Influence of social determinants of health and county vaccination rates on machine learning models to predict COVID-19 case growth in Tennessee

Lukasz S. Wylezinski, PhD[1-2,5], Coleman R. Harris, BS[1-3], Cody N. Heiser, BS[1-2,4], Jamieson D. Gray, BS[1-2], and Charles F. Spurlock, III, PhD[1-2,5-6†]

[1]Decode Health, Inc., Nashville, TN, 37203

[2]IQuity Labs, Inc., Nashville, TN 37203

[3]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, 37232

[4]Program in Chemical and Physical Biology, Vanderbilt University School of Medicine, Nashville, TN, 37232

[5]Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, 37232

[6]Wagner School of Public Health, New York University, New York, NY, 10012

† Address correspondence to CFS**:** Charles F. Spurlock, III, PhD, 111 10th Ave South, Suite 102, Nashville, TN, USA, chase@decodehealth.ai

Keywords: SARS-CoV-2, health equity, vaccines, social determinants of health, machine learning, demographics

Manuscript Word Count: 648

Pages: 8

Abstract

The SARS-CoV-2 (COVID-19) pandemic has exposed health disparities throughout the United States, particularly among racial and ethnic minorities. As a result, there is a need for data-driven approaches to pinpoint the unique constellation of clinical and social determinants of health (SDOH) risk factors that give rise to poor patient outcomes following infection in US communities.

We combined county-level COVID-19 testing data, COVID-19 vaccination rates, and SDOH information in Tennessee. Between February-May 2021, we trained machine learning models on a semi-monthly basis using these datasets to predict COVID-19 incidence in Tennessee counties. We then analyzed SDOH data features at each time point to rank the impact of each feature on model performance.

Our results indicate that COVID-19 vaccination rates play a crucial role in determining future COVID-19 disease risk. Beginning in mid-March 2021, higher vaccination rates significantly correlated with lower COVID-19 case growth predictions. Further, as the relative importance of COVID-19 vaccination data features grew, demographic SDOH features such as age, race, and ethnicity decreased while the impact of socioeconomic and environmental factors, including access to healthcare and transportation, increased.

Incorporating a data framework to track the evolving patterns of community-level SDOH risk factors could provide policymakers with additional data resources to improve health equity and resilience to future public health emergencies.

Introduction

The SARS-CoV-2 (COVID-19) pandemic exacerbated health inequities throughout the United States disproportionately affecting at-risk populations.[1] Identifying social determinants of health (SDOH) risk factors within US communities that contribute to poor outcomes following infection can improve health equity and strengthen community readiness for future public health emergencies.[2, 3] Following vaccine rollouts in 2021, we predicted Tennessee COVID-19 case growth using machine learning models and investigated the influence of SDOH factors on COVID-19 incidence to quantify and track opportunities to improve health equity.

Methods

Our approach combined publicly available COVID-19 testing, vaccination, hospitalization, and death metrics with county-specific SDOH and demographic data.[4, 5] We employed feature engineering and selection to identify novel county-level predictors such as offset case counts, summed and averaged case growth statistics, and days since the k^{th} COVID-19 case to best capture trends in Tennessee county COVID-19 incidence between February and May 2021. We aggregated data from multiple sources, including the Tennessee Department of Health, John's Hopkins Coronavirus Research Center, and the U.S. Census database to minimize any implicit bias, and removed or ignored missing values depending on the model type. We trained and tested multiple machine learning models using a grid search approach, including neural networks and tree-based methods, with four to six weeks of historical COVID-19 case data to

generate COVID-19 case predictions at thirteen timepoints. We selected optimal models using cross-validation and holdout metrics (e.g., Tweedie deviance, mean absolute error, R^2).[6] We analyzed the impact of all features using permutation importance to quantify and rank particular SDOH by their relative influence on COVID-19 case growth predictions.[7] Finally, we generated linear regression models of county COVID-19 case growth using vaccination rates as regressors to calculate correlation coefficients that quantify the association between local vaccination rates and COVID-19 incidence.

Results

Machine learning models across all timepoints were more than 90% accurate when comparing model predictions to actual cases (Supplementary Figure 1A & C). The top models demonstrated an average R^2 value of 0.99, mean absolute error of 0.21, and 0.001 mean Tweedie deviance (Supplementary Figure 1B).

Highly predictive SDOH features changed in importance over time. Categorically, demographic SDOH were most important in February 2021, but socioeconomic and environmental SDOH became increasingly more influential towards May. Health outcome SDOH features remained largely consistent during the study period. Individually, the female and under-18 age demographic features ranked highest in February and then declined while African American poverty and health infrastructure features, such as the number of hospital beds and community provider access statistics, increased in importance by mid-April. Lastly, COVID-19 vaccination data features grew in relative importance by May compared to the other SDOH factors (Figure 1).

As Tennessee vaccination rates increased, counties with the lowest vaccination rates exhibited the highest COVID-19 case growth (Supplemental Figure 2A). Initially, vaccination rates were not correlated with COVID-19 risk, but by mid-March, a statistically significant correlation with low risk of COVID-19 case growth emerged (Supplemental Figure 2B).

Discussion

Efforts to curtail the health and economic impact of the SARS-CoV-2 pandemic illuminate the need to define specific risk factors that catalyze future case growth, worsen health disparities, and adversely impact the public health response across US communities. Addressing these challenges, we constructed a real-time predictive framework to discover and rank county-level SDOH risk factors that drive machine learning predictions of future COVID-19 case growth (Figure 1).

In Tennessee, we found that communities with rapid vaccine rollout were at lower risk for case growth (Supplemental Figure 2). As vaccination levels began to rise, demographic SDOH features such as age, race and ethnicity declined in relative importance while socioeconomic and environmental risk factors such as poverty, access to transportation and healthcare infrastructure increased significantly. Measures promoting health equity rely on constant assessment of risk mitigation effectiveness. Real-time knowledge of community specific SDOH risk factors empowers healthcare organizations and local governments to improve policy and resource allocation to mitigate outbreaks and enhance resilience to future public health threats.

Acknowledgements

This work was supported by Decode Health, Inc., IQuity Labs, Inc., and grants from the National Institutes of Health (AI124766, AI129147 and AI145505). Dr. Spurlock had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Dr. Spurlock devised the concept and study design. All authors took part in acquisition, analysis and interpretation of the data along with drafting and revising the manuscript.

Conflict of interest statement

Authors Wylezinski, Gray and Spurlock are shareholders in IQuity Labs, Inc. (Nashville, TN) and Decode Health, Inc. (Nashville, TN). IQuity Labs develops blood-based tools using RNA to aid in the diagnosis and treatment of human disease. Decode Health develops artificial intelligence approaches to predict chronic and infectious disease risk in patient populations.

References

1. Alberti PM, Lantz PM, Wilkins CH. Equitable Pandemic Preparedness and Rapid Response: Lessons from COVID-19 for Pandemic Health Equity. *J Health Polit Policy Law*. 12 2020;45(6):921-935. doi:10.1215/03616878-8641469

2. Paremoer L, Nandi S, Serag H, Baum F. Covid-19 pandemic and the social determinants of health. *BMJ*. 01 2021;372:n129. doi:10.1136/bmj.n129

3. Seligman B, Ferranna M, Bloom DE. Social determinants of mortality from COVID-19: A simulation study using NHANES. *PLoS Med*. 01 2021;18(1):e1003490. doi:10.1371/journal.pmed.1003490

4. Johns Hopkins University Coronavirus Resource Center. COVID-19 United States Cases by County. https://coronavirus.jhu.edu/us-map. Accessed February 1, 2021,

5. Vest JR, Ben-Assuli O. Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. *Int J Med Inform*. 09 2019;129:205-210. doi:10.1016/j.ijmedinf.2019.06.013

6. Muhlestein WE, Akagi DS, Chotai S, Chambless LB. The impact of presurgical comorbidities on discharge disposition and length of hospitalization following craniotomy for brain tumor. *Surg Neurol Int*. 2017;8:220. doi:10.4103/sni.sni_54_17

7. Breiman L. *Random Forests*. vol 45. Machine Learning. Kluwer Academic Publishers; 2001.

Figure Legends

Figure 1. Social determinants of health (SDOH) linked to COVID-19 case growth in Tennessee dynamically shift in importance over time. SDOH include social, physical and environmental factors that impact community health such as age, race, gender, access to transportation, access to primary care and community vaccination rates. Twelve of these SDOH features demonstrated the highest feature importance across all predictive models during the study period. Size and color are used to emphasize SDOH feature importance at each timepoint. Large, red (●) bubbles connote the top ranked SDOH feature while small dark blue (•) bubbles signify least importance of a given feature at each timepoint. Black bubbles (•) represent the least important feature at each time point compared to the other top ranked SDOH data elements.

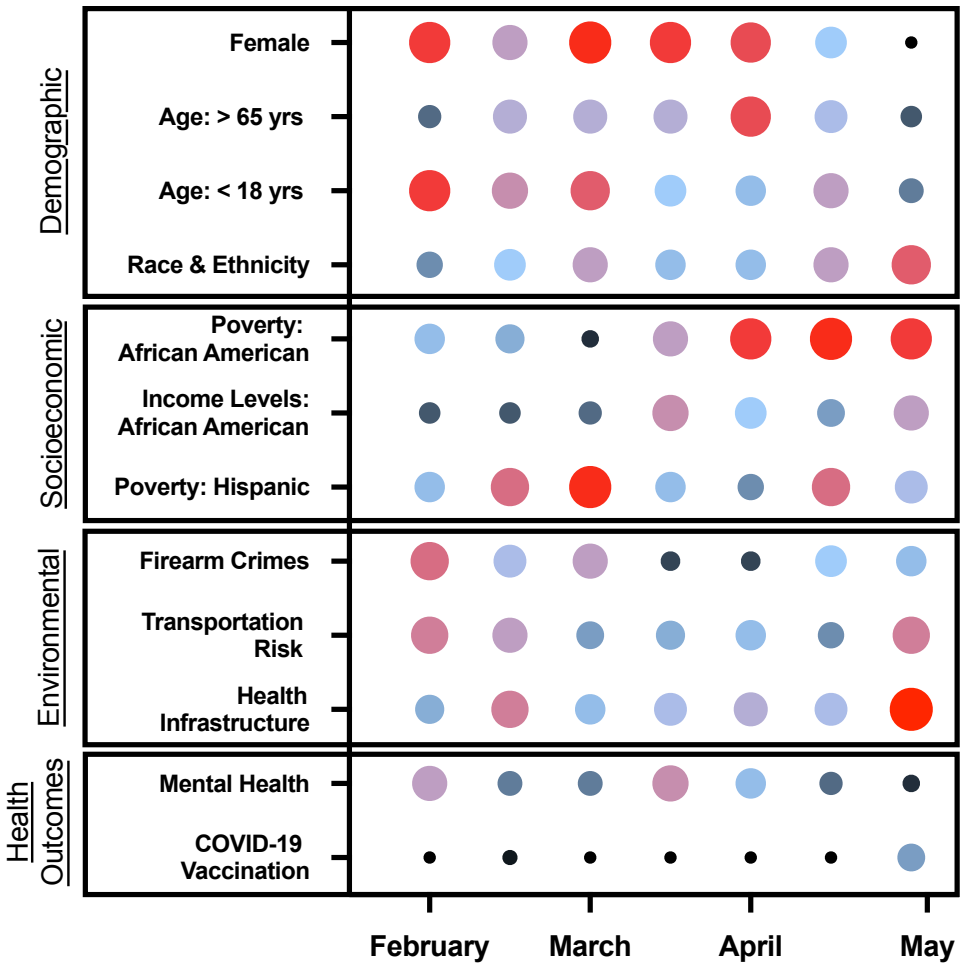

Figure 1: Social determinants of health (SDOH) linked to COVID-19 case growth in Tennessee dynamically shift in importance over time. SDOH include social, physical and environmental factors that impact community health such as age, race, gender, access to transportation, access to primary care and community vaccination rates. Twelve of these SDOH features demonstrated the highest feature importance across all predictive models during the study period. Size and color are used to emphasize SDOH feature importance at each timepoint. Large, red (●) bubbles connote the top ranked SDOH feature while small dark blue (•) bubbles signify least importance of a given feature at each timepoint. Black bubbles (•) represent the least important feature at each time point compared to the other top ranked SDOH data elements.