a mechanistic understanding of neural system dysfunction in schizophrenia, *Frontiers in Psychiatry* **4**, p. 169 (2013).
55. A. Kottaram, L. A. Johnston, L. Cocchi, E. P. Ganella, I. Everall, C. Pantelis, R. Kotagiri and A. Zalesky, Brain network dynamics in schizophrenia: Reduced dynamism of the default mode network, *Human brain mapping* **40**, 2212 (2019).
56. J. Li, R. Kong, R. Liégeois, C. Orban, Y. Tan, N. Sun, A. J. Holmes, M. R. Sabuncu, T. Ge and B. T. Yeo, Global signal regression strengthens association between resting-state functional connectivity and behavior, *Neuroimage* **196**, 126 (2019).

Multi-modal Imaging-based Pseudotime Analysis of Alzheimer progression*

Bing He

Biomedical Engineering and Informatics, Indiana University Indianapolis, 535 W Michigan St., Indianapolis, Indiana 46202 USA
Email: hebing@iu.edu

Shu Zhang

Department of Computer Science, University of California Los Angeles 404 Westwood Plaza Engineering IV, Los Angeles, CA 90095
Email: shuzhang4@g.ucla.edu

Shannon L. Risacher and Andrew J. Saykin

Department of Radiology and Imaging Sciences, Indiana University School of Medicine, 355 W 15th St., Indianapolis, Indiana 46202 USA
Email: srisache@iu.edu and asaykin@iu.edu

Jingwen Yan

Biomedical Engineering and Informatics, Indiana University Indianapolis, 535 W Michigan St., Indianapolis, Indiana 46202 USA
Email: jingyan@iu.edu

Alzheimer's disease (AD) is a neurodegenerative disorder that results in progressive cognitive decline but without any clinically validated cures so far. Understanding the progression of AD is critical for early detection and risk assessment for AD in aging individuals, thereby enabling initiation of timely intervention and improved chance of success in AD trials. Recent pseudotime approach turns cross-sectional data into "faux" longitudinal data to understand how a complex process evolves over time. This is critical for Alzheimer, which unfolds over the course of decades, but the collected data offers only a snapshot. In this study, we tested several state-of-the-art pseudotime approaches to model the full spectrum of AD progression. Subsequently, we evaluated and compared the pseudotime progression score derived from individual imaging modalities and multi-modalities in the ADNI cohort. Our results showed that most existing pseudotime analysis tools do not generalize well to the imaging data, with either flipped progression score or poor separation of diagnosis groups. This is likely due to the underlying assumptions that only stand for single cell data. From the only tool with promising results, it was observed that all pseudotime, derived from either single imaging modalities or multi-modalities, captures the progressiveness of diagnosis groups. Pseudotime from multi-modality, but not the single modalities, confirmed the hypothetical temporal order of imaging phenotypes. In addition, we found that multi-modal pseudotime is mostly driven by amyloid and tau imaging, suggesting their continuous changes along the full spectrum of AD progression.

Keywords: AD progression; Neuroimaging; Pseudotime Analysis.

* This work is supported by NIH grants R01AG081951, R21 AG072101, U19 AG074879, U01 AG068057, and NSF CAREER 1942394.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder that results in progressive cognitive decline but without any clinically validated cures so far. Understanding the progression of AD is critical for early detection and risk assessment for AD in aging individuals, thereby enabling initiation of timely intervention and improved chance of success in AD trials. Current ATN framework (A for amyloid, T for tau, and N for neurodegeneration) used for AD classification and progression [1], however, relies on the dichotomous classification of individuals based on biomarker evidence of pathology (i.e., amyloid positive vs negative). Therefore, it is limited in capturing the full spectrum of AD progression, with early-stage individuals all treated as amyloid negative without differentiation.

Recent pseudotime approach has achieved tremendous success in modeling the dynamic process of cell differentiation [2-4]. It turns cross-sectional data into "faux" longitudinal data to understand how a complex process evolves over time. This is critical for Alzheimer, which unfolds over the course of decades, but the collected data offers only a snapshot. Pseudotime analysis has only been recently applied to AD on gene expression data [5] and tau imaging data [6]. Both studies aim to order subjects on the trajectory curve, where the relative position on the trajectory, known as pseudotime, was leveraged as the progression score. Despite some promising results, pseudotime-based imaging progression modeling remains less explored in the imaging data.

In this study, we tested several state-of-the-art pseudotime approaches to model the full spectrum of AD progression using multi-modal imaging data in the ADNI cohort. Progression scores were generated from both single modalities and multi-modal imaging data for comparison. Our results showed that existing pseudotime tools designed for single cell analysis don't generalize well to imaging data likely due to several underlying assumptions. Based on the most promising results yielded from PHATE, we found that all pseudotime, derived from either individual modalities or multi-modalities, well captured the progressiveness of diagnosis groups. Pseudotime from multi-modality, together with those found in the single modalities, confirmed the hypothetical temporal order of imaging phenotypes, like the tau spreading across Braak regions and structural brain changes observed mostly in late stage. In addition, we observed that multi-modal pseudotime is mostly driven by amyloid and tau imaging, suggesting their continuous changes along the full spectrum of AD progression. In particular, pseudotime derived from multi-modal imaging data recapitulated the acceleration point of amyloid clock [7]. Taken together, our results indicate the great potential of pseudotime and imaging in modeling AD progression, and the necessity of novel pseudotime approaches for imaging data.

2. Materials and Methods

2.1. *Participants*

Data used in this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/). The ADNI is a longitudinal study that was launched in 2003 to track the progression of AD by using clinical and cognitive tests, MRI, FDG-PET,

amyloid PET, CSF, and blood biomarkers. This study was approved by each participating site's institutional review board. For more details about this study, see the previous report[8-14].

The study population was composed of participants from the ADNI-1, ADNI-2, and ADNI-GO stages [15]. Participants were classified as cognitively normal controls (CN), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) or AD. In total, we have 1684 participants (455 CN, 321 EMCI, 551 LMCI and 357 AD) for MRI, 1041 participants (279 CN, 330 EMCI, 231 LMCI, and 201 AD) for Amyloid-PET, and 540 participants (195 CN, 152 EMCI, 106 LMCI, 87 AD) for Tau-PET. Out of those, 223 participants (95 CN, 58 EMCI, 40 LMCI, 30 AD) were found with complete set of multi- modal imaging data sets. Detailed demographic information of all participants is presented in Table. 1.

Table. 1. Demographic information of the ADNI participants

		CN	EMCI	LMCI	AD
MRI	Number	455	321	551	357
	Gender(M/F)	213/242	184/137	342/209	202/155
	Age(mean±sd)	73.96±5.98	72.10±7.22	74.26±7.44	75.17±7.79
	Educ(mean±sd)	16.46±2.60	16.05±2.67	15.94±2.87	15.27±2.88
Amyloid-PET	Number	279	330	231	201
	Gender(M/F)	131/148	188/142	136/95	121/80
	Age(mean±sd)	75.31±6.96	72.02±7.29	74.34±8.23	75.44±7.86
	Educ(mean±sd)	16.51±2.60	16.08±2.63	16.23±2.77	15.81±2.66
Tau-PET	Number	195	152	106	87
	Gender(M/F)	90/105	94/58	66/40	54/33
	Age(mean±sd)	74.24±789	74.00±7.77	74.64±8.33	77.00±8.76
	Educ(mean±sd)	16.93±2.34	16.25±2.70	16.08±2.69	15.71±2.36
Multi- modality	Number	95	58	40	30
	Gender(M/F)	38/57	37/21	24/16	16/14
	Age(mean±sd)	74.41±7.59	75.63±7.05	75.20±7.99	76.79±7.51
	Educ(mean±sd)	16.60±2.20	16.37±2.80	16.10±2.45	15.57±2.57

2.2. *Imaging data*

All available baseline structural MRI scans were downloaded from ADNI for participants. Then Freesurfer version 5.1 was used to process MRI scans and extract whole-brain and region-of-interest (ROI)-based neuroimaging endophenotypes including volumes and cortical thickness determined by automated segmentation and parcellation [16-19]. Florbetapir PET scans for amyloid and flortaucipir PET scans for tau were downloaded from the ADNI and processed as described in previous report for acquisition and processing of PET scans [20-23]. Summary ROI SUVRs from amyloid-PET scans were intensity normalized using a whole cerebellum reference region to create standardized uptake value ratio (SUVR) images. Summary ROI SUVRs from Tau-PET scans were intensity normalized using an inferior cerebellar reference region to create uptake value ratio (SUVR) images. All imaging summary measures were further adjusted for age, sex, years of education and additionally intracranial volume for MRI, using the linear regression weights derived from cognitive normal patients. The adjustment was performed for each imaging modality

separately. In total, we have 84 volume and thickness measures from MRI, amyloid SUVR of 68 cortical ROIs and tau burden in 72 ROIs. Subcortical regions were excluded for amyloid analysis since their amyloid burden has been commonly considered as non-specific and not related to AD risk.

2.3. *Pseudotime analysis*

Cross-sectional imaging data only provide a snapshot of brain traits but could not directly reflect the disease progression process. Recent emerging pseudotime analysis makes it possible to recapitulate the temporal changes from cross-sectional data. Briefly, pseudotime analysis is a dimension reduction technique built on top of patient similarity networks, instead of raw features [24]. It can return a 2D dimension disease progression trajectory following a principal curve, and the relative position of individuals on the curve is known as pseudotime, ranging from 0 to 1, which can be leveraged as progression score. Taking PHATE [25] as an example, it starts with the similarity network, followed by a diffusion process to learn global relationships. Next, it encodes these learned relationships using potential distance, which will finally go through eigen decomposition to generate a two-dimensional trajectory. A principal curve along the trajectory will be generated and all subjects were projected onto the curve for estimation of pseudotime. Subjects with similar imaging profiles (in single modalities or multi-modality data) are well connected in the similarity network and therefore are expected to stay close in the trajectory curve, leading to similar pseudotime (or progression score). The pseudotime is expected to be low for cognitive normal individuals and early-stage patients, and high for late-stage patients.

We conducted an imaging-based pseudotime analysis utilizing four established tools specifically designed for single-cell analysis. For simplicity, we did not model branches representing progression subtypes within this analysis. We tested Slingshot [2], Monocle3 [3], PHATE [25], Destiny [4] using summary measures from baseline MRI (N=1634), Amyloid-PET (N=843), Tau-PET data (N=306) and multi-modalities (N=223) respectively. However, out of 4 tools, Slingshot and Destiny often generated pseudotime with flipped directions, with AD patients much lower than cognitive normal individuals. Similarly, disease progression trajectory from Monocle3 showed extremely poor separation of disease stages. Such poor generalizability to imaging data is likely due to some assumptions underlying those tools that only stand for single cell data. Consequently, the subsequent analysis was performed only on the results from PHATE.

3. Results

3.1. *Imaging-based pseudotime captures the progressiveness of diagnosis groups*

Shown in Fig. 1 top are the disease progression trajectories (i.e., 2D embedding) generated from PHATE. For simplicity, we didn't model the branches (i.e., progression subtypes) in this study. Thereby, all disease progression trajectories follow one principal curve, with diagnosis groups mostly separated. The principal curve is overall smooth across MRI and amyloid SUVR, but not for tau and multi-modalities, likely due to small sample size. Examining each single imaging modality individually, we observed that pseudotime progression score can well capture the progressiveness of diagnosis groups, with CN group having lowest pseudotime and AD group overall having the

highest pseudotime (Fig. 1). In addition, we also observed slight variation in pseudotime patterns along progression across modalities. MRI derived pseudotime differentiates diagnosis groups well after EMCI stage, but not earlier. For amyloid pseudotime, we observed clear progression pattern across all stages. For tau, pseudotime change is subtle between EMCI and LMCI, but becomes significant after progressing to AD. Taken together, it suggests that 1) amyloid change starts early and continues until late stage, 2) there are some changes of tau in early stages, which becomes very significant in late stage, and 3) brain atrophy starts around mid-to-late stages.

We further examined the contributions of brain functional circuits to the final pseudotime progression scores. For each functional groups defined in Yeo Atlas [26], we averaged the normalized imaging measure of all member ROIs and tested its Pearson's correlation with the progression score. For amyloid and tau (Fig. 2 top), all brain regions within default mode network (DMN) made significant contributions to the progression scores, with high correlation and small variation. But for MRI (Fig. 2 top), DMN regions showed variable correlations with progression score indicating their potential inconsistent atrophy rate and pattern along progression. Across all single modalities, all brain regions consistently showed the highest correlation with amyloid progression score. This suggests the continuous and consistent amyloid changes from CN to AD, which is in line with our observation in the progression score distribution across diagnosis groups. For multi-modal pseudotime progression score (Fig. 2 bottom), it was found most associated with amyloid and then tau, and default mode network regions remain the top contributors. Compared to single modality, we observed significant change in correlation of multi-modal progression score with brain atrophy, but not with amyloid and tau. This is likely due to the fact that significant brain atrophy doesn't occur until the mid-to-late stage. Differences between patients in the early-to-mid stages are mostly captured by amyloid and tau changes, and therefore the overall multi-modal progression score is likely driven by the amyloid and tau measures.

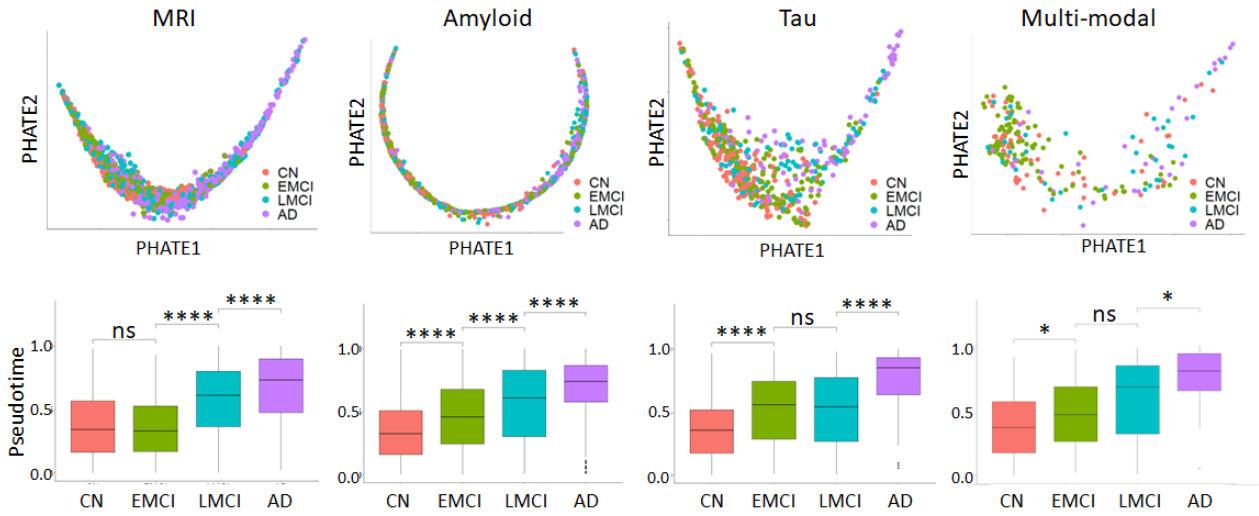


Figure. 1. Disease progression trajectory (top) and progression score distribution (bottom) derived from MRI, amyloid, tau and multi-modal brain imaging data respectively. ns: not significant with p>0.05; *: p<=0.05; **: p<=0.01; ***: p<=0.001; ****: p<=0.0001.

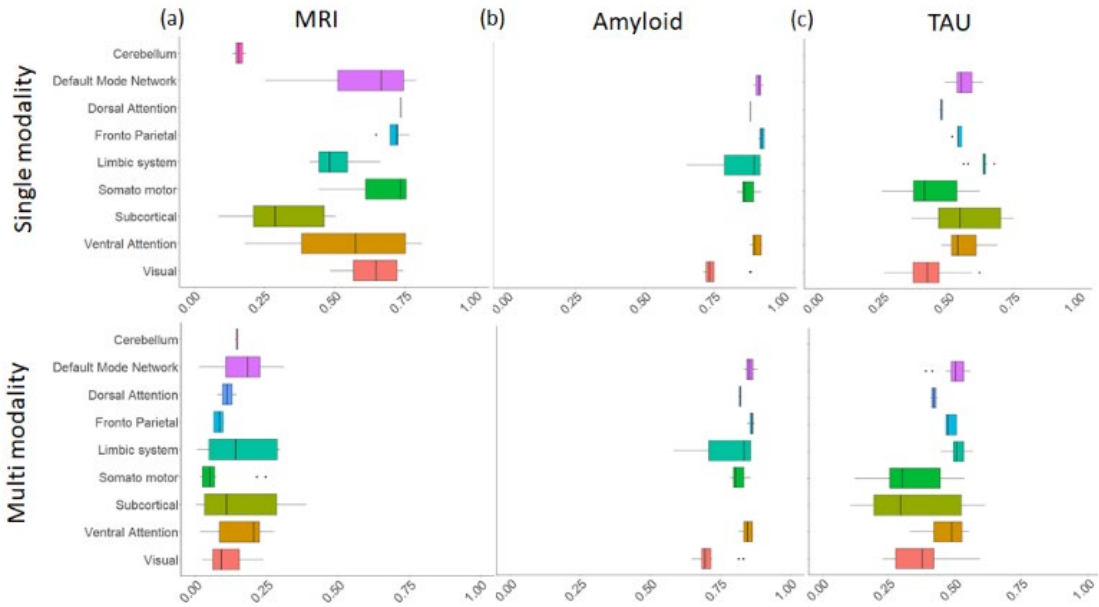


Figure 2. Correlation of brain regions with progression scores derived from single modalities (Top) and multi-modalities (Bottom). Brain regions were categorized into functional groups based on Yeo atlas.

3.2. *Multi-modal pseudotime captures the turning point of amyloid clock*

We further examined the progression of amyloid composite SUVR along the estimated pseudotime progression score. With amyloid pseudotime, we observed an approximately linear relationship, which is expected as the pseudotime is built on top of patient similarities. Interestingly, when plotted against the multi-modal pseudotime, amyloid composite SUVR showed a sigmoid like progression pattern, which is frequently observed in recent amyloid clock studies [27]. The amyloid clock maps the accumulation of amyloid-beta plaques as individuals age and helps estimate how quickly amyloid pathology progresses from initial detectable levels to more advanced stages. In line with existing findings, Fig. 3b showed a turning point after which we observed a rapid increase in the rate of amyloid-beta plaque accumulation and the turning point is around the threshold to determine amyloid positivity [27]. In addition, we examined the impact of APOE e4 status on the progression. Shown in Fig. 3c is the age distribution of subgroups stratified by APOE e4 status and multi-modal pseudotime around the turning point. Here, X axis is the pseudotime estimated from multimodal imaging data using PHATE, and Y axis is the age when subjects progress to that severity stage. APOE e4 positive groups tend to reach the amyloid turning point at a much younger age than APOE e4 negative groups, suggesting e4 allele contributes to the accelerated disease progression.

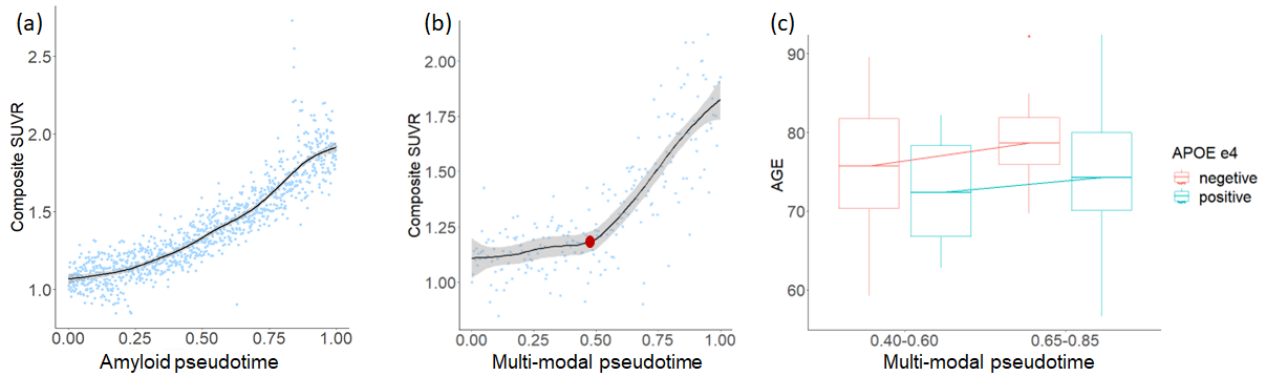


Figure. 3 Association of Amyloid composite SUVR (Y axis) with Amyloid-specific pseudotime (a) and multi-modalities-derived pseudotime (b). (c) Age distribution of subgroups stratified by *APOE* e4 status and multi-modal-derived pseudotime progression score around the turning point.

3.3. *Multi-modality pseudotime confirmed the hypothetical temporal order of imaging phenotypes.*

We further examined the pseudotime from combined MRI, Amyloid-PET and Tau-PET measures. In Fig. 4 is the fitted curve showing the progression pattern of brain-wide atrophy, amyloid and Tau deposition along the estimated pseudotime derived from multi-modal imaging data. Brain-wide atrophy was calculated as the average of normalized thickness and volume across the brain, and similarly for brain-wide amyloid deposition. For Tau, which is typically localized, we averaged the Tau deposition across brain regions associated with Braak stage 1, 3/4, and 5/6 respectively (1 for early stage and 6 for late stage). Hippocampal regions associated with Braak stage 2 were contaminated with off binding in the Tau-PET and therefore not included in the results. Fig. 4 confirmed our findings from single modalities. It shows that noticeable changes in brain atrophy (MRI) only start after mid-to-late stages. Acceleration of amyloid accumulation (pseudotime around 0.5) starts earlier than that of tau accumulation and atrophy. It also validated the tau spreading pattern inside the brain, where tau starts to accumulate in Braak 1 regions, then spread to Braak 3/4 and Braak 5/6 regions. In line with previous findings that Tau-PET mirrors the regional patterns of neurodegeneration (i.e., brain atrophy observed on MRI) [28], we also found that overall tau pathology progression is tightly linked to atrophy with a similar acceleration point (multi-modal pseudotime around 0.75). However, early-stage changes observed in tau deposition, though subtle, did not occur in MRI. Findings from both single modality and multi-modality pseudotime align perfectly with existing hypothetical progression models, suggesting the great potential of pseudotime analysis in AD progression modeling [29-32].

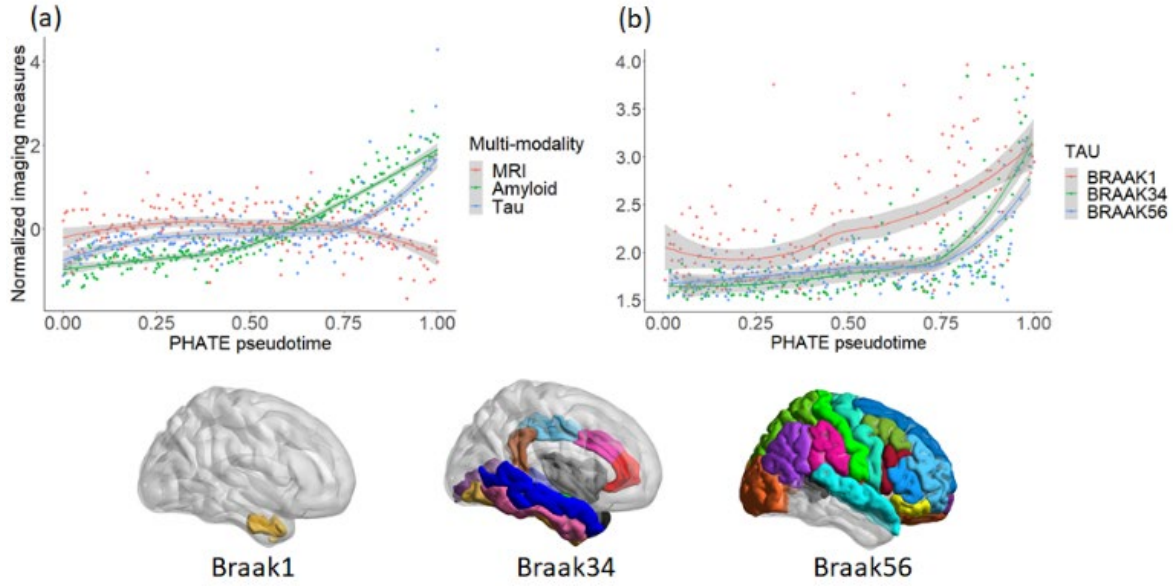


Figure. 4. Temporal ordering of multi-modal summary measures (Left) and Braak regions (Right) with pseudotime derived from multi-modal imaging data. Bottom: brain regions associated with different Braak stages. Hippocampal regions associated with Braak stage 2 were contaminated with off binding in the Tau-PET and therefore not included in the results.

4. Discussion

In this study we investigated imaging-based continuous disease progression using state-of-the-art pseudotime analysis tools. We examined the progression patterns captured from single imaging modalities and multi-modalities respectively. Imaging-based pseudotime progression score could well capture the progressiveness of diagnosis groups and hypothetical temporal order of dynamic brain changes, which were built on top of clinical observations. Notably, the multi-modal progression scores not only reflected the severity of the disease but also the rate of progression, revealing an amyloid acceleration point consistent with recent amyloid clock studies. Additionally, we validated the impact of APOE e4 status on AD progression, showing that e4-positive individuals reach the amyloid acceleration point at a significantly younger age compared to e4-negative individuals. These results suggested the potential of pseudotime approaches to model AD progression as a continuous process and could be utilized to supplement the current A/T/N framework for AD progression quantification.

Despite these encouraging results, imaging-based pseudotime progression modeling remains underexplored. Several limitations warrant further investigation and improvement. Firstly, our study relied on cross-sectional data, not able to capture the short-term progression patterns that longitudinal data could provide. Secondly, linking estimated pseudotime to chronological age or years remains challenging, which limits the interpretation and clinical utility of the progression scores. Our results also demonstrated the overall poor generalizability of existing pseudotime analysis tools to imaging data, underscoring the need for the development of new tools tailored to imaging applications.

References

1. Ebenau, J.L., T. Timmers, L.M.P. Wesselman, I.M.W. Verberk, S.C.J. Verfaillie, R.E.R. Slot, et al., *ATN classification and clinical progression in subjective cognitive decline: The SCIENCe project.* Neurology, 2020. **95**(1): p. e46-e58.
2. Street, K., D. Risso, R.B. Fletcher, D. Das, J. Ngai, N. Yosef, et al., *Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics.* BMC Genomics, 2018. **19**(1): p. 477.
3. Qiu, X., Q. Mao, Y. Tang, L. Wang, R. Chawla, H.A. Pliner, et al., *Reversed graph embedding resolves complex single-cell trajectories.* Nat Methods, 2017. **14**(10): p. 979-982.
4. Angerer, P., L. Haghverdi, M. Buttner, F.J. Theis, C. Marr, and F. Buettner, *destiny: diffusion maps for large-scale single-cell data in R.* Bioinformatics, 2016. **32**(8): p. 1241-3.
5. Mukherjee, S., L. Heath, C. Preuss, S. Jayadev, G.A. Garden, A.K. Greenwood, et al., *Molecular estimation of neurodegeneration pseudotime in older brains.* Nat Commun, 2020. **11**(1): p. 5781.
6. Hong, J., S.K. Kang, I. Alberts, J. Lu, R. Sznitman, J.S. Lee, et al., *Image-level trajectory inference of tau pathology using variational autoencoder for Flortaucipir PET.* Eur J Nucl Med Mol Imaging, 2022. **49**(9): p. 3061-3072.
7. Schindler, S.E., Y. Li, V.D. Buckles, B.A. Gordon, T.L.S. Benzinger, G. Wang, et al., *Predicting Symptom Onset in Sporadic Alzheimer Disease With Amyloid PET.* Neurology, 2021. **97**(18): p. e1823-e1834.
8. Jack Jr, C.R., M.A. Bernstein, B.J. Borowski, J.L. Gunter, N.C. Fox, P.M. Thompson, et al., *Update on the magnetic resonance imaging core of the Alzheimer's disease neuroimaging initiative.* Alzheimer's & Dementia, 2010. **6**(3): p. 212-220.
9. Jagust, W.J., D. Bandy, K. Chen, N.L. Foster, S.M. Landau, C.A. Mathis, et al., *The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core.* Alzheimer's & Dementia, 2010. **6**(3): p. 221-229.
10. Trojanowski, J.Q., H. Vandeerstichele, M. Korecka, C.M. Clark, P.S. Aisen, R.C. Petersen, et al., *Update on the biomarker core of the Alzheimer's Disease Neuroimaging Initiative subjects.* Alzheimer's & dementia, 2010. **6**(3): p. 230-238.
11. Petersen, R.C., P. Aisen, L.A. Beckett, M. Donohue, A. Gamst, D.J. Harvey, et al., *Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization.* Neurology, 2010. **74**(3): p. 201-209.
12. Saykin, A.J., L. Shen, T.M. Foroud, S.G. Potkin, S. Swaminathan, S. Kim, et al., *Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans.* Alzheimer's & dementia, 2010. **6**(3): p. 265-273.
13. Weiner, M.W., P.S. Aisen, C.R. Jack Jr, W.J. Jagust, J.Q. Trojanowski, L. Shaw, et al., *The Alzheimer's disease neuroimaging initiative: progress report and future plans.* Alzheimer's & Dementia, 2010. **6**(3): p. 202-211. e7.
14. Saykin, A.J., L. Shen, X. Yao, S. Kim, K. Nho, S.L. Risacher, et al., *Genetic studies of quantitative MCI and AD phenotypes in ADNI: progress, opportunities, and plans.* Alzheimer's & Dementia, 2015. **11**(7): p. 792-814.
15. Weiner, M.W., D.P. Veitch, P.S. Aisen, L.A. Beckett, N.J. Cairns, R.C. Green, et al., *The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception.* Alzheimer's & Dementia, 2013. **9**(5): p. e111-e194.

16. Risacher, S.L., S. Kim, K. Nho, T. Foroud, L. Shen, R.C. Petersen, et al., *APOE effect on Alzheimer's disease biomarkers in older adults with significant memory concern.* Alzheimer's & Dementia, 2015. **11**(12): p. 1417-1429.
17. Dale, A.M., B. Fischl, and M.I. Sereno, *Cortical surface-based analysis: I. Segmentation and surface reconstruction.* Neuroimage, 1999. **9**(2): p. 179-194.
18. Fischl, B., M.I. Sereno, and A.M. Dale, *Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system.* Neuroimage, 1999. **9**(2): p. 195-207.
19. Nho, K., A. Kueider-Paisley, S. Ahmad, S. MahmoudianDehkordi, M. Arnold, S.L. Risacher, et al., *Association of altered liver enzymes with Alzheimer disease diagnosis, cognition, neuroimaging measures, and cerebrospinal fluid biomarkers.* JAMA network open, 2019. **2**(7): p. e197978-e197978.
20. Joshi, A.D., M.J. Pontecorvo, C.M. Clark, A.P. Carpenter, D.L. Jennings, C.H. Sadowsky, et al., *Performance characteristics of amyloid PET with florbetapir F 18 in patients with Alzheimer's disease and cognitively normal subjects.* Journal of Nuclear Medicine, 2012. **53**(3): p. 378-384.
21. Landau, S., B. Thomas, L. Thurfjell, M. Schmidt, R. Margolin, M. Mintun, et al., *Amyloid PET imaging in Alzheimer's disease: a comparison of three radiotracers.* European journal of nuclear medicine and molecular imaging, 2014. **41**(7): p. 1398-1407.
22. Maass, A., S. Landau, S.L. Baker, A. Horng, S.N. Lockhart, R. La Joie, et al., *Comparison of multiple tau-PET measures as biomarkers in aging and Alzheimer's disease.* Neuroimage, 2017. **157**: p. 448-463.
23. Schöll, M., S.N. Lockhart, D.R. Schonhaut, J.P. O'Neil, M. Janabi, R. Ossenkoppele, et al., *PET imaging of tau deposition in the aging human brain.* Neuron, 2016. **89**(5): p. 971-982.
24. Campbell, K.R. and C. Yau, *Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data.* Nature communications, 2018. **9**(1): p. 1-12.
25. Moon, K.R., D. van Dijk, Z. Wang, S. Gigante, D.B. Burkhardt, W.S. Chen, et al., *Visualizing structure and transitions in high-dimensional biological data.* Nature biotechnology, 2019. **37**(12): p. 1482-1492.
26. Yeo, B.T., F.M. Krienen, J. Sepulcre, M.R. Sabuncu, D. Lashkari, M. Hollinshead, et al., *The organization of the human cerebral cortex estimated by intrinsic functional connectivity.* J Neurophysiol, 2011. **106**(3): p. 1125-65.
27. Therneau, T.M., D.S. Knopman, V.J. Lowe, H. Botha, J. Graff-Radford, D.T. Jones, et al., *Relationships between β-amyloid and tau in an elderly population: An accelerated failure time model.* Neuroimage, 2021. **242**: p. 118440.
28. Sirkis, D.W., L.W. Bonham, T.P. Johnson, R. La Joie, and J.S. Yokoyama, *Dissecting the clinical heterogeneity of early-onset Alzheimer's disease.* Mol Psychiatry, 2022. **27**(6): p. 2674-2688.
29. Bejanin, A., D.R. Schonhaut, R. La Joie, J.H. Kramer, S.L. Baker, N. Sosa, et al., *Tau pathology and neurodegeneration contribute to cognitive impairment in Alzheimer's disease.* Brain, 2017. **140**(12): p. 3286-3300.
30. Berron, D., J.W. Vogel, P.S. Insel, J.B. Pereira, L. Xie, L.E.M. Wisse, et al., *Early stages of tau pathology and its associations with functional connectivity, atrophy and memory.* Brain, 2021. **144**(9): p. 2771-2783.

31. Jack, C.R., D.S. Knopman, W.J. Jagust, R.C. Petersen, M.W. Weiner, P.S. Aisen, et al., *Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers.* Lancet Neurology, 2013. **12**(2): p. 207-216.
32. Jack, C.R., Jr., D.S. Knopman, W.J. Jagust, L.M. Shaw, P.S. Aisen, M.W. Weiner, et al., *Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade.* Lancet Neurol, 2010. **9**(1): p. 119-28.

A Dynamic Model for Early Prediction of Alzheimer's Disease by Leveraging Graph Convolutional Networks and Tensor Algebra

Cagri Ozdemir[1,3,4,†],Mohammad Al Olaimat[1,3,4], Serdar Bozdag[1,2,3,4,†], and Alzheimer's Disease Neuroimaging Initiative*

[1]*Department of Computer Science and Engineering, University of North Texas, TX, USA*

[2]*Department of Mathematics, University of North Texas, TX, USA*

[3]*BioDiscovery Institute, University of North Texas, TX, USA*

[4] *Center for Computational Life Sciences, University of North Texas, TX, USA*
[†]*E-mail: cagri.ozdemir@unt.edu, serdar.bozdag@unt.edu*

Data used in preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

Alzheimer's disease (AD) is a neurocognitive disorder that deteriorates memory and impairs cognitive functions. Mild Cognitive Impairment (MCI) is generally considered as an intermediate phase between normal cognitive aging and more severe conditions such as AD. Although not all individuals with MCI will develop AD, they are at an increased risk of developing AD. Diagnosing AD once strong symptoms are already present is of limited value, as AD leads to irreversible cognitive decline and brain damage. Thus, it is crucial to develop methods for the early prediction of AD in individuals with MCI. Recurrent Neural Networks (RNN)-based methods have been effectively used to predict the progression from MCI to AD by analyzing electronic health records (EHR). However, despite their widespread use, existing RNN-based tools may introduce increased model complexity and often face difficulties in capturing long-term dependencies. In this study, we introduced a novel **Dy**namic deep learning model for **E**arly **P**rediction of **AD** (DyEPAD)* to predict MCI subjects' progression to AD utilizing EHR data. In the first phase of DyEPAD, embeddings for each time step or visit are captured through Graph Convolutional Networks (GCN) and aggregation functions. In the final phase, DyEPAD employs tensor algebraic operations for frequency domain analysis of these embeddings, capturing the full scope of evolutionary patterns across all time steps. Our experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) and National Alzheimer's Coordinating Center (NACC) datasets demonstrate that our proposed model outperforms or is in par with the state-of-the-art and baseline methods.

Keywords: Alzheimer's disease, early prediction, dynamic graphs, tensor algebra.

*The source code is available at `https://github.com/bozdaglab/DyEPAD`

1. Introduction

Throughout the past few decades Alzheimer's disease (AD), once thought to be a rare disorder, has gained recognition as a major public health concern.[1,2] According to the survey data,[3] AD affected more than 30 million people in 2015, and it is estimated that this number could surpass 114 million by 2050. AD causes an irreversible decline in memory, mood, and behavior, along with difficulties with everyday tasks and other cognitive challenges. Mild Cognitive Impairment (MCI) is a condition characterized by observable cognitive decline, which is generally considered as not sufficient to effect patients' daily functioning. More importantly, MCI serves as a critical stage for identifying individuals at risk of developing AD. Individuals with MCI are at a higher risk of progressing to AD, with an annual progression rate between 10% and 20%.[4] Although, to date, there is no complete cure for AD, there are treatments to slow AD-related symptoms at their early stages. Therefore, to slow down AD progression and avoid its worst effects, it is crucial to develop methods for the early prediction of AD in individuals with MCI.

Many early prediction tools for AD mainly rely on image data, making use of advanced imaging technologies like MRI, PET scans, and CT scans.[2,5–8] However, even though image-based approaches provide valuable insights, imaging is an expensive method and is not easily accessible particularly for people in developing countries. Electronic health records (EHR) consist of temporal sequences of clinical features. The longitudinal nature of EHR enables the examination of patients' medical history trajectories. These records have been utilized to train machine learning (ML) models for classifying and clustering patient data, enhancing clinical decision-making.[9,10] However, traditional ML methods (e.g., Random Forest and SVM) fail to account for the temporal dependencies in the data sequences.[11] An effective method for capturing the temporal patterns in sequential data is Recurrent Neural Networks (RNN). However, irregular time intervals between consecutive inputs (i.e., clinical visits of patients), a common occurrence in EHR, pose a challenge for RNN models.[12] When intervals vary, it disrupts the model's ability to effectively capture temporal dependencies and may lead to suboptimal performance. To address this challenge, Time-aware long short-term memory (T-LSTM)[13] has been introduced. T-LSTM modifies LSTM architecture to address challenges arising from irregular time intervals between clinical visits. Another computational tool, named Predicting Progression of Alzheimer's Disease (PPAD),[14] utilizes an RNN component where patients' ages at the time of clinical visits were utilized to handle varying time intervals between clinical visits. More recently, time-aware RNN (TA-RNN) has been presented for early prediction of AD.[15] TA-RNN utilizes a time embedding layer that incorporates elapsed time between consecutive visits to address lack of consideration of irregular time intervals between consecutive inputs by RNN models.

To enhance graph analysis, Graph Convolutional Networks (GCN)[16] has been introduced as a more efficient variant of Graph Neural Networks (GNN). GCN utilizes a convolution operation that aggregates information from multiple hops of neighbors. While GCN treats all neighboring nodes equally during aggregation, Graph Attention Networks (GAT)[17] employs an attention mechanism to learn the importance of each neighboring node. In the context of early diagnosis of AD, a GNN-based method has been introduced that constructs patient-patient

graphs using image features from both MRI and PET scans.[18] In addition, an interpretable dynamic graph convolutional networks (IDGCN) integrates dynamic graph leaning into a GCN architecture to improve the performance of personalized diagnosis for AD and provide interpretable results.[19]

Even though these existing tools offer RNN-based solutions for the early diagnosis of AD using EHR, they have some limitations. In all these approaches, RNN parameters are updated using the entire sequence of time steps. While this methodology can capture sequential patterns, it often leads to increased complexity and potential issues with learning long-term dependencies due to RNN's structural design. As sequences get longer, they tend to forget earlier information, making it hard to capture patterns over extended periods. Additionally, longitudinal data may have hierarchical temporal structures, such as monthly and yearly patterns. RNN units often struggle to effectively capture these hierarchies. Furthermore, for long sequences, RNN layers often encounter vanishing or exploding gradient problems, which complicate the training process.

To address these limitations, in this study, we introduced a novel **Dy**namic deep learning model for **E**arly **P**rediction of **AD** (DyEPAD). DyEPAD consists of a two-phase training process. In the first phase, DyEPAD learns latent representations (i.e., embeddings) of patients at each clinical visit. For this, for each visit a patient similarity network is constructed. In each graph, the nodes represent patients with attributes derived from the corresponding clinical visit, and the edges capture the similarities between patients based on these visit attributes. At each time step, node embeddings of the corresponding graph are learned using a GCN layer and an aggregation function, which incorporates the node embeddings of the graph in the previous visit. In the second phase of DyEPAD, spatiotemporal tensor is built by stacking the embeddings learned in the first phase. Then, tensorial functions are employed to capture full scope of evolutionary pattern in the data by mapping it into a non-linear feature space and utilizing frequency domain representations.

We presented our experimental results on the Alzheimer's Disease Neuroimaging Initiative (ADNI)[20] and National Alzheimer's Coordinating Center (NACC)[21] datasets to predict AD diagnosis at the next visit and multiple visits ahead. Our experimental results show that our proposed model outperforms or is in par with the state-of-the-art and baseline methods.

2. Methods

2.1. *Preliminaries: Overview of Tensor Algebra*

Multidimensional data is defined as arrays of numbers organized in more than two dimensions, commonly known as *tensors*.[22] Unlike traditional data structures such as vectors (1D) or matrices (2D), tensors extend to higher dimensions, allowing for more complex data representations. The dimensions of a tensor are called ways or modes. The number of modes determines the order of a tensor. If, for example, $\mathcal{A} \in \mathbb{R}^{m \times n \times \ell}$, then $\mathcal{A}$ is a third-order tensor. Here, m could represent time, n could represent different patients, and ℓ could represent various clinical measurements. The definitions presented in this section are fundamentally based on recent advancements in Fourier theory and the algebra of circulants,[23–28] which provide powerful tools for analyzing multidimensional data. By using tensor-based methods, we can

better capture and model the intricate relationships across multiple modes, which are often lost in simpler, lower-dimensional representations like vectors or matrices.

It will be useful to divide a third-order tensor $\mathcal{A}$ into different slices and tubal elements as shown in Fig. 1. In Python notation, $\mathcal{A}^{(i)} \equiv \mathcal{A}[i,:,:]$ refers to the ith frontal slice; $\mathcal{A}_{(i)} \equiv \mathcal{A}[:,i,:]$ refers to the $i$the horizontal slice; and $\vec{\mathcal{A}}_{(i)} \equiv \mathcal{A}[:,:,i]$ refers to the ith lateral slice.

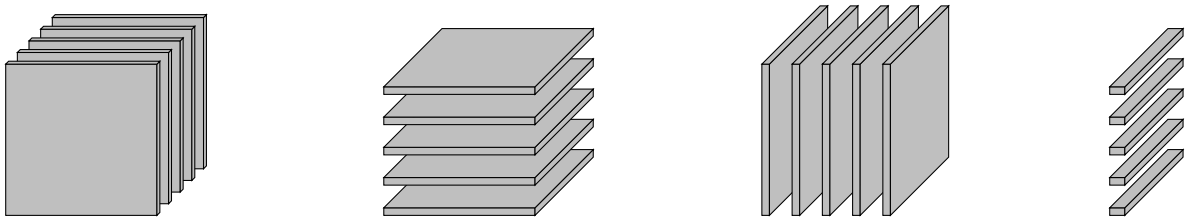


Fig. 1. (a) Frontal, (b) horizontal, (c) lateral slices, and (d) tubal scalars of a third-order tensor.

Definition 1. An element $\vec{\mathbf{a}}_i^j \in \mathbb{R}^{n\times 1\times 1}$ is called a **tubal scalar** of length n. $\vec{\mathbf{a}}_i^j \equiv A[:,j,i]$ refers to the jth tubal scalar of ith lateral slice.

Definition 2. Let $\mathcal{A} \in \mathbb{R}^{m_1\times m_2\times m_3}$ be a third-order tensor. Then $\texttt{unfold}(\mathcal{A})$ maps the tensor $\mathcal{A}$ into a $m_1 \times (m_2 \cdot m_3)$ matrix by stacking all the tubal scalars as the columns of the resultant matrix. The operation that takes $\texttt{unfold}(\mathcal{A})$ back to tensor form is the `fold` command:

$$\mathcal{A} = \texttt{fold}\big(\texttt{unfold}(\mathcal{A})\big).$$

Definition 3. Let $\mathcal{A} \in \mathbb{R}^{m_1\times m_2\times m_3}$ be a third-order tensor, and let $\mathbf{B} \in \mathbb{R}^{m_1\times m_1}$ be a matrix. The mode-1 product of $\mathcal{A}$ and $\mathbf{B}$, denoted $\mathcal{A} \times_1 \mathbf{B}$, is defined as:

$$\mathcal{A} \times_1 \mathbf{B} = \texttt{fold}\big(\mathbf{B} * \texttt{unfold}(\mathcal{A})\big),$$

where $*$ denotes matrix multiplication.

Definition 4. The face-wise product multiplies each of the frontal slices of two tensors. Let $\mathcal{A} \in \mathbb{R}^{m\times n\times k}$ and $\mathcal{B} \in \mathbb{R}^{m\times k\times \ell}$ be third-order tensors. Then the face-wise product $\mathcal{C} = \mathcal{A}\Delta\mathcal{B} \in \mathbb{R}^{m\times n\times \ell}$ is defined by performing matrix multiplication between the corresponding frontal slices of $\mathcal{A}$ and $\mathcal{B}$ as follows:

$$\begin{aligned}&\mathcal{C} = \mathcal{A}\Delta\mathcal{B},\\ &\mathcal{C}^{(i)} = \mathcal{A}^{(i)} * \mathcal{B}^{(i)} \ \text{ for } i = 1,\ldots,m.\end{aligned}$$

2.2. *Fundamental Tensor Operations*

From a theoretical perspective, it is well known that block circulant matrices can be block diagonalized by using the Fourier transform.[29] Therefore, the multiplication, transpose, and inverse operations on tensors were defined based on block circulant matrices and the Fourier transform.[23–26,30] Most recently, it was shown that these tensor operators can be effectively

defined by performing an invertible linear transform along all tubal scalars of tensors, conducting pair-wise matrix multiplications for all frontal slices of the tensors in the transform domain.[31] To this end, we will define tensor operators in the so-called transform domain. The "$_L$" subscript is used to represent any invertible linear transformation.

Definition 5. Let $\mathcal{A} \in \mathbb{R}^{m\times n\times k}$ and $\mathcal{B} \in \mathbb{R}^{m\times k\times \ell}$ be third-order tensors. The tensor-tensor product based on L transform, denoted $\mathcal{A} \circ_L \mathcal{B} \in \mathbb{R}^{m\times n\times \ell}$, is defined as:

$$\begin{aligned}
&\tilde{\mathcal{A}} = \mathcal{A} \times_1 \mathbf{L},\\
&\tilde{\mathcal{B}} = \mathcal{B} \times_1 \mathbf{L},\\
&\mathcal{A} \circ_L \mathcal{B} = (\tilde{\mathcal{A}} \Delta \tilde{\mathcal{B}}) \times_1 \mathbf{L}^{-1},
\end{aligned}$$

where $\mathbf{L}$ is an $m \times m$ invertible transformation matrix. $\mathbf{L}^{-1}$ is the inverse of the transformation matrix. "$\times_1$" is the mode-1 product and "Δ" is the face-wise product given in **Definition 3** and **Definition 4**, respectively.

Definition 6. If $\mathcal{A} \in \mathbb{R}^{m\times n\times k}$, then the tensor transpose, denoted $transpose_L(\mathcal{A}) \in \mathbb{R}^{m\times k\times n}$, is defined by taking matrix transpose of each frontal slice of $\mathcal{A}$ in the transform domain as following:

$$\begin{aligned}
&\mathcal{B} = transpose_L(\mathcal{A}),\\
&\tilde{\mathcal{A}} = \mathcal{A} \times_1 \mathbf{L},\\
&\tilde{\mathcal{B}}^{(i)} = (\tilde{\mathcal{A}}^{(i)})^T \text{ for } i = 1, \ldots, m,\\
&\mathcal{B} = \tilde{\mathcal{B}} \times_1 \mathbf{L}^{-1},
\end{aligned}$$

where "T" denotes matrix transpose.

2.3. *The Proposed Method*

Our proposed method, DyEPAD (Fig. 2), employs GCN layers to extract node (patient) embeddings from graph-structured EHR data. Each time step (visit) is trained in static network manner, meaning that the GNN parameters are updated independently based on a specific loss function for each time step. By doing so, the complexities of training an RNN unit are avoided. Furthermore, our model incorporates advanced designs, such as dropout, batch normalization, and mini- batch, which are present in static GNN-based learning methods.[32–34] To capture the evolutionary patterns of the patient embeddings across all time points in EHR, these embeddings are subsequently subjected to tensor algebraic operations for frequency domain analysis.

2.4. *Graph Convolutional Networks and Embedding Aggregation*

In DyEPAD, we construct a patient similarity network for each clinical visit. In this network, nodes represent patients, and edges encode the similarities between patients based on their EHR for that specific visit. Additionally, their EHR are also assigned as node features. For

each clinical visit at time t, DyEPAD utilizes a GCN module to learn node embeddings as follows:

$$H_t = \sigma(D_t^{-1/2} A_t D_t^{-1/2} X_t W_t), \tag{1}$$

for $t = 1, 2, ..., n$, where n is the total number of graphs (visits), and $X_t \in \mathbb{R}^{m \times d}$ is the feature matrix of nodes (m is the number of nodes and d is the feature size). $A_t \in \mathbb{R}^{m \times m}$ and $D_t \in \mathbb{R}^{m \times m}$ are the adjacency and the node degree matrices, respectively. $W_t \in \mathbb{R}^{d \times \ell}$ is the learnable weight matrix, σ is the activation function, and $H_t \in \mathbb{R}^{m \times \ell}$ is the node embeddings matrix. It is important to note that the hidden layer size ℓ is determined by the column size of W_t. The adjacency matrix $A_t \in \mathbb{R}^{m \times m}$ was constructed using k-nearest neighbors (k was set to 5) based on cosine similarities between patients' EHR.

GCN layers of DyEPAD capture embeddings in a given graph-structured EHR at time step (visit) t. To update the embeddings learned by the GCN layer based on the embeddings

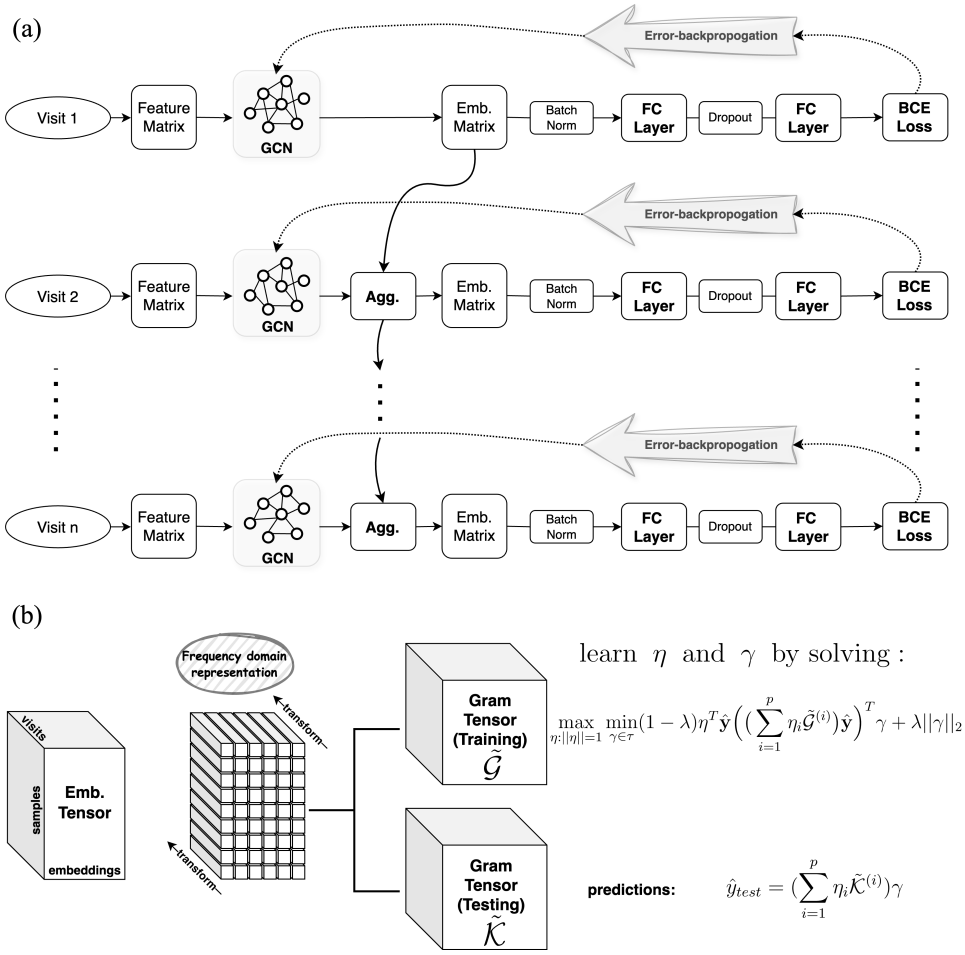


Fig. 2. Graphical illustration of DyEPAD. (a) DyEPAD utilizes GCN layers to derive node embeddings from graph-structured EHR data, and subsequently employs aggregation layers (**Agg.**) to aggregate the current embedding with those from the most recent previous visit. (b) The derived embeddings are then processed through tensor algebraic operations for frequency domain analysis, capturing the complete range of evolutionary patterns.

of the most recent previous visit, DyEPAD employs an aggregation function. This function makes the model dynamic by allowing it to adapt and update based on the most recent visit, thus effectively capturing and reflecting the evolving nature of the data over time. In traditional RNN architectures, Gated Recurrent Units (GRU) is used to process sequential data by updating hidden states across all time steps. In our work, GRU was used in a non-traditional way. Instead of processing sequences as part of an RNN, we applied GRU to aggregate the current embedding with those from the most recent previous visit as follows:

$$\begin{aligned} R_t &= sigmoid(H_t W_{ir} + \bar{H}_{t-1} W_{hr}), \\ U_t &= sigmoid(H_t W_{iz} + \bar{H}_{t-1} W_{hz}), \\ N_t &= tanh\big(H_t W_{in} + R_t \otimes (\bar{H}_{t-1} W_{hn})\big), \\ \bar{H}_t &= \big((1 - U_t) \otimes N_t\big) + \big(U_t \otimes \bar{H}_{t-1}\big), \end{aligned} \tag{2}$$

where $\otimes$ denotes element-wise multiplication, $R_t \in \mathbb{R}^{m \times \ell}$ is the reset gate, $U_t \in \mathbb{R}^{m \times \ell}$ is the update gate, and $N_t \in \mathbb{R}^{m \times \ell}$ is the new state matrices for a given time step t. W_{ir} and W_{hr} are the parameters for the reset gate. W_{iz} and W_{hz} are the parameters for the update gate. W_{in} and W_{hn} are the parameters for the new state. $\bar{H}_t \in \mathbb{R}^{m \times \ell}$ is the updated embeddings for given input embeddings $H_t \in \mathbb{R}^{m \times \ell}$ (current state embeddings) and $\bar{H}_{t-1} \in \mathbb{R}^{m \times \ell}$ (previous updated state embeddings).

We then employ two fully connected layers as follows:

$$p = sig\big(\sigma(\bar{H}_t \bar{W}_1)\bar{W}_2\big), \tag{3}$$

where σ is the activation function for the first layer, and sig denotes the sigmoid activation function.

To learn model parameters, we use binary cross entropy (BCE) loss function for each time point. The loss function for a single prediction can be defined as:

$$Loss = -\big(ylog(p) + (1 - y)log(1 - p)\big), \tag{4}$$

where y is the ground truth binary label (0 denotes MCI, and 1 represents AD labels), and p is the predicted probability. Adam optimization[35] is used as the state-of-the-art for stochastic gradient descent algorithm.

2.5. *Spatiotemporal Tensor Representation of Embeddings*

The patient embeddings learned by GCN and GRU in the previous step can be structured as a spatiotemporal tensor, where the dimensions correspond to time, patients, and patient embeddings as shown in Fig. 2(b). This spatiotemporal tensor preserves the intrinsic correlations present in the data while enabling to capture complex patterns across multiple dimensions.

$$\mathcal{H}^{(t)} = \bar{H}_t \;\text{ for }\; t = 1, 2, \ldots, n. \tag{5}$$

$\mathcal{H} \in \mathbb{R}^{n \times m \times \ell}$ is a spatiotemporal tensor where each frontal slice ($\mathcal{H}^{(t)}$) is the patient embedding matrix at time t (see Eq. (2)). Since there is no aggregation unit at time step 1 (Fig. 2(a)),

$\bar{H}_1$ is equal to H_1 (see Eq. (1)). To process the spatiotemporal tensor, we utilize gram tensors as described in **Theorem 1** as follow:

Theorem 1. *Let $\mathcal{H}_{(i)} \in \mathbb{R}^{n\times 1\times \ell}$ and $\mathcal{H}_{(j)} \in \mathbb{R}^{n\times 1\times \ell}$ be horizontal slices of the spatiotemporal tensor $\mathcal{H} \in \mathbb{R}^{n\times m\times \ell}$ (Eq. (5)). The gram tensor $\mathcal{G} \in \mathbb{R}^{n\times m\times m}$ for third-order tensors is constructed as follows:*

$$\mathbf{k}\big(\mathcal{H}_{(i)}, \mathcal{H}_{(j)}\big) = \big(\mathcal{H}_{(i)} \circ_L transpose_L(\mathcal{H}_{(j)})\big)^d,$$
$$\vec{\mathbf{g}}_i^j = \mathbf{k}\big(\mathcal{H}_{(i)}, \mathcal{H}_{(j)}\big).$$

*The kernel function applied to $\mathcal{H}_{(i)}$ and $\mathcal{H}_{(j)}$, denoted as $\mathbf{k}\big(\mathcal{H}_{(i)}, \mathcal{H}_{(j)}\big)$, results in $\vec{\mathbf{g}}_i^j$ which is a tubal scalar of $\mathcal{G}$ defined in **Definition 1**. This tensorial polynomial function was built upon on the inner product of two horizontal slices. We used the tensor-tensor multiplication (**Definition 5**) and the tensor transpose operation (**Definition 6**).*

Proof. As each frontal slice of the gram tensor is a kernel matrix in the transform domain.

$$\tilde{\mathcal{G}} = \mathcal{G} \times_1 \mathbf{L} \text{ (transform domain representation)},$$

It has been demonstrated that the quadratic form $\alpha^T \tilde{\mathcal{G}}^{(t)} \alpha$ is non-negative for all vectors $\alpha \in \mathbb{R}^m$.[36,37] As this implies that the frontal slices $(\tilde{\mathcal{G}}^{(t)})$ are positive semi-definite in the transform domain, $\tilde{\mathcal{G}}$ is a collection of positive-definite gram matrices. □

Each horizontal slice of a spatiotemporal tensor represents a patient. Each tubal scalars of a spatiotemporal tensor provides a sequence of embeddings over time. In matrix algebra, inner product of two vector gives a scalar. Each sample (horizontal slice) of the spatiotemporal tensor can be vectorized and a downstream task can be applied using matrix algebra. However, vectorizing samples destroys the spatial and temporal correlation within each sample. In our case, inner product of two horizontal slices provides a tubal scalar, as shown in **Theorem 1**. By doing so, we keep all the spatial and temporal correlation within each sample while computing the inner product between samples.

To learn non-linear patterns by implicitly mapping data into a higher-dimensional space, we need to construct gram tensors. Let $\mathcal{G}_{(1)}, \mathcal{G}_{(2)}, \cdots, \mathcal{G}_{(q)}$ be selected q horizontal slices of the spatiotemporal tensor $\mathcal{H}$ (Eq. (5)) for training. Similarly, let $\mathcal{K}_{(1)}, \mathcal{K}_{(2)}, \cdots, \mathcal{K}_{(w)}$ be selected w horizontal slices of the spatiotemporal tensor $\mathcal{H}$ for testing. Gram tensors for training, $\mathcal{G} \in \mathbb{R}^{n\times q\times q}$, and for testing, $\mathcal{K} \in \mathbb{R}^{n\times w\times q}$, can be constructed as follows:

$$\mathcal{G} = \begin{bmatrix} \mathbf{k}(\mathcal{G}_{(1)}, \mathcal{G}_{(1)}) & \mathbf{k}(\mathcal{G}_{(2)}, \mathcal{G}_{(1)}) & \cdots \mathbf{k}(\mathcal{G}_{(q)}, \mathcal{G}_{(1)}) \\ \mathbf{k}(\mathcal{G}_{(1)}, \mathcal{G}_{(2)}) & \mathbf{k}(\mathcal{G}_{(2)}, \mathcal{G}_{(2)}) & \cdots \mathbf{k}(\mathcal{G}_{(q)}, \mathcal{G}_{(2)}) \\ \vdots & \vdots & \vdots \\ \mathbf{k}(\mathcal{G}_{(1)}, \mathcal{G}_{(q)}) & \mathbf{k}(\mathcal{G}_{(2)}, \mathcal{G}_{(q)}) & \cdots \mathbf{k}(\mathcal{G}_{(q)}, \mathcal{G}_{(q)}) \end{bmatrix},$$

$$\mathcal{K} = \begin{bmatrix} \mathbf{k}(\mathcal{G}_{(1)}, \mathcal{K}_{(1)}) & \mathbf{k}(\mathcal{G}_{(2)}, \mathcal{K}_{(1)}) & \cdots \mathbf{k}(\mathcal{G}_{(q)}, \mathcal{K}_{(1)}) \\ \mathbf{k}(\mathcal{G}_{(1)}, \mathcal{K}_{(2)}) & \mathbf{k}(\mathcal{G}_{(2)}, \mathcal{K}_{(2)}) & \cdots \mathbf{k}(\mathcal{G}_{(q)}, \mathcal{K}_{(2)}) \\ \vdots & \vdots & \vdots \\ \mathbf{k}(\mathcal{G}_{(1)}, \mathcal{K}_{(w)}) & \mathbf{k}(\mathcal{G}_{(2)}, \mathcal{K}_{(w)}) & \cdots \mathbf{k}(\mathcal{G}_{(q)}, \mathcal{K}_{(w)}) \end{bmatrix}.$$

As outlined in **Theorem 1**, each frontal slice of a gram tensor is a kernel matrix in the transform domain. Therefore, $\tilde{\mathcal{G}}$ and $\tilde{\mathcal{K}}$ represent the transform domain representation of the gram tensors $\mathcal{G}$ and $\mathcal{K}$, respectively. To perform a classification task, we first need to integrate the multiple kernel matrices (frontal slices) in the gram tensor. To this end, we utilized Easy Multiple Kernel (EasyMKL) learning algorithm.[38] EasyMKL employs a minimization–maximization learning criterion to determine the optimal weighting of multiple kernel matrices, thereby maximizing the margin between two classes and improving classification performance. We want to find the best parameter combinations for the frontal slices of the gram tensor $\tilde{\mathcal{G}}$ by solving:

$$\max_{\eta:||\eta||=1} \min_{\gamma\in\tau} (1-\lambda)\eta^T y\Big((\sum_{i=1}^{n} \eta_i \tilde{\mathcal{G}}^{(i)})y\Big)^T \gamma + \lambda||\gamma||_2, \tag{6}$$

where y represents the target labels corresponding to visit $n+1$. We note that the spatiotemporal and gram tensors need to be constructed based on the first n visits. To learn the learnable parameters in Eq. (6), the target labels of the training set at visit $n+1$ were utilized along with the gram tensor $\tilde{\mathcal{G}}$ constructed for training. η is a learnable vector used to weight the linear combination of the kernel matrices (frontal slices $\tilde{\mathcal{G}}^{(i)}$). η is constrained to lie on the unit sphere. Mathematically, this is expressed as $||\eta|| = 1$. γ is another learnable parameter that is adjusted to minimize the objective function for a given η. The values of γ need to be optimized within a constrain set τ. λ $(0 \leq \lambda \leq 1)$ represents a regularization term that penalizes the magnitude of γ.

As shown in **Algorithm 1**, to solve the min-max problem, we decomposed the problem into two stages: first, we addressed the inner minimization problem over γ, and and then we solved the outer maximization problem over η. We employed convex optimization techniques to solve the inner maximization problem, utilizing the solvers available in the CVXPY library for Python. The outer maximization problem was solved using gradient ascent algorithms. We updated η iteratively (**Algorithm 1 line:8**) while ensuring it satisfies the unit norm constraint (**Algorithm 1 line:9**).

After learning the learnable parameters η and γ, the predictions can be computed for the test set $\tilde{\mathcal{K}}$ as following:

$$\hat{y} = (\sum_{i=1}^{n} \eta_i \tilde{\mathcal{K}}^{(i)})\gamma,$$
$$p_j = \frac{1}{1 + exp(\hat{y}_j)},$$

where $\hat{y}$ is the vector of raw prediction scores, and p_j is the probability of the raw score $\hat{y}_j$.

Algorithm 1 Min-Max Optimization Algorithm

Input: Initial vector η, regularization parameter λ, vector y, matrices $\{\tilde{\mathcal{G}}^{(i)}\}_{i=1}^{n}$, constraint set τ, step size α, tolerance ϵ
Output: Optimized vectors η and γ
Normalize $\eta \leftarrow \frac{\eta}{\|\eta\|}$
while true **do**
  $\eta_{prev} = \eta$
  Compute $\gamma(\eta)$ by solving

$$\gamma(\eta) = \arg\min_{\gamma \in \tau} \left[(1-\lambda)\eta^T y \left(\left(\sum_{i=1}^{n} \eta_i \tilde{\mathcal{G}}^{(i)} \right) y \right)^T \gamma + \lambda \|\gamma\|_2 \right]$$

  Calculate gradient:

$$\nabla_\eta J(\eta) = (1-\lambda)\nabla_\eta \left[\eta^T y \left(\left(\sum_{i=1}^{n} \eta_i \tilde{\mathcal{G}}^{(i)} \right) y \right)^T \gamma(\eta) \right]$$

  Update η:

$$\eta \leftarrow \eta + \alpha \nabla_\eta J(\eta)$$

  Project onto unit sphere:

$$\eta \leftarrow \frac{\eta}{\|\eta\|}$$

  if $\|\eta - \eta_{prev}\| < \epsilon$ **then**
    Break
  end if
end while
Return: η and γ

3. Results

3.1. *Datasets*

In this study, we utilized longitudinal data from two large AD databases: the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the National Alzheimer's Coordinating Center (NACC) database.

The ADNI (adni.loni.usc.edu) was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Since it has been launched, the public–private cooperation has contributed to significant achievements in AD research by sharing data to researchers from all around the world. The NACC, a comprehensive repository of data from several research sites across the United States, was specifically designed to aid research focused on understanding, diagnosing, and treating AD.

ADNI and NACC databases include data from cognitive performance tests, MRI scans, CSF analysis, demographic information, and diagnostic labels. However, only longitudinal

data were considered in this study, including cognitive performance tests, MRI scans, CSF analysis, and diagnostic labels that are AD and MCI. We preprocessed the data following the steps in PPAD method .[14] Missing values were imputed using k-NN algorithm, where we used average values from the nearest k neighbors with the same diagnosis (i.e., MCI or AD), employing the Euclidean as the distance metric and setting k to 5. To maintain data quality, we removed visits and features with $\geq 40\%$ and $\geq 60\%$ missing rate, respectively.

For the ADNI dataset, the original dataset comprised 15,087 records for 2,288 distinct patients, with each record representing a patient visit with 115 features. After preprocessing, the dataset had 20 longitudinal features for 1,169 patients across 5,759 visits, derived from cognitive performance tests and MRI scans. CSF analysis features were excluded due to significant missing data. For the NACC dataset, the original dataset comprised 172,026 records (i.e., visits) with 1024 features for 46,513 distinct patients. After preprocessing, the dataset had 5 longitudinal features for 8,121 patients across 35,423 visits, derived from cognitive performance tests. MRI scans and CSF analysis features were excluded due to significant missing data. Finally, we focused on patients with at least seven visits for training and model evaluation. This resulted in a final ADNI dataset of 20 longitudinal features for 250 patients and a final NACC dataset with 5 longitudinal features for 1,414 patients.

3.2. *Next Visit Prediction*

We trained DyEPAD on longitudinal EHR data to predict conversion of MCI patients to AD at the next visit. The experiments were performed on ADNI and NACC datasets separately and the results were compared with the state-of-the-art methods (i.e., GCN, GAT, T-LSTM, PPAD, and TA-RNN) as well as baseline methods, namely Random Forest (RF) and Support Vector Machine (SVM). For both datasets, the first six visits were considered to train the models. We measured the performance of all methods based on the visit number 7. Since RF and SVM cannot handle longitudinal data, we stacked all six visits' feature matrices to train these models. For GNN-based methods, the adjacency matrices were constructed using k-NN method (see **Section 2.4**). We evaluated all the methods on ten different randomly generated training and test splits. The number of patients selected for training was 80% of the total number of patients, while 20% were used for testing.

Table 1 demonstrates that our proposed approach DyEPAD outperformed all the state-of-the-art and baseline methods for both datasets for all three evaluation metrics. As outlined in **Section 2.1** and **2.2**, any invertible linear transformation can be used for our tensor operations. In our experiments, we used the discrete Fourier transform (DFT)[39] and discrete Hartley transform (DHT).[40] Although, both transformations resulted similar predictive performance, compared to the DFT, the DHT has the advantage of converting real functions into real functions, without the need for complex numbers. Therefore, running DyEPAD with DHT is more computationally efficient than running DyEPAD with DFT.[40]

3.3. *Multiple Visits Ahead Prediction*

In this subsection, we evaluated how well our proposed DyEPAD model performed at predicting conversion to AD in multiple visits ahead. We compared our results to TA-RNN and

Table 1. The reported values on ADNI and NACC datasets represent the averages along with standard deviations, based on ten runs, for three performance measures, namely: Accuracy, Macro F1 and Area Under the ROC Curve (AUCROC). Best values are shown in bold.

Dataset	Method	Accuracy	Macro F1	AUROC
ADNI	SVM	0.594 ± 0.040	0.570 ± 0.045	0.572 ± 0.045
	RF	0.566 ± 0.048	0.550 ± 0.047	0.550 ± 0.046
	GCN	0.661 ± 0.032	0.642 ± 0.033	0.642 ± 0.032
	GAT	0.685 ± 0.044	0.659 ± 0.035	0.660 ± 0.037
	PPAD	0.896 ± 0.035	0.893 ± 0.036	0.895 ± 0.034
	TA-RNN	0.883 ± 0.043	0.880 ± 0.043	0.885 ± 0.040
	T-LSTM	0.819 ± 0.145	0.778 ± 0.212	0.806 ± 0.157
	DyEPAD (DFT)	$\mathbf{0.900 \pm 0.035}$	$\mathbf{0.895 \pm 0.035}$	$\mathbf{0.896 \pm 0.035}$
	DyEPAD (DHT)	0.898 ± 0.038	0.893 ± 0.035	0.894 ± 0.038
NACC	SVM	0.773 ± 0.032	0.710 ± 0.032	0.690 ± 0.037
	RF	0.754 ± 0.030	0.688 ± 0.033	0.674 ± 0.036
	GCN	0.797 ± 0.032	0.742 ± 0.036	0.723 ± 0.044
	GAT	0.789 ± 0.035	0.735 ± 0.037	0.717 ± 0.042
	PPAD	0.950 ± 0.010	0.892 ± 0.023	0.878 ± 0.033
	TA-RNN	0.944 ± 0.011	0.880 ± 0.024	0.867 ± 0.031
	T-LSTM	0.935 ± 0.033	0.824 ± 0.153	0.812 ± 0.135
	DyEPAD (DFT)	0.950 ± 0.009	0.901 ± 0.014	0.900 ± 0.020
	DyEPAD (DHT)	$\mathbf{0.952 \pm 0.007}$	$\mathbf{0.902 \pm 0.014}$	$\mathbf{0.905 \pm 0.018}$

PPAD only, as the other methods are not designed to predict multiple visits ahead. All these methods were trained using the first three, four, and five visits of ADNI and NACC datasets and evaluated the performance on the seventh visit. The results in Figure 3 illustrate that DyEPAD performs comparably to the top state-of-the-art methods. In DyEPAD, tensorial functions operate via a linear transform to capture evolutionary characteristic in the data. However, applying a transform to a very short discrete signal may fail to capture periodic components, trends, and other evolutionary characteristics in the data. Therefore, not surprisingly, we observed that the performance of DyEPAD increased as the interval between the visits decreased. While this finding highlights the potential of DyEPAD, it also underscores a limitation that the model's effectiveness may be constrained by the granularity of the input data. In cases where patient visits are infrequent, the model may struggle to capture the dynamic nature of the underlying processes, potentially affecting its predictive accuracy.

3.4. *Ablation Study*

To assess the impact of deactivating various components of the proposed architecture on the model's performance, we conducted an ablation study. Specifically, to examine the impact of the tensorial functions and the aggregation layers, we compared the performance of the proposed DyEPAD architecture with two variants of DyEPAD: 1) we used an identity transformation, instead of the DHT; 2) we disabled the GRU layers and the embeddings were computed based on the GCN layer at that time point. We conducted the experiments on both ADNI and NACC datasets according to the experimental settings outlined in **Section 3.2**. The results given in Table 2 show that both the frequency domain representation and aggregation functions were crucial for capturing full scope of evolutionary patterns, as demonstrated by

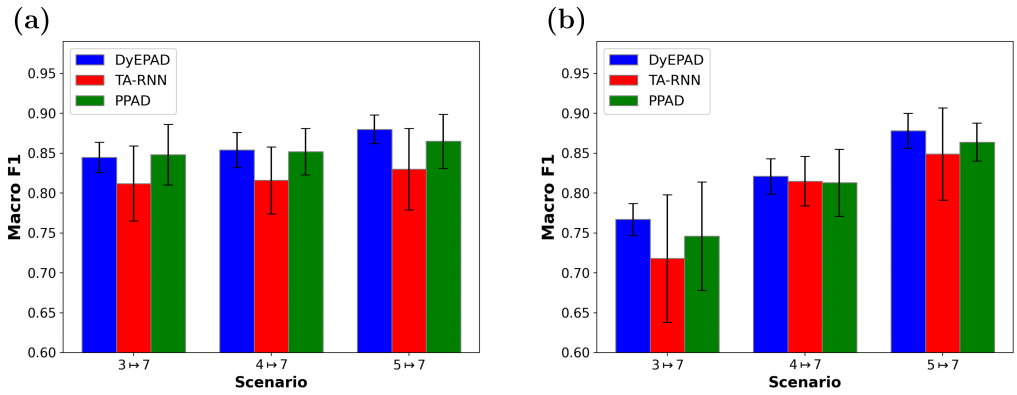


Fig. 3. Comparison of the performance of DyEPAD for different scenarios on (a) ADNI and (b) NACC datasets. Scenario $a \mapsto b$ denotes that the first a consecutive visits were utilized to train the model and bth visit was used for testing.

the superior performance of the original DyEPAD architecture for both datasets. We observed that for both ADNI and NACC datasets, the effect of the transform domain representation on the predictive performance was higher than the effect of the aggregation component. An identity transformation maps any element to itself. This means that we cannot capture any frequency components or patterns within the data because no change is applied to transform the data into the frequency domain or any other domain that might highlight such features. We can still used tensor operators, however the data remains in its original state, preserving its initial structure and values without revealing underlying periodicity or frequency information that a transform like the DHT would provide. This suggests that the considering a feasible invertible linear transformation is crucial in DyEPAD.

Table 2. The average Macro F1 and AUROC scores of different variants of DyEPAD on ADNI and NACC datasets over ten runs. Best values are shown in bold.

Variants of DyEPAD	**ADNI**		**NACC**	
	Macro F1	AUROC	Macro F1	AUROC
DyEPAD without any transformation	0.651 ± 0.071	0.654 ± 0.072	0.691 ± 0.019	0.718 ± 0.024
DyEPAD (DHT) without GRU	0.757 ± 0.074	0.754 ± 0.073	0.872 ± 0.028	0.868 ± 0.033
Proposed DyEPAD (DHT) architecture	$\mathbf{0.893 \pm 0.035}$	$\mathbf{0.894 \pm 0.038}$	$\mathbf{0.902 \pm 0.014}$	$\mathbf{0.905 \pm 0.018}$

4. Conclusions and Future Work

This paper presents a novel approach for predicting the progression of MCI subjects to AD using longitudinal EHR. Our proposed method, DyEPAD, captures latent space representations of EHR at each time step by utilizing GCN and GRU layers. We also use tensor algebraic operations for frequency domain analysis of these embeddings, capturing the complete range of evolutionary patterns across all time steps. The experimental outcomes reveal a notable superiority of DyEPAD over both state-of-the-art and baseline methods for most cases. Future work will aim to assess DyEPAD's performance on additional longitudinal biomedical datasets and examine the impact of different transformations and aggregation functions on its performance.

References

1. H. Hampel, D. Prvulovic, S. Teipel, F. Jessen, C. Luckhaus, L. Frölich, M. W. Riepe, R. Dodel, T. Leyhe, L. Bertram *et al.*, The future of alzheimer's disease: the next 10 years, *Progress in neurobiology* **95**, 718 (2011).
2. R. Cui, M. Liu, A. D. N. Initiative *et al.*, Rnn-based longitudinal analysis for diagnosis of alzheimer's disease, *Computerized Medical Imaging and Graphics* **73**, 1 (2019).
3. X. Wang, J. Qi, Y. Yang and P. Yang, A survey of disease progression modeling techniques for alzheimer's diseases, in *2019 IEEE 17th International Conference on Industrial Informatics*, (vol.1, pp. 1237–1242).
4. K. M. Langa and D. A. Levine, The diagnosis and management of mild cognitive impairment: a clinical review, *Jama* **312**, 2551 (2014).
5. X. W. Gao, R. Hui and Z. Tian, Classification of ct brain images based on deep learning networks, *Computer methods and programs in biomedicine* **138**, 49 (2017).
6. H. Li, M. Habes, D. A. Wolk, Y. Fan, A. D. N. Initiative *et al.*, A deep learning model for early prediction of alzheimer's disease dementia based on hippocampal magnetic resonance imaging data, *Alzheimer's & Dementia* **15**, 1059 (2019).
7. J. Guo, W. Qiu, X. Li, X. Zhao, N. Guo and Q. Li, Predicting alzheimer's disease by hierarchical graph convolution from positron emission tomography imaging, in *2019 IEEE international conference on big data (big data)*, (pp. 5359–5363, 2019).
8. M. Nguyen, T. He, L. An, D. C. Alexander, J. Feng, B. T. Yeo, A. D. N. Initiative *et al.*, Predicting alzheimer's disease progression using deep recurrent neural networks, *NeuroImage* **222**, p. 117203 (2020).
9. B. Shickel, P. J. Tighe, A. Bihorac and P. Rashidi, Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, *IEEE journal of biomedical and health informatics* **22**, 1589 (2017).
10. S. S. Prabhu, J. A. Berkebile, N. Rajagopalan, R. Yao, W. Shi, F. Giuste, Y. Zhong, J. Sun and M. D. Wang, Multi-modal deep learning models for alzheimer's disease prediction using mri and ehr, in *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*, (pp. 168–173, 2022).
11. M. Tanveer, B. Richhariya, R. U. Khan, A. H. Rashid, P. Khanna, M. Prasad and C.-T. Lin, Machine learning techniques for the diagnosis of alzheimer's disease: A review, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **16**, 1 (2020).
12. P. Yadav, M. Steinbach, V. Kumar and G. Simon, Mining electronic health records (ehrs) a survey, *ACM Computing Surveys (CSUR)* **50**, 1 (2018).
13. I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain and J. Zhou, Patient subtyping via time-aware lstm networks, in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, (pp. 65–74, 2017).
14. M. Al Olaimat, J. Martinez, F. Saeed, S. Bozdag and A. D. N. Initiative, Ppad: A deep learning architecture to predict progression of alzheimer's disease, *Bioinformatics* **39**, i149 (2023).
15. M. Al Olaimat, S. Bozdag and A. D. N. Initiative, Ta-rnn: an attention-based time-aware recurrent neural network architecture for electronic health records, *Bioinformatics* **40**, i169 (2024).
16. T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, *arXiv preprint arXiv:1609.02907* (2016).
17. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, Graph attention networks, *arXiv preprint arXiv:1710.10903* (2017).
18. Y. Zhang, X. He, Y. H. Chan, Q. Teng and J. C. Rajapakse, Multi-modal graph neural network for early diagnosis of alzheimer's disease from smri and pet scans, *Computers in Biology and Medicine* **164**, p. 107328 (2023).

19. Y. Zhu, J. Ma, C. Yuan and X. Zhu, Interpretable learning based dynamic graph convolutional networks for alzheimer's disease analysis, *Information Fusion* **77**, 53 (2022).
20. L. Jiang, H. Lin, Y. Chen, A. D. N. Initiative *et al.*, Sex difference in the association of apoe4 with cerebral glucose metabolism in older adults reporting significant memory concern, *Neuroscience Letters* **722**, p. 134824 (2020).
21. L. Besser, W. Kukull, D. S. Knopman, H. Chui, D. Galasko, S. Weintraub, G. Jicha, C. Carlsson, J. Burns, J. Quinn *et al.*, Version 3 of the national alzheimer's coordinating center's uniform data set, *Alzheimer Disease & Associated Disorders* **32**, 351 (2018).
22. T. G. Kolda and B. W. Bader, Tensor decompositions and applications, *SIAM review* **51**, 455 (2009).
23. K. Braman, Third-order tensors as linear operators on a space of matrices, *Linear Algebra and its Applications* **433**, 1241 (2010).
24. M. E. Kilmer, K. S. Braman, N. Hao and R. C. Hoover, Third order tensors as operators on matrices: A theoretical and computational framework with applications in imaging, *SIAM Journal on Matrix Analysis and Applications (SIMAX)* **34**, 148 (Feb. 2013).
25. C. Ozdemir, R. C. Hoover and K. Caudle, 2DTPCA: A new framework for multilinear principal component analysis, in *2021 IEEE International Conference on Image Processing (ICIP)*, (pp. 344-348, 2021).
26. C. Ozdemir, R. C. Hoover and K. Caudle, Fast tensor singular value decomposition using the low-resolution features of tensors, in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, (pp. 527–533, 2021).
27. C. Ozdemir, R. C. Hoover, K. Caudle and K. Braman, Kernelization of tensor discriminant analysis with application to image recognition, in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, (pp. 183–189, 2022).
28. C. Ozdemir, R. C. Hoover, K. Caudle and K. Braman, Tensor discriminant analysis on grassmann manifold with application to video based human action recognition, *International Journal of Machine Learning and Cybernetics* , 1 (2024).
29. R. N. Bracewell, The fourier transform, *Scientific American* **260**, 86 (1989).
30. N. Hao, M. E. Kilmer, K. Braman and R. C. Hoover, Facial recognition using tensor-tensor decompositions, *SIAM Journal on Imaging Sciences* **6**, 437 (2013).
31. E. Kernfeld, M. Kilmer and S. Aeron, Tensor–tensor products with invertible linear transforms, *Linear Algebra and its Applications* **485**, 545 (2015).
32. G. Li, M. Müller and G. Qian, Itzel carolina delgadillo perez, abdulellah abualshour, ali kassem thabet, and bernard ghanem. deepgcns: Making gcns go as deep as cnns, *IEEE transactions on pattern analysis and machine intelligence* **6**, p. 10 (2021).
33. J. You, Z. Ying and J. Leskovec, Design space for graph neural networks, *Advances in Neural Information Processing Systems* **33**, 17009 (2020).
34. C. Ozdemir, M. A. Olaimat, Y. Vashishath, S. Bozdag and A. D. N. Initiative, IGCN: Integrative graph convolutional networks for multi-modal data, *arXiv preprint arXiv:2401.17612* (2024).
35. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
36. V. N. Vapnik, V. Vapnik *et al.*, Statistical learning theory, *wiley New York* (1998).
37. T. Hofmann, B. Schölkopf and A. J. Smola, Kernel methods in machine learning, *The Annals of Statistics* (2008).
38. F. Aiolli and M. Donini, Easymkl: a scalable multiple kernel learning algorithm, *Neurocomputing* **169**, 215 (2015).
39. P. Duhamel and M. Vetterli, Fast fourier transforms: a tutorial review and a state of the art, *Signal processing* **19**, 259 (1990).
40. R. N. Bracewell, Discrete hartley transform, *JOSA* **73**, 1832 (1983).

All Together Now: Data Work to Advance Privacy, Science, and Health in the Age of Synthetic Data

Lindsay Fernández-Rhodes

College of Health and Human Development, Department of Biobehavioral Health; Social Science Research Institute; Population Research Institute; Clinical and Translational Science Institute; 219 Biobehavioral Health Building, 296 Henderson Drive, Pennsylvania State University, University Park, PA 16802, USA
Email: fernandez-rhodes@psu.edu

Jennifer K. Wagner

School of Engineering Design and Innovation; Department of Anthropology; Department of Biomedical Engineering; Institute for Computational and Data Sciences; Huck Institutes of the Life Sciences; Rock Ethics Institute; Pennsylvania State University, University Park, PA 16802 USA and Penn State Law, University Park, PA 16802 USA
Email: jkw131@psu.edu

There is a disconnect between data practices in biomedicine and public understanding of those data practices, and this disconnect is expanding rapidly every day (with the emergence of synthetic data and digital twins and more widely adopted Artificial Intelligence (AI)/Machine Learning tools). Transparency alone is insufficient to bridge this gap. Concurrently, there is an increasingly complex landscape of laws, regulations, and institutional/ programmatic policies to navigate when engaged in biocomputing and digital health research, which makes it increasingly difficult for those wanting to "get it right" or "do the right thing." Mandatory data protection obligations vary widely, sometimes focused on the type of data (and nuanced definition and scope parameters), the actor/entity involved, or the residency of the data subjects. Additional challenges come from attempts to celebrate biocomputing discoveries and digital health innovations, which frequently transform fair and accurate communications into exaggerated hype (e.g., to secure financial investment in future projects or lead to more favorable tenure and promotion decisions). Trust in scientists and scientific expertise can be quickly eroded if, for example, synthetic data is perceived by the public as "fake data" or if digital twins are perceived as "imaginary" patients. Researchers appear increasingly aware of the scientific and moral imperative to strengthen their work and facilitate its sustainability through increased diversity and community engagement. Moreover, there is a growing appreciation for the "data work" necessary to have scientific data become meaningful, actionable information, knowledge, and wisdom–not only for scientists but also for the individuals from whom those data were derived or to whom those data relate. Equity in the process of biocomputing and equity in the distribution of benefits and burdens of biocomputing both demand ongoing development, implementation, and refinement of embedded Ethical, Legal and Social Implications (ELSI) research practices. This workshop is intended to nurture interdisciplinary discussion of these issues and to highlight the skills and competencies all too often considered "soft skills" peripheral to other skills prioritized in traditional training and professional development programs. Data scientists attending this workshop will become better equipped to embed ELSI practices into their research.

Keywords: bioethics, data privacy, data work, health research, synthetic data

1. Introduction

The breadth of this workshop is deliberate, intended to bring together scholars from diverse areas of expertise and to promote interdisciplinary understandings foundational to the development and use of digital twins for biomedical research.[1] This effort responds to growing recognition of need for an interdisciplinary workforce prepared to seize opportunities and overcome challenges for digital twins, exemplified by the recent recommendations of the National Academies of Sciences, Engineering, and Medicine.[2,3]

The title of this workshop—*All Together Now: Data Work to Advance Privacy, Science, and Health in the Age of Synthetic Data*—itself is significant, with multiple levels of meaning to shine light on areas in which biocomputing can be enhanced. The reference to "all together now" refers not only to the importance of interdisciplinary, multidisciplinary, transdisciplinary collaboration but also to participatory, community-engaged research and collaborative governance. "Data work" draws attention to recent anthropological scholarship[4,5] as well as the novel biocomputing approaches that are now possible (e.g., synthetic data and digital twins, see, e.g., Foraker et al.[6], Moore et al.[7]). The explicit mention of "privacy, science, and health" is intended to draw attention to three distinct but interconnected international human rights (Articles 12, 27, and 25(1) of the Universal Declaration of Human Rights, respectively) that underlie ongoing debates about AI governance around the world and influence Fair Information Practice Principles (FIPPs), FAIR guiding principles for scientific data,[8] CARE principles[9] and more. The scope of this workshop is further intended to help attendees situate their biocomputing research more deliberately within the revised National Institute on Minority Health and Health Disparities (NIMHD) research framework for digital health equity.[10]

In this workshop, we will begin by highlighting scholars who routinely utilize synthetic data, digital twining, 'fake' data, simulations, or other obfuscation of data to ensure data privacy. They will present on how this limits the utility of data and/or their explainability. To address the tension these data present for public engagement, we will have community, implementation scientists, and communication scholars present on best practices on how these new data technologies can (and should) be incorporated into community engagement activities, to ensure that all populations have access to these new scientific approaches and insights. Workshop attendees will learn approaches to manage the gap between a) public expectations of science or assumptions of how biocomputing is performed, and b) the reality of the modern healthcare system, methodologic innovations within the biomedical research data ecosystem.

2. Workshop Topics and Presenters

2.1. *Opportunities and Challenges of Synthetic Data, Digital Twins, and Data Governance*

2.1.1. *Expanding Information Accessibility through Synthetic Data*

Presented by: Randi Foraker, PhD, MA, FAHA, FAMIA, FACMI (University of Missouri)

Synthetic healthcare data allow informaticians, data scientists, and clinicians to unlock siloed data and provide access to clinical researchers and consumers (e.g., students, citizen scientists) for improving the health of patients. The core benefit of synthetic data in medicine is that they can address obstacles to rapid research, methods development, and data sharing by representing the trends and relationships in the data without exposing the individual patients, and data — and therefore knowledge — can be shared while protecting individuals' privacy. This talk will explore the current and future state of synthetic data, highlighting its ability to support data sharing, to address privacy and confidentiality, and to advance national and international initiatives. The presenter will share their own work with synthetic data, which spans statistical validation (comparing results of analyses between real and computationally derived data); national and international research partnerships; and leveraging synthetic data for informatics, biostatistics, and data science education.

2.1.2. *The Role of Synthetic Data in Patient Privacy, Healthcare, and Biomedical Research*

Presented by: Jason Moore, PhD, FACMI, FIAHSI, FASA (Cedars-Sinai Medical Center)

Paramount to healthcare and biomedical research is the protection of patient privacy and the security of their data. Synthetic data may address these concerns by providing artificial data points that preserve the correlation structure and patterns of the original patient data. The presenter will review artificial intelligence methods the generation of synthetic data and their use in clinical and biomedical research. These will include deep learning and large language model approaches. They will highlight several use cases from the literature and will discuss the use of synthetic data for creating digital twins that might improve the prediction of clinical outcomes. Limitations and challenges of these methods will be discussed.

2.1.3. *The Long View on Emerging Data Science Technologies*

Presented by: Anjali Deshmukh, MD, JD (Georgia State University College of Law)

This talk will examine emerging data technologies in children, focusing on privacy, longitudinal impacts, and FDA regulation. Real world data of children's health outcomes are difficult to obtain and analyze, and technologies including digital twins have the potential to solve the current limitations of claims data. Yet, the potential benefits must be considered against the risk. Current data policy choices will impact children's privacy rights and drug safety over the long-term. Therefore, understanding current FDA regulations and creating regulatory flexibilities to optimize outcomes for children over their lives is important.

2.1.4. *Technical Approaches to Balance Patient Privacy and Shared Analytic Utility*

Presented by: John Wilbanks (Astera Institute)

National biobanks such as the UK Biobank, the All of Us Research Program, and similar emergent state-level efforts around the world hold the promise of driving novel research on large, diverse participant cohorts. Many such biobanks simultaneously center ideals of aggressive participant empowerment and Open/FAIR science, which can create tensions between protecting an individual patient's privacy while seeking thousands of researchers to generate analysis and insights. Some privacy-enhancing technology approaches can enable multi-party computation, or create synthetic data sets, which can introduce other tensions between data availability and data trustability. This talk will explore how intentional choices in cloud architecture can address these tensions, with specific examples drawn from the All of Us Researcher Workbench and the Broad Institute's Data Science Platform.

2.2. *Data Work from the Perspective of Scholars in Community Engagement, Ethics, and Science Communication*

2.2.1. *Biorepositories and Group Harm: A Choice Architecture for Researchers*

Presented by: Meg Doerr, MS LGC (Sage Bionetworks)

This talk will help workshop participants (1) distinguish between individual and group harm from research; (2) appreciate why group harm should be a primary consideration of AI researchers and those that enable AI research including data access committees, ethics boards, and funders; (3) renew their understanding of current regulations on individual and group harm in research; and (4) learn about new tools (created in a project funded by the Robert Wood Johnson Foundation) to aid researchers, data access committees, ethics boards, and funders in enabling responsible AI-driven research.

2.2.2. *Nothing About Us, Without Us Leading*

Presented by: Maile Tauali‘i, PhD MPH (Hawaii Permanente Medical Group)

Indigenous Peoples are often the target of research and not the owners of the research process. We are also domestically dependent under nations and are often subjected to rules and decisions made about us and not with us. So, when we speak about “own-voice” research, we are speaking in opposition to colonial settler science which subjects us to decisions made without us. We want our voices heard. Learning objectives: 1) Participants will be able to identify 3 laws that uphold Indigenous ownership of data 2) Participants will have 3 strategies to respectfully engage with Indigenous Peoples 3) Participants will learn 3 examples where Indigenous Peoples rights were violated by scientists.

2.2.3. *Bounded Justice, the Performance of Trust, and Anti-Racism in Biocomputing*

Presented by: Melissa C. Creary, PhD, MPH (University of Michigan)

Drawing upon her prior published works on bounded justice, the public performativity of trust[11] and the application of anti-racism in informatics,[12] the speaker will discuss best practices for fostering community engagement while simultaneous embracing new data practices for biocomputing.

2.2.4. *Communication Data, Communicating Science*

Presented by: Jasmine McNealy, PhD, JD (University of Florida)

The hallmark of any interaction between scientists and the public is communication. Communication is important for developing and sustaining relationships, for building trust, and enhancing partnerships. Effective communication is important for interactions with marginalized and/or vulnerable communities, especially those whose distrust of biomedical researchers is born of past missteps and harmful programs. Therefore, scientists should be able to communicate with both media and publics beyond those connected to academia and scholarly research. This is particularly important for helping the public to understand novel data practices.

2.2.5. *Soulful Innovation: A New Framework to Create Responsible Technologies of the Future*

Presented by: Samira Kiani, MD (University of Pittsburgh)

The speaker proposes a new framework for innovation by revisiting our relationship with ourselves, our relationship with the impact we create, the spaces in which innovation happens and our collective. Through this framework– called “soulful innovation”--we ask how we can move away from the culture of “be first” and “star is born” to a culture that celebrates our “collectiveness” and puts inner human values at the core of innovation.

3. Conclusion

The workshop will conclude with a discussion panel facilitated by the organizers involving all of the workshop presenters to what they would like to see from data scientists who use synthetic data in the near future and to address questions and comments from workshop attendees.

By attending in this workshop, participants will gain 1) expertise in understanding how new data technologies that use obfuscation are being implemented in the biomedical sciences, 2) awareness of the potential opportunities and concerns related to these practices with respect to participant and community engagement, and 3) familiarity with the best practices for fostering community engagement and science communication while simultaneously embracing these new data practices.

4. Acknowledgments

We would like to the acknowledge the Penn State Social Science Research Institute, Population Research Institute, and the Department of Biobehavioral Health for their support of LFR.

References

1.National Academies of Sciences, Engineering, and Medicine. Opportunities and Challenges for Digital Twins in Biomedical Research: Proceedings of a Workshop–in Brief. *The National Academies Press* (2023) doi:https://doi.org/10.17226/26922.
2.National Academies of Sciences, Engineering, and Medicine. *Briefing Slides: Foundational Research Gaps and Future Directions for Digital Twins*. https://nap.nationalacademies.org/resource/26894/briefing-slides-digital-twins.pdf (2023).
3.National Academies of Sciences, Engineering, and Medicine. *Foundational Research Gaps and Future Directions for Digital Twins*. https://doi.org/10.17226/26894 (2024).
4.Fiske, A., Prainsack, B. & Buyx, A. Data Work: Meaning-Making in the Era of Data-Rich Medicine. *J Med Internet Res* **21**, e11672 (2019).
5.Fiske, A., Degelsegger-Márquez, A., Marsteurer, B. & Prainsack, B. Value-creation in the health data domain: a typology of what health data help us do. *Biosocieties* 1–25 (2022) doi:10.1057/s41292-022-00276-6.
6.Foraker, R., Mann, D. L. & Payne, P. R. O. Are Synthetic Data Derivatives the Future of Translational Medicine? *JACC Basic Transl Sci* **3**, 716–718 (2018).
7.Moore, J. H. *et al.* SynTwin: A graph-based approach for predicting clinical outcomes using digital twins derived from synthetic patients. *Pac Symp Biocomput* **29**, 96–107 (2024).
8.Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
9.Carroll, S. R., Herczog, E., Hudson, M., Russell, K. & Stall, S. Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Sci Data* **8**, 108 (2021).
10.Richardson, S., Lawrence, K., Schoenthaler, A. M. & Mann, D. A framework for digital health equity. *NPJ Digit Med* **5**, 119 (2022).
11.Creary, M. & Gerido, L. H. The Public Performativity of Trust. *Hastings Cent Rep* **53 Suppl 2**, S76–S85 (2023).
12.Platt, J. *et al.* Applying anti-racist approaches to informatics: a new lens on traditional frames. *J Am Med Inform Assoc* **30**, 1747–1753 (2023).

Command line to pipeLine: Cross-biobank analyses with Nextflow

Anurag Verma
Department of Pathology and Laboratory Medicine,
University of Pennsylvania
Philadelphia, PA 19104, USA
Email: anurag.verma@pennmedicine.upenn.edu

Zachary Rodriguez
Department of Pathology and Laboratory Medicine,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: zachary.rodriguez@pennmedicine.upenn.edu

Lindsay Guare
Department of Pathology and Laboratory Medicine,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: lindsay.guare@pennmedicine.upenn.edu

Katie Cardone
Department of Genetics,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: katie.cardone@pennmedicine.upenn.edu

Christopher Carson
Department of Pathology and Laboratory Medicine,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: christopher.carson@pennmedicine.upenn.edu

Biobanks hold immense potential for genomic research, but fragmented data and incompatible tools slow progress. This workshop equipped participants with Nextflow, a powerful workflow language to streamline bioinformatic analyses across biobanks. We taught participants to write code in their preferred language and demonstrated how Nextflow handles the complexities, ensuring consistent, reproducible results across different platforms. This interactive session was ideal for beginner-to-intermediate researchers who want to (1) Leverage biobank data for genomic discoveries, (2) Build portable and scalable analysis pipelines, (3) Ensure reproducibility in their findings, (4) Gain hands-on experience through presentations, demonstrations, tutorials, and discussions with bioinformatics experts.

Keywords: bioinformatics, genomics, phenome, biobanks

1. Introduction, Background, and Motivation

The field of genomics has entered a transformative era fueled by the rapid expansion of biobanks. These repositories, including public entities like the UK Biobank (Sudlow et al., 2015) and the NIH's All of Us Research Program (Ramirez et al., 2022), along with numerous institutional biobanks such as Million Veteran Program (Gaziano et al., 2016) and Penn Medicine BioBank (Penn Medicine BioBank, n.d.), have been instrumental in accelerating genomic discovery at an unprecedented pace. By bringing together extensive collections of biological samples and rich clinical data, biobanks have been a goldmine for medical research. We can leverage biobanks to pinpoint genetic variations linked to diseases and unravel the complexities of various phenotypes.

Despite the move towards cloud computing to share data, biobanks face significant technical hurdles that slow down their potential. Data is often kept in isolated pockets, and researchers have to navigate a technical maze to use different platforms and tools. This not only hinders the speed of research but also leads to the same work being repeated and a mix of data analysis practices that can cast doubt on findings and make it challenging to scale up genomic studies.

As bioinformatic analyses grow in scale and popularity, the methodology and best practices are becoming more standardized. Rather than introducing redundant code by copying commands between projects, pipeline managers offer a way to re-configure the code while recycling it from the same source. Avoiding redundancy is important because each new copy-paste of a section of code results in propagating errors which will take additional time to track and fix (Leitão, 2004). Furthermore, highly parallel computational work on university computing clusters has often looked like manually watching the queue to wait for all jobs of a particular step to finish. Pipeline managers have automated interfaces which work with multiple platforms, allowing them to track jobs and submit dependent ones as they finish.

Workflow languages like Nextflow play a pivotal role in the development of scalable and reproducible genomic pipelines by offering a platform-agnostic framework for seamless data analysis across diverse computing environments (Figure 1). By abstracting the complexities of platform-specific hardware/software configurations, Nextflow enables researchers to focus on the scientific logic of their analyses and interpretation of results. This abstraction allows researchers to create workflows from their pre-existing code written in any language that can be easily deployed on local servers, high-performance computing clusters, or cloud-based platforms without modification. Further, Nextflow’s containerization support through technologies like Docker and Singularity ensures analyses can be deployed and parallelized across different computing architectures without risk of data conflicts, dependency issues, or concurrent data access and processing.

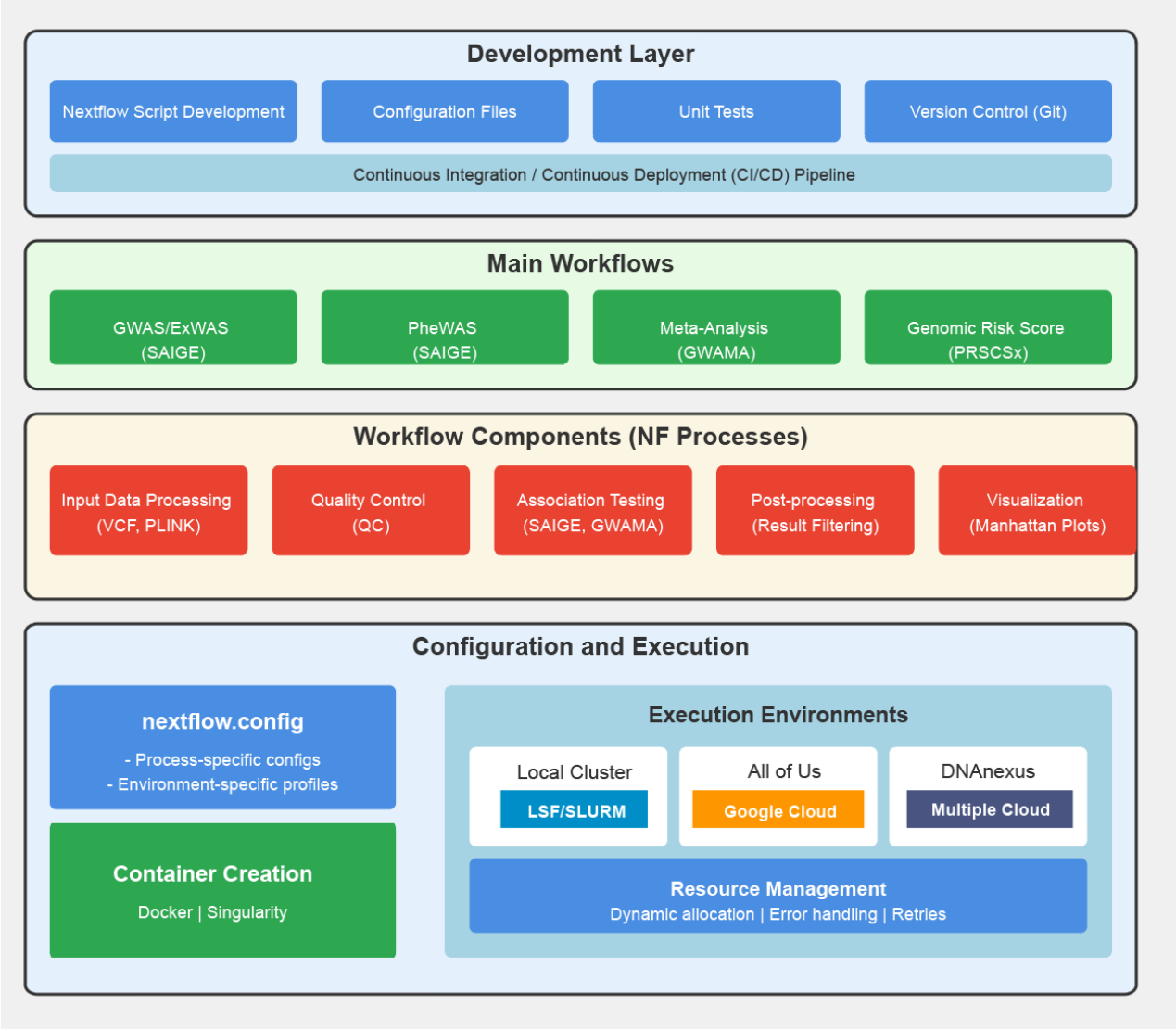


Figure 1. This figure illustrates the multi-layered architecture of genomic workflow development and execution using Nextflow. It encompasses the development layer (including script development, configuration, testing, and version control), main workflows (such as GWAS/ExWAS, PheWAS, Meta-Analysis, and Genomic Risk Score), workflow components (NF processes for various stages of analysis), and the configuration and execution layer (including environment-specific configurations and diverse execution environments). This architecture demonstrates the scalability and flexibility of the workflows across different computing infrastructures, from local clusters to cloud platforms.

Through this workshop, our goal is to address a critical need within the genetic research and bioinformatics community. The rapid expansion of biobank data availability marks a significant milestone in human genetics research, offering unparalleled opportunities to study the genetic predisposition of complex diseases. Although there are platforms and tools for effectively utilizing these datasets for complex, multimodal analysis, there remains an unmet need to develop educational workshops. These workshops are essential to equip participants with the necessary skills and knowledge to fully exploit biobank resources, effectively bridging the gap between the abundance of available data and the capacity for research innovation.

We provided attendees with hands-on workflows to develop and deploy existing tools from institutional biobanks to cloud-based platforms such as the UK Biobank and All of Us. We recognize a strong demand for proficiency in integrating omics data with genetic findings and a growing interest in conducting cross-biobank analyses for more extensive and robust research applications. By focusing on these areas, our workshop directly addresses these educational needs, offering content that builds on past experiences while also anticipating future research trends.

2. Workshop Presentations and Tutorials

To enable communication and discussion between experimental scientists and our expert developers, each module in this workshop included presentations that provided brief introductions to key topics before the demonstrations and hands-on exercises. The goal of these presentations was to educate participants on the foundational principles of developing genomic workflows using existing tools and resources. The workshop format featured demonstrations, hands-on tutorials, exercises, and discussions led by our five Bioinformatics experts. Demonstrations included pre-recorded vignettes showing how to configure and run large-scale genomic pipelines, with step-by-step explanations and Q&A sessions. Hands-on tutorials offered guided introductions to Nextflow workflows, while exercises allowed attendees to practice independently and in group settings, with on-demand assistance from our team. Throughout the workshop, we highlighted our: Case study of analysis on local and cloud platforms such as UK Biobank and All of Us.

- **Genomic Pipelines for Biobanks: Development and Deployment. (Speaker: Anurag Verma)**: Overview of current biobank landscapes; Challenges in developing scalable genomic pipelines;
- **PMBB Toolkit: GWAS and PRS (Speaker: Chris):** Demonstration on how to utilize and understand genome-wide association study and polygenic score pipelines built in our PMBB Genomic Toolkit.
- **Command Line to Pipeline (Speaker: Lindsay and Zach):** Introduction to cloud-agnostic workflow languages with a focus on demystifying Nextflow pipeline management concept so participants can write their own Nextflow pipeline with the help of our experts.
- **Overcoming Limitations of Working Across Biobanks & Cloud Platforms (Speaker: Katie):** Deploying a workflow across cloud environments and coding collaboratively with Google Cloud Shell.

3. Conclusion

Through this workshop, participants gained the essential tools and expertise to harness the full potential of biobank data, ultimately accelerating the pace of genomic research and discovery. By the end of this workshop, participants were equipped with the knowledge and skills to develop and deploy scalable and reproducible genomic workflows, navigate the complexities of cloud-based platforms, and conduct meaningful cross-biobank analyses to advance their research projects. This workshop provided a platform not only as a repository of knowledge but also as a forum for academic exchange. Throughout the workshop, scientists discussed (1) The challenges of conducting bioinformatic analyses across different cloud platforms, (2) Best practices for integrating different biobanks with an emphasis on reproducibility, interpretability, and scalability, and (3) How to use GitHub for transparency, version control, and collaboration.

4. Speakers

Anurag Verma, PhD, University of Pennsylvania. Anurag is an Assistant Professor in the Department of Medicine at the University of Pennsylvania, and he also serves as Associate Director of Clinical Informatics and Genomics for Penn Medicine BioBank. His research has focused on the study of the genetic basis of complex diseases using big data techniques with the main focus on studying the genetic architecture of multimorbidity, the phenotypic architecture of common genetic risk, polygenic risk scores, and phenome-wide association studies to identify the complex phenotypic and genomic interactions that lead to complex disease. In his capacity at PMBB, Anurag leads a team called CodeWorks that develops scalable workflows and harnesses both in-house and cloud computing resources for advancements in genetic research. His team's efforts are in expanding the boundaries of how data informatics can be applied to keep pace with the rapidly changing landscape of large-scale biobanks.

Lindsay Guare, University of Pennsylvania. Lindsay is a second-year PhD student in the Genomics and Computational Biology Program at UPenn with a focus in Biomedical Informatics. She has been involved in many large-scale genetic association study collaborations, but her research will be focused on leveraging innovative computational data science approaches to explore clinical and genetic heterogeneity in endometriosis. Her interdisciplinary background includes computer science, contributing to her leadership in CodeWorks.

Katie Cardone, BS, University of Pennsylvania. Katie is a Research Specialist in the Department of Genetics at the University of Pennsylvania and is a Graduate Student in the University of Pennsylvania's Master of Biomedical Informatics Program. In her role, Katie executes a wide range of bioinformatic analyses, including genome-wide association studies, phenome-wide association studies, exome-wide rare variant association studies, and polygenic scores on large biobanks, including the Penn Medicine BioBank, the eMERGE network, and the All of Us research program. She also develops Nextflow pipelines for polygenic score tools.

Christopher Carson, MS, University of Pennsylvania. Chris is a Bioinformatician at the University of Pennsylvania Institute for Biomedical Informatics. His role in the Verma lab covers an extensive range of workflow pipeline development, conducting genetic analysis requests for the Penn Medicine Biobank (PMBB), and producing bioinformatics software for analyzing large-scale genomic and phenomic datasets. He has experience conducting genome-wide, phenome-wide, and exome-wide association studies using the large-scale datasets retained in the PMBB with the use of SAIGE.

Zachary Rodriguez, PhD, University of Pennsylvania. Zach is a Bioinformatician at the University of Pennsylvania's Perelman School of Medicine. His research has focused on the study of the genetic basis of complex diseases using big data techniques with the focus on studying the genetic architecture of multimorbidity, the phenotypic architecture of common genetic risk, polygenic risk scores, and phenome-wide association studies to identify the complex phenotypic and genomic interactions that lead to complex disease. He has informatics expertise in machine learning, natural language processing, and pipeline development, with extensive experience in analyzing large-scale genomic data, electronic health records (EHR), and biobank datasets, including Penn Medicine BioBank.

5. Acknowledgements

We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the patient- participants of Penn Medicine who consented to participate in this research program. We would also like to thank the

Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. We would like to thank the PMBB leadership team: Daniel J. Rader, M.D., Marylyn D. Ritchie, Ph.D; The PMBB Patient Recruitment and Regulatory Oversight Team: oEllen Weaver, Nawar Naseer, Ph.D., M.P.H., Giorgio Sirugo, M.D., P.h.D., Afiya Poindexter, Yi-An Ko, Ph.D., Kyle P. Nerz; The PMBB Clinifal Informatics Team: Anurag Verma, Ph.D., Colleen Morse Kripke, M.S. DPT, MSA, Marjorie Risman, M.S., Renae Judy, B.S., Colin Wollack, M.S.; The PMBB Genome Informatics Team: Anurag Verma Ph.D., Shefali S. Verma, Ph.D., Scott Damrauer, M.D., Yuki Bradford, M.S., Scott Dudek, M.S., Theodore Drivas, M.D., Ph.D. Lastly, we would like to thank all the bioinformaticians, developers, and testers on our Codeworks team that contributed to the PMBB Geno-Pheno Toolkit Github: Zachary Rodriguez, Lindsay Guare, Chris Carson, Lannawill Caruth, Katie M. Cardone, Aude Ikuzwe, Michael Condiff, Alexis Garofalo, Xueqiong Li, Karl Keat, Rachit Kumar, Trust Odia, Colleen Morse Kripke, Hritvik Gupta, Theodore Drivas, Shefali Setia-Verma, and Anurag Verma.

References

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., Guarino, P., Aslan, M., Anderson, D., LaFleur, R., Hammond, T., Schaa, K., Moser, J., Huang, G., Muralidhar, S., … O'Leary, T. J. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, *70*, 214–223.

Leitão, A. M. (2004). Detection of Redundant Code Using R 2 D 2. *Software Quality Journal*, *12*(4), 361–382.

Penn Medicine BioBank. (n.d.). Retrieved September 30, 2024, from https://pmbb.med.upenn.edu/

Ramirez, A. H., Sulieman, L., Schlueter, D. J., Halvorson, A., Qian, J., Ratsimbazafy, F., Loperena, R., Mayo, K., Basford, M., Deflaux, N., Muthuraman, K. N., Natarajan, K., Kho, A., Xu, H., Wilkins, C., Anton-Culver, H., Boerwinkle, E., Cicek, M., Clark, C. R., … All of Us Research Program. (2022). The All of Us Research Program: Data quality, utility, and diversity. *Patterns (New York, N.Y.)*, *3*(8), 100570.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779.

Leveraging Foundational Models in Computational Biology: Validation, Understanding, and Innovation*

Brett Beaulieu-Jones
Department of Medicine, University of Chicago, 5841 South Maryland Avenue, MC 6092 Chicago, IL, USA
Email: beaulieujones@uchicago.edu

Steven Brenner
Department of Plant and Microbial Biology, 111 Koshland Hall, Berkeley, CA, USA
Email: brenner@compbio.berkeley.edu

Large Language Models (LLMs) have shown significant promise across a wide array of fields, including biomedical research, but face notable limitations in their current applications. While they offer a new paradigm for data analysis and hypothesis generation, their efficacy in computational biology trails other applications such as natural language processing. This workshop addresses the state of the art in LLMs, discussing their challenges and the potential for future development tailored to computational biology. Key issues include difficulties in validating LLM outputs, proprietary model limitations, and the need for expertise in critical evaluation of model failure modes.

Keywords: Generative AI, Large Language Models, Foundational Models, Computational Biology

1. Background

Large Language Models (LLMs) have demonstrated immense potential[1–9] within and outside of the biomedical domain but currently have substantial limitations when applied to biomedical research.[10,11] These models promise a new paradigm for data analysis, interpretation and hypothesis generation, but it is not clear how fully this promise will be fulfilled. LLMs are just one class of foundational models, and while they have already made a significant impact to computational biology, it is unlikely that a singular architecture geared at processing natural language will be the ideal framework for general learning in computational biology. This workshop aims to provide an understanding of the state of the art today, current challenges in the application or development of models tailored to computational biology, as well as to start a discussion of what the future holds for our community.

At present, LLMs are commonly used in attempt to directly answer complex problems in ways that are difficult to validate. Existing methods for interpretation are limited, and it is difficult without a ground truth to tell whether an answer is accurate or a "hallucination".[12] These challenges contrast

* This work is partially supported by NIH grant R00NS114850

with typical goals in biomedical research where researchers aim to understand the underlying system. Issues with LLM hallucination have been well documented and approaches for dealing with uncertainty within generative models are nascent. Proprietary models create challenges to reproducibility, privacy, and present barriers to finetuning and open sharing. The successful use of LLMs for research still requires a high degree of expertise in order to "red team", or critically interrogate and evaluate failure modes of LLMs. This process is currently poorly defined with best practices not yet widely agreed upon.

Most prior work has focused on either training LLMs or using available models (locally or via vendor provided APIs) for related tasks. A critical issue with the status quo is that the field is rapidly evolving, meaning building upon any one model is a risk and there is a constant need to retrain models and update workflows based on newly released models. Additionally, the majority of innovation has come either through using large general-purpose models (e.g., GPT4), or in training models derived from architectures designed for natural language processing. Increasingly we are seeing the development of foundational models for multimodal data in addition to more specific subfields. As a new state of the art model is released, within a relatively short period of time, researchers have developed smaller, domain or task specific models that appear to achieve comparable or slightly worse performance despite having access to vastly fewer resources. Recently, we have seen the emergence of novel architectures for foundational models trained on electronic medical record data[13,14] and multimodal models for medical-imaging and text.[15–19] While these models have demonstrated early promise, their impact does not yet compare to that of LLMs.

Topics around foundational models, specifically LLMs, have been widely covered at academic journals, conferences, and in a wide variety of other settings. However, the majority of discussions around these models have focused on the low hanging fruit, posing questions like how GPT-4 can be used as a knowledge integration tool for hypothesis generation or evaluating its capabilities against professional exams or clinical case diagnostics. There has been decidedly less attention paid to the methodological side of tailoring these models to workflows in computational biology through techniques like the programmatic generation of prompts and labels for supervised and even weakly supervised instruction fine-tuning, interpretation and/or explanation leveraging expert knowledge-based uncertainty exploration, retrieval-augmented generation strategies with "-omics" style data, multimodal approaches to include assets like clinical notes and medical imaging for phenotyping. Finally, with the rapid advancement of the larger field of foundational models, it is nearly impossible for the transdisciplinary scientists who typically attend PSB to keep up with all of the literature in a critical but separate field from their primary research.

2. Leveraging Foundational Models in Computational Biology: Workshop

LLM's and the broader field of generative AI are in period of rapid evolution. This workshop aims to help attendees of PSB differentiate between the signal and the noise. What are the breakthrough ideas, technologies, and applications that are already or are poised to have substantial impacts on the field of computational biology.

This workshop aims to provide:

1. Provide an understanding of the current state of the art for foundational models both in general and specifically within computational biology
2. Understand common failure modes and survey methods to validate results
3. Explore recent innovations in foundational models and LLMs that address prior challenges most relevant to computational biology (e.g., novel approaches for tokenization, representation of modalities outside of natural language, uncertainty estimation and explanation)
4. Showcase innovative uses of LLMs in computational biology through a "year-in-review" overview of the past years most interesting works in this area
5. Plan for the future based on invited talks by researchers on the strategies for development and utilization of the next generation of LLMs.

To do this, the workshop will be composed of three invited talks covering, “What is the current state of the art?”, “What are the Strategies for recognizing and Mitigating Failure Modes”, and a “Year-in-Review” talk based on extensive literature review. Our aim with this is to help the PSB audience determine what is worth paying attention to and which developments are simply “shining objects” that are potential distractions. Additionally, there will be a panel discussion covering the challenges and shortcomings of current approaches and what does the future look like?

3. Conclusion

LLMs hold immense potential for transforming biomedical research, but their current limitations, such as hallucinations and challenges in reproducibility, necessitate careful scrutiny. The field is evolving rapidly, with new foundational models being introduced frequently, requiring constant retraining and workflow updates. It is essential to develop methodologies specifically suited to computational biology, as general-purpose models may not be optimal for this domain. The workshop seeks to guide researchers in discerning between valuable advancements and distractions in this rapidly changing environment.

4. Acknowledgements

We would like to thank the Organizing committee of the Pacific Symposium for Bioinformatics 2024 for giving us the opportunity of organizing the proposed workshop.

References

1. Bubeck, S. *et al.* Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv [cs.CL]* (2023).
2. Singhal, K. *et al.* Large Language Models Encode Clinical Knowledge. *arXiv [cs.CL]* (2022).
3. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).

4. Eriksen Alexander V., Möller Sören & Ryg Jesper. Use of GPT-4 to Diagnose Complex Clinical Cases. *NEJM AI* **1**, AIp2300031 (2023).
5. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv [cs.CL]* (2023).
6. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
7. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).
8. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
9. Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
10. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* **6**, 120 (2023).
11. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *NPJ Digit Med* **6**, 195 (2023).
12. Ji, Z. *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
13. Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit. Med.* **6**, 135 (2023).
14. Thapa, R., Steinberg, E. & Fries, J. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Adv. Neural Inf. Process. Syst.* (2024).
15. Azad, B. *et al.* Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision. *arXiv [cs.CV]* (2023).
16. Moor, M. *et al.* Med-Flamingo: a Multimodal Medical Few-shot Learner. in *Proceedings of the 3rd Machine Learning for Health Symposium* (eds. Hegselmann, S. et al.) vol. 225 353–367 (PMLR, 2023).
17. Jeong, J. *et al.* Multimodal Image-Text Matching Improves Retrieval-based Chest X-Ray Report Generation. *arXiv [cs.CL]* (2023).
18. Willemink, M. J., Roth, H. R. & Sandfort, V. Toward Foundational Deep Learning Models for Medical Imaging in the New Era of Transformer Networks. *Radiol Artif Intell* **4**, e210284 (2022).
19. Chambon, P., Bluethgen, C., Langlotz, C. P. & Chaudhari, A. Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains. *arXiv [cs.CV]* (2022).

Opportunities and Pitfalls with Large Language Models for Biomedical Annotation

Cecilia Arighi
Department of Computer and Information Sciences, University of Delaware, Ammon-Pinizzotto Biopharmaceutical Innovation Building, 590 Avenue 1743, Newark, DE19713, US
Email: arighi@udel.edu

Jin-Dong Kim[1]
Database Center for Life Science (DBCLS), DS-ROIS, ROIS, 178-4-4 Wakashiba Kashiwa, Chiba 277-0871, Japan
Email: jdkim@dbcls.rois.ac.jp

Zhiyong Lu[2]
NCBI, NLM, NIH, Bethesda, MD 20894 Bethesda, MD20894, US
Email: zhiyong.lu@nih.gov

Fabio Rinaldi
IDSIA USI-SUPSI, Polo universitario Lugano - Campus Est, Via la Santa 1, CH-6962 Lugano - Viganello, Switzerland
Email: fabio.rinaldi@idsia.ch

Large language models (LLMs) and biomedical annotations have a symbiotic relationship. LLMs rely on high-quality annotations for training and/or fine-tuning for specific biomedical tasks. These annotations are traditionally generated through expensive and time-consuming human curation. Meanwhile LLMs can also be used to accelerate the process of curation, thus simplifying the process, and potentially creating a virtuous feedback loop. However, their use also introduces new limitations and risks, which are as important to consider as the opportunities they offer. In this workshop, we will review the process that has led to the current rise of LLMs in several fields, and in particular in biomedicine, and discuss specifically the opportunities and pitfalls when they are applied to biomedical annotation and curation.

Keywords: large language model; LLM; biomedical curation; generative AI; biomedicine and health; education; ethics.

[1] Work supported by the Database Integration Coordination Program Funced by NDBC of JST.
[2] This research is partly supported by the NIH Intramural Research, National Library of Medicine.

1. Background

High-quality, well-annotated biomedical data is crucial for training LLMs to understand and process scientific information. These annotations can include labeling entities (genes, proteins), relations (interactions), and other relevant information. By incorporating annotated data, LLMs can learn specific domain knowledge and improve their accuracy in tasks like information extraction, knowledge base creation, and text summarization. Diverse and unbiased annotations can help mitigate bias in LLMs, ensuring their outputs are fair and representative of the underlying data. At the same time, LLMs can be used to automate some aspects of annotation, such as identifying potential entities or suggesting relevant relations. This can significantly reduce the workload for human annotators. LLMs can identify areas of uncertainty in the data and suggest which annotations would be most valuable for improving their performance. This creates a feedback loop where LLMs guide the annotation process for optimal results. Finally, LLMs can be used to check the consistency and accuracy of annotations, identifying potential errors or inconsistencies.

A recent survey of LLMs for data annotation [1] describes how advanced large language models (LLMs), like GPT-4, can transform data annotation by automating and improving accuracy in this traditionally labor-intensive process. It categorizes the methods used for LLM-based data annotation, explores the effectiveness of LLM-generated annotations, and discusses learning strategies incorporating these annotations. The paper also highlights the challenges and limitations of using LLMs in this field, offering guidance for future research and development in automating data annotation. Goel et al [2] proposes a method that uses Large Language Models (LLMs) combined with human expertise to speed up medical text annotation for information extraction, significantly reducing human labor while maintaining high accuracy in generating labeled datasets. Several recent approaches exploit the in-context learning capabilities of LLMs based on a limited number of examples (few-shot) to create annotations, using suitably engineered prompts [3,4,5]. Other recent works discuss the usage of LLMs for knowledge distillation [6,7], or even how LLMs could themselves be used as evaluators [8]. Finally, several studies evaluate the reliability of the annotations generated by LLMs [9,10].

While opportunities with LLMs are actively being explored, it is equally important to be aware of the potential pitfalls that may arise during their use. The limitations and risks associated with using LLMs have been thoroughly examined in previous studies [11]. Some research has explored these challenges within the contexts of biology and medicine [12,13], offering more specific case studies and proposing mitigation strategies. These insights provide invaluable guidance that should be shared with researchers in the field to help avoid unnecessary risks and complications.

2. Workshop

The years 2022 and 2023 marked the emergence of Large Language Models (LLMs). Reflecting this pivotal shift, PSB2024 organized a workshop entitled "Large Language Models (LLMs) and ChatGPT for Biomedicine," aimed at providing introductory insights into LLMs within the realm of Biomedicine. In the meantime, a wealth of diverse experiences with LLMs has been accumulated, and the emphasis of the workshop will be on sharing these varied encounters. As such, presentations showcasing a spectrum of application cases of LLMs have been considered, encompassing both successful implementations and instances where expectations were not met.
The intention is to focus in particular on the impact of LLMs on biomedical annotation and curation. Some of the issues and questions to be addressed in the workshop include but not limited to:

- Are annotation and curation still necessary in the age of LLMs?
- Can LLMs replace those completely?
- How can we assess the quality of automated annotations?
- What are the limitations?

By addressing these challenges this workshop aims to clarify the potential and limits of LLMs in advancing biomedical research and knowledge discovery.

3. Conclusions

LLMs are already making major inroads in our social fabric, rapidly changing the way several highly skilled activities are performed, and leading to serious challenges to societal organization and profound questions about how to best make use of their capabilities for the advantage of humanity. We hope that this workshop will offer a valuable contribution to this ongoing discussion.

4. Acknowledgments

We are grateful to the PSB 2025 organizing committee for enabling us to organize this workshop.

References

1. Tan, Z., Beigi, A., Wang, S., Guo, R., Bhattacharjee, A., Jiang, B., … Liu, H. (2024). Large Language Models for Data Annotation: A Survey. doi:10.48550/arXiv.2402.13446
2. Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L.H., Hao, X., Jaber, B., Reddy, S., Kartha, R., Steiner, J., Laish, I. & Feder, A.. (2023). LLMs Accelerate Annotation for Medical Information Extraction. Proceedings of the 3rd Machine Learning for Health Symposium, in Proceedings of Machine Learning Research, 225:82-100 doi:10.48550/arXiv.2312.02296

3. Choi, J., Lee, E., Jin, K., & Kim, Y. (2024, March). GPTs Are Multilingual Annotators for Sequence Generation Tasks. In Y. Graham & M. Purver (Eds). Findings of the Association for Computational Linguistics: EACL 2024 (pp. 17–40). https://aclanthology.org/2024.findings-eacl.2
4. He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., … Chen, W. (2024, June). AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. In Y. Yang, A. Davani, A. Sil, & A. Kumar (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 165–190). doi:10.18653/v1/2024.naacl-industry.15
5. Smith, R., Fries, J. A., Hancock, B., & Bach, S. H. (2024). Language Models in the Loop: Incorporating Prompting into Weak Supervision. ACM / IMS J. Data Sci., 1(2). doi:10.1145/3617130
6. Tan, S., Tam, W. L., Wang, Y., Gong, W., Zhao, S., Zhang, P., & Tang, J. (2023, July). GKD: A General Knowledge Distillation Framework for Large-scale Pre-trained Language Model. In S. Sitaram, B. Beigman Klebanov, & J. D. Williams (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track) (pp. 134–148). doi:10.18653/v1/2023.acl-industry.15
7. Gu, Y., Dong, L., Wei, F., & Huang, M. (2024). MiniLLM: Knowledge Distillation of Large Language Models, The Twelfth International Conference on Learning Representations, https://openreview.net/forum?id=5h0qf7IBZZ.
8. Chiang, C.-H., & Lee, H.-Y. (2023). Can Large Language Models Be an Alternative to Human Evaluations?, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, pages 15607–15631, https://aclanthology.org/2023.acl-long.870.
9. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30). doi:10.1073/pnas.2305016120
10. Honovich, O., Scialom, T., Levy, O., & Schick, T. (2022). Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14409–14428,, https://aclanthology.org/2023.acl-long.806
11. OpenAI. (2023) GPT-4 System Card. OpenAI Research Papers. https://cdn.openai.com/papers/gpt-4-system-card.pdf
12. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. (2024) Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. Ann Intern Med. 2024

Feb;177(2):210-220. doi: 10.7326/M23-2772. Epub 2024 Jan 30. https://www.acpjournals.org/doi/10.7326/M23-2772

13. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, Yang Y, Chen Q, Kim W, Comeau DC, Islamaj R, Kapoor A, Gao X, Lu Z. (2023) Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. 2023 Nov 22; 25(1):bbad493. doi:10.1093/bib/bbad493. https://academic.oup.com/bib/article/25/1/bbad493/7505071

Earth Friendly Computation: Applying Indigenous Data Lifecycles in Medical and Sovereign AI

Keolu Fox[1,2,5], Krystal Tsosie[2,3], Alex Ioannidis[4], Kaja Wasik[5], Alec Calac[1] & Eric Dawson[6]

[1]University of California San Diego, [2]Native BioData Consortium, [3]Arizona State University, [4]University of California Santa Cruz, [5]J. Craig Venter Institute, [6]Nvidia

1. Overview

The rapid expansion of medical artificial intelligence (AI) is generating vast amounts of data, leading to increased energy consumption and a growing environmental footprint. While this growth is advancing healthcare, it also presents the risk of worsening the climate crisis, which could impact global health. The session, "Earth Friendly Computation: Applying Indigenous Data Lifecycles in Medical AI" at PSB 2025, will explore how Indigenous communities have historically approached technology development with a focus on sustainability and long-term resource management, in contrast to Western societies' focus on resource extraction and maximizing profit through exponential growth.

Indigenous communities have long optimized technologies in ways that prioritize harmony, balance, and the maintenance of abundance rather than short-term gains. This approach stands in stark contrast to the Western model of technological advancement, which prioritizes profit and efficiency, often leading to environmental degradation. The EFC 574 initiative embodies this Indigenous approach by proposing sustainable, renewable-powered data centers on tribal lands, promoting data sovereignty while addressing climate resilience. Complementary to building this infrastructure, the AI in Point-of-Care (POCT) project harnesses edge computing to deliver AI-driven diagnostics to remote areas, reducing energy consumption and supporting healthcare in resource-limited regions.

The projects featured in this session are supported by initiatives like the Univeristy of California Systems, *California Cares* Policy initiative, the Canadian government's *Abundant Intelligences* initiative and the University of Cambridge's, *Green Algorithms* initiative, which aim to create transparency around the environmental impacts of AI through carbon footprint calculators and incentivization programs. Together, these efforts demonstrate how Indigenous communities are positioned not just to participate in the AI revolution, but to lead it by prioritizing sustainable development that maintains harmony with the planet's health. Rather than optimizing every advancement for profit and exponential growth, this framework advocates for a future driven by industrial symbiosis and long-term relationships with the earth, ensuring technology serves both people and the environment for generations to come.

2. Background & key terms

"We live in the future. Come join us."
—Prof. Bryan Kamaoli Kuwada

For the past five centuries, Indigenous communities have directly witnessed seismic shifts in the integrated global economy, from natural resource extraction such as that of spices, fur, wood, oil, uranium, and tourism, to the rise of movies, gaming, casinos, and other parts of the entertainment and attention economy. These industries have served as vital sources of income, yet they are also subject to the winds of change, as generations pass and the global economy evolves. Despite the rapid advancement of technology the underlying economic models and values often remain static, unresponsive to the shifting technological landscape or the rights and needs of the communities participating in them. As we stand on the brink of a new era, it is clear that data is emerging as the next pivotal economic base. The ten largest companies on earth no longer deal in steel, railroads, or oil, but in vast amounts of data and the tools needed to process it. This shift demands not only technological acumen to overcome the pressing challenges of scaling these new technologies but also a reevaluation of our values and the way we envision progress for future generations.

Tired stereotypes of Indigenous communities like - "*Hawaiian people have 1,000 ways to describe the rain*" - obscure the truth that in these high resolution environmental or natural observations are also "*our data*," informed by thousands of years of observations, input, and action. This collective knowledge, and the practices developed from it, will be critical to spurring the transition into an economy that incorporates artificial intelligence in a way that is sustainable for the planet and individual and collective sovereignty.

The right of Indigenous peoples to own, control, access, and possess data that pertains to them, their lands, and their cultures is called *Indigenous data sovereignty*. It is rooted in the principle that Indigenous nations hold the inherent authority to govern themselves and manage their own affairs. *Indigenous data governance*, on the other hand, refers to the mechanisms, processes, and systems through which Indigenous communities control and manage this sovereign data. Functionally, this means developing policies, standards, and practices for data collection, storage, access, and dissemination that align with the community's values and needs.

Indigenous peoples have historically been separated from their resources using violent means. In 1965, Fairchild Semiconductors - which would soon lose founding engineers that would go on to found Intel and other semiconductor behemoths opened their assembly plant in Shiprock, New Mexico on the Navajo nation reservation. At its peak, the plant employed over a thousand Navajos, the majority of whom were women. Yet despite simultaneously drawing on Navajo women's expertise and exploiting their labor, never once were Navajo or Indigenous knowledge systems considered, integrated, or acknowledged as guiding forces in the development of integrated circuit architecture and what would later become an entire economy that today spans cloud computation, data center architecture, e-waste management, and parallel computing (*See, Figure 1*). Until now.

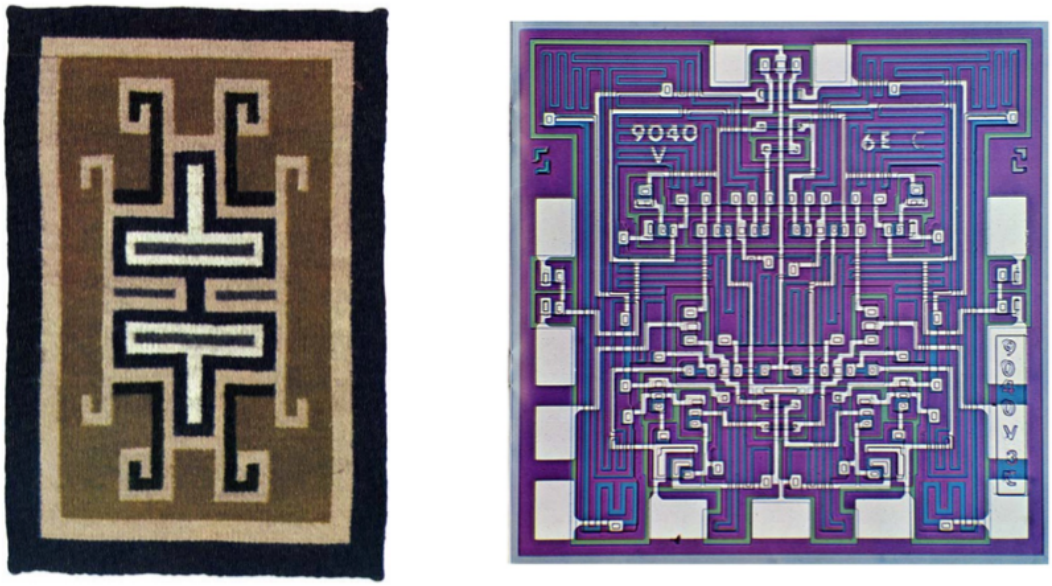


Fig 1. Left: A traditional Navajo rug; Right: the die layout of a Fairchild 9040 integrated circuit. These images are adapted from a Fairchild commemorative brochure for the 1965 opening of a new manufacturing plant on the Navajo Nation Reservation in Shiprock, NM. At this facility, Navajo women were specifically recruited to assemble early integrated circuits because their exceptional dexterity and precision, developed through generations of intricate rug-weaving, were invaluable for handling delicate components. The rug-weaving traditions, which involved complex pattern recognition and fine motor skills, directly translated into their ability to work with early microelectronics.

3. Earth Friendly Computation 574: Indigenous Data Sovereignty, Circular Systems, and Solarpunk Solutions for a Sustainable Future

Indigenous communities are not only envisioning but actively leading efforts to reshape AI through Earth-friendly principles. *Alipio et al.* present a policy and technology development proposal rooted in the idea of Indigenous Data Sovereignty and the establishment of decentralized, sustainable computing systems built on sovereign Indigenous lands. The number "574" refers to the 574 federally recognized Indigenous tribes in the United States, each uniquely positioned to lead this initiative.

This approach draws inspiration from Indigenous knowledge systems that promote circular economies and sustainable land stewardship, integrating these values with modern technology solutions. For example, the Hawaiian ahupua'a land management system, which operates on closed-loop cycles of resource renewal, provides a model for designing data centers that reuse resources, minimize waste, and generate less heat.

Here are some key recommendations for the development of policy and technology to decentralize sustainable access to machine intelligence among Indigenous communities:
- *Indigenous Data Sovereignty:* EFC 574 empowers the 574 federally recognized tribal nations to build and maintain control over their own data infrastructure, ensuring data governance that aligns with cultural values, autonomy, and the unique legal frameworks of sovereign tribal lands.
- *Sustainable Data Center Development:* This initiative focuses on creating environmentally sustainable data centers on tribal lands by integrating renewable energy sources like solar, wind,

and hydroelectric power. These centers will serve as models for climate-resilient and sovereign data management.
- *GIS-Based App for Sustainable Planning:* EFC 574 will develop an app that overlays GIS data onto Indigenous reservation topography and coordinate data, combined with renewable energy availability (solar, wind, hydroelectric), megawatt capacity, and biodiversity data. This tool will determine the most sustainable locations for data center architecture on sovereign tribal lands, promoting responsible and efficient development.

By leveraging ideas from *Solarpunk*—a speculative genre that imagines harmonious relationships between humans, technology, and the environment—Indigenous-led initiatives can guide the development of *circular data systems*. These systems would recycle electronic components, repurpose waste heat for energy generation, and use renewable power sources like solar and wind energy to fuel AI development.

4. AI in Point-of-Care: A Sustainable Healthcare Revolution at the Edge

AI-enhanced *Point-of-Care Testing (POCT)* is transforming healthcare by bringing diagnostics closer to patients, particularly in underserved areas. AI-driven POCT offers rapid, real-time diagnostic insights, reducing the need for centralized laboratory testing and improving healthcare outcomes in regions with limited access to medical infrastructure. However, this progress comes with a cost: the increased use of AI in healthcare systems exacerbates energy consumption and contributes to e-waste through the frequent upgrading of diagnostic devices.

To address these concerns, *Rajput et al.* present *edge computing* as a promising solution. By processing data locally on devices closer to the point of care—rather than relying on cloud-based infrastructure—edge computing reduces energy consumption and lowers the environmental impact of AI-driven healthcare technologies. This localized processing significantly reduces latency and energy costs associated with transmitting large amounts of data to distant data centers.

Here are some key recommendations for the utilization of AI in edge computing settings:
- *AI-Enhanced POCT:* AI reduces diagnostic latency in hospitals, delivering real-time results crucial for urgent care and improving access in underserved areas.
- *Scalability vs. Sustainability:* Balancing AI scalability with sustainability is key, favoring energy-efficient models over large, resource-intensive LLMs.
- *LLM Limitations:* LLMs aren't always ideal for point-of-care; careful selection of AI models is necessary to meet healthcare needs without increasing environmental impact.

In healthcare settings, edge computing can support *AI-driven genome sequencing*, *disease diagnostics*, and *patient monitoring systems*, enabling faster and more accurate medical decisions while minimizing energy use. For instance, edge-powered AI devices have shown success in reducing diagnostic times for infectious diseases like COVID-19 and improving personalized care in intensive care units (ICUs). These innovations illustrate that sustainable AI is not only possible but can also enhance healthcare delivery, making it more accessible, eco-friendly, and efficient.

5. Conclusion: The Future of Earth Friendly Computation

The future of Earth-friendly computation is one where AI development coexists with environmental sustainability and technological sovereignty. Indigenous communities, leading with principles of data sovereignty, circular systems, and land stewardship, are charting a path toward a

more sustainable digital future. From reducing the carbon footprint of algorithms to promoting secondhand markets for GPUs, these innovative solutions offer a blueprint for mitigating the environmental impact of AI.

In the coming years, it is essential to prioritize the development of green algorithms, sustainable hardware, and decentralized computing systems that emphasize energy efficiency and waste reduction. By incorporating Indigenous knowledge and practices into the fabric of AI development, the world can harness the power of AI for the greater good, while preserving the health of the planet. Sovereign AI should be informed by the knowledge hard-won by Indigenous data scholars over many generations.

The integration of edge computing into healthcare systems, the expansion of secondhand GPU markets, and the implementation of policies like Earth Friendly Computation 574 represent critical steps in achieving this vision. As demand for AI, its technological capabilities, and the systems available to train and use these tools continue to grow, so too must our commitment to ensuring that this growth aligns with the needs of both the planet and future generations. Aligning AI development and deployment with Indigenous data lifecycles and principles makes it possible to advance these technologies while preserving our planet's past, present, and future (*See, figure 2*).

Fig 2. Art by Wally Dion (Canadian and Yellow Quill First Nation/Saulteaux, born 1976). Left: "*Green Star Quilt* (*2019)*." E-waste, circuit boards, brass wire, copper tube. Right: "*Caterpillar, Egg, Chrysalis, Moth (2018).*" E-waste, circuit boards on plywood, nails. Serves as a powerful commentary on humanity's hidden environmental toll, symbolizing the lifecycle of AI and data-driven technologies. The artwork, crafted from discarded e-waste and circuit boards, evokes the transformation of a moth, paralleling the unseen extraction of rare earth metals and natural resources required to sustain our addiction to data centers, AI, and cloud computation. It highlights the environmental cost and the unsustainable hunger for energy that drives the digital age.

6. Acknowledgments

We extend our gratitude to the anonymous reviewers who contributed to the peer review process for this session's submissions, and to Indigenous communities worldwide, whose data we are committed to safeguarding.

References

Alipio, K., García-Colón, J., Boscarino, N., & Fox, K. (2025). Earth Friendly Computation 574: Indigenous Data Sovereignty, Circular Systems, and Solarpunk Solutions for a Sustainable Future. In *Pacific Symposium on Biocomputing*.

Rajput, Y., Tarif, T., Wolfe, A., Dawson, E., & Fox, K. (2025). AI in Point-of-Care: A Sustainable Healthcare Revolution at the Edge. In *Pacific Symposium on Biocomputing.*

Indigenous Data Sovereignty, Circular Systems, and Solarpunk Solutions for a Sustainable Future

Ka'ulawena Alipio
Department of Anthropology, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, United States
Email: Kalipio@ucsd.edu

Javier García-Colón
Department of Anthropology, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, United States
Email: jgarciacolon@ucsd.edu

Nima Boscarino
Department of Communication, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, United States
Email: nboscarino@ucsd.edu

Keolu Fox
Department of Anthropology, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, United States
Email: pkfox@ucsd.edu

Recent advancements in Artificial Intelligence (AI) and data center infrastructure have brought the global cloud computing market to the forefront of conversations about sustainability and energy use. Current policy and infrastructure for data centers prioritize economic gain and resource extraction, inherently unsustainable models which generate massive amounts of energy and heat waste. Our team proposes the formation of policy around earth-friendly computation practices rooted in Indigenous models of circular systems of sustainability. By looking to alternative systems of sustainability rooted in Indigenous values of *aloha 'āina*, or love for the land, we find examples of traditional ecological knowledge (TEK) that can be imagined alongside Solarpunk visions for a more sustainable future. One in which technology works with the environment, reusing electronic waste (e-waste) and improving data life cycles.

Keywords: sustainability, cloud computing, Indigenous Data Sovereignty, environmental policy, heat waste, solarpunk, Indigenous futurism

1. Data Center Infrastructure and Environmental Impact

The energy industry is the top contributor to the growing climate crisis, with fossil fuel burning being the main culprit for decades. However, in recent years the rapid growth of data center infrastructure — driven by advancements in Artificial Intelligence (AI) and Graphics Processing Unit (GPU) technology — has become a significant contributor to global heat emissions and electronic waste (e-waste). The GPU, originally designed for rendering 3D graphics, has accelerated AI training, contributing to immense concentrations in heat waste from data centers. Modern bioinformatics, like other disciplines reliant on large-scale computation, is contributing to the environmental impact of data centers (Grealey et al., 2022). This rise in heat production poses critical questions about the environmental cost of accelerated computing in both the AI sector and biocomputing more broadly, necessitating new policies and sustainable infrastructure to mitigate these impacts.

As the AI sector has continued to boom in recent years, several large corporations have cornered the market in key areas including NVIDIA's monopoly as the sole producer of the GPU, and Amazon Web Services as they have the money and reach to build numerous data centers around the globe. This should concern communities who wish to retain control and ownership over their data. NVIDIA's monopoly over the GPU led to a "141% increase in the company's data center segment compared to the previous quarter" indicating the massive interest in AI technology (De Vries, 2023).

Additionally, companies like Microsoft and Google's parent company Alphabet have led the push to integrate generative AI into basic web searches, creating their chatbots Bing Chat and Bard, after seeing the success of Open AI's ChatGPT, leading to a significant increase in energy use and heat waste (De Vries, 2023). In a 2023 study, researchers estimated that if generative AI was integrated into every Google search, it would result in an energy use increase as high as 29.3 TWh per year which is the equivalent of the yearly energy use of an entire country the size of Ireland (De Vries, 2023). Despite the fact that this study was only published last year and the authors predicted that Google would not integrate AI into their searches due to the significant energy and monetary costs – an estimated 100 billion U.S. dollars for a year of server space – at the time of writing this article, Google currently has generative AI answers at the top of most Google searches. When considering the fact that 2023 and 2024 have consistently broken heat records, the need to investigate the environmental impacts of AI and data center infrastructure is more pressing than ever.

1.2 Current infrastructure model

The current infrastructure model for deciding where to build data centers is based around Western ideologies of economic gain, terraformation, and extraction of resources. These values are inherently opposed to sustainability and in direct conflict with environmental protection laws

such as the Executive Orders on Climate issued by President Joe Biden in January 2021 that seek to reduce emissions to net zero economy-wide by 2050 and emphasize the transition to clean energy (U.S. Council on Environmental Quality, 2021). Despite these types of laws, current environmental oversight on how data center infrastructure contributes to these emissions does not do enough. Current policies surrounding the environmental impact of AI call for "voluntary reporting" from data centers on the amounts of energy used and how their infrastructure affects the environment (Markey, et.al. 2024). Voluntary reporting is flawed in its conception because statistics are often underreported and there is no oversight or enforcement. Instead, these bills serve as a smokescreen for companies to hide behind while pretending that they are altruistically reporting their emissions. Therefore, the most popular cloud service companies including Amazon Web Services, Microsoft Azure, and Google Cloud, hold a majority of the power in the cloud computing market and are projected to continue making "Year-over-Year" (YoY) increases in profit revenue (AAG IT, 2024).

Data center locations are notoriously secretive and the largest companies within the sector, Amazon Web Services, Google, and Microsoft, often go to great lengths to not advertise their locations. Buildings are usually non-descript with high security measures as the only indication that there is valuable data being held inside. However, although these data centers may be inconspicuous to the average passerby, they are not invisible under infrared visualization. Data centers can be located through heat mapping GIS sensing technology which highlights the central problem with the current "status quo" – the immense amount of heat being generated by computational action (Johnson-Zafiris, 2024).

1.3 Proposed Policies: Earth Friendly Computation (EFC 574)

In this paper, we propose the formation of policies around the construction and location of data center infrastructure, entitled "EFC 574" which stands for Earth Friendly Computation among the 574 federally recognized Indigenous tribes located in the so-called United States (*See Figure 1*). Proposed policies would be structured around Indigenous values of land stewardship, circular systems of sustainability, and data sovereignty. Drawing upon lessons from the past and applying them to the future, we can begin to imagine a world in which technology and nature are intertwined harmoniously, rather than at odds with one another.

Building upon discourse from environmental justice and Indigenous futurism through the lens of the genre Solarpunk, we propose alternative solutions rooted in tangible decolonial actions of sovereignty. Inspired by the wisdom of Robin Wall Kimmerer's *Braiding Sweetgrass*, we hope to show how Indigenous values of sustainability and working to serve nature through cycles of renewal instead of linear extraction can be visualized in the data center industry (Kimmerer, 2013). In doing so, we hope to provide a blueprint for Indigenous Data Sovereignty rooted in the sovereignty of our lands.

Figure 1: Indian Lands of Federally Recognized Tribes of the United States
Map of the 574 federally recognized tribes and their regional designations within the United States from the Bureau of Indian Affairs.

2. Indigenous Leadership in Earth-Friendly Computation

Current models for data center construction contribute to the already devastating effects of climate change and colonial extraction on Indigenous communities. From diverting and polluting water tables, to mining and extracting dangerous elements for technological components, extractive processes currently threaten the already precariously limited resources in areas like drought-prone Nevada, or the rising sea levels and sand erosion of the California coast (Fernández‑Llamazares et.al., 2020). As highlighted by Potawatomi scholar Kyle Whyte, climate mitigation strategies often fail to consider how they will impact Indigenous communities, who disproportionately experience the impacts of climate change, leading to a dilemma in which urgent races to adapt to climate change and prevent increases in temperature are implemented without the necessary "kin relationships" to sustain such changes (Whyte, 2019). Instead, EFC 574 provides a pathway for Indigenous communities to further empower Indigenous Data Sovereignty principals through decentralized data center infrastructure. Indigenous communities are poised to lead the way in earth-friendly computation policies by shifting the focus away from Western ideologies of linear consumption and toward sustainable practices inspired by closed loop or circular economy systems like the Hawaiian ahupua'a, a land division system based on the value of aloha ʻāina, the Hawaiian term for circular systems of care in which the people and the land are continuously reciprocating care for each other (Vitousek and Beamer, 2013; Beamer, 2023; Smith, 2021). In Hawaiian resource management, every system is connected to feeding

each other; fish feed the people, leftover food waste feeds the fish, the fish ponds make the soil fertile, and the plants grow strong, which in turn feed the people and the fish (Beamer, 2023). This sort of closed loop circular economy allowed island communities throughout the pacific to thrive for centuries, developing nuanced techniques for recycling resources (Vitousek and Beamer, 2013) . Importantly, we emphasize the weaving of Traditional Ecological Knowledge (TEK) like the ahupua'a system and aloha 'āina values with Western technology and imaginative concepts like Solarpunk, rather than absorbing one into the other (Whyte, 2013).

Applying such an approach to data center infrastructure, we suggest the creation of a federated decentralized network where Indigenous communities can choose to retain control and possession of their data rather than sending it off to a centralized cloud at a large corporation like Amazon Web Services (Boscarino et al., 2022). By creating a decentralized network of data centers where servers are built on Indigenous sovereign land, with sustainable technologies that aim to work with the environment rather than against it, we aim to enact "good relations with the land" as described by Métis scholar Max Liboiron in *Pollution is Colonialism* (Liboiron, 2021). Liboiron's use of anticolonial science to critique the plastic industry provides a framework for moving away from Western ideologies of extraction and unsustainable profit. Similarly, we hope to show how being in good relations with the land is important for sustainable data computation. Land is central to this argument; the concept of data sovereignty and its connection to the land is important since data can not be considered sovereign if it is not on sovereign land.

2.2 Addressing Heat Waste through Circular Systems

In addition to data sovereignty concerns, the heat waste generated by centralized cloud computing contributes to the current climate crisis through the demands of air conditioning, water extraction for liquid cooling, and the creation of heat waste as a byproduct. Several solutions have been proposed including Sea Water Air Conditioning (SWAC) a method of cooling which relies on a network of cooling tubes that run continuous seawater through in order to cool down components (Elahee and Jugoo, 2013). This is one solution that could be helpful in areas along coastlines, particularly in communities where sea levels are rising quickly and forcing the abandonment of structures. If these structures could be reinforced for SWAC and repurposed into data centers this would be one example of earth friendly computation where technologies work with nature rather than against it. However, it is important to consider the ways in which these technologies can be misappropriated. While utilizing water to cool components might be a helpful tool, we must be careful not to reproduce the same kinds of water exploitation and pollution that are already occurring throughout Indigenous lands (Fernández‐Llamazares et.al., 2020).

Inspired by our Indigenous and Islander roots, we propose looking to circular systems of sustainability like "the *moku* system" of "biocultural resource management" in the Hawaiian islands (Winter et al., 2018). This pre-contact system divided the islands into socio-ecological sections from mauka to makai (mountain to ocean) called an ahupua'a, which are often referred

to as models of sustainability (Winter et al., 2018). This system was unique because it took into account all the different regions, weather patterns, resources, and conservation needs that were unique to various parts of the island. The Hawaiian islands include a diverse range of climate types, with the Koppen classification system being used as early as 1937 to highlight the five different types of climate commonly experienced including; humid tropical, arid and semi-arid, temperate cold continental, and ice or alpine climates (Jones and Bellaire, 1937; Peel et al., 2007). Given the diversity of climate and the limited landmass, the *moku* system was developed to be sustainably integrated into the landscape demonstrating how watershed systems can be implemented in data center construction without having to shape the earth to bend to the will of current infrastructure models. Drawing from these systems of knowledge we propose looking to speculative visions of the future for inspiration on policy regarding sustainable data center infrastructure and earth-friendly computation.

2.3 E-waste and Data Centers

In addition to concerns around heat waste, electronic waste (e-waste) is also a major issue in current data center infrastructure. One proposed solution is to repurpose old GPU technology to build out these sovereign data centers. Technological components under the current standard are often built with planned obsolescence in mind, pushing for constant upgrades in order to maintain performance. However, there are still many less intensive computing tasks that these older components could be repurposed to handle. Therefore, in addition to rethinking the locations of data centers, our team also proposes the formation of policy around recycling GPU and CPU components into these new sovereign data centers in order to maintain the cyclical pattern of reuse and renewable energy. This is where envisioning futures through the lens of the Solarpunk genre can be helpful, as it allows us to step outside the current energy landscape and imagine what a stronger investment in solar energy use could look like for data centers.

2.4 Indigenous Futurism and Solarpunk Solutions

Solarpunk is a term that describes the utopian vision of a futuristic society in which technology is intertwined with nature and facilitates sustainable innovation and development (Wagner and Wieland, 2022). Indigenous communities around the world have been inspired to envision futures through the lens of these Solarpunk themes coupled with already established networks of Indigenous sovereignty (Reina-Rozo et al., 2024). Indigenous futurism is about taking the values that our communities are built around – such as sustainability, circular closed loop systems, and TEK practices – and forming policies that will begin to push back against the centuries of colonial extraction and destruction of ecosystems. In turning to such solutions, we begin to imagine a world that unsettles the status quo, in which data centers are thought of as inert and isolated ‘heat objects’. Instead, data centers are ontologically understood as non-human kin (TallBear, 2017; Lewis et al., 2018), whose heating, cooling, and other infrastructural needs are always already entangled with their landscape in harmonious circular systems of reciprocal renewal.

For example, in the case of heat waste, excess heat generated by machine components inside a data center could be harnessed and redirected into other appliances that would otherwise need to generate more energy for heating. Imagine a data center located next to a gym or a salon where the excess heat could be funneled into a heating, ventilation and air conditioning (HVAC) system attached to heated floors or hooded hair drying stations. In a cold environment it could be redirected into district heating for homes (Ding et al., 2024). The applications are endless if we just shift our thinking from "we must create more energy" to "how do we re-use the energy we already have" and the first step in this paradigm shift is to re-think the way data centers are constructed and located.

Our team proposes reforming the data center industry by emphasizing these values through public policy. Building on the already established networks throughout Indigenous communities we hope to build policy around the establishment of data center nodes located on sovereign Indigenous lands. Tribal Nations would be asked if they would like to establish their own independent data centers on their land and should they choose to share any pieces of information with other communities, organizations, or other entities, they would have full control and autonomy to do so without having to go through third party companies that could exploit or endanger their data.

3. Importance of Sovereign Indigenous Nations Controlling Their Data

Current models for cloud computing rely on centralized networks that force users to surrender consent, ownership, and other rights in order to use those services (Boscarino et al., 2022; Mackey et al., 2022). This should be considered particularly alarming in the context of genomic data and the future of precision medicine. Standard Practices for storing data do not take into account the massive multiplicity of data being generated through genetic research, both by for-profit companies like 23 and Me, and by health industries (Mackey et al., 2022). Indigenous communities can be leaders in this space by implementing decentralized federated networks for genetic research at places like The Native BioData Consortium (NBDC) (Boscarino et al., 2022). Located in The Cheyenne River Reservation, NBDC, would be the first demonstration of one of these data center "nodes" serving as an example of the power of Indigenous data sovereignty applications in the medical field.

This is important for many reasons, first through the implementation of edge computing, which aims to bring technology out of the lab and into the field, data could be kept closer to the point of collection, creating less computational energy (Raith et al., 2023). Secondly, this would allow Indigenous communities to choose with whom and under what circumstances they would like to share such data giving them more autonomy over their data (Mackey et al., 2022). Additionally, this would also be an opportunity for communities to engage with meaningful medical research that is urgently needed in their specific communities. As mentioned by Tuck and Yang in their highly cited paper, *Decolonization is not a metaphor,* the decolonization of the data and tech services needs to be grounded in sovereignty and that is rooted in the land (Tuck and Yang, 2012). Ownership and control over Indigenous peoples' data is integral to the fight for

sovereignty and decolonization (Walter and Carroll, 2020). When a community is forced to use a large corporation for their data processing services, they have to sign away at least some part of ownership, consent, and/or rights to that data. All too often "green computing" solutions are offered in ways that systemically harm Indigenous communities by clustering benefits around the most wealthy and pushing the harm on to marginalized communities. In the context of data center infrastructure, the result is building centers in low income areas, diverting water away from the community for machine cooling, and venting heat waste and other harmful emissions into the surrounding air. Instead, we propose a solution that is oriented in Indigenous epistemologies of sustainability and circular systems, based on TEK maintained by Indigenous communities around the world since time immemorial (Vitousek and Beamer, 2013; Whyte, 2013; Smith, 2021).

3.1 Digital Anthropology and Indigenous Data Sovereignty

Frameworks from the newly rising field of digital anthropology allow us to conceptualize the massive amounts of data being collected and stored in cloud servers. In recent years scholars have highlighted the need for digital archivists and anthropologists who can apply their archaeological excavation skills to the digital realm (Geismar and Knox, 2021). This should raise concern for communities who have a history of being exploited and their data stolen. As more and more information is stored on cloud servers and housed in data centers, this build up of digital clutter generates massive amounts of heat waste driving an ever rising need for additional cooling systems. This in turn leads to more exploitation of Indigenous lands and values for the sake of maintaining unsustainable systems. Through the formation of policy around data center construction and infrastructure and by emphasizing the need for Indigenous perspectives in shaping future sustainable policy, we hope to provide an alternative path for communities to opt out of such digital excavation and retain full sovereignty and control over their data.

3.2 Biological Data

The importance of Indigenous Data Sovereignty and the need for a decentralized data center network can be seen in examples of genomic data related to human health as well as environmental data. Since the inception of the Human Genome Project scientists and entrepreneurs have been racing to mine and map the human genome in order to commodify and control specific genes for the sake of drug development (Sunder Rajan, 2006). This has created a dangerous mainstream framework for economic value to be the main driver behind decisions about data use including whose genome gets studied and for what purposes.

Similarly, in the case of environmental data, the need for bioremediation solutions to clean up pollution through the use of synthetic biology and metagenomics, presents a precarious dilemma. Researchers who want to develop solutions, but fear having those genetic discoveries patented by large corporations which in turn sell the solutions back to the communities in need for a profit are in a difficult position. Synthetic biology is poised to harness the power of

biotechnology to create bioremediation solutions (Rylott, 2020), but at what cost to the communities which are most affected by climate change and pollution? Instead, through the lens of a circular economy, we propose protections for Indigenous Data Sovereignty that go hand in hand with our proposed policies for more sustainable data center infrastructure.

3.3 New GIS Tools to Guide the Development of Policy

Advancements in Geographic Information Systems (GIS) technology have decentralized the power of map making, allowing more communities to have a say in their territories' self-determination. For example, GIS mapping technology has been used to track the repercussions of the Morrill Act of 1862, which granted land to U.S. colleges and universities by expropriating nearly 11 million acres of Indigenous land. This land, scattered across 24 Western states, became the financial foundation for many of today's top universities. The money raised from land sales remains on university ledgers today, and some states still hold unsold parcels and mineral rights, which continue generating revenue for higher education institutions. The act masked a massive wealth transfer, contributing to the violent history of North American colonization. Scholars have used GIS tools to map how these land transfers occurred showing the money trail and corruption through story maps (Ahtone and Lee, 2020). Additionally, Scholars are already implementing GIS technology to combat environmental risks with algorithms that analyze and predict complex wildfire patterns (He, 2022).

As the demand for data centers escalates, so does the urgency around identifying environmentally favorable locations for their construction. We propose the development of policy utilizing tools to assess and identify the most appropriate way to integrate data center infrastructure into existing landscapes. Considering variables such as climate vulnerability, natural resource availability, biodiversity hubs and other unique environmental variables, we plan to build out a prototype resource tool which Indigenous communities can be in control of for strategic guiding of sustainable data center development (*See, Figure 2*). Looking at the Agua Caliente Band of Cahuilla Indian (ACBCI) Reservation as an example, due to its unique reservation boundaries which form a sort of checkerboard pattern, we can see that there are solar and wind resources which could be beneficial in building out sustainable data centers, as well as areas with high biodiversity which should be considered and protected as infrastructure is developed. Looking at the Intersection of Resources map, we see that the northwestern quadrant of the ACBCI Reservation where solar and wind resources are abundant but biodiversity hubs are not present, would be the most ideal place to construct a data center. Our aim is to guide decision-makers, including policymakers and industry leaders, in coming up with informed choices that balance the need for data center expansion with environmental responsibility. Our suitability map tool would include geolocating information to make these sites easy to identify. Ultimately, we strive to promote sustainable development in the digital infrastructure sector, ensuring a greener and more efficient future for data centers worldwide.

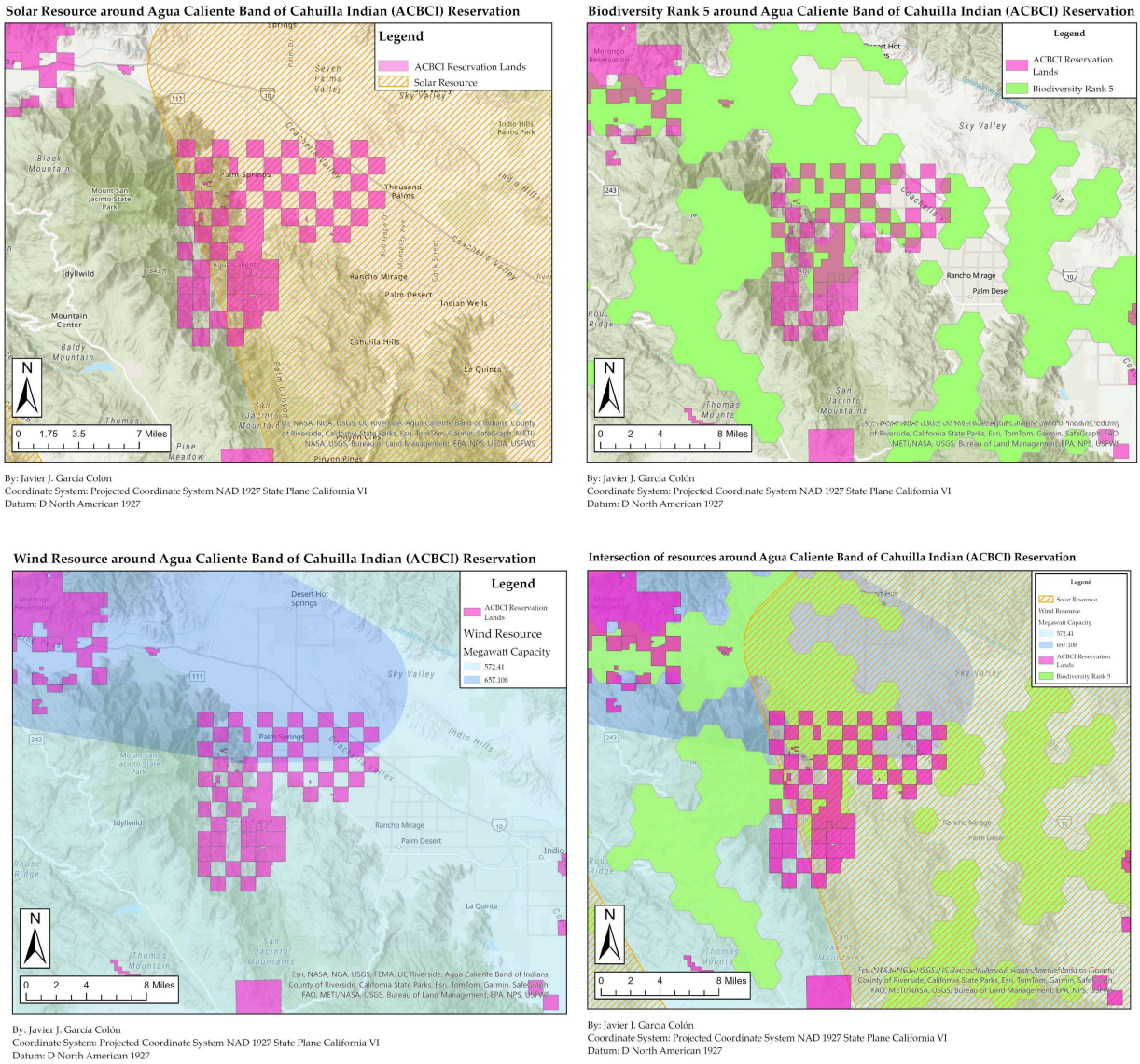


Figure 2: Suitability Maps for Data Center Locations within The Agua Caliente Band of Cahuilla Indians (ACBCI) Reservation *(see appendix for larger maps)*

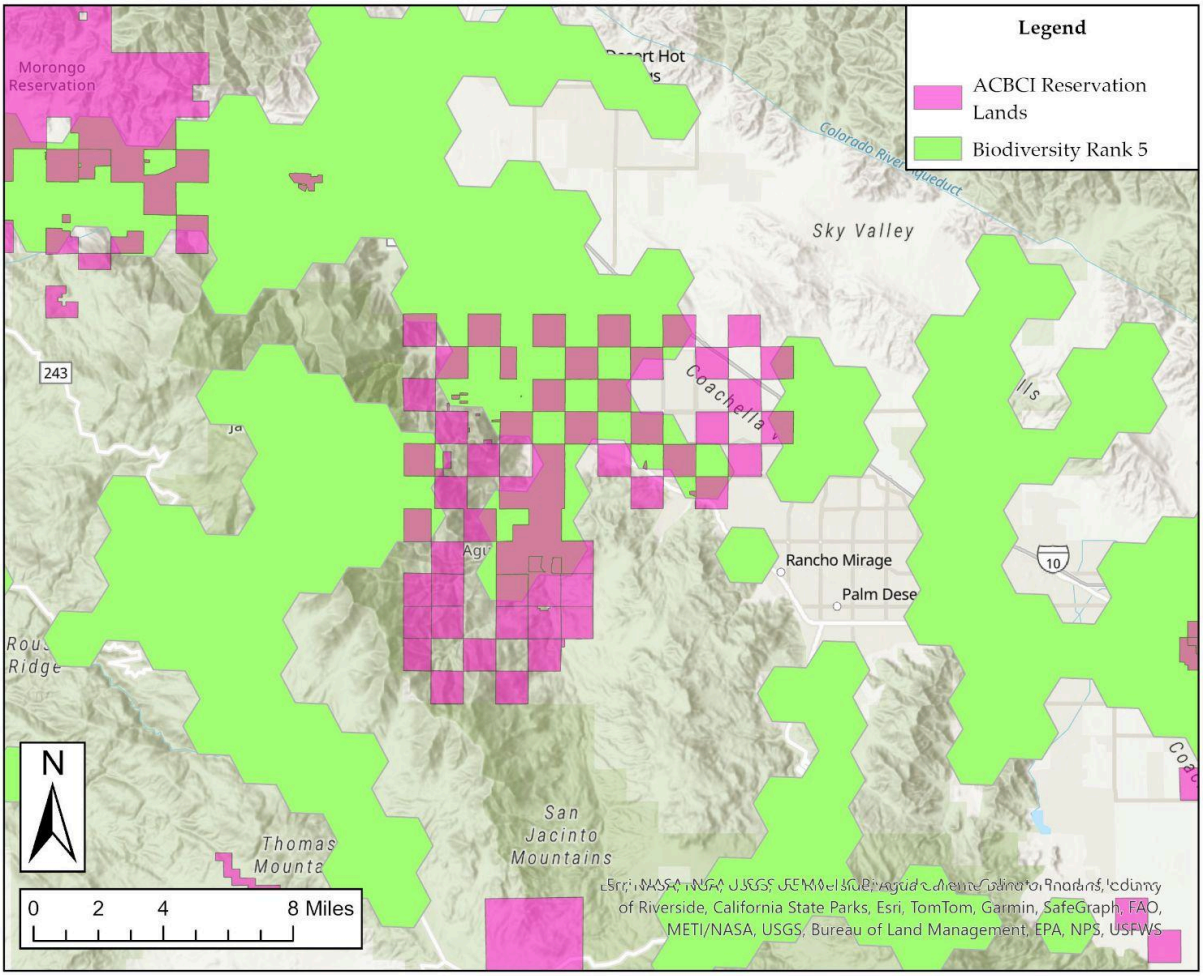
4. Conclusion

In considering the future of AI and data center policy, we are concerned about the exponential increases in data center construction, the implementation of generative AI into basic web searches, and heat waste production driven by current models of western expansion and extraction. By looking to alternative systems of sustainability rooted in Indigenous values of *aloha ʻāina*, or love for the land, we find examples of TEK that can be integrated into Solarpunk visions of a future that integrates technology with the environment, reusing electronic waste (e-waste) and improving data life cycles for a more sustainable future.

Acknowledgements: The authors of this paper are collaborating scholars as part of the Indigenous Futures Institute with Indigenous and settler backgrounds ranging from Kānaka Maoli (Native Hawaiian), Boricua (Puerto Rican), and Iranian identities as well as others. We acknowledge the differences and similarities that our unique perspectives bring to this topic.

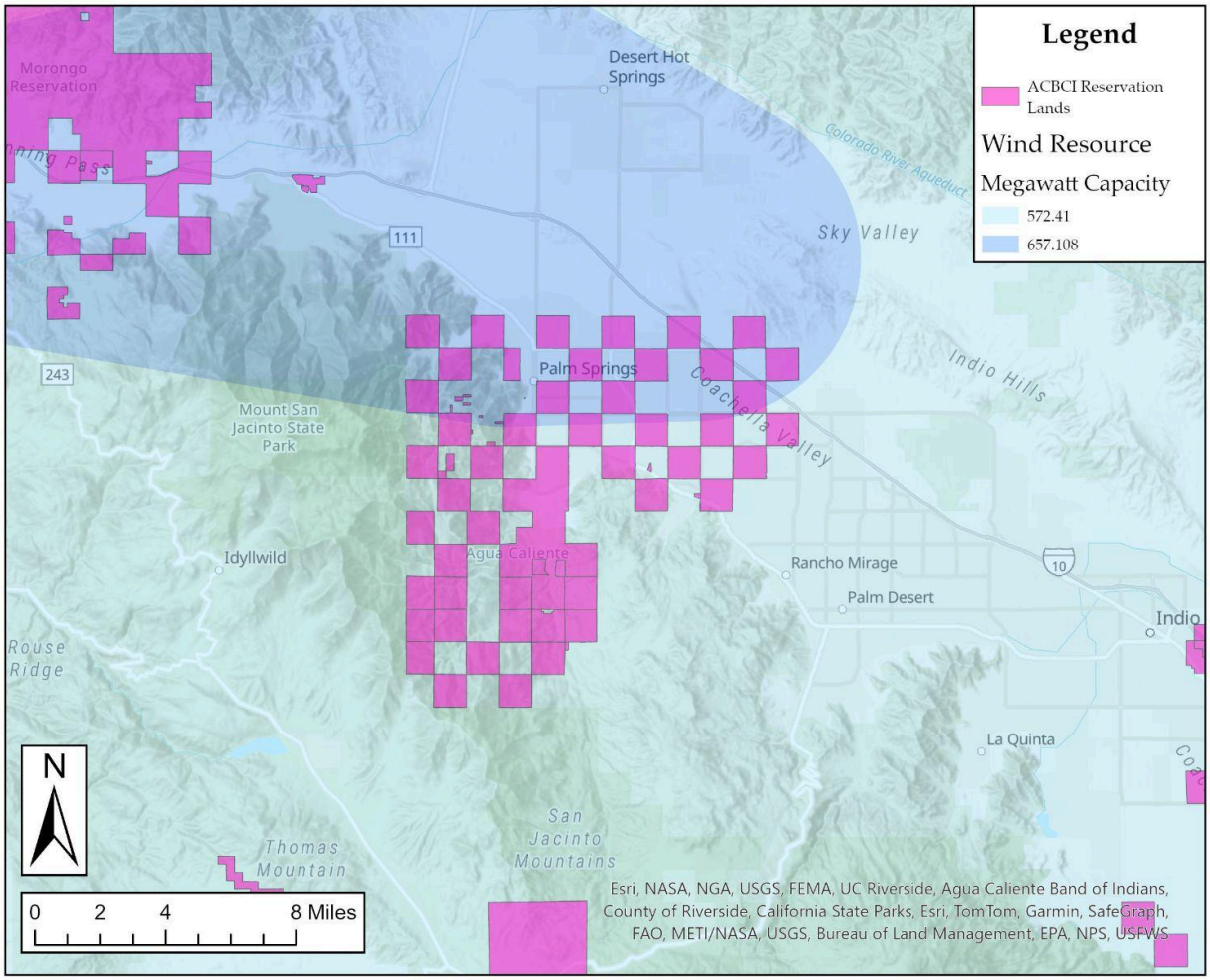
Appendix

Biodiversity Rank 5 around Agua Caliente Band of Cahuilla Indian (ACBCI) Reservation



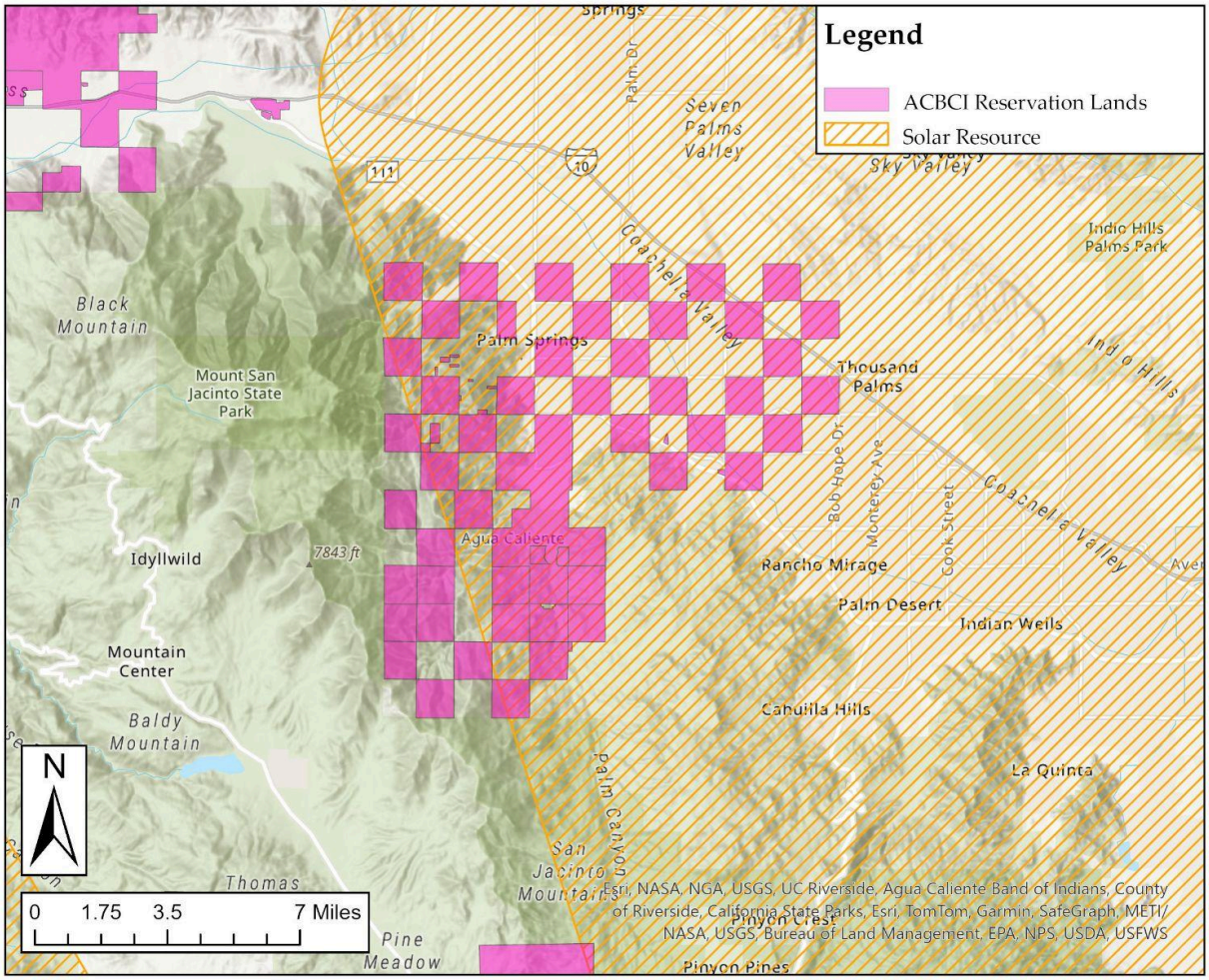
By: Javier J. García Colón
Coordinate System: Projected Coordinate System NAD 1927 State Plane California VI
Datum: D North American 1927

Wind Resource around Agua Caliente Band of Cahuilla Indian (ACBCI) Reservation



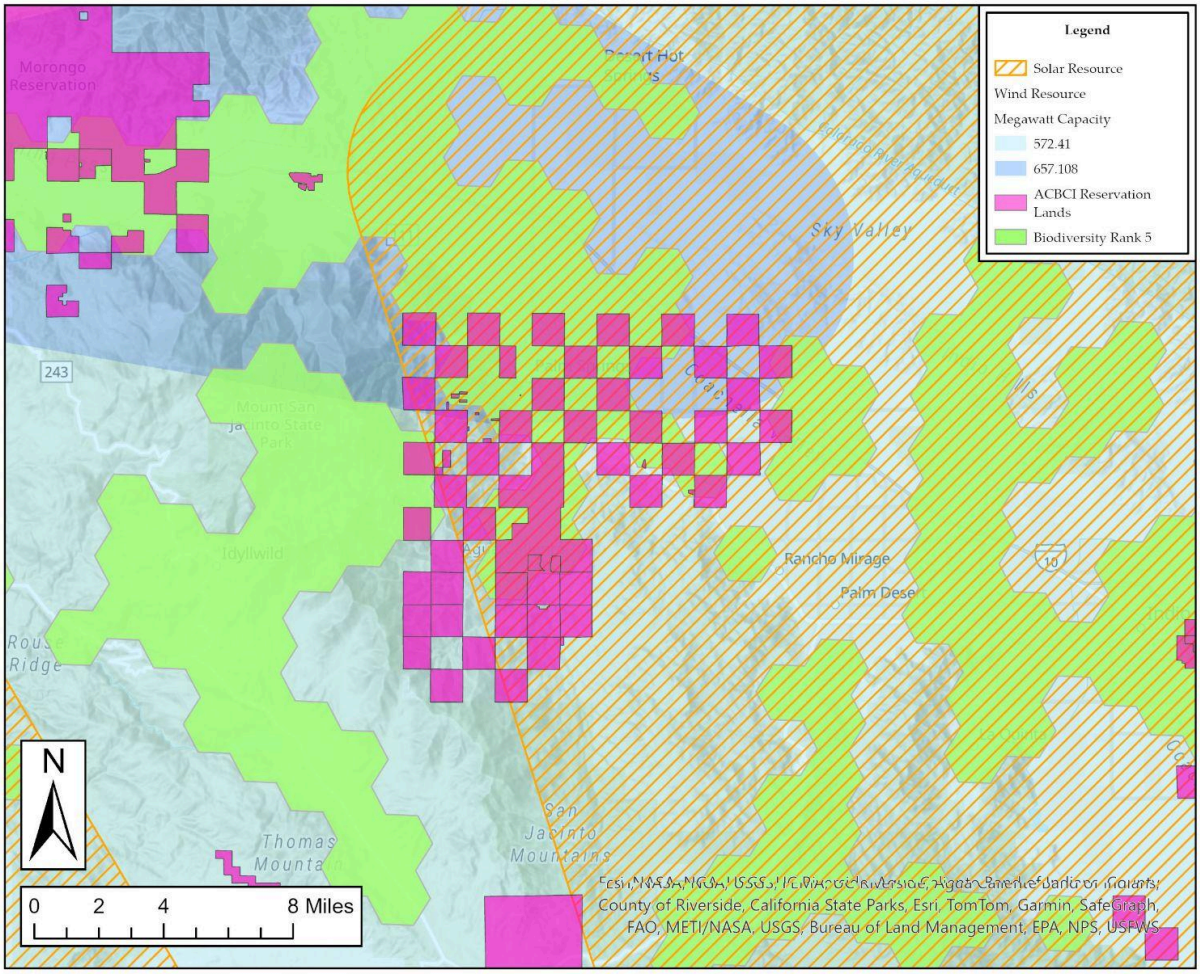
By: Javier J. García Colón
Coordinate System: Projected Coordinate System NAD 1927 State Plane California VI
Datum: D North American 1927

Solar Resource around Agua Caliente Band of Cahuilla Indian (ACBCI) Reservation



By: Javier J. García Colón
Coordinate System: Projected Coordinate System NAD 1927 State Plane California VI
Datum: D North American 1927

Intersection of resources around Agua Caliente Band of Cahuilla Indian (ACBCI) Reservation



By: Javier J. García Colón
Coordinate System: Projected Coordinate System NAD 1927 State Plane California VI
Datum: D North American 1927

References

1. Ahtone, T. & Lee, R. (2020, March 30). *Land-grab universities*. High Country News. http://www.hcn.org/issues/52-4/indigenous-affairs-education-land-grab-universities/
2. Beamer, K., Elkington, K., Souza, P., Tuma, A., Thorenz, A., Köhler, S., Kukea-Shultz, K., Kotubetey, K., & Winter, K. (2023). Island and Indigenous systems of circularity: How Hawai‘i can inform the development of universal circular economy policy goals. Ecology and Society, 28(1), art9. https://doi.org/10.5751/ES-13656-280109
3. Boscarino, N., Cartwright, R. A., Fox, K., & Tsosie, K. S. (2022). Federated learning and Indigenous genomic data sovereignty. Nature Machine Intelligence, 4(11), 909–911. https://doi.org/10.1038/s42256-022-00551-y

4. De Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004
5. Ding, W., Ebrahimi, B., Kim, B.-D., Devenport, C. L., & Childress, A. E. (2024). Analysis of Anthropogenic Waste Heat Emission from an Academic Data Center. *Energies*, *17*(8), 1835. https://doi.org/10.3390/en17081835
6. Elahee, K., & Jugoo, S. (2013). Ocean Thermal Energy for Air-conditioning: Case Study of a Green Data Center. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *35*(7), 679–684. https://doi.org/10.1080/15567036.2010.504941
7. *Engineering for Heat Waves*. (2023, October 4). American Scientist. https://www.americanscientist.org/article/engineering-for-heat-waves
8. Fernández‐Llamazares, Á., Garteizgogeascoa, M., Basu, N., Brondizio, E. S., Cabeza, M., Martínez‐Alier, J., McElwee, P., & Reyes‐García, V. (2020). A State‐of‐the‐Art Review of Indigenous Peoples and Environmental Pollution. Integrated Environmental Assessment and Management, 16(3), 324–341. https://doi.org/10.1002/ieam.4239
9. Fuchs, H., Shehabi, A., Ganeshalingam, M., Desroches, L.-B., Lim, B., Roth, K., & Tsao, A. (2020). Comparing datasets of volume servers to illuminate their energy use in data centers. *Energy Efficiency*, *13*(3), 379–392. https://doi.org/10.1007/s12053-019-09809-8
10. Grealey, J., Lannelongue, L., Saw, W.-Y., Marten, J., Méric, G., Ruiz-Carmona, S., & Inouye, M. (2022). The Carbon Footprint of Bioinformatics. Molecular Biology and Evolution, 39(3), msac034. https://doi.org/10.1093/molbev/msac034
11. He, W., Shirowzhan, S., & Pettit, C. J. (2022). GIS and Machine Learning for Analysing Influencing Factors of Bushfires Using 40-Year Spatio-Temporal Bushfire Data. *ISPRS International Journal of Geo-Information,* 11(6), Article 6. https://doi.org/10.3390/ijgi11060336
12. Johnson-Zafiris, M. (2024, April 11). Reclaiming Akwesasne's Pollution Data [powerpoint presentation]. U.S. Indigenous Data Sovereignty & Governance Summit 2024, Tucson, Arizona. usindigenousdatanetwork.org/2024/01/11/reclaiming-akwesasnes-pollution-data/
13. Jones, S. B., & Bellaire, R. (1937). The Classification of Hawaiian Climates: A Comparison of the Koppen and Thornthwaite Systems. *Geographical Review*, *27*(1), 112. https://doi.org/10.2307/209664
14. Kimmerer, R. W. (2013). *Braiding sweetgrass : Indigenous wisdom, scientific knowledge, and the teachings of plants* (First edition.). Milkweed Editions.
15. Lewis, J. E., Arista, N., Pechawis, A., & Kite, S. (2018). Making Kin with the Machines. Journal of Design and Science. https://doi.org/10.21428/bfafd97b
16. Mackey, T. K., Calac, A. J., Chenna Keshava, B. S., Yracheta, J., Tsosie, K. S., & Fox, K. (2022). Establishing a blockchain-enabled Indigenous data sovereignty framework for genomic data. *Cell*, *185*(15), 2626–2631. https://doi.org/10.1016/j.cell.2022.06.030
17. Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Koppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.*

18. Raith, P., Nastic, S., & Dustdar, S. (2023). Serverless Edge Computing—Where We Are and What Lies Ahead. *IEEE Internet Computing*, *27*(3), 50–64. https://doi.org/10.1109/MIC.2023.3260939
19. Reina-Rozo, J. D., Castro, A., Zambrano-Caviedes, F., & Epieyu, A. N. (2024). Technologies to embrace the sun: solarpunk-based project as an exploration for a just energy transition. Revista Iberoamericana de Estudios de Desarrollo, 13(1), 162–187. https://doi.org/10.26754/ojs_ried/ijds.871
20. Rylott, E. L., & Bruce, N. C. (2020). How synthetic biology can help bioremediation. Current Opinion in Chemical Biology, 58, 86–95. https://doi.org/10.1016/j.cbpa.2020.07.004
21. Smith, L. T. (2021). Decolonizing methodologies : research and Indigenous peoples (Third edition.). Zed Books. https://doi.org/10.5040/9781350225282
22. Sunder Rajan, K. (2006). Biocapital : the constitution of postgenomic life. Duke University Press.
23. TallBear, K. (2017). Beyond the Life/Not-Life Binary: A Feminist-Indigenous Reading of Cryopreservation, Interspecies Thinking, and the New Materialisms. In J. Radin & E. Kowal (Eds.), Cryopolitics (pp. 179–202). The MIT Press. https://doi.org/10.7551/mitpress/10456.003.0015
24. *The Latest Cloud Computing Statistics (updated June 2024) | AAG IT Support*. (n.d.). Retrieved August 1, 2024, from https://aag-it.com/the-latest-cloud-computing-statistics/
25. Tuck, E., & Yang, K. W. (n.d.). *Decolonization is not a metaphor*.
26. U.S. Council on Environmental Quality. (2021). *Update to the regulations implementing the procedural provisions of the National Environmental Policy Act*. Regulations.gov. https://www.regulations.gov/docket/CEQ-2021-0002/unified-agenda
27. Vitousek, P., & Beamer, K. (2013). Traditional Ecological Values, Knowledge, and Practices in Twenty-First Century Hawai‘i. In R. Rozzi, S. T. A. Pickett, C. Palmer, J. J. Armesto, & J. B. Callicott (Eds.), *Linking Ecology and Ethics for a Changing World* (pp. 63–70). Springer Netherlands. https://doi.org/10.1007/978-94-007-7470-4_5
28. Wagner, P., & Wieland, B. C. (2022). *Almanac for the Anthropocene : A Compendium of Solarpunk Futures*. (1st ed.). West Virginia University Press.
29. Wahlroos, M., Pärssinen, M., Rinne, S., Syri, S., & Manner, J. (2018). Future views on waste heat utilization – Case of data centers in Northern Europe. *Renewable and Sustainable Energy Reviews*, *82*, 1749–1764. https://doi.org/10.1016/j.rser.2017.10.058
30. Whyte, K. (2020). Too late for indigenous climate justice: Ecological and relational tipping points. WIREs Climate Change, 11(1), e603. https://doi.org/10.1002/wcc.603
31. Whyte, K. P. (2013). On the role of traditional ecological knowledge as a collaborative concept: A philosophical study. *Ecological Processes*, *2*(1), 7. https://doi.org/10.1186/2192-1709-2-7
32. Winter, K. B., Beamer, K., Vaughan, M. B., Friedlander, A. M., Kido, M. H., Whitehead, A. N., Akutagawa, M. K. H., Kurashima, N., Lucas, M. P., & Nyberg, B. (2018). The Moku

System: Managing Biocultural Resources for Abundance within Social-Ecological Regions in Hawai‘i. *Sustainability*, *10*(10), 3554. https://doi.org/10.3390/su10103554

AI in Point-of-Care - A Sustainable Healthcare Revolution at the Edge

Yousuf Rajput and Tarek Tarif

Department of Computer Science & Indigenous Futures Institute, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, United States
E-mail: yrajput@ucsd.edu
ttarif@ucsd.edu

Akira Wolfe

Department of Cognitive Science, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, United States
E-mail: akwolfe@ucsd.edu

Eric Dawson, Ph.D

Nvidia Corporation — Genomics / AI Division
E-mail: eric.t.dawson@gmail.com

Keolu Fox, Ph.D

Indigenous Futures Institute, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, United States
E-mail: pkfox@ucsd.edu

This paper examines the integration of artificial intelligence (AI) in point-of-care testing (POCT) to enhance diagnostic speed, accuracy, and accessibility, particularly in underserved regions. AI-driven POCT is shown to optimize clinical decision-making, reduce diagnostic times, and offer personalized healthcare solutions, with applications in genome sequencing and infectious disease management. The paper highlights the environmental challenges of AI, including high energy consumption and electronic waste, and proposes solutions such as energy-efficient algorithms and edge computing. It also addresses ethical concerns, emphasizing the reduction of algorithmic bias and the need for equitable access to AI technologies. While AI in POCT can improve healthcare and promote sustainability, collaboration within the POCT ecosystem—among researchers, healthcare providers, and policymakers—is essential to overcome the ethical, environmental, and technological challenges.

Keywords: Artificial Intelligence (AI), Point-of-Care Testing (POCT), Diagnostics, Sustainability, Energy Efficiency, Genome Sequencing, Electronic Waste (E-Waste), Edge Computing, Ethical AI, Personalized Healthcare

1. Introduction

POCT enables rapid diagnostics and treatment at the patient's location, revolutionizing healthcare delivery, particularly in remote or resource-limited settings. However, traditional POCT systems face challenges like slow diagnostic times and limited reach. The integration of AI offers a promising solution, enhancing POCT capabilities while addressing sustainability concerns.

AI's potential in healthcare is clear in its ability to improve diagnostic precision and efficiency. By leveraging AI advancements, POCT technologies can deliver rapid diagnostics and improve clinical decision-making, making them essential in modern healthcare, especially in underserved regions.

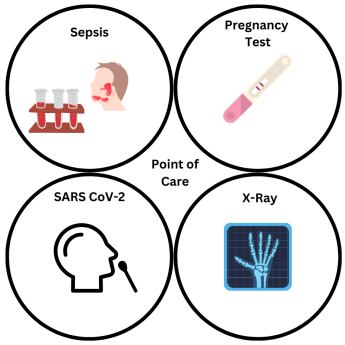
This paper explores AI's role in enhancing POCT through sustainable means, showing how AI can boost accuracy, accessibility, and eco-friendliness. Through real-world examples, the paper demonstrates AI-enhanced POCT applications in medical scenarios, such as infectious disease management.

The paper also addresses potential barriers to AI integration with POCT systems and proposes solutions to ensure seamless adoption. By incorporating sustainable practices, AI in POCT aims to reduce the ecological footprint of diagnostics, promoting eco-friendly healthcare solutions.

2. AI in Clinical Setting & POCT: Where do we begin?

While there is skepticism about the role of supervised learning in identifying pathogenic variants in clinical settings, recent advancements in rapid genome sequencing highlight its growing utility as a decision-support tool. In critical care settings, for example, ultrarapid nanopore genome sequencing was used to diagnose genetic conditions in as little as 7 hours, enabling immediate treatment decisions for critically ill patients.[1]

A similar application can be seen in Pediatric Intensive Care Units (PICUs), where rapid whole genome sequencing (rWGS) has significantly impacted patient care, providing molecular diagnoses that influenced clinical management in 76% of cases. In both settings, AI-driven supervised learning models could be employed to prioritize genetic variants for review by genetic counselors and clinicians, thereby streamlining the diagnostic process. While human expertise remains critical, supervised learning can enhance the efficiency of this process, especially in urgent cases where time is of the essence.



POCT has been a pivotal tool in healthcare, especially in rural and remote areas where access to hospitals and trained staff is limited. Early POCT technologies focused on simple diagnostics, such as lateral flow immunoassays (LFIAs), which are user-friendly and could deliver results within minutes. These systems were designed to be portable and easy to use

without specialized training. However, challenges with accuracy, sensitivity, and quality control persist, particularly in low-resource settings, which limits the reliability of traditional POCT systems.[2,3]

3. Building the POCT Ecosystem: A Holistic Approach to Enhanced Healthcare Delivery

The POCT ecosystem is an interconnected network that integrates technology, healthcare providers, manufacturers, regulatory bodies, and patients. Its aim is to create a multifaceted framework that enhances patient outcomes and reduces healthcare disparities worldwide.

By fostering collaboration among key stakeholders—such as technology developers, healthcare professionals, policymakers, and patient advocacy groups—the POCT ecosystem seeks to leverage AI innovations. This collaboration aims to streamline processes, facilitate personalized medicine, and empower patients with timely and precise health information.

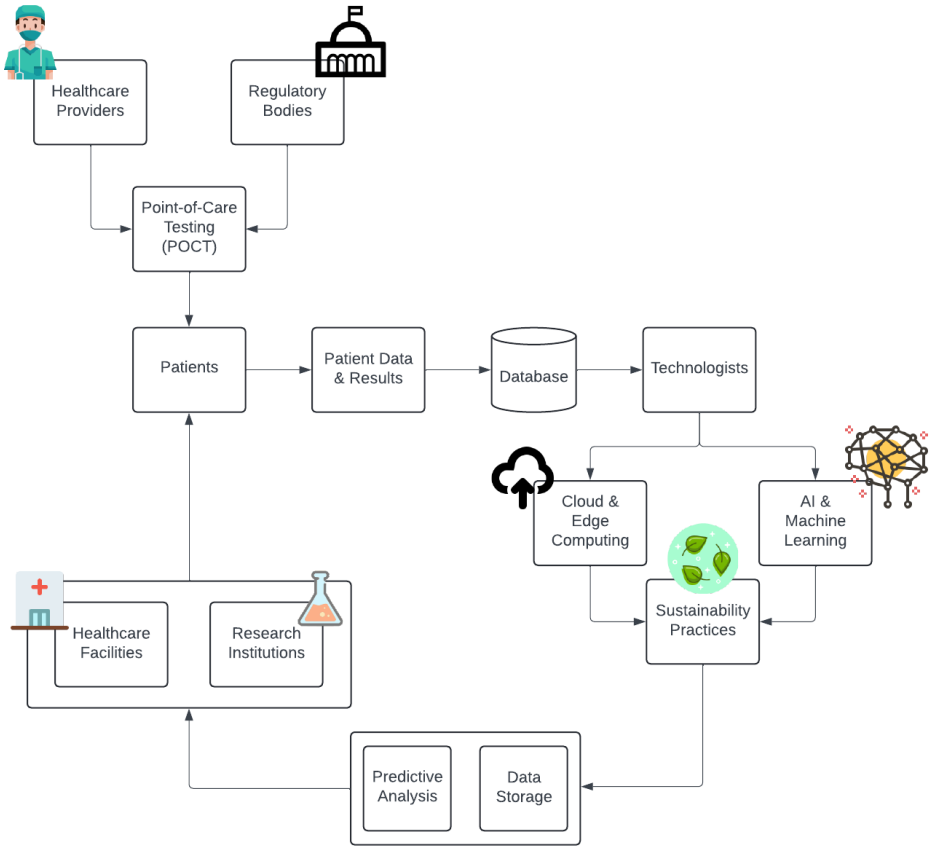


Fig. 1: This figure illustrates the cyclical approach that AI will introduce into the healthcare ecosystem. It shows the interaction among stakeholders, technological contributions, and patient outcomes, highlighting how AI creates a continuous improvement cycle that benefits communities as scientific advancements progress.

Collaboration between healthcare providers and regulatory bodies is essential for designing POCT systems that meet clinical and regulatory standards. The use of POCT by patients generates valuable data, enabling technologists to apply AI and advanced computing to enhance these tools continuously. Additionally, eco-friendly practices and energy-efficient AI contribute to reducing the environmental impact of increased technology use in healthcare. AI-enhanced POCT offers predictive diagnostic analyses, with stored data available for future research, creating a feedback loop that ultimately benefits patients.

3.1. *Attributes that Make POCT Effective in Healthcare*

POCT has evolved significantly, offering critical benefits that enhance healthcare delivery. Its primary goal is to improve healthcare provision, enabling healthcare providers to deliver better treatment to patients in both critical and non-critical settings. The growing demand for POCT is driven by its attributes, which are continuously refined to meet the evolving needs of the healthcare industry.

The effectiveness of POCT in healthcare is significantly influenced by several key factors, with accuracy, ease of use, and availability standing out. Each of these factors plays a crucial role in ensuring that POCT achieves its intended benefits of rapid diagnostics, improved patient outcomes, and enhanced healthcare efficiency.

Table 1: Attributes of POCT and Their Importance

Attributes	**Importance**	**Example**	**AI/Cloud Integration**
Accuracy	Prevents misdiagnosis and inappropriate treatment.	42% success rate in nanopore sequencing.	Classification algorithms improve detection.
Availability	Ensures accessibility in diverse settings.	Mobile POCT units for rural areas.	Edge computing enhances access.
Ease of Use	Usable by professionals and patients.	User-friendly devices for COVID-19 tests.	Cloud platforms simplify data management.

These attributes are not only fundamental to the success of POCT but also set the stage for the transformative impact of AI in this field. AI's capabilities in data analysis, ML, and automation can significantly enhance the accuracy of POCT by providing rapid and precise diagnostic insights.

By enhancing the speed and precision of diagnostics, AI has the potential to further amplify the effectiveness of POCT, leading to more personalized and data-driven care while reducing the dependency on centralized lab testing.

3.2. *AI's Transformative Power in POCT*

AI has also showcased its transformative potential, enabling rapid diagnostic feedback. For instance, an AI-based POCT device for glucose monitoring can achieve a mean absolute relative error (MARE) of approximately 9.5%, indicating high accuracy in managing blood glucose levels for diabetes patients.[4]

Similarly, AI-enhanced cardiac biomarker tests offer real-time analysis with high sensitivity and specificity, enabling healthcare providers to diagnose acute coronary syndrome with up to 94% sensitivity and 89% specificity.[4] The use of AI in POCT extends to infectious disease diagnostics as well. For example, AI-driven diagnostic tests for COVID-19 achieved diagnostic accuracies of up to 98% in detecting SARS-CoV-2, which was crucial for managing public health during the pandemic.[5] By providing real-time, actionable diagnostic information, AI-powered POCT devices significantly enhance clinical decision-making processes and improve patient outcomes.

3.3. *AI in POCT Use Case - Genome Sequencing*

Ultra-rapid nanopore sequencing has proven to have profound impacts on diagnosing genetic conditions and variations.[1] Distinguishing between benign and pathogenic genetic variants relies heavily on the genetic sequencing and variant classification scheme used[6] and can play a significant role in the patient's overall wellbeing. Supervised and unsupervised algorithms can, therefore, play an impactful role in this diagnostic process.

By leveraging cloud-based bioinformatics, researchers have achieved significant reductions in processing times, transforming genetic diagnostics from a lengthy process into one that can be completed in hours. Nanopore genome sequencing demonstrates how technological contributions enhance our ability to interpret complex genetic data.[1] With the further integration of AI, we will be able to pave the way for advancements in genomic medicine to further heights.

The link between nanopore sequencing and AI is particularly evident in improving base calling accuracy and managing large genomic datasets generated by sequencing technologies. Nanopore sequencing, while fast, can produce noisy data, but AI algorithms have been shown to improve base calling by learning from large datasets and correcting errors in raw signals.[1]

In Pediatric Intensive Care Unit (PICU) settings, where rapid diagnoses can make the difference between life and death, AI-driven analysis of nanopore-generated data has proven invaluable. The combination of AI with cloud-based bioinformatics not only reduced processing times by 93% but also helped to more effectively mine and interpret vast genomic datasets.[1,7] This demonstrates the critical role of AI in enhancing the speed and accuracy of genomic diagnostics in PICUs, where timely and accurate information is essential.

Additionally, one of the main challenges in genome sequencing lies in the ability to quickly and accurately classify genetic variants. Machine learning (ML) algorithms, mainly supervised and unsupervised learning algorithms, have shown to be effective in aiding diagnostic results in medicine and healthcare.[8] Supervised learning algorithms in AI have the potential to play a pivotal role in classifying genetic variants due to their ability to categorize known inputs into discrete categories.[8] These algorithms can be used to analyze sequencing data and identify pathogenic variants that could explain the patients' critical conditions.

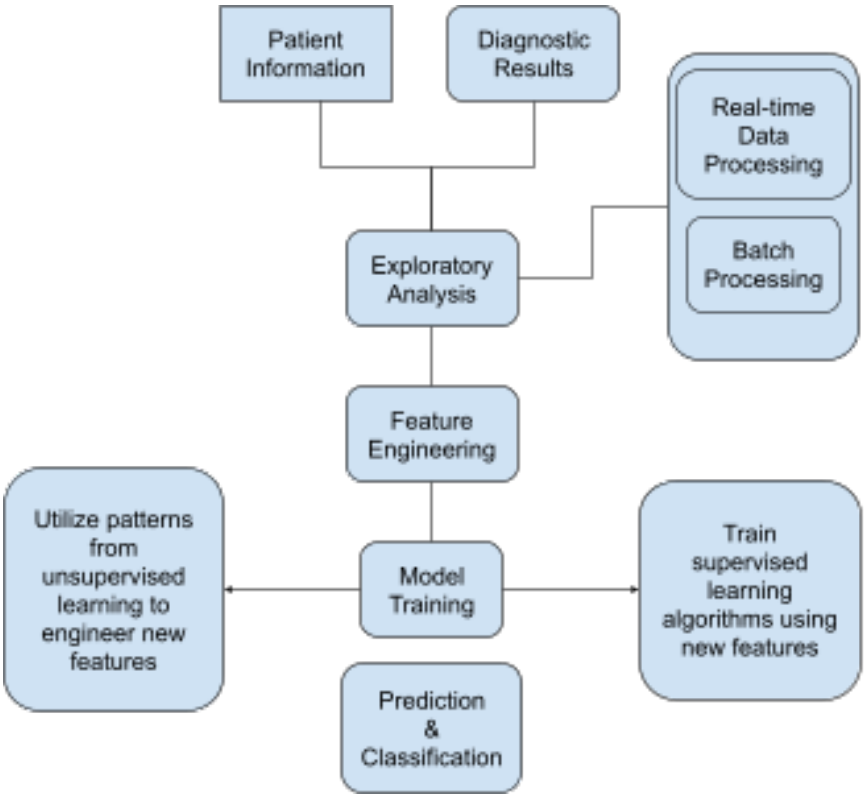


Fig. 2: This diagram illustrates the general pipeline of an AI-driven system for POCT, highlighting the integration of various machine learning methodologies from data collection to diagnostic output.

3.3.1. *Pipeline Overview*

The patient data analysis pipeline begins with collecting clinical metrics and medical history. This data undergoes exploratory data analysis (EDA) using unsupervised learning techniques to identify patterns and prepare the dataset. Key features are engineered to enhance model performance before training various machine learning algorithms.

Post-training, the models are evaluated for accuracy and effectiveness, allowing for the classification of new patient data and delivery of diagnostic results. These results can be processed in real-time for urgent care or through batch processing for less critical analysis. Ultimately, the AI-driven system provides healthcare providers with rapid and precise diagnostic outcomes, facilitating informed medical decision-making.

Table 2: Comparison of ML Methods and Edge Computing in Genomic Diagnostics and POCT

Technique	**Type**	**Application**	**Benefits**
Random Forests (RF)	Supervised Learning	Classifies genetic variants by analyzing large genomic datasets to identify pathogenic patterns.	Reduces manual review time; effectively handles complex interactions and non-linear relationships.
Supervised Learning Models	Integrated with POCT	Provides immediate diagnostic insights for informed decision-making in critical care settings.[8]	Quickly distinguishes between disease-causing and benign variants, guiding the diagnostic process.[6]
K-means Clustering	Unsupervised Learning	Identifies clusters of genetic variants based on features such as nucleotide changes or genomic position.	Uncovers new patterns and subtypes, refining diagnostic frameworks, and processing large datasets.[8]
Edge Computing	Data Processing	Enhances genomic diagnostics at the point of care, especially in rural areas with limited infrastructure.	Ensures timely access to advanced diagnostics, bypassing extensive cloud infrastructure; supports collaborations like NVIDIA and SoftBank's AI and 5G performance.[9]

4. Environmental Challenges of AI in POCT

4.1. *Energy Consumption*

As POCT tools continue to evolve with AI integration, they introduce significant environmental sustainability challenges. These challenges primarily stem from the energy-intensive processes required to train and operate AI models, as well as the associated data storage demands. These factors, coupled with issues like heat emissions, e-waste generation, and the ethical concerns of using personal health data, necessitate a careful consideration of the environmental impact of AI-enhanced POCT systems.

Training AI models using significant amounts of data to accurately diagnose medical conditions demands vast amounts of computational power, resulting in substantial energy consumption.[10] For instance, tech companies such as Amazon and NVIDIA have noted that inference processing after training a model makes up to 80–90% of the energy cost of neural networks.[4] Inference consumes the greatest amount of energy but is also responsible for accuracy. There is a correlation of higher accuracy and higher energy consumption, further increasing the challenge of increased climate and environmental impact with further technological advancements.

Furthermore, as these models scale to meet the demands of real-time data analysis in POCT, the environmental impact increases. The reliance on large datasets for training AI models necessitates significant energy use, contributing to the carbon footprint of these systems.[11]

In order to minimize energy consumption, while also maintaining essential health data, we can look into byte pair encoding. This encoding practice adds shortcuts in text or records that

compress the data while retaining the same information. This can indirectly lead to decreased energy usage by simplifying tokens, a practice that has already been utilized in the medical field for large language models (LLM) and record management.

4.2. *Electronic Waste (E-Waste)*

E-waste is another critical concern associated with AI-integrated POCT tools. The hardware used in AI applications, such as GPUs, TPUs, FPGAs, and CPUs, becomes obsolete as newer, more efficient models are developed.[12] This results in a continuous cycle of hardware disposal, contributing to the growing issue of electronic waste.

Medical providers can look into EPEAT (Electronic Product Environmental Assessment Tool) to help them choose greener AI devices. EPEAT is a metric of sustainability for electronics. For them to be rated highly, they must fall under 75% of their criteria. The goal is to promote green products and sustainability for electronic life cycles. This metric has already been used to showcase greener computers, displays, imaging equipment, mobile phones, photovoltaic modules and inverters, servers, and televisions.[13]

4.3. *Data Centers and Cooling Systems*

In addition to energy consumption, heat emissions from data centers pose a challenge, as cooling systems account for approximately 30% of their power consumption.[14,15]

There have been initiatives that work with nature to provide natural cooling. Data centers, for example, have been built in the Arctic, which require less energy that would otherwise be allocated towards cooling.[16]

A downside is that there will need to be new infrastructure built in harsher climates and higher latency due to their location. A different approach by Microsoft was started in 2013 under the name Project Natick.[17] They submerged a data center off the coast of Scotland and utilized the seawater's more consistent temperature as a form of cooling.

Reliable information on the energy consumption and emissions of data centers is often fragmented and difficult to authenticate. This lack of transparency has led to accusations of "greenwashing," where companies might exaggerate their environmental efforts while avoiding genuine sustainability improvements.

4.4. *Sustainable AI and Community Engagement*

Addressing the environmental challenges posed by AI-powered POCT tools, the focus must shift towards developing more sustainable AI models and technologies that prioritize efficiency without sacrificing accuracy. Emerging AI algorithms with reduced computational demands, such as model compression, quantization, and efficient hardware utilization, offer the potential to significantly cut energy consumption and reduce the environmental footprint of these systems.

However, technological advancements alone are insufficient. Engaging the broader community is essential to ensuring that these innovations translate into meaningful change. Medical providers, technology developers, and even patients need to be educated on the sustainable use of AI-driven tools. Community knowledge-sharing initiatives could play a pivotal role

in promoting awareness of energy consumption, the life cycle of electronic devices, and the importance of greener alternatives like EPEAT-rated products.

An impactful approach could be creating an online platform where healthcare professionals and technology users can share best practices, resources, and case studies focused on sustainable AI use. This platform could feature interactive content such as webinars, forums, and sustainability toolkits, enabling users to learn from real-world examples and collaborate on solutions for reducing energy consumption and electronic waste in their AI-driven practices.

5. Scalability of Sustainability

Although there are a number of sustainable practices that can enhance the use of POCT devices in the context of environmental impact, the scalability of AI in POCT faces several explicit barriers, but there are also key facilitators that can drive its widespread adoption. One major barrier is the need for robust quality assurance protocols and the availability of trained staff to ensure reliable interpretation of test results, particularly in remote or resource-limited settings.[2,3]

Additionally, regulatory challenges and the slow pace of adopting AI-driven tools in clinical environments create significant hurdles. On the other hand, facilitators include the increasing reliability of AI in improving diagnostic accuracy, particularly in technologies such as lateral flow immunoassays and hematology analyzers, where AI-driven tools have shown enhanced sensitivity and specificity.[3]

Finally, advances in mobile and edge computing can enable faster data processing at the point of care, improving access to high-quality diagnostics even in underserved regions.[3] By addressing these barriers and leveraging the facilitators, AI-driven POCT can become a more integral part of global healthcare systems.

6. Edge Computing vs. Centralized Data Centers

Edge computing has emerged as a crucial technology for enhancing the effectiveness and accessibility of POCT in rural and remote areas. Unlike traditional cloud computing, which relies on centralized data centers for processing, edge computing facilitates data processing at or near the data source, such as a mobile POCT device or a local healthcare facility.

6.1. *Energy Consumption and Efficiency*

Energy Savings: Edge computing can significantly reduce energy consumption by processing data closer to where it is generated, thereby reducing the need to transmit large volumes of data to centralized locations. According to a study by Gartner, edge computing can reduce energy consumption by 40-60% compared to traditional data centers.[18] For example, edge computing for video analytics could reduce the amount of data transmitted by 95%, resulting in significant energy savings.[19,20]

6.2. *Data Transmission and Latency*

Lower Latency: Edge computing can reduce latency by up to 90%, as data does not need to travel as far to be processed. This is crucial for real-time health diagnostics, where ev-

ery millisecond counts.[21] Additionally, mobile POCT units equipped with edge computing capabilities could perform real-time analyses of patient samples and deliver results within 10 minutes, compared to the 24 to 48 hours required for traditional lab-based testing methods.[22]

6.3. *Environmental Impact*

Reduced Resource Use: Smaller, localized edge devices require fewer materials to build and maintain, which can reduce the environmental impact associated with construction and maintenance. In a rural healthcare setting, edge computing devices can manage patient information and perform data analysis directly at the point of care. For example, edge computing solutions can handle up to 90% of data processing tasks locally, significantly reducing latency and improving the speed of data access.[23]

While edge computing reduces energy use by minimizing data transfers, it's not inherently more energy-efficient than centralized data centers, which optimize power consumption at scale. In healthcare, however, the efficiency of edge computing is enhanced when paired with smaller, less energy-intensive AI models designed for localized tasks. Smaller AI models, designed to perform well with fewer computational resources, are suited for edge computing environments and better suited for specialized tasks.[24] These models not only reduce the energy consumption, heat emissions, and cooling requirements that are typically associated with traditional centralized data centers, but they allow real-time processing in decentralized POCT systems.

Table 3: Comparison of Edge Computing and Centralized Data Centers

Parameter	**Centralized Data Centers**	**Edge Computing**
Energy Consumption	High	40-60% reduction
Carbon Footprint	2% of global emissions	Significantly lower
Data Transmission Energy Use	Up to 5% of total energy use	Minimal
Latency	High	Up to 90% reduction
Cooling Requirements	Up to 40% of energy use	Lower
Resource Use	High	Lower

7. Application in Healthcare

A practical example can be drawn from the deployment of edge computing in healthcare. In a pilot program, rural clinics in India used edge computing devices to perform real-time diagnostics for diseases like tuberculosis and malaria. These edge devices reduced the need for data transmission to central labs. This led to reduced energy consumption of approximately 50% and allowed healthcare providers in remote areas to conduct advanced diagnostics locally, significantly reducing the time and resources needed for traditional lab analysis.[25]

This reduction of data transmission time is not only effective in reducing energy usage but also leads to reduced latency, enabling healthcare providers to make faster informed decisions. By reducing the dependency on central data centers, edge computing not only conserves

energy but also empowers local healthcare facilities to deliver timely and accurate diagnostics, improving patient outcomes.

8. Limitations of Specialization

The integration of modern technological advancements in POCT underscores the importance of thorough diagnosis and understanding the complexities of patient narratives. While technology can provide invaluable data and diagnostic support, it should complement rather than replace the critical human element in healthcare.[26]

By streamlining routine diagnostic processes, these technologies empower clinicians to be better at their jobs, ensuring that the art of medicine—listening to and understanding the patient's story—remains at the forefront of healthcare. This synergy between advanced diagnostics and personalized care enhances overall patient outcomes and reinforces the essential role of healthcare providers in the diagnostic process.

For example, in a pilot program, AI-assisted diagnostic tools in emergency departments reduced the average patient wait time by 20%, allowing doctors more time to engage with patients and discuss treatment options.[27]

Incorporating the principles of the 4Ps of medicine—Personalized, Preventive, Predictive, and Participatory—further emphasizes the importance of this balance: personalized medicine tailors treatment to the individual, preventive approaches aim to ward off diseases before they occur, predictive analytics forecast potential health risks, and participatory care involves patients in their own health decisions.

Another critical limitation of AI in POCT is the introduction of bias, which can arise when AI algorithms are trained on non-diverse data sets.[28] If the training data lacks representation from various demographic groups, the AI system may perform well for some populations but poorly for others, exacerbating existing health disparities.

This bias not only limits the effectiveness of AI in diverse clinical settings but also highlights the need for inclusive data collection and algorithm development. Addressing these biases is crucial for ensuring that AI technologies in POCT contribute to equitable healthcare improvements rather than reinforcing existing inequalities.

9. Discussion: Challenges, Future Research, & Call to Action

AI in POCT enhances diagnostic accuracy and healthcare accessibility but poses environmental and ethical challenges. High energy consumption from AI model training and centralized data centers, along with AI hardware obsolescence, contribute to carbon emissions and e-waste. Addressing these requires energy-efficient algorithms and sustainable hardware innovations, such as biodegradable components.

The substantial carbon emissions and e-waste resulting from outdated AI hardware highlight the need for energy-efficient algorithms and sustainable hardware innovations, such as biodegradable components. Ethically, AI must complement human decision-making in healthcare. Future research should focus on improving AI's energy efficiency, addressing biases, and ensuring equitable access to POCT technologies. Achieving sustainability requires concerted efforts from researchers, healthcare providers, and policymakers to responsibly integrate AI

without compromising environmental and social values.

Additionally, the reliance on edge computing in POCT introduces its own set of challenges. While edge computing can reduce latency and improve efficiency, it may also raise concerns related to data security, privacy, and the potential for increased hardware obsolescence. The environmental impact of the widespread deployment of edge devices, particularly in resource-limited settings, must also be considered.

The integration of AI in POCT enhances diagnostic accuracy and healthcare accessibility but also presents significant environmental, ethical, and technological challenges. While AI offers the potential to improve healthcare outcomes, it also carries the risk of biases in algorithmic decision-making. These biases can lead to unequal access to diagnostics and misinterpretation of data, particularly for underrepresented populations. Ensuring that AI systems are developed and validated using diverse datasets is essential to mitigate these biases.

By optimizing AI for sustainability and equity, we can transform healthcare with more accessible diagnostics and personalized care, aligning with global sustainability goals.

Achieving sustainability in healthcare requires a collaborative effort among researchers, healthcare providers, and policymakers. By responsibly integrating AI while upholding environmental and social values, we can optimize these technologies for greater sustainability and equity. By focusing on these areas, we can transform healthcare delivery, making diagnostics more accessible and personalized while aligning with global sustainability goals.

References

1. John E. Gorzynski et al. Ultrarapid nanopore genome sequencing in a critical care setting. *New England Journal of Medicine*, 386(7), January 2022.
2. Raj M. Ratwani et al. Patient safety and artificial intelligence in clinical care. *JAMA Health Forum*, 5(2), February 2024.
3. Adil I. Khan et al. Artificial intelligence in point-of-care testing. *Annals of Laboratory Medicine*, 43(5):401–407, April 2023.
4. Radosvet Desislavov et al. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857, April 2023.
5. David A. Paltiel et al. Sars-cov-2 screening strategies to permit the safe reopening of college campuses in the united states. *JAMA Network Open*, July 2020.
6. Blueprint Genetics. A guide to understanding variant classification. Accessed: 30 July 2024.
7. Erica Sanford Kobayashi et al. Cost efficacy of rapid whole genome sequencing in the pediatric intensive care unit. *Frontiers in Pediatrics*, 9, January 2022.
8. Q. An, S. Rahman, J. Zhou, and J. J. Kang. A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges. *Sensors*, 23(9), April 2023.
9. NVIDIA Newsroom. Nvidia collaborates with softbank corp. to power softbank's next-gen data centers using grace hopper superchip for generative ai and 5g/6g, June 2024.
10. Alex De Vries. The growing energy footprint of artificial intelligence. *Joule*, 7(10):2191–2194, October 2023.
11. Sasha Luccioni et al. Power hungry processing: Watts driving the cost of ai deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, June 2024.
12. Roberto Verdecchia et al. A systematic review of green ai. *WIREs Data Mining and Knowledge Discovery*, 13(4), June 2023.
13. EPEAT Registry. About epeat: Epeat registry. Accessed: 1 Aug. 2024.
14. Richard Evans. Deepmind ai reduces google data centre cooling bill by 40
15. Thomas Liquori. Cooling costs - data center energy efficiency, March 2024.
16. Juha SAUNAVAARA and Antti LAINE. Research, development, and education: Laying foundations for arctic and northern data centers. *Arctic and North*, (42):145–169, March 2021.
17. John Roach. Microsoft finds underwater datacenters are reliable, practical and use energy sustainably, July 2023.
18. Katie Costello. Gartner predicts the future of cloud and edge infrastructure, February 2021.
19. Adam Zewe. Technique enables ai on edge devices to keep learning over time, November 2023.
20. Kylie Foy. New tools are available to help reduce the energy that ai models devour, October 2023.
21. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, October 2016.
22. Seyedeh Shabnam Jazaeri et al. Composition of caching and classification in edge computing based on quality optimization for sdn-based iot healthcare solutions. *The Journal of Supercomputing*, May 2023.
23. Rohit Nishant et al. Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *International Journal of Information Management*, 53:102104, August 2020.
24. Melissa Heikkila. Why bigger is not always better in ai, October 2024.
25. World Health Organization. Digital health technology. Accessed: 30 July 2024.
26. Lisa Sanders. *Every Patient Tells a Story: Medical Mysteries and the Art of Diagnosis*. Broadway Books, 2009.

27. Kaio Jia Bin et al. The impact of artificial intelligence on waiting time for medical care in an urgent care service for covid-19: Single-center prospective study. *JMIR Formative Research*, February 2022.
28. Sachil Singh. Racial biases in healthcare: Examining the contributions of point of care tools and unintended practitioner bias to patient treatment and diagnosis. *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine*, 27(5):829–846, December 2021.
29. Silvana Secinaro et al. The role of artificial intelligence in healthcare: A structured literature review. *BMC Medical Informatics and Decision Making*, 21(1), April 2021.
30. Taylor Orwig et al. Sampling of healthcare professionals' perspective on point-of-care technologies from 2019–2021: A survey of benefits, concerns, and development. *PLOS ONE*, 19(3), March 2024.
31. Sam Rowe et al. Evaluation of point-of-care tests for identification of pathogens to inform clinical mastitis treatment decisions in pasture- and confinement-managed dairy cows in australia. *Journal of Dairy Science*, May 2024.
32. Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2):94–98, June 2019.
33. Arman Shehabi et al. Data center growth in the united states: Decoupling the demand for services from electricity use. *Environmental Research Letters*, 13(12):124030, December 2018.
34. Akeeb Adepoju Fawole et al. Climate change implications of electronic waste: Strategies for sustainable management. *Bulletin of the National Research Centre*, 47(1), October 2023.
35. Qingxia Zhang et al. A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization. *Journal of Systems Architecture*, 119:102253, October 2021.
36. Dlzar Al Kez et al. Exploring the sustainability challenges facing digitalization and internet data centers. *Journal of Cleaner Production*, 371:133633, October 2022.
37. Xi Yang et al. A large language model for electronic health records. *NpJ Digital Medicine*, 5(1), December 2022.
38. Kirsten Bibbins-Domingo et al. The us preventive services task force 2017 draft recommendation statement on screening for prostate cancer. *JAMA*, 317(19):1949, May 2017.
39. A. Esteva, A. Robicquet, B. Ramsundar, et al. A guide to deep learning in healthcare. *Nature Medicine*, 25:24–29, 2019.
40. Natural Resources Defense Council. America's data centers are wasting huge amounts of ..., August 2014.
41. IEA. Data centres & networks. Accessed: 30 July 2024.
42. Kuba Smolorz. How to make edge computing more sustainable, 2024. Accessed: 30 July 2024.
43. Eric J. Topol. High-performance medicine: The convergence of human and artificial intelligence. *Nature News*, January 2019.
44. Bin Cheng et al. Fogflow: Easy programming of iot services over cloud and edges for smart cities, August 2017.
45. Zi-Hang Chen et al. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Communications (London, England)*, November 2021.
46. Yancong Zhou et al. Transfer learning and its application research. *Journal of Physics: Conference Series*, 1920(1):012058, May 2021.

ERRATUM

Polygenic risk scores for cardiometabolic traits demonstrate importance of ancestry for predictive precision medicine

Rachel L. Kember*
Department of Psychiatry, University of Pennsylvania, 3535 Market Street
Philadelphia, PA 19104, USA
Email: rkember@pennmedicine.upenn.edu

Shefali S. Verma*
Department of Pathology and Laboratory Medicine, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: shefali.setiaverma@pennmedicine.upenn.edu

Anurag Verma*
Department of Medicine, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: anurag.verma@pennmedicine.upenn.edu

Brenda Xiao
Graduate Program in Genomics and Computational Biology, University of Pennsylvania,
Philadelphia, PA 19104, USA
Email: brendax@pennmedicine.upenn.edu

Anastasia Lucas
Graduate Program in Genomics and Computational Biology, University of Pennsylvania,
Philadelphia, PA 19104, USA
Email: anastasia.lucas@pennmedicine.upenn.edu

Colleen M. Kripke
Institute for Translational Medicine and Therapeutics, University of Pennsylvania,
Philadelphia, PA 19104, USA.
Email: colleen.morse@pennmedicine.upenn.edu

Renae Judy
Department of Surgery, Division of Vascular Surgery and Endovascular Therapy, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: rjudy@pennmedicine.upenn.edu

Jinbo Chen
Department of Biostatistics and Epidemiology, University of Pennsylvania,
203 Blockley Hall, Philadelphia, PA 19104, USA
Email: jinboche@pennmedicine.upenn.edu

Scott M. Damrauer
Department of Surgery, Division of Vascular Surgery and Endovascular Therapy, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: scott.damrauer@pennmedicine.upenn.edu

Daniel J. Rader
Department of Medicine and Genetics, Institute for Translational Medicine and Therapeutics, University of Pennsylvania, 3801 Filbert St
Philadelphia, PA 19104, USA
Email: rader@pennmedicine.upenn.edu

Marylyn D. Ritchie
Department of Genetics, Institute for Biomedical Informatics, University of Pennsylvania, Perelman School of Medicine, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: marylyn@pennmedicine.upenn.edu

In the above PSB article published in Biocomputing 2024: Proceedings of the Pacific Symposium, pp. 611-626; PMCID: PMC10947742; PMID: 38160310
The following correction has been made.

Polygenic risk scores for cardiometabolic traits demonstrate importance of ancestry for predictive precision medicine

Rachel L. Kember*
Department of Psychiatry, University of Pennsylvania, 3535 Market Street
Philadelphia, PA 19104, USA
Email: rkember@pennmedicine.upenn.edu

Shefali S. Verma*
Department of Pathology and Laboratory Medicine, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: shefali.setiaverma@pennmedicine.upenn.edu

Anurag Verma*
Department of Medicine, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: anurag.verma@pennmedicine.upenn.edu

Brenda Xiao
Graduate Program in Genomics and Computational Biology, University of Pennsylvania,
Philadelphia, PA 19104, USA
Email: brendax@pennmedicine.upenn.edu

Anastasia Lucas
Graduate Program in Genomics and Computational Biology, University of Pennsylvania,
Philadelphia, PA 19104, USA
Email: anastasia.lucas@pennmedicine.upenn.edu

Colleen M. Kripke
Institute for Translational Medicine and Therapeutics, University of Pennsylvania,
Philadelphia, PA 19104, USA.
Email: colleen.morse@pennmedicine.upenn.edu

Renae Judy
Department of Surgery, Division of Vascular Surgery and Endovascular Therapy, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: rjudy@pennmedicine.upenn.edu

Jinbo Chen
Department of Biostatistics and Epidemiology, University of Pennsylvania,
203 Blockley Hall, Philadelphia, PA 19104, USA
Email: jinboche@pennmedicine.upenn.edu

Scott M. Damrauer
Department of Surgery, Division of Vascular Surgery and Endovascular Therapy, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: scott.damrauer@pennmedicine.upenn.edu

Daniel J. Rader
Department of Medicine and Genetics, Institute for Translational Medicine and Therapeutics, University of Pennsylvania, 3801 Filbert St
Philadelphia, PA 19104, USA
Email: rader@pennmedicine.upenn.edu

Marylyn D. Ritchie
Department of Genetics, Institute for Biomedical Informatics, University of Pennsylvania, Perelman School of Medicine, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: marylyn@pennmedicine.upenn.edu

Polygenic risk scores (PRS) have predominantly been derived from genome-wide association studies (GWAS) conducted in European ancestry (EUR) individuals. In this study, we present an in-depth evaluation of PRS based on multi-ancestry GWAS for five cardiometabolic phenotypes in the Penn Medicine BioBank (PMBB) followed by a phenome-wide association study (PheWAS). We examine the PRS performance across all individuals and separately in African ancestry (AFR) and EUR ancestry groups. For AFR individuals, PRS derived using the multi-ancestry LD panel showed a higher effect size for four out of five PRSs (DBP, SBP, T2D, and BMI) than those derived from the AFR LD panel. In contrast, for EUR individuals, the multi-ancestry LD panel PRS demonstrated a higher effect size for two out of five PRSs (SBP and T2D) compared to the EUR LD panel. These findings underscore the potential benefits of utilizing a multi-ancestry LD panel for PRS derivation in diverse genetic backgrounds and demonstrate overall robustness in all individuals. Our results also revealed significant associations between PRS and various phenotypic categories. For instance, CAD PRS was linked with 18 phenotypes in AFR and 82 in EUR, while T2D PRS correlated with 84 phenotypes in AFR and 78 in EUR. Notably, associations like hyperlipidemia, renal failure, atrial fibrillation, coronary atherosclerosis, obesity, and hypertension were observed across different PRSs in both AFR and EUR groups, with varying effect sizes and significance levels. However, in AFR individuals, the strength and number of PRS associations with other phenotypes were generally reduced compared to EUR individuals. Our study underscores the need for future research to prioritize 1) conducting GWAS in diverse ancestry groups and 2) creating a cosmopolitan PRS methodology that is universally applicable across all genetic backgrounds. Such advances will foster a more equitable and personalized approach to precision medicine.

Keywords: Polygenic risk scores, multi-ancestry GWAS, cardiometabolic phenotypes, precision medicine

1. Introduction

The era of precision medicine has been marked by significant efforts to identify the genetic and environmental factors that influence the risk of disease as well as the disease prognosis and treatment. Advance knowledge of these factors can provide a major health benefit to individuals, as preventative strategies and tailored therapies can be targeted toward individuals at higher risk. Results from genome-wide association studies (GWAS) have highlighted the polygenic nature of

most common, complex diseases in that they have identified a large number of loci with small genetic effects[1,2]. The polygenic risk score (PRS) has thus emerged as a promising factor for predicting disease risk. PRS is the cumulative, mathematical aggregation of risk derived from the contributions of many DNA variants across the genome[3].

Recent studies have shown the high prevalence of cardiometabolic conditions among adults in the United States[4], and together they are the leading cause of mortality around the world[5,6]. GWAS have identified hundreds of loci associated with common diseases such as coronary artery disease (CAD)[7], obesity[8], hypertension[9] (measured using systolic blood pressure [SBP] and diastolic blood pressure [DBP]), and type 2 diabetes (T2D)[10]. Among the individuals that are diagnosed with one disease (for example, T2D), the prevalence of comorbidities such as hypertension, CAD, heart failure, and chronic kidney disease is also increased. To fully evaluate disease risk in an individual, it is therefore essential to also consider comorbid or secondary conditions related to the primary disease. There are several GWAS that have identified shared genetic associations between cardiometabolic conditions, demonstrating similarity in the underlying genetic architecture[11,12]. Pathophysiology of these conditions also shows the *cross-talk* between organ systems and its effect on disease progression, such as hemodynamic interaction between heart and kidney in heart failure[13]. With PRS, it is possible to derive an individuals' disease risk for each cardiometabolic condition using GWAS summary statistics. PRS represents an aggregate measure of the cumulative effect of numerous genetic variants on a particular disease, capturing an individual's genetic predisposition. As such, PRS can be instrumental in assessing the genetic interplay among coexisting or comorbid conditions.

Numerous methodologies exist for constructing PRS targeted at specific diseases. Conventionally, genetic risk scores (GRS) were derived using the genome-wide significant SNPs from a GWAS; however, recent studies show that using association results with much lower p-value significance ($p<0.05$) segregate individuals risk with better accuracy[1]. The development and clinical utility of PRS is under active investigation, especially in globally diverse populations[14–16]. Most large-scale GWAS have been conducted in individuals from European ancestry populations and most PRS are derived from these studies. Subsequently, the majority of PRS investigations published to date have been conducted in populations of European ancestry[17]. There can be several differences such as linkage disequilibrium (LD) structure and allele frequency of the variants, which can lead to inaccurate PRS for non-European populations[17]. This is not unique to PRS studies, and the majority of human genetic research suffers from this same phenomenon[18]. To ensure the successful clinical implementation of PRS, it is imperative to evaluate its performance in diverse global populations that closely reflect the healthcare population being treated. Moreover, for PRS to become a truly inclusive and effective tool for precision medicine, they must be applicable to individuals of all genetic backgrounds, including those with mixed ancestral backgrounds. Achieving this level of equity and broad usability will contribute significantly to the advancement of personalized healthcare practices.

In this study, we investigated the implementation of PRS for cardiometabolic conditions in individuals in the Penn Medicine BioBank (PMBB). PMBB is a cohort of >250,000 individuals

established for genomic and precision medicine research. Approximately 45,000 of the individuals have genetic data imputed using the Trans-Omics for Precision Medicine (TOPMed) v2 dataset[19]. 20% of the PMBB study population is classified as African (AFR) ancestry based on genetic similarity to the 1000 genome (1KGP)[20] AFR superpopulation group. We calculated PRS in the PMBB based on GWAS summary statistics generated in multi-ancestry data to evaluate 1) risk prediction accuracy among all individuals, and among AFR and European (EUR) subpopulations; and 2) the utility of PRS in determining genetic overlap among cardiometabolic conditions.

2. Methods

2.1. Penn Medicine BioBank

The Penn Medicine BioBank (PMBB) recruits participants through the University of Pennsylvania Health System by enrolling at the time of appointment[21]. Patients participate by donating either blood or a tissue sample and allowing researchers access to their electronic health record (EHR) information. This academic biobank provides researchers with centralized access to a large number of blood and tissue samples with extensive health information from the EHR. The facility banks both blood specimens (i.e., whole blood, plasma, serum, buffy coat, and DNA isolated from leukocytes) and tissues (i.e., formalin-fixed paraffin-embedded, fresh, and flash frozen).

2.2. Genotyping and Quality Control and Imputation

The DNA extracted from blood samples was genotyped using the Illumina Global Screening Array. To ensure data integrity, we conducted quality control measures, excluding SNPs with a marker call rate of less than 95% and samples with a call rate of less than 90%. Additionally, individuals with sex discrepancies were removed from the analysis. Imputation was carried out using the Michigan Imputation server, leveraging the TOPMed Reference panel[19]. To determine genetic ancestry, we employed principal component analysis (PCA) using the smartpca tool[22] and the 1KGP dataset[20]. Genetic ancestry was inferred through a k-means clustering approach, utilizing the 1KGP super populations as genetic ancestry labels.

2.3. Polygenic Risk Scores

To derive PRS, we used the multi-ancestry summary statistics from the largest and/or most recent GWAS studies for each trait (See Table 1).

Table 1. Multi-ancestry GWAS

Phenotype	Sample size (N cases)	PMID
BMI	241,258	28443625[8]
CAD	547,261 (122,733)	29212778[7]
Hypertension (DBP, SBP)	318,891	30578418[9]
T2D	1,407,282 (228,499)	32541925[10]

Weights for each SNP were calculated using PRS-CS[23] (version from April 24, 2020), a method that performs Polygenic Prediction via Bayesian regression and continuous shrinkage priors. PRS-CS requires a reference panel that matches the ancestry distribution of the target data set. We generated multiple reference panels for analyses: a multi-ancestry LD reference panel using the HapMap SNPs from the entire 1KGP populations (2504 individuals), an African-only reference panel from the 1KGP African ancestry population, and a European-only reference panel from 1KGP European ancestry population. We identified LD patterns within the 1KGP population by using PLINK (version 1.90) to determine LD blocks and calculate the LD between the SNPs in each block. For PRS-CS, the global shrinkage parameter φ was fixed to 0.01, and default values were selected for all other parameters. PRSs were then calculated using the weights with PLINK. Only the SNPs in the target data set, summary statistics, and LD reference panel were included in the PRSs.

2.4. Phenotypes

We focused on four primary phenotypes to derive and evaluate the PRS association: CAD, hypertension (for DBP and SBP PRS), T2D, and BMI. Cases and controls for each binary phenotype were defined using International Classification of Diseases (ICD-9 and ICD-10) diagnosis codes (CAD: 414.0*, I25.1*; T2D: 250*, E11*; hypertension: 401*, I10*). Participants were coded as cases of a given phenotype if their records contained at least 1 of the corresponding ICD-9 or ICD-10 codes. The median value for BMI was extracted from the EHR.

For Phenome-wide Association Study (PheWAS) analysis, we derived phenotypes using ICD-9 and ICD-10 data from individuals from the Penn Medicine EHR. ICD-9 codes were aggregated to phecodes using the phecode ICD-9 map 1.2[24,25]; ICD-10 codes were aggregated to phecodes using the phecode ICD-10 map 1.2 (beta)[26]. Individuals are considered cases for the phenotype if they have at least 2 instances of the phecode on unique dates, controls if they have no instance of the phecode, and ‘other/missing’ if they have one instance of the phecode or a related phecode.

2.5. Statistical Analysis

PRS were normalized (mean of 0 and standard deviation of 1) for each analysis separately (stratified by ancestry and overall). Logistic or linear regression models accounting for age, sex, and the first 5 within-ancestry principal components (PCs) were used to test for association of PRS with each of the primary phenotypes (T2D, BMI, hypertension, and CAD). Area under the receiver operator curve (AUC) and DeLong test was determined using the R package pROC, using the full logistic regression model as above. AUC was also calculated for a reduced logistic regression model including covariates alone (age, sex, and the first 5 PCs). The DeLong test[27] is a non-parametric approach used to compare the AUCs of two correlated ROC curves, especially when the models are applied to the same set of samples. This test was used to compare null model and full model that includes PRS and obtain a p-value indicating the statistical significance of the

difference between the two AUCs. For BMI, we treated it as a continuous trait and provided the R^2 value for all analyses.

A PheWAS was performed using logistic regression models with each PRS as the independent variable, phecodes as the dependent variables, and age, sex, and the first 10 PCs as covariates. A phenome-wide Bonferroni significance threshold of 4.2×10^{-5} (0.05/1190) in AFR and 3.6×10^{-5} (0.05/1377) in EUR was applied to account for multiple testing.

3. Results

3.1. Penn Medicine BioBank (PMBB) Demographics

PMBB currently consists of >250,000 consented individuals. Approximately 45,000 of these participants have been genotyped to date. Demographics of the sample included in this study are shown in Table 2.

Table 2. Demographics of PMBB sample

	All	AFR	EUR
Total patients	43,530	11,189	30,094
% Female	50.1%	62.8%	44.9%
Mean age	55.2	51.7	57.3
% CAD	23.8%	18.8%	26.4%
% Hypertension	54.4%	65.2%	51.7%
% T2D	23.5%	35.1%	19.3%
Patients with BMI data	40,043	10,619	27,489
% Female	50.4%	63.4%	44.9%
Mean age	55.6	51.9	57.7

3.2. Determining the effect of linkage disequilibrium panel on PRS in the overall sample

Using publicly available multi-ancestry GWAS data (Table 1), we generated a PRS for each primary phenotype of interest: type 2 diabetes, body mass index, hypertension (SBP and DBP), and coronary artery disease. We assessed the impact of using a multi-ancestry LD panel, akin to the GWAS data, and compared it with an AFR LD panel (in all PMBB individuals and in AFR PMBB individuals) and an EUR LD panel (in all PMBB individuals and in EUR PMBB individuals). AUC values were computed for each binary phenotype PRS in all individuals (Table 3) and contrasted between the full model (AUC, covariates + PRS) and the model containing covariates alone (AUC Null). The addition of PRS consistently improved the covariate model for all phenotypes, showing an average AUC improvement of 0.014. Across the entire dataset, the PRS created with the multi-ancestry LD panel (DBP, BMI) or the EUR LD panel (CAD, SBP, T2D) demonstrated the strongest association with their respective primary phenotypes (Table 3).

Table 3. Comparison of LD panel for PRS in all

PRS	LD Panel	AUC[1] Null	AUC[1]	DeLong P	Model OR	Model P-value
CAD	Multi-ancestry	0.795	0.808	1.22E-53	1.495	5.82E-186
	AFR		0.807	1.22E-52	1.472	7.11E-182
	EUR		0.807	2.33E-52	1.515	1.00E-184
DBP	Multi-ancestry	0.770	0.773	8.90E-06	1.236	1.65E-49
	AFR		0.772	1.32E-15	1.219	1.59E-49
	EUR		0.772	6.15E-14	1.226	6.32E-43
SBP	Multi-ancestry	0.770	0.775	4.47E-23	1.365	2.48E-83
	AFR		0.775	3.74E-22	1.338	2.78E-80
	EUR		0.775	7.40E-23	1.376	2.31E-83
T2D	Multi-ancestry	0.695	0.730	5.41E-88	2.223	1.24E-286
	AFR		0.727	2.68E-79	2.095	3.18E-266
	EUR		0.731	2.44E-91	2.263	1.46E-297

PRS	LD Panel	R^2 Null	R^2	R^2 difference	Model Beta	Model P-value
BMI	Multi-ancestry	0.067	0.110	0.043	2.205	0
	AFR		0.110	0.043	2.125	0
	EUR		0.108	0.042	2.198	0

3.3. Determining the effect of linkage disequilibrium panel on PRS within ancestry

In both AFR (Table 4) and EUR (Table 5) individuals, the addition of PRS to the covariate model enhances model performance. However, it is noteworthy that PRS performance was relatively stronger in EUR individuals compared to AFR individuals. In AFR, the full model shows a somewhat smaller improvement over the covariate-based model (average improvement in AUC=0.011) compared to the improvement observed in EUR (average improvement in AUC=0.021).

Notably, in AFR individuals, the PRS calculated using the multi-ancestry LD panel exhibited a higher effect size in four out of the five PRSs (DBP, SBP, T2D, and BMI) compared to the AFR LD panel (Table 4). This indicates the potential benefits of using a multi-ancestry LD panel to derive PRS in populations with diverse genetic backgrounds.

[1] AUC rounded to three decimal points

Table 4. Comparison of LD panel for PRS in AFR individuals

PRS	LD Panel	AUC Null	AUC	DeLong P	Model OR	Model P-value
CAD	AFR	0.764	0.770	1.33E-06	1.261	2.75E-18
	Multi-ancestry		0.770	4.52E-06	1.253	2.45E-17
DBP	AFR	0.793	0.797	1.72E-05	1.208	4.56E-15
	Multi-ancestry		0.797	1.25E-05	1.214	2.56E-15
SBP	AFR	0.793	0.797	3.82E-06	1.252	3.00E-18
	Multi-ancestry		0.797	1.11E-06	1.277	9.65E-20
T2D	AFR	0.681	0.710	3.03E-25	1.630	5.73E-77
	Multi-ancestry		0.711	4.21E-26	1.689	1.73E-79
PRS	LD Panel	R^2 Null	R^2	R^2 difference	Model Beta	Model P-value
BMI	AFR	0.041	0.065	0.024	1.449	1.02E-59
	Multi-ancestry		0.063	0.022	1.462	6.84E-56

In EUR individuals, the PRS calculated using the multi-ancestry LD panel demonstrated a higher effect size in two out of the five PRSs (SBP and T2D) when compared to the EUR LD panel (Table 5). This observation highlights the potential advantages of leveraging a multi-ancestry LD panel in deriving PRS for certain phenotypes in populations with European ancestry.

Table 5. Comparison of LD panel for PRS in EUR individuals

PRS	LD Panel	AUC Null	AUC	DeLong P	Model OR	Model P-value
CAD	EUR	0.796	0.812	9.49E-48	1.533	5.65E-166
	Multi-ancestry		0.812	2.38E-48	1.531	5.73E-165
DBP	EUR	0.747	0.750	6.17E-11	1.173	9.17E-34
	Multi-ancestry		0.750	1.51E-12	1.158	9.43E-29
SBP	EUR	0.747	0.753	6.64E-21	1.251	1.49E-64
	Multi-ancestry		0.753	1.61E-20	1.255	2.40E-66
T2D	EUR	0.651	0.708	8.26E-87	1.721	5.68E-243
	Multi-ancestry		0.710	1.12E-82	1.757	8.59E-258
PRS	LD Panel	R^2 Null	R^2	R^2 difference	Model Beta	Model P-value

BMI	EUR	0.006	0.076	0.070	1.637	0
	Multi-ancestry		0.075	0.069	1.626	0

3.4 PheWAS of polygenic risk scores

We conducted a PheWAS of each multi-ancestry LD panel PRS in AFR and EUR individuals, identifying additional phenotypes associated with the PRS for our primary phenotypes (Figure 1, full results in Supplemental Tables Online: https://shorturl.at/uBDSX). The results reveal significant associations between the PRS and various phenotypic categories, shedding light on the potential implications of PRS in predicting disease susceptibility. All PRS exhibited associations with other phenotypes. However, in AFR individuals, the strength and number of PRS associations with other phenotypes were generally reduced compared to EUR individuals.

In our analysis, the CAD PRS in AFR individuals was associated with 18 distinct phenotypes, including notable associations with hyperlipidemia (OR=1.12, $p=1.1x10^{-6}$) and renal failure (OR=1.12, $p=1.0x10^{-5}$). In contrast, EUR individuals exhibited associations with a broader range of 82 phenotypes, with hyperlipidemia (OR=1.23, p=7.3x10-45) and renal failure (OR=1.10, $p=2.1x10^{-8}$) being among them.

For the DBP and SBP PRS, AFR individuals showed associations with 9 and 20 phenotypes respectively. Specific associations of interest included atrial fibrillation for DBP (OR=1.20, $p=1.4x10^{-5}$) and both coronary atherosclerosis (OR=1.20, $p=3.7x10^{-7}$) and T2D (OR=1.12, $p=3.2x10^{-5}$) for SBP. EUR individuals, on the other hand, had DBP and SBP PRS associated with 12 and 27 phenotypes, respectively. This encompassed associations like coronary atherosclerosis for both DBP (OR=1.09, $p=4.9x10^{-7}$) and SBP (OR=1.13, $p=1.6x10^{-13}$), and T2D specifically for SBP (OR=1.17, $p=1.0x10^{-17}$).

The T2D PRS in AFR individuals was linked with a vast array of 84 phenotypes. Key associations here were hyperlipidemia (OR=1.30, $p=6.0x10^{-16}$), obesity (OR=1.20, $p=6.6x10^{-10}$), and hypertension (OR=1.22, $p=4.5x10^{-9}$). EUR individuals had a slightly lesser range with 78 phenotypes, but with significant associations like hyperlipidemia (OR=1.31, $p=9.2x10^{-17}$), obesity (OR=1.29, $p=9.9x10^{-57}$), and hypertension (OR=1.22, $p=3.2x10^{-38}$). Lastly, the BMI PRS in AFR was associated with 19 phenotypes, including T2D (OR=1.17, $p=1.6x10^{-8}$) and hypertension (OR=1.18, $p=8.6x10^{-8}$). In EUR individuals, this PRS was linked with a more extensive 72 phenotypes, with notable associations being T2D (OR=1.26, $p=4.6x10^{-39}$) and hypertension (OR=1.19, p=2.2x10-32).

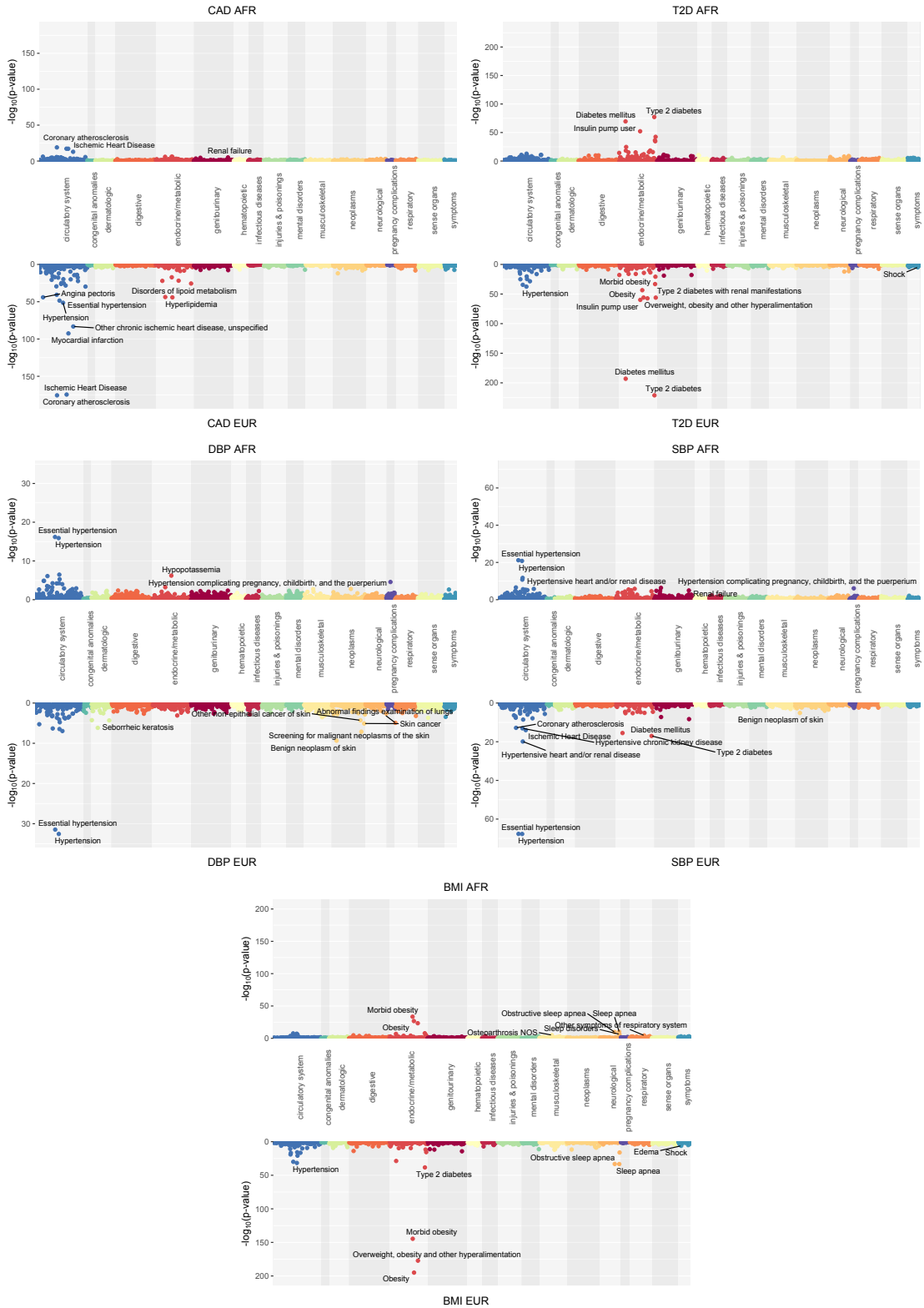


Figure 1. Phenome-wide Association Study (PheWAS) Results for Polygenic Risk Scores (PRS) for coronary artery disease (CAD), Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), Type 2 Diabetes (T2D), and Body Mass Index (BMI). The x-axis represents the phecode categories, and the y-axis shows the -log10 p-values, color-coded by category.

4. Discussion

We generated five polygenic risk scores representing genetic liability for cardiometabolic diseases and assessed their performance across different ancestry groups in the Penn Medicine BioBank (PMBB), a biobank including DNA linked with electronic health records. For all PRS tested, we identified a statistically significant association with the primary phenotype in both ancestry groups, as validated by the DeLong test comparing the null and the full model.

Type 2 diabetes consistently exhibited the highest effect size, reflecting the large number of cases in the GWAS used to generate this PRS and the PMBB dataset. Contrarily, the hypertension PRSs (DBP and SBP) showed a weaker effect size, even with a larger GWAS and over 50% of PMBB patient participants with hypertension. These observations suggest that factors beyond sample size, such as disease heterogeneity, prevalence, and non-additive effects, influence PRS associations. Consequently, understanding the interplay of these factors will be pivotal in refining and optimizing the application of PRS in disease prediction and risk assessment.

Our PheWAS analyses were conducted to explore the broader phenotypic landscape associated with each PRS with an EHR-linked biobank. Many of the identified phenotypes could be linked to broader effects of known disease risk factors and established comorbidities. For instance, risk for Type 2 diabetes was associated with hypertension, a known commonly co-occurring trait[28]. Similarly, the BMI PRS was associated with sleep apnea, diabetes, and hypertension, all of which are known to be more prevalent in individuals with higher BMI[29–32]. However, these associations don't necessarily imply causality. The high prevalence of comorbidities among these phenotypes complicates the task of discerning whether the genetic risk for one condition directly influences the onset of another.

Our findings underscore a significant challenge in the future implementation of PRS into routine clinical care. While PRS derived from multi-ancestry GWAS can be associated with phenotypes in individuals of African ancestry (AFR), their impact is not as pronounced as those generated in European ancestry (EUR). This observation, although expected, has been a topic of extensive discussion in recent years, emphasizing a notable disparity in genetic research[15,17]. Our results here affirm that these expectations persist even in large-scale, diverse ancestry datasets. Furthermore, our study suggests that PRS for cardiometabolic diseases based on multi-ancestry GWAS data might not perform as robustly for the primary disease and its associated secondary cardiometabolic traits.

Our utilization of a multi-ancestry LD panel to compute PRS for all individuals from multi-ancestry GWAS demonstrated robust performance across all populations. This was especially true for African ancestry individuals, emphasizing the potential advantages of leveraging a multi-ancestry reference panel in PRS generation. As the field of precision medicine continues to evolve, advocating for the adoption of such panels becomes increasingly important. By addressing these challenges, we can pave the way for more inclusive and accurate personalized healthcare strategies.

One notable limitation of our study is the modest gain in predictive performance over the null model across all categories, as reflected in the AUC values. While we observed differences in AUC between the ancestry groups, the absolute increase in AUC over the null model was relatively small. This underscores the need for further refinement in PRS methodologies to achieve more substantial improvements in predictive performance. Additionally, in our PheWAS approach, there are inherent challenges when comparing results between AFR and EUR groups. The difference in sample sizes between these groups can lead to variations in statistical power, potentially influencing the observed associations. Moreover, the generally lower PRS performance in the AFR group, as highlighted in our results, can further compound these challenges. It's essential to interpret the PheWAS results with these considerations in mind.

In conclusion, while there's considerable enthusiasm surrounding PRS in clinical care, there remains a significant amount of research to be conducted to determine its optimal implementation. It is essential to explore how PRS can be incorporated alongside other commonly used predictors[33], such as family history, clinical comorbidities, and environmental/lifestyle factors. By combining PRS with established clinical guidelines, we can aim for a more comprehensive risk assessment, leading to personalized interventions. Another important issue to address is whether we will ultimately need ancestry-specific PRS models or if we can develop the statistical framework to integrate global and local LD patterns into the PRS model to produce a cosmopolitan PRS approach. For clinical implementation, a cosmopolitan PRS approach will be easier for clinicians to adopt; however, it is unclear how this can be done effectively, given the heterogeneity in LD patterns, effect sizes, and causal variants in different ancestry groups. Our work here suggests that the use of multi-ancestry GWAS and LD panels may be a step towards this goal. The ultimate success of PRS in precision medicine lies in integrating it seamlessly with published clinical guidelines and incorporating an individual's ancestry within the PRS framework. This integration will empower clinicians to make informed decisions based on a comprehensive and personalized risk profile for each patient. By addressing these key aspects and enhancing our understanding of PRSs role in precision medicine, we can unlock its full potential as a transformative tool in healthcare, facilitating early interventions and preventive measures that cater to each individual's unique genetic makeup and health needs.

5. Acknowledgements

We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the patient-participants of Penn Medicine who consented to participate in this research program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. The authors thank Million Veteran Program (MVP) staff, researchers, and volunteers, who have contributed to MVP, and especially

participants who previously served their country in the military and now generously agreed to enroll in the study. (See https://www.research.va.gov/mvp/ for more details). The citation for MVP is Gaziano, J.M. et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol 70, 214-23 (2016). This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration, and was supported by the Veterans Administration (VA) Million Veteran Program (MVP) award #000. We have accessed the MVP summary statistics via dbGaP phs001672.

References

1. Abraham, G. *et al.* Genomic prediction of coronary heart disease. *Eur. Heart J.* **37**, 3267–3278 (2016).
2. Tada, H. *et al.* Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. *Eur. Heart J.* **37**, 561–567 (2016).
3. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
4. Arnold, S. V. *et al.* Burden of cardio-renal-metabolic conditions in adults with type 2 diabetes within the Diabetes Collaborative Registry. *Diabetes Obes. Metab.* **20**, 2000–2003 (2018).
5. Wang, H. *et al.* Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* **388**, 1459–1544 (2016).
6. Ogurtsova, K. *et al.* IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Res. Clin. Pract.* **128**, 40–50 (2017).
7. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).
8. Justice, A. E. *et al.* Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nat. Commun.* **8**, 14977 (2017).
9. Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **51**, 51–62 (2019).
10. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691 (2020).
11. Ma, R. C. Genetics of cardiovascular and renal complications in diabetes. *J. Diabetes Investig.* **7**, 139–154 (2016).

12. Regele, F. *et al.* Genome-wide studies to identify risk factors for kidney disease with a focus on patients with diabetes. *Nephrol. Dial. Transplant.* **30**, iv26–iv34 (2015).
13. Rangaswami, J. *et al.* Cardiorenal Syndrome: Classification, Pathophysiology, Diagnosis, and Treatment Strategies: A Scientific Statement From the American Heart Association. *Circulation* **139**, (2019).
14. Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J. & Lachance, J. Genetic disease risks can be misestimated across global populations. *Genome Biol.* **19**, (2018).
15. De La Vega, F. M. & Bustamante, C. D. Polygenic risk scores: a biased prediction? *Genome Med.* **10**, (2018).
16. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
17. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
18. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* (2019) doi:10.1038/s41576-019-0144-0.
19. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
20. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
21. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *J. Pers. Med.* **12**, 1974 (2022).
22. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
23. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
24. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

25. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).

26. Wu, P. *et al.* Developing and Evaluating Mappings of ICD-10 and ICD-10-CM Codes to PheCodes. *bioRxiv* (2019) doi:10.1101/462077.

27. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).

28. Sun, D. *et al.* Type 2 Diabetes and Hypertension: A Study on Bidirectional Causality. *Circ. Res.* **124**, 930–937 (2019).

29. Romero-Corral, A., Caples, S. M., Lopez-Jimenez, F. & Somers, V. K. Interactions Between Obesity and Obstructive Sleep Apnea. *Chest* **137**, 711–719 (2010).

30. Dua, S., Bhuker, M., Sharma, P., Dhall, M. & Kapoor, S. Body mass index relates to blood pressure among adults. *North Am. J. Med. Sci.* **6**, 89 (2014).

31. Xiang, B.-Y. *et al.* Body mass index and the risk of low bone mass–related fractures in women compared with men: A PRISMA-compliant meta-analysis of prospective cohort studies. *Medicine (Baltimore)* **96**, e5290 (2017).

32. Gray, N., Picone, G., Sloan, F. & Yashkin, A. Relation between BMI and Diabetes Mellitus and Its Complications among US Older Adults. *South. Med. J.* **108**, 29–36 (2015).

33. Arnett, D. K. *et al.* 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* (2019) doi:10.1161/CIR.0000000000000678.