INCORPORATING EXPERT TERMINOLOGY AND DISEASE RISK FACTORS INTO CONSUMER HEALTH VOCABULARIES

Michael Seedorff*

University of Iowa
240 Schaeffer Hall, Iowa City, IA, USA
Email: michael-seedorff@uiowa.edu

Kevin J. Peterson, BS

Department of Information Technology, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: peterson.kevin@mayo.edu

Laurie A. Nelsen, MS

Mayo Clinic Global Business Solutions, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: nelsen.laurie@mayo.edu

Cristian Cocos, PhD

Mayo Clinic Global Business Solutions, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: cocos.cristian@mayo.edu

Jennifer B. McCormick, PhD, MPP

Department of General Internal Medicine, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: mccormick.jb@mayo.edu

Christopher G. Chute, MD, DrPH

Department of Health Sciences Research, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: chute@mayo.edu

Jyotishman Pathak, PhD

Department of Health Sciences Research, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: pathak.jyotishman@mayo.edu

It is well-known that the general health information seeking lay-person, regardless of his/her education, cultural background, and economic status, is not as familiar with—or comfortable using—the technical terms commonly used by healthcare professionals. One of the primary reasons for this is due to the differences in perspectives and understanding of the vocabulary used by patients and providers even when referring to the same health concept. To bridge this "knowledge gap," consumer health vocabularies are presented as a solution. In this study, we introduce the Mayo Consumer Health

*This work was done while the author was an undergraduate summer intern at Mayo Clinic.

Vocabulary (MCV)—a taxonomy of approximately 5,000 consumer health terms and concepts—and develop text-mining techniques to expand its coverage by integrating disease concepts (from UMLS) as well as non-genetic (from deCODEme) and genetic (from GeneWiki+ and PharmGKB) risk factors to diseases. These steps led to adding at least one synonym for 97% of MCV concepts with an average of 43 consumer friendly terms per concept. We were also able to associate risk factors to 38 common diseases, as well as establish 5,361 Disease:Gene pairings. The expanded MCV provides a robust resource for facilitating online health information searching and retrieval as well as building consumer-oriented healthcare applications.

Keywords: Information Extraction; Consumer Health Vocabularies; Disease Risk Factors

1. Introduction

In the age of individualized medicine, it is becoming increasingly evident that more and more consumers are using the Internet and the World Wide Web to seek medical and health related information.[1,2] According to surveys by the Jupiter Organization and Harris Interactive, in 2007, 71% of people who used the Internet, also used it to seek health information (an increase by 37% since 2005).[3] Furthermore, it has been reported that 70% of people who obtain health information online say that it has influenced a decision about their treatment.[4] However, often due to various educational, economical, cultural, and language differences between patients and healthcare professionals, there exists a barrier in the process of gathering and interpreting health related information. One of the primary reasons for this is due to differences in perspectives and understanding of healthcare between patients and providers, as well as a significant disconnect in the vocabulary used even when they are referring to the same health concept.

Since various aspects of healthcare outcomes, including empowering consumers to make better-informed decisions and increasing patient compliance, can be affected due to this information disconnect, addressing the consumer health vocabulary problem has emerged as an important research activity in the recent past[5–7] as evidenced by services such as MedLine Plus[8] provided by the NIH. Cole et al.[9] proposed using a standardized biomedical terminology, SNOMED-CT,[10] and a commercially developed consumer health vocabulary, Intelligent Medical Object's Personal Health Terminology (PHT^{TM}), to assist patients and physicians who use common language terms to find specialist physicians with a particular clinical expertise. In particular, based on a user's input string, PHT was searched for term matching to acquire the SNOMED-CT codes (via PHT^{TM}–SNOMED map) that were in turn used to find physicians with the appropriate clinical specialty. In more recent work, the Open Access Collaboratory Consumer Health Vocabulary (OAC-CHV[11]) developed at the University of Utah contains more than 150,000 consumer health terms that are mapped to clinically oriented terms from the UMLS.[12] OAC-CHV has also been demonstrated in successfully translating clinical text from electronic medical records to consumers.[13]

While the above research has shown promising outcomes, there are several limitations hindering widespread adoption and application of these results. First, excluding OAC-CHV, the existing vocabularies for consumer health, such as PHT^{TM}, are either closed sourced, or have a commercial license. This not only prevents them from being leveraged in consumer health applications and tools, but also limits further development and community input. Second, with the recent advances in genomic medicine, the science and the role of non-genetic and

genetic risk factors in disease etiology is becoming clearer. Consequently, there is an increasing need to incorporate such information within consumer health vocabularies–a requirement not adequately met by existing vocabularies. Finally, best practices in modeling vocabularies require explicit specification of relationships between the terms and concepts, as well as providing appropriate metadata (e.g., synonyms, definitions, provenance). This impacts the vocabulary management and development to semantics-based querying and navigation leveraging the vocabulary. Our preliminary findings indicate that none of the existing consumer health vocabularies, including freely available OAC-CHV, adopt such methodologies, and are developed using ad-hoc vocabulary modeling formalisms. For example, OAC-CHV is modeled and maintained using Microsoft Excel files, instead of a more formal knowledge representation language, such as OWL (Web Ontology Language).[14]

In this study, we attempt to address the first two limitations. Specifically, we introduce the Mayo Consumer Health Vocabulary (MCV) developed and maintained by the ontology team at Mayo Clinic Global Products and Services to support annotation on MayoClinic.com (`http://www.mayoclinic.com`) health portal initially launched in 1995. Currently, MCV comprises approximately 5,000 consumer health terms arranged in a taxonomy, and includes mappings to SNOMED-CT and ICD-9[15] for some of the core concepts. The terminology extends beyond the typical medical terminologies to include lifestyle terms representing consumer health concepts related to nutrition, exercise and other lifestyle behaviors that influence a persons health. While successfully used to annotate health related information (articles, documents, blog entries, multimedia etc.) within the MayoClinic.com portal[b], MCV currently lacks the coverage for several disease concepts as well as relevant disease risk factors. The current study addresses these requirements by developing text mining approaches for integrating disease concepts (from OAC-CHV[16]) as well as non-genetic (from deCODEme[17]) and genetic (from GeneWiki+[18] and PharmGKB[19]) risk factors to diseases. The integration led to adding at least one synonym for 97% of MCV concepts with an average of 43 consumer friendly terms per concept, an important step in increasing search result coverage for future versions of MayoClinic.com. We were also able to associate non-genetic risk factors to 38 common diseases, as well as establish 5,361 Disease:Gene pairings. We discuss the details of our methods and findings in the remainder of this manuscript.

2. Resources and Tools

The following resources and tools were leveraged to conduct this study.

2.1. *Open Access Collaboratory Consumer Health Vocabulary*

The Open Access Collaboratory Consumer Health Vocabulary (OAC-CHV[16]) is created and maintained by the Consumer Health Vocabulary Initiative. It is a relationship file that links commonly used real-world vocabulary to associated medical terminology. Additionally, it provides the associated UMLS CUIs as well as understandability scores for each term and whether

[b]Our recent Web analytics statistics indicate that the MayoClinic.com portal is, on average, visited by more than 22 million unique visitors every month.

a term is disparaged (has an abnormality, such as a misspelling). In total there are 158,519 terms and 57,819 unique UMLS CUIs (2.7 terms per UMLS CUI). We used this file for finding near-matching terms to those in MCV and retrieving the connected terms based on common UMLS CUIs. We also used the UMLS CUIs connected to retrieved terms for comparison of similarity between MCV and OAC-CHV terms.

2.2. *SNOMED-CT*

The Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT[10]) was created by the College of American Pathologists and is maintained by the International Health Terminology Standards Development Organisation. It is a hierarchical ontology of medical terms. Similarity information can be gathered to compare two items in SNOMED-CT using several ontology-based algorithms such as Wu-Palmer.[20] We used SNOMED-CT as the UMLS source for comparison of UMLS CUIs and retrieval of UMLS CUI synonyms within the UMLS::Similarity and UMLS::Interface modules, respectively (see below).

2.3. *PharmGKB*

The Pharmacogenomics Knowledge Base (PharmGKB) is managed at Stanford University and focuses on maintaining information about gene:drug relationships and the corresponding gene variations, but also includes limited information on gene:disease relationships.[19] The data is collected from literature and other databases that report study results having to do with gene:drug interactions. It uses its own ID system for genes and diseases but provides data sets that allow for translation of genes into Entrez Gene IDs and diseases into SNOMED or UMLS IDs. We retrieved all Disease:Entrez Gene ID relationships and used this as a basis for our list of genetic risk factors by disease.

2.4. *deCODE genetics*

deCODE genetics[17] is a pharmaceutical company with an interest in genetic effects on disease and medicine. They sell a Direct-to-Consumer genetic testing service, called deCODEme, for sequencing a portion of an individual's genome to estimate genetic risk of various diseases. deCODEme has a website that contains information on the 47 diseases that are being tested, including information on both non-genetic and genetic factors that increase an individual's risk. We use the non-genetic factors portion of these disease pages to mine risk factors.

2.5. *UMLS::Interface*

UMLS::Interface is a Perl module that retrieves the position of UMLS CUIs from a UMLS ontology source (i.e. SNOMED-CT).[21] It provides tools for translating medical terms given as strings into the corresponding UMLS CUIs, getting positions in the ontology based on the UMLS CUI, and returning related UMLS CUIs and associated medical terms. Position in UMLS can be retrieved using a UMLS CUI or, if no UMLS CUI is available, one can be estimated based on an input string. It requires UMLS be loaded into a MySQL Database for access. We used this module to retrieve sister nodes (synonyms) for each MCV term.

2.6. *UMLS::Similarity*

UMLS::Similarity is a Perl module that retrieves a similarity score between two concepts based on their positioning in the hierarchical UMLS source (i.e. SNOMED-CT).[21] It has several options for evaluating either similarity or relatedness for two UMLS CUIs. Eight similarity measures, based on location in the ontology, were incorporated into the module (including Wu-Palmer, the similarity measure used in this study) as well as various relatedness measures that were not used in this study. This module was used for computing the similarity of MCV and OAC-CHV terms to indicate whether the relationship was valid (should be maintained) or invalid (should be deleted).

2.7. *MetaMap*

MetaMap is a program designed to extract biomedical terminology from text and map it to appropriate UMLS concepts.[22] It splits input text into minimal phrases and provides potential UMLS matches for the terms, indicating a score from 0-1000 with a higher score meaning a better match, as well as the semantic type (i.e. disease, substance, ...), UMLS source, and UMLS CUI. We used this program to extract non-genetic risk factors from plain text with the ability to divide sentences into phrases and indicate the semantic type being crucial.

3. Materials and Methods

3.1. *Materials*

The primary materials used in this study are the following:

- The February 4, 2011 OAC-CHV data set, available for download via `http://consumerhealthvocab.org`. The data set contains 158,519 mappings between medical concepts and terms along with several measures of understandability for each term. There is a one–to–many relationship between UMLS CUIs and OAC-CHV terms.
- The July 3, 2012 GeneWiki+ relationships data set, available for download via `http://genewikiplus.org/wiki/GeneWiki:Data`. The data set contains 18,230 relationships between genes and diseases, referencing the diseases using a Disease Ontology ID (DOID).[23]
- The June 13, 2012 Human Disease Ontology data set, available for download via `http://obofoundry.org`. The data set contains 8,631 entries, each with at least one DOID, and a total of 14,311 SNOMED IDs mapped to the entries.
- The July 5, 2012 PharmGKB relationships data set, available by request via `http://www.pharmgkb.org/downloads.jsp`. The data set contains 11,706 unique relationships between drugs, diseases, genes, haplotypes, and gene variant locations (see Table 1). It includes information on whether pharmacokinetic and pharmacodynamic effects play a part in the relationship as well as PubMed IDs for articles that provide evidence supporting the relationship. Also available are gene and disease data sets, providing mappings between genes and Entrez Gene IDs, and diseases and SNOMED-CT IDs, respectively.
- The MCV data set and MCV-SNOMED relationship data set, not publicly available for this study but, in the future, will be made available for public use. MCV includes a list of around 5,000 medical terms, 2,126 of which are considered core terms (directly associated with

Table 1. PharmGKB Relationships (Highlighted fields indicate relationships studied in this work)

	Haplotype	Gene	Variant Location	Drug	Disease	Entrez Gene ID	SNOMED-CT
Haplotype	0	0	0	762	169	0	0
Gene		684	0	2,578	1,541	27,421	0
Variant Location			0	3,147	2,053	0	0
Drug				0	772	0	0
Disease					0	0	4,348
Entrez Gene ID						0	0
SNOMED-CT							0

clinical concepts) and were the basis of this effort. These core terms are identified by MCV IDs and divided into 4 groups: diseases (1,443), first aid (63), symptoms (102), and test procedures (518). The MCV-SNOMED relationship data set contains 1,476 relationships between MCV IDs and SNOMED IDs.

3.2. *Methods for integrating disease concepts*

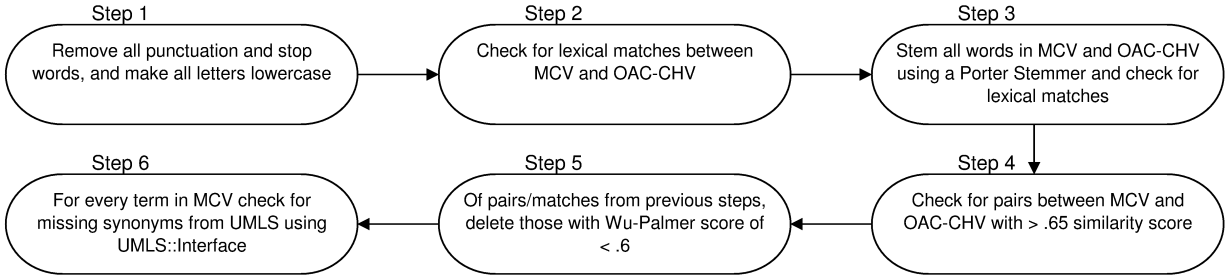

Fig. 1. Outline for linking MCV and OAC-CHV terms

For this study, we compared biomedical terms in MCV and OAC-CHV to expand the list of word alternatives for MCV. Note that traditional methods for ontology matching and alignment are not applicable here because they rely primarily on relationships between concepts as well as the hierarchical structure in the source and target ontologies (which are "metadata–based"), whereas both OAC-CHV and MCV are at present a nearly flat list of terms with minimal relationships and hierarchies. A general outline for integrating MCV and OAC-CHV is given in Fig. 1. For the strings in MCV and OAC-CHV, we removed all punctuation and stop words and made all letters lowercase. We used a specific subset of stop words that showed up often in the data to avoid deleting good words (i.e. 'a' in "vitamin a deficiency"). Because every term in OAC-CHV was paired with a UMLS CUI and a medically preferred term, we were able to create sets of potential phrases for each UMLS CUI which allowed us to retrieve a list of synonyms quickly for any entry in OAC-CHV. We began by simply seeing if any terms in MCV were exact matches to terms in OAC-CHV. This was followed by stemming all words in every term using a Porter Stemmer[24] and checking for exact matches between the two sets.

All matches were added to a matched list.

We then created a similarity score for each pair of terms between OAC-CHV and MCV. This score was calculated by giving one point to each word that was in the other term and .75 points to each stemmed word that was in the other stemmed term, summing these points, and dividing by the total number of words between the two terms. For example, the terms 'knee knees injury' and 'knee injuries' would receive a score of $(.75 + 1 + .75 + 1 + .75)/5 = .85$ (Fig. 2). Based on outcome observations, an empirical threshold of .65 was set where any pair that achieved a score equal to or over this threshold was considered to be matching and was added to the matched list.

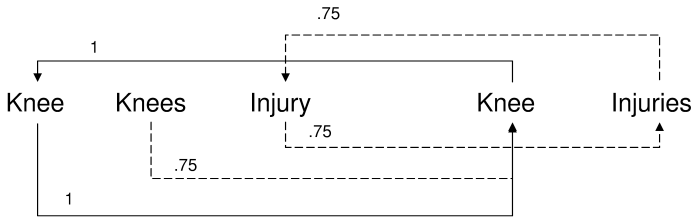

Fig. 2. Comparison scoring of two example terms

The next step was to get UMLS CUI codes for every term that had been paired in the matched list. For OAC-CHV terms, that information was already included in the file. For MCV terms, we used a relationships file developed by Mayo to get the connected SNOMED IDs. With those SNOMED IDs, we queried the BioPortal REST service which returned the appropriate UMLS CUIs.[25]

Next we evaluated the strength of the SNOMED relationship between each pair, using their UMLS CUI codes and the UMLS::Similarity module. MCV terms that were connected to multiple UMLS CUIs had the highest similarity score counted for each pairing. MCV terms which were connected to no UMLS CUIs did not go through this step. The similarity measure used was the Wu-Palmer Similarity score,[20] a measure that ranged from 0 (exclusive) to 1 (inclusive) with a larger number indicating two UMLS CUIs being more similar. Based on output observations, we set a threshold of .6 where any pair scoring below that would be deleted from the list of pairings.

Once all pairings had been computed, we began gathering synonyms for MCV terms. For every pairing between MCV and OAC-CHV, the OAC-CHV term was connected to a group of terms with the same UMLS CUI. For every pairing, this group of OAC-CHV terms was added to the correct MCV. UMLS::Interface was then queried for equivalent terms to every MCV term. These two groups of synonyms were combined for each MCV term and duplicate synonyms were deleted.

3.3. *Methods for integrating non-genetic and genetic disease risk factors*

For the second part of this study, we integrated non-genetic and genetic risk factors to diseases in MCV. Non-genetic factors were obtained by mining information from deCODEme's website

for most of the 47 medical conditions that they do genetic testing on. The text mining algorithm was implemented using the XML and RCurl packages in R.[26–28] First, a list of diseases was queried from the "about deCODEme" page of their website. The page for each individual disease was then accessed and the associated factors were retrieved. Because the information on non-genetic risk factors was stored in consistent locations within deCODEMe's website templates (usually in bold text; as seen in Fig. 3), our retrieval algorithm processed just the relevant text area. For factors that included ambiguous terms such as 'age,' 'ethnicity,' and 'gender,' we developed the following heuristics based on typical structures of the paragraphs that followed the highlighted function:

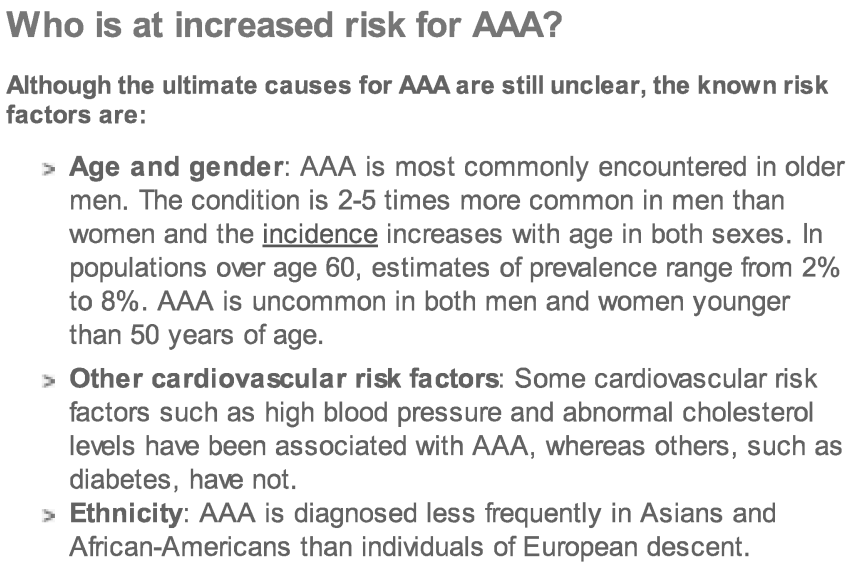

Fig. 3. Sample of risk factor portion of deCODEme site

- Gender – Typically the first gender to show up in the paragraph was at higher risk. When no gender was at higher risk, then either no gender was named or the first instance of a gender in the paragraph was accompanied by a conjunction and the opposite gender. For instance, in Fig. 3, "AAA is most commonly encountered in older men" would give us 'men' as the higher risk group, because it is spotted first in the paragraph. However, if the sentence were to instead say "AAA is most commonly encountered in older men and women," we would not assume a higher risk group.
- Age – There were many different structures for ages being described. We made a list of the typical ones for querying the text such as "over age ##," "between the ages of ## and ##," and "in their ##s." For instance, in Fig. 3, the phrase "over age 60" indicates that 60+ is a high risk group.
- Ethnicity – Typically there were many ethnicities mentioned and there was a rough ordering indicated by the comparison words used. Words such as 'more,' 'highest,' and 'fourfold'

indicated that the earliest ethnicities in the paragraph were at higher risk while the word 'less' indicated that the earliest ethnicities following the word 'than' represented high risk groups. Our method deleted all words that did not have to do with this ordering and were not ethnicities, allowing us to extract ethnicities based on locations of comparison words. For example, in Fig. 3 the ethnicity sentence is reduced to "less Asians African-Americans than European" and European would be chosen as the high risk group.

- 'Other' Categories – Categories that included the word 'other' in their title often listed many risk factors but did not have a uniform structure, making it much more difficult to extract the factors. To solve this problem we ran the paragraphs through MetaMap, a biomedical terminology extraction tool which split the paragraph up into concepts and provided expected semantic categories as well as goodness-of-fit scores. We took the terms which were substance, disease, or injury related, based on their semantic categories, and, if they had a perfect fit score of 1000, added them to the non-genetic factors list. In addition, if the words 'smoking,' 'alcohol,' or 'cocaine' were found, they were added to the factors list, even without a perfect goodness-of-fit score.

The second type of factor that we looked at was genetic. Initially we extracted all SNOMED IDs that were linked to each MCV ID by processing MCV's relationships file. The Human Disease Ontology[23] holds relationships between SNOMED IDs and DOIDs, allowing us to extend our connections between MCV IDs and DOIDs (Disease Ontology IDs). Using these relationships, we queried the GeneWiki+ data set to retrieve genes that were correlated to each DOID, and by extending that relationship to MCV and accumulating the genes, we created a relationship file between MCV IDs and Entrez Gene IDs.

In addition to using the GeneWiki+ data set, we also had access to PharmGKB relationships files which, among other things, linked diseases and genes through their PharmGKB Accession IDs. Subsequently, by using the PharmGKB genes relationships file, we replaced the listed genes with their Entrez Gene IDs. Similarly, by using the PharmGKB diseases relationships file, we replaced the diseases in the relationships with the connected SNOMED-CT IDs. We then replaced these SNOMED-CT IDs with the connected MCV IDs from MCV's relationships file and added any MCV:Entrez Gene ID pairs that were missing from GeneWiki+ to our list of MCV ID:Entrez Gene ID relationships.

4. Results

The MCV file we began with included 2,126 terms. After just looking for exact matches or stemmed perfect matches, 1,677 terms had found matches in OAC-CHV and 449 had not. When we did not use UMLS::Similarity to evaluate matches, we had 2,092 terms that found matches and 34 that did not. After using UMLS:Simliarity to eliminate weak or incorrect matches we had 2,069 terms that had matches and 57 that did not. Table 2 shows a summary of these findings.

On average, each term in MCV had 50.2 synonyms when not checking against UMLS::Similarity, but just 38.5 synonyms after incorporating this extra measure. UMLS::Interface averaged adding 4.5 synonyms to each term in MCV with a final average output of 43 synonyms per MCV term.

Table 2. Summary of MCV terms mapping results

	MCV Terms mapped to OAC-CHV	MCV Terms *not* mapped to OAC-CHV
Perfect Matches	1,646	480
Perfect Matches after stemming	1,677	449
Close matches using algorithm	2,092	34
Matches after UMLS::Similarity	2,069	57

deCODEme contained information on 47 diseases or conditions. Of these, five either did not have non-genetic factor information in the usual area (in lists within the main text area) or did not have any non-genetic factor information at all. Of the 42 that did contain non-genetic factor information, 38 matched either an MCV name or one of the synonyms previously created. On average each of these 38 diseases had 6.7 non-genetic factors gathered from deCODEme.

GeneWiki+ contained information on 18,230 Gene:Disease relationships and a total of 10,084 unique Entrez Gene ID:DOID relationships. There were a total of 361 diseases and seven symptoms from MCV that mapped to at least one gene and a total of 4,884 mappings between MCV entries and Entrez Gene IDs (once the MCV IDs had been processed into SNOMED IDs and then DOIDs).

The PharmGKB relationships file contained a total of 11,706 unique relationships, but only 1,541 of those were between diseases and genes. There were 570 MCV ID:Entrez Gene ID relationships recorded after tracking the DOIDs to the corresponding SNOMED-CT IDs and then MCV IDs. Of these, 93 already existed in the GeneWiki+ information and 477 were new. See Table 3 for a summary of these results. After including the PharmGKB information, coverage of MCV terms was the same (361 diseases and seven symptoms).

Table 3. Matching between Diseases and Genes

	MCV:Entrez Gene Pairs
Only in GeneWiki+	4,791
In both GeneWiki+ and PharmGKB	93
Only in PharmGKB	477
Total	5,361

5. Discussion

The principle goal of this study was to map terms and concepts from MCV to synonyms or near-synonyms from publicly available sources. Connecting similar terms from OAC-CHV, checking the quality of these matches using UMLS::Similarity, and extracting close relations from UMLS::Interface expanded the base list of terms by more than 43 times and over 97% of terms in MCV added at least one synonym. Having such a list will allow for improved search results that minimize the difficulty of finding an exact phrase to retrieve information on an

expected medical concept.

Our extraction of genetic factors was also very helpful in adding to MCV. GeneWiki+ and PharmGKB each added a valuable amount of gene:disease matchings with GeneWiki+ contributing somewhat more, reasonable considering PharmGKB specializes in gene:drug relationships. A large number of relationships presented in these files were unable to be mapped to any diseases in MCV due to either MCV lacking the disease or one of the ID relationship files being incomplete. With only 42 diseases from deCODEme having non-genetic risk information, it may have been more valuable to just manually edit those relationships. Extraction of ethnicity, gender, and age information was valuable but many factors were included in the 'other' categories and were not always correctly retrieved by MetaMap. It may be worthwhile to map these to a database of risk factors at some point, but that was not considered in this study.

6. Conclusion

In this study we integrated synonyms for medical terminologies as well as both non-genetic and genetic risk factors for diseases into MCV. Bringing this information into medical query services oriented towards consumers is an important step to providing better results and risk information that is growing in importance, especially as genetic risks become better known. The expanded version of MCV created in this exercise provides a solid basis for creation of consumer-oriented healthcare applications and online health information searching. With MCV becoming publicly available in the future, current limitations due to many consumer health vocabulary sources being closed source should be reduced.

7. Acknowledgments

This study was funded in part by the Mayo-NIH Relief Fund Award (FP00068486) and Mayo Clinic Early Career Development Award (FP00058504).

References

1. D. Borzekowski and V. Rickert. Adolescent Cybersurfing for Health Information: A New Resource That Crosses Barriers. *Archives of Pediatrics and Adolescent Medicine*, 155(7):813–817, 2001.
2. R. J. W. Cline and K. M. Haynes. Consumer Health Information Seeking on the Internet: The State of the Art. *Health Education Research*, 16(6):671–692, 2001.
3. Harris Interactive: Consumer Health Care Survey Reveals Mixed Bag of Results. Last accessed: 6th October, 2009.
4. Robert J. Bensley and Jodi Brookins Fisher. *Community Health Education Methods: A Practical Guide*. Jones and Bartlett Publishers, 2008.
5. Rita D. Zielstorff. Controlled Vocabularies for Consumer Health. *Journal of Biomedical Informatics*, 36(4-5):326–333, 2003.
6. Catherine Smith and P. Stavri. Consumer health vocabulary. *Consumer Health Informatics: Informing Consumers and Improving Health Care*, pages 122–128, 2005.
7. Q.T. Zeng and T. Tse. Exploring and Developing Consumer Health Vocabularies. *Journal of American Medical Informatics Association*, 13(1):24–29, 2006.
8. N. Miller, E. M. Lacroix, and J. E. Backus. MEDLINEplus: Building and Maintaining the Na-

tional Library of Medicine's Consumer Health Web Service. *Bulletin of the Medical Library Association*, 88(1):11–17, 2000.
9. Curtis L. Cole, Andrew S. Kanter, Michael Cummens, Sean Vostinara, and Frank Naeymi-Rad. Using a Terminology Server and Consumer Search Phrases to Help Patients Find Physicians with Particular Expertise. *Proceedings of the 11th World Congress on Medical Informatics: MEDINFO*, 107:492–496, 2004.
10. SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms. Last accessed: 10th July, 2012.
11. Q.T. Zeng, T. Tse, G. Divita, A. Keselman, and et al. Term Identification Methods for Consumer Health Vocabulary Development. *Journal of Medical Internet Research*, 9(1):e4, 2007.
12. Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32(Database issue):267–270, 2004.
13. Qing T. Zeng, S. Goryachev, H. Kim, A. Keselman, and S. Rosendale. Making Texts in Electronic Health Records Comprehensible to Consumers: A Prototype Translator. In *AMIA Annual Syposium*, pages 846–850, 2007.
14. Deborah L. McGuinness, Frank van Harmelen, and et al. OWL: Web Ontology Language. In *http://www.w3.org/2004/OWL/*, 2004.
15. World Health Organization International Classification of Diseases (ICD-9) Clinical Modification. Last accessed: 7th October, 2009.
16. Q. T. Zeng, T. Tse, G. Divita, A. Keselman, J. Crowell, A. C. Browne, S. Goryachev, and L. Ngo. Term identification methods for consumer health vocabulary development. *J. Med. Internet Res.*, 9(1):e4, 2007.
17. Decodeme. `http://www.decodeme.com/`.
18. Genewiki+. `http://genewikiplus.org/`.
19. R. B. Altman. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat. Genet.*, 39(4):426, Apr 2007.
20. Z. Wu and M. Palmer. Verb semantics and lexical selection. In *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pages 133–138, Las Cruces, NM, 1994.
21. McInnes, Pedersen, and Pakhomov. Umls-interface and umls-similarity : Open source software for measuring paths and semantic similarity. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 431–435, San Francisco, CA, November 2009.
22. A. R. Aronson and F. M. Lang. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3):229–236, 2010.
23. J. D. Osborne, J. Flatow, M. Holko, S. M. Lin, W. A. Kibbe, L. J. Zhu, M. I. Danila, G. Feng, and R. L. Chisholm. Annotating the human genome with Disease Ontology. *BMC Genomics*, 10 Suppl 1:S6, 2009.
24. Kurt Hornik. *Snowball: Snowball Stemmers*, 2012. R package version 0.0-8.
25. P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, 39(Web Server issue):W541–545, Jul 2011.
26. Duncan Temple Lang. *XML: Tools for parsing and generating XML within R and S-Plus.*, 2012. R package version 3.9-4.1.
27. Duncan Temple Lang. *RCurl: General network (HTTP/FTP/...) client interface for R*, 2012. R package version 1.91-1.1.
28. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.