How Much Does the (Social) Environment Matter? Using Artificial Intelligence to Predict COVID-19 Outcomes with Socio-demographic Data*

Christos A. Makridis
Arizona State University, MIT Sloan School of Management, Department of Veterans Affairs
Washington, DC, 20005
Email: christos.a.makridis@gmail.com

Anish Mudide
Phillips Exeter Academy
Exeter, NH 03833
Email: amudide@gmail.com

Gil Alterovitz
Brigham and Women's Hospital/Harvard Medical School, Boston, MA 02115 and
Department of Veterans Affairs, Washington, DC, 20005
Email: ga@alum.mit.edu

While the coronavirus pandemic has affected all demographic brackets and geographies, certain areas have been more adversely affected than others. This paper focuses on Veterans as a potentially vulnerable group that might be systematically more exposed to infection than others because of their co-morbidities, i.e., greater incidence of physical and mental health challenges. Using data on 122 Veteran Healthcare Systems (HCS), this paper tests three machine learning models for predictive analysis. The combined LASSO and ridge regression with five-fold cross validation performs the best. We find that socio-demographic features are highly predictive of both cases and deaths—even more important than any hospital-specific characteristics. These results suggest that socio-demographic and social capital characteristics are important determinants of public health outcomes, especially for vulnerable groups, like Veterans, and they should be investigated further.

Keywords: Artificial Intelligence, Coronavirus, COVID-19, Machine Learning, Veterans.

1. Introduction

Following months of the coronavirus ("COVID-19") pandemic, a large body of research has emerged quantifying the contribution of individual characteristics towards exposure of the virus (Britton et al., 2020; Martin et al., 2020). Moreover, there is also increasing evidence that certain vulnerable groups have been affected more adversely than others, especially minorities (Pan et al., 2020). However, researchers have struggled to obtain bias-free, reliable, and externally-valid predictions on representative datasets (Wynants et al., 2020).

* All replication files are available here: https://github.com/amudide/COVID-Sociodemographics-AI

The majority of studies have focused on the role of individual-level factors, but a separate vein of research in computational social science has found that socio-economic factors also play an important role in mediating the spread of the virus (Makridis and Wu, 2020; Ding et al., 2020; Barrios et al., 2020). For example, the Joint Economic Committee (JEC) in the United States Congress has focused on quantifying social capital and its important implications for economic outcomes and well-being (JEC, 2018). Moreover, social capital has also been associated with community-level health outcomes (Gordeev and Egan, 2015; Kolak et al., 2020).

We use machine learning (ML) and artificial intelligence (AI) methods on a combination of socio-demographic and social capital data to investigate the importance of local factors in explaining coronavirus health outcomes among Veterans. Given that Veterans are a vulnerable group and exhibit more mental and physical health challenges than non-Veterans, even within the same organization (Schult et al., 2019), community characteristics may play an important role in mediating the effects of the pandemic. For example, Veterans in communities with greater social capital may engage in more preventative health investments, which would bolster their immunity and recovery to viruses.

Using three different estimators—naïve ordinary least squares (OLS), ridge with cross validation (CV), and least absolute shrinkage and selection operator (LASSO) with ridge and CV—we predict coronavirus case and mortality outcomes using data on 122 Veteran Healthcare Systems (HCS). While our OLS specification performs well in-sample, it exhibits weak out-of-sample behavior. Instead, the ridge regression with LASSO for feature selection performs the best with an R-squared of 0.617 (0.471) when we predict coronavirus cases (deaths). We show that socio-demographic features matter more than any standard hospital features, such as its patient satisfaction or the number of services provided. These results are important for at least two reasons. First, we can obtain reasonable model accuracy on such a small sample. Second, we show that socio-demographic features matter even more than hospital features. This suggests that further research and predictive modeling on infectious diseases must incorporate socio-demographic and social capital characteristics if these models are going to be useful for policymakers and clinicians.

This paper contributes to a growing literature about the importance of socio-economic factors for understanding health outcomes for the spread of infectious diseases (Amarasingham et al., 2010; Navathe et al., 2018; Bejan et al., 2018; Makridis et al., 2020). The inclusion of socio-demographic variables at a geographic-level can improve the performance of otherwise standard ML models of the virus, but disaggregating between county and ZIP code does not make much of a difference. However, disaggregating between state and county does make a significant difference. The result is evidence that the community and the resulting healthcare infrastructure is determined at more of a county-level, rather than a ZIP code-level. Even if residential decisions take place at a more granular level, public health interventions may reside at a more aggregate level.

Our paper also contributes directly to an existing and timely research agenda on the effects of COVID-19 and the identification of individuals who are more exposed to it than others. For

example, age has emerged as one of the most important comorbidities (Zhou et al., 2020; Richardson et al., 2020). Similarly, Makridis and Wu (2020) show that social capital---that is, measures that describe the quality and strength of ties and relationships within a community---plays an important mediating role over the duration of the pandemic: counties with higher social capital have systematically lower infections and a slower spread of infection even after controlling for demographic characteristics and population density.

2. Data and Measurement

2.1. *Location-specific demographic characteristics*

Our socio-demographic data comes from the Census Bureau's 2014-2018 American Community Survey. The Census provides demographic characteristics, e.g.: the race distribution, the population density, the share male, the age distribution (the share under age 18, age 25-44, age 45-64, and 65+), the share married, the education distribution (the share with less than a high school degree, some college, and college or more), the income distribution (the share with less than $15,000, $15-29,000, $30-39,000, $40-49,000, $50-59,000, $60-99,000, $100-149,000, over $150,000), and the poverty rate (the share of people living in poverty under age 18, age 18-64, and 65+).[a]

2.2. *Hospital coronavirus cases, mortality, and features*

We use the Department of Veterans Affairs (VA) Facilities API.[b] We observe the number of services that a hospital provides for Veterans and the average satisfaction. We also observe the logged number of coronavirus cases and deaths within an HCS.[c] While we observe 1,297 VA health facilities, our coronavirus cases and deaths are available at only 122 HCS, which consist of multiple VA medical facilities. We map VA health facilities into an HCS by taking the weighted average of features in each health facility in an HCS using the number of Veterans in the area as our weight.

3. Methods

We use three standard statistical estimators: naïve ordinary least squares (OLS), ridge with cross validation (CV), and least absolute shrinkage and selection operator (LASSO) with ridge CV. A ridge estimator is given by the following:

$$\hat{\beta}^{RIDGE} = \arg\min\{(y - X\beta)^2 + \lambda||\beta||^2\} \quad (1)$$

This, unlike a standard OLS estimator, which only minimizes the sum of square error, inserts an additional term, λ, that biases certain features over others in the regression. While this will lead to "biased" parameter estimates, the fit is better because more important features are given additional weight. Moreover, the regularization term, λ, prevents overfitting since the model cannot adjust too

[a] We also include county data from the Joint Economic Committee (JEC) social capital index (JEC, 2018).
[b] See: https://developer.va.gov/explore/facilities/docs/facilities.
[c] See: https://www.accesstocare.va.gov/Healthcare/COVID19NationalSummary.

many feature weights without the $||\beta||^2$ term getting too big. We also experiment with a LASSO estimator for feature selection followed by ridge.

4. Results

We begin by reporting the performance of our ridge regression with CV and ridge regression with LASSO feature selection and CV in Table 1. We omit the performance of our OLS regression: since we did not do CV on it, the in-sample fit is artificially high because of the common problem of overfitting. Note that our measure of model performance is R-squared, rather than the more common Regression Receiver Operating Characteristic (RROC) curve plot that is more common for predicting continuous variables (Hernandez-Orallo, 2013). We find that the models for predicting mortality perform worse than for infections. One reason for this stems from the fact that deaths are relatively infrequent, so there is less variation available for prediction.

We also observe that the combined LASSO and ridge CV regression performs better than a standard ridge CV regression regardless of the performance metric that we focus on. For example, Panel A shows that the R-squared for LASSO and ridge is 0.617 and it is 0.564 for ridge both when the outcome variable is logged coronavirus cases. The R-squared values for logged deaths for our two respective models are 0.471 and 0.401. Turning towards Panel B, the RMSE for the LASSO and ridge CV regression is 6.6% (7.1%) lower when predicting cases (deaths). Note that since we are predicting $\log_2(cases)$, being 0.9 off, for example, translates to being a factor of $2^{0.9} = 1.86$ off. We find similar patterns in Panel C, which shows the MAE by model.

These differences between the two estimators emerge because of at least two reasons: (i) our sample of HCS is small ($N = 122$), (ii) and we have a large and multi-dimensional feature set, especially for demographic characteristics. Although ridge regressions allow features to contribute differently to the RMSE, it may still not perform optimally when there are nearly as many variables as there are observations. By applying LASSO, we can select only the most predictive variables and include them in a subsequent ridge regression. This performs the best.

Table 1: COVID-19 Model Performances

Outcome Variable =	log(Coronavirus Cases)	log(Coronavirus Deaths)
Panel A: Coefficient of Determination (R-squared)		
Ridge CV	0.564	0.401
LASSO + Ridge CV	0.617	0.471
Panel B: Root Mean Square Deviation (RMSE)		
Ridge CV	0.958	1.115
LASSO + Ridge CV	0.895	1.035
Panel C: Mean Absolute Error (MAE)		
Ridge CV	0.770	0.903
LASSO + Ridge CV	0.707	0.838

Given that the LASSO and ridge CV regression performs the best, we now treat this model as the baseline and examine the most important features for the coronavirus cases and deaths prediction problems in Figures 1 and 2, respectively. Note that all variable importance coefficients are measured in absolute value so that we can focus on the relative magnitudes.

We find that the share of the population between ages 55 and 64 is the most predictive, followed by the share of the population working in professional services and in sales occupations. The poverty rate for those over the age of 65, the male unemployment rate, the employment share in construction, agriculture / mining, and the employment share in education / health all enter as important predictors too. We find similar results when our outcome variable is coronavirus deaths, but several notable differences emerge. While certain variables, like dropping out of high school, were highly associated with deaths, they were not with cases (see Figure 1 and 2). There may be hidden variables associated with this, and other social capital variables, that lead to such worse health outcomes for such individuals.

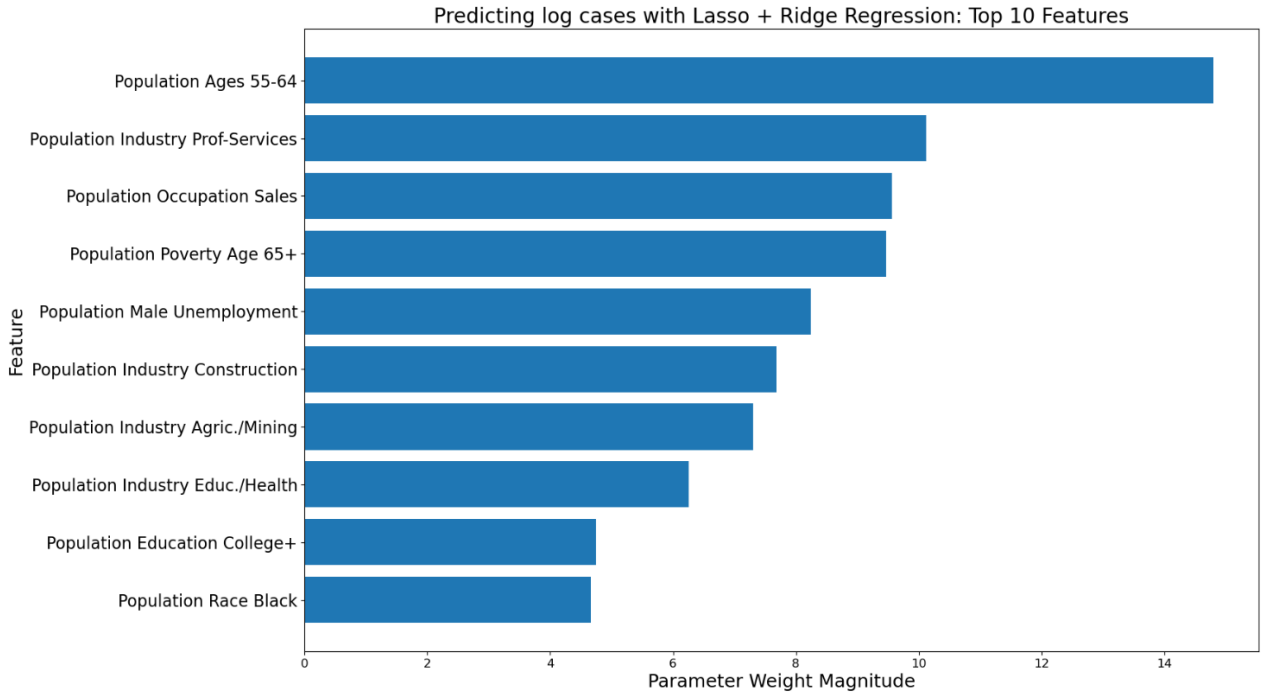

Fig. 1. Predicting log cases with Lasso + Ridge Regression: Top 10 Features.

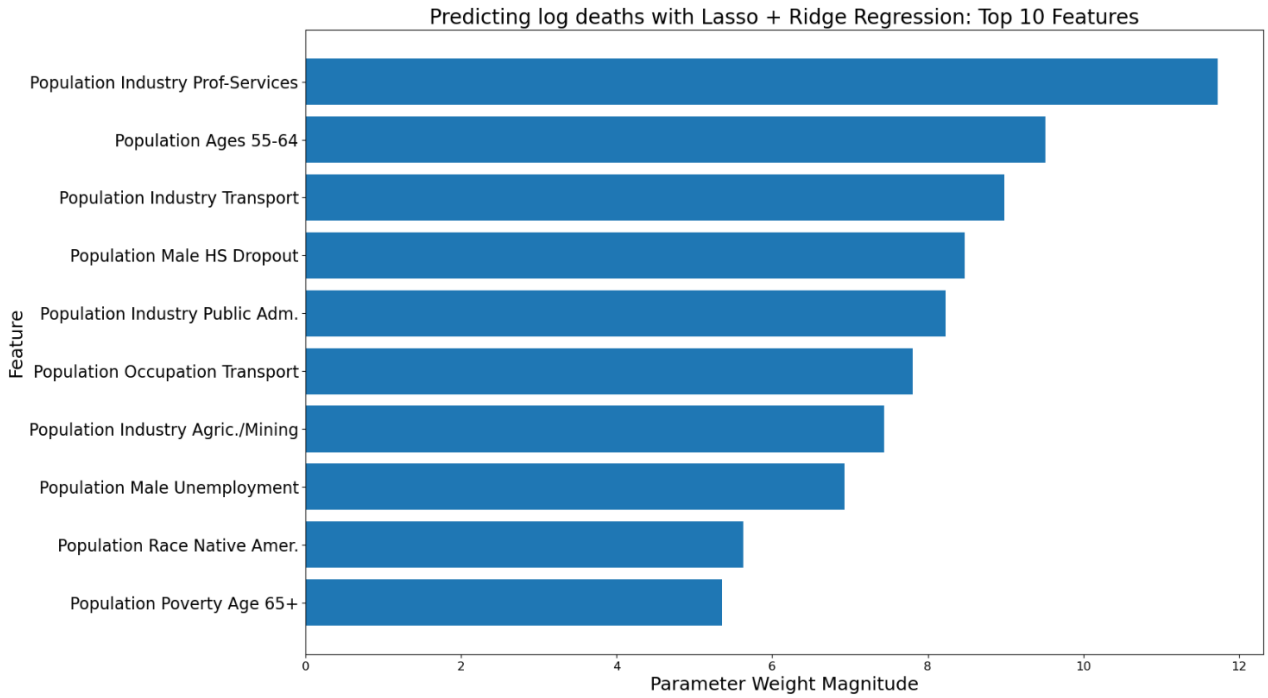

Fig. 2. Predicting log deaths with Lasso + Ridge Regression: Top 10 Features.

5. Conclusion

While there are now many predictive models aimed at understanding the role that co-morbidities play in explaining coronavirus outcomes, there is little research that explores the outcomes among Veterans and the specific role that social capital and other socio-economic local factors play as mediating forces. We estimate three predictive models for coronavirus cases and deaths at a healthcare system (HCS) level, aggregating across 1,297 VA medical facilities. We find that the combined LASSO and ridge with CV regression performs the best. Importantly, while HCS characteristics matter, socio-demographic characteristics also matter greatly and are more important than any of the hospital features. This suggests that public health interventions, especially towards vulnerable groups, must account for the role of an individual's environment and surrounding.

References

1. Amarasingham, R., Moore, B. J., Tabak, Y. P., Drazner, M. H., Clark, C. A., Zhang, S., Reed, W. G., Swanson, T. S., Ma, Y., and Halm, E. A. (2010). An Automated Model to Identify Heart Failure Patients at Risk for 30-day Readmission or Death Using Electronic Medical Record Data. Medical Care, 48(11):981–988.
2. Barrios, J. M., Benmelech, E., Hochberg, Y. V., and Zingales, L. (2020). Civic capital and social distancing during the covid-19 pandemic. NBER working paper.
3. Bejan, C. A., Angiolillo, J., Conway, D., Nash, R., Shirey-Rice, J. K., Lipworth, L., Cronin, R. M., Pulley, J., Kripalani, S., Barkin, S., Johnson, K. B., and Denny, J. C. (2018). Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. Journal of the American Medical Informatics Association, 25(1):61–71.
4. Britton, T., Ball, F., and Trapman, P. (2020). A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. Science, 23.
5. Ding, W., Levine, R., Lin, C., and Xie, W. (2020). Social Distancing and Social Capital: Why U.S. Counties Respond Differently to COVID-19. NBER working paper.
6. Dingel, J. I. and Neiman, B. (2020). How many jobs can be done at home. Journal of Public Economics.
7. Gordeev, V. S., and Egan, M. (2015). Social cohesion, neighbourhood resilience, and health: evidence from New Deal for Communities programme. Lanclet, 386(2): S39.
8. Hernandez-Orallo, J. (2013). ROC curves for regression. Pattern Recognition, 46.
9. Joint Economic Committee (JEC). (2018). “The geography of social capital in America.” SCP Report No. 1-18.
10. Kolak, M., Bhatt, J., Park, Y. H., Padron, N. A., and Molefe, A. (2020). Quantification of Neighborhood-Level Social Determinants of Health in the Continental United States. Journal of the American Medical Association Network Open, 3(1).
11. Makridis, C. A. and Wu, C. (2020). Ties that Bind (and Social Distance): How Social Capital Helps Communities Weather the COVID-19 Pandemic. SSRN working paper.
12. Makridis, C. A., Zhao, D., Bejan, A. C., and Alterovitz, G. (2020b). Leveraging Machine Learning to Characterize the Role of Socio-economic Determinants of Physical Health and Well-being Among Veterans. SSRN working paper.
13. Martin, C. A., Jenkins, D. R., Minhas, J. S., Gray, L. J., Tang, J., Williams, C., Sze, S., Pan, D., Jones, W., Verma, R., Knapp, S., Major, R., Davies, M., Brunskill, N., Wiselka, M., Brightling, C., Khunti, K., Haldar, P., and Pareek, M. (2020). Socio-demographic heterogeneity in the prevalence of COVID-19 during lockdown is associated with ethnicity and household size: Results from an observational cohort study. The Lanclet.
14. Navathe, A. S., Zhong, F., Lei, V. J., Chang, F. Y., Sordo, M., Topaz, M., Navathe, S. B., Rocha, R. A., and Zhou, L. (2018). Hospital readmission and social risk factors identified from physician notes. Health Services Research, 53(2):1110–1136.

15. Pan, D., Sze, S., Minhas, J. S., Bangash, M. N., Pareek, N., Divall, P., Williams, C. M., Oggioni, M. R., Squire, I. B., Nellums, L. B., Hanif, W., Khunti, K., and Pareek, M. (2020). The impact of ethnicity on clinical outcomes in COVID-19: A systematic review. Lanclet, 23(100404).
16. Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., and Northwell, C.-. R. C. (2020). Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. Journal of the American Medical Association, 323(20):2052–2059.
17. Schult, T. M., Schmunk, S. K., Marzolf, J. R., and Mohr, D. R. (2019). The Health Status of Veteran Employees Compared to Civilian Employees in Veterans Health Administration. Military Medicine, 184(7-8): e218–e224.
18. Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Damen, J. A. A., Debray, T. P. A., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., Navarro, C. L., Reitsma, J. B., Sergeant, J. C., Shi, C., Skoetz, N., Smits, L. J. M., Snell, K. I. E., Sperrin, M., Spijker, R., Steyerberg, E. W., Takada, T., van Kuijk, S. M. J., van Royen, F. S., Wallisch, C., Hooft, L., Moons, K. G. M., and van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. British Medical Journal, 369.
19. Zhou, F., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., and Cao, B. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lanclet, 395(10299):1054–1062.