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 Spurlock1,2,5

ABSTRACT

Introduction The SARS-CoV-2 (COVID-19) pandemic has exposed health inequities 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.
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. 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.

**Twitter** Coleman R Harris @colemanharris, Cody N Heiser @codyheiser, 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 otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

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

**REFERENCES**
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.
Supplementary Figure 1. Machine learning models that incorporate historical COVID-19 case growth statistics and SDOH data accurately predict future COVID-19 case growth in Tennessee. (A) Model accuracy at each time point for predicting counties that will experience the highest COVID-19 case growth normalized to population. We set future population-normalized COVID-19 case growth as the target for predictive modeling and performed a grid search of generalized linear and tree-based machine learning models. Case growth predictions were compared to actual case numbers to determine accuracy. (B) Cross-validation metrics of top models at each time point. Shown are $R^2$ values, mean absolute error, and Tweedie deviance. (C) Representative illustration of counties predicted (■) or not predicted (□) for future highest case growth versus those that recorded the highest case growth normalized to population at the predicted timepoint. The top third of Tennessee counties with the highest case growth are depicted (■).
Supplementary Figure 2. Tennessee county vaccination rates correlate with risk for future COVID-19 case growth. (A) Counties at each time point with the highest population normalized vaccination rate and lowest vaccination rate are plotted as a proportion of the top third of Tennessee counties with the highest predicted COVID-19 case growth. (B) Depiction of the line of best fit for each time point (colored from February to May 2021) using each county’s vaccination rate as a regressor for county COVID-19 case count rankings. The slope represents the correlation between county COVID-19 cases and county vaccination rates. The corresponding intercepts depict changes in vaccination rate and COVID-19 cases over time. By the fourth timepoint, vaccination rate is significantly correlated with decreasing COVID-19 case growth. This significance was measured by an ANOVA test performed on the coefficient for vaccination rate. *p-values less than 0.05 were considered significant: * = p < 0.05; ** = p < 0.01. (C) Heatmap illustration of the top third of Tennessee counties with the highest predicted COVID-19 case growth at the final timepoint overlayed with county vaccination rates. Tennessee counties with the highest COVID-19 case growth risk are enriched for counties with lower overall vaccination rates.