BMJ Health & Care Informatics

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,[1,2] Coleman R Harris,[1,3] Cody N Heiser,[1,4] Jamieson D Gray,[1]
Charles F Spurlock [1,2,5]

To cite: Wylezinski LS, Harris CR, Heiser CN, *et al*. Influence of social determinants of health and county vaccination rates on machine learning models to predict COVID-19 case growth in Tennessee. *BMJ Health Care Inform* 2021;**28**:e100439. doi:10.1136/bmjhci-2021-100439

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100439).

Received 01 July 2021
Accepted 06 September 2021

[1]Decode Health, Inc. and IQuity Labs, Inc, Nashville, Tennessee, USA
[2]Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[3]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[4]Program in Chemical and Physical Biology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[5]Wagner School of Public Health, New York University, New York, New York, USA

Correspondence to
Dr Charles F Spurlock;
chase.spurlock@vanderbilt.edu

ABSTRACT

Introduction The SARS-CoV-2 (COVID-19) pandemic has exposed health disparities throughout the USA, 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.

Methods We combined county-level COVID-19 testing data, COVID-19 vaccination rates and SDOH information in Tennessee. Between February and May 2021, we trained machine learning models on a semimonthly 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.

Results 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.

Conclusion Incorporating a data framework to track the evolving patterns of community-level SDOH risk factors could provide policy-makers 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 USA, 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 roll-outs 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] Data sources included the Tennessee Department of Health, Johns Hopkins Coronavirus Research Center and the US Census database. We employed feature engineering and feature selection to identify novel predictors such as offset case counts to best represent changes in Tennessee county COVID-19 incidence between February and May 2021. We aggregated data from multiple sources to minimize implicit bias and removed or ignored missing values depending on the model type. An ensemble of generalized linear and tree-based machine learning models was built in parallel, each trained and tested with 4–6 weeks of historical COVID-19 case data to generate predictions from 40 to 50 models at 13 time points. Optimal models were selected using cross-validation metrics (eg, mean absolute error, R^2) and prediction accuracy for future relative case growth normalized to county population.[6] We analyzed the impact of all features from top performing models to quantify and rank SDOH by their influence on COVID-19 incidence predictions. Finally, we calculated Pearson coefficients to quantify associations between vaccination rates and county COVID-19 case growth over time.

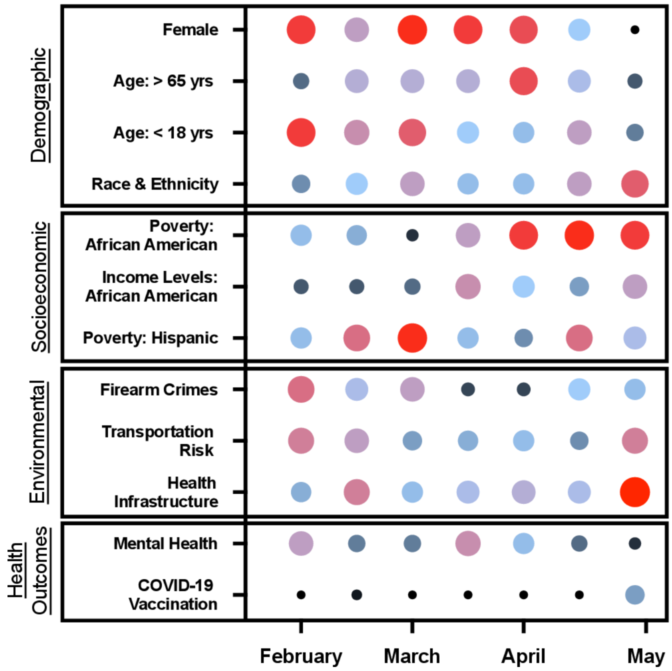

Figure 1 Social determinants of health (SDOH) linked to COVID-19 case growth in Tennessee dynamically shift in importance over time. SDOH include social, physical and environmental factors that impact community health such as age, race, gender, access to transportation, access to primary care and community 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 time point. large, red (●) bubbles connote the top ranked SDOH feature while small dark blue (●) bubbles signify least importance of a given feature at each time point. Black bubbles (●) represent the least important feature at each time point compared with the other top ranked SDOH data elements.

RESULTS

Machine learning models across all time points were more than 90% accurate when comparing model predictions to actual cases (online supplemental figure 1A and 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 (online supplemental 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 with the other SDOH factors (figure 1).

As Tennessee vaccination rates increased, counties with the lowest vaccination rates exhibited the highest COVID-19 case growth (online 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 (online 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.[7] 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 incidence (figure 1).

In Tennessee, we found that communities with rapid vaccine roll-out were at lower risk for case growth (online 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, enhance resilience to future public health threats, and capture evolving risk profiles as novel virus variants emerge.[8]

Twitter Coleman R Harris @colemanrharris, Cody N Heiser @cody_heiser, Jamieson D Gray @jamiesongray and Charles F Spurlock @cfspurlock

Contributors CFS 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. CFS 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.

Funding This work was supported by Decode Health, IQuity Labs and grants from the National Institutes of Health (AI124766, AI129147 and AI145505). CFS 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. CFS 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.

Competing interests LSW, JDG and CFS are shareholders in IQuity Labs (Nashville, Tennessee, USA) and Decode Health (Nashville, Tennessee, USA). IQuity Labs develops blood-based RNA tools 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.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; internally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines,

terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Charles F Spurlock http://orcid.org/0000-0001-9015-6321

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* 2020;45:921–35.

2 Paremoer L, Nandi S, Serag H, *et al*. Covid-19 pandemic and the social determinants of health. *BMJ* 2021;372:n129.

3 Seligman B, Ferranna M, Bloom DE. Social determinants of mortality from COVID-19: a simulation study using NHANES. *PLoS Med* 2021;18:e1003490.

4 Johns Hopkins University Coronavirus Resource Center. COVID-19 United States cases by County. Available: https://coronavirus.jhu.edu/us-map [Accessed 1 Feb 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* 2019;129:205–10.

6 Muhlestein WE, Akagi DS, Chotai S, *et al*. The impact of presurgical comorbidities on discharge disposition and length of hospitalization following craniotomy for brain tumor. *Surg Neurol Int* 2017;8:220.

7 Karmakar M, Lantz PM, Tipirneni R. Association of social and demographic factors with COVID-19 incidence and death rates in the US. *JAMA Netw Open* 2021;4:e2036462.

8 Shadmi E, Chen Y, Dourado I, *et al*. Health equity and COVID-19: global perspectives. *Int J Equity Health* 2020;19:104.