BMJ Health & Care Informatics

Predictive modeling of COVID-19 case growth highlights evolving racial and ethnic risk factors in Tennessee and Georgia

Jamieson D Gray,[1] Coleman R Harris,[1,2] Lukasz S Wylezinski,[1,3] Charles F Spurlock, III [1,3]

To cite: Gray JD, Harris CR, Wylezinski LS, *et al*. Predictive modeling of COVID-19 case growth highlights evolving racial and ethnic risk factors in Tennessee and Georgia. *BMJ Health Care Inform* 2021;**28**:e100349. doi:10.1136/bmjhci-2021-100349

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100349).

Received 25 February 2021
Accepted 19 July 2021

[1]Decode Health, Inc. and IQuity Labs, Inc, Nashville, Tennessee, USA
[2]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[3]Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

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

ABSTRACT

Introduction The SARS-CoV-2 (COVID-19) pandemic has exposed the need to understand the risk drivers that contribute to uneven morbidity and mortality in US communities. Addressing the community-specific social determinants of health (SDOH) that correlate with spread of SARS-CoV-2 provides an opportunity for targeted public health intervention to promote greater resilience to viral respiratory infections.

Methods Our work combined publicly available COVID-19 statistics with county-level SDOH information. Machine learning models were trained to predict COVID-19 case growth and understand the social, physical and environmental risk factors associated with higher rates of SARS-CoV-2 infection in Tennessee and Georgia counties. Model accuracy was assessed comparing predicted case counts to actual positive case counts in each county.

Results The predictive models achieved a mean R^2 of 0.998 in both states with accuracy above 90% for all time points examined. Using these models, we tracked the importance of SDOH data features over time to uncover the specific racial demographic characteristics strongly associated with COVID-19 incidence in Tennessee and Georgia counties. Our results point to dynamic racial trends in both states over time and varying, localized patterns of risk among counties within the same state. For example, we find that African American and Asian racial demographics present comparable, and contrasting, patterns of risk depending on locality.

Conclusion The dichotomy of demographic trends presented here emphasizes the importance of understanding the unique factors that influence COVID-19 incidence. Identifying these specific risk factors tied to COVID-19 case growth can help stakeholders target regional interventions to mitigate the burden of future outbreaks.

INTRODUCTION

In January 2021, Tennessee and Georgia reported over 1,637,000 cases and 23,848 deaths due to COVID-19. Hispanic individuals comprise 14% of the states' population but represent 25% of confirmed cases, suggesting race and ethnicity are associated with case growth.[1] To explore this association, we sought to combine publicly available COVID-19 data and proprietary social determinants of health (SDOH), which measure certain physical, social, economic and demographic characteristics, to build and tune machine learning models predicting COVID-19 incidence in Tennessee and Georgia. Our objective was to accurately predict COVID-19 case growth, while investigating model-based relevance of racial demographic features influencing these predictions over time. We hypothesised that SDOH features significantly influence COVID-19 incidence and that underlying risk patterns associated with county-level race and ethnicity data features influence prediction accuracy.

METHODS

Our approach combined publicly available COVID-19 case, hospitalization and death metrics with county-specific SDOH data.[2 3] Information sources for the study included the State of Tennessee Department of Health, State of Georgia Department of Health, the Johns Hopkins Coronavirus Research Center and the US Census database. Feature engineering and feature selection, including interaction terms, were employed to define the data inputs that best represent changes in COVID-19 incidence over time. We developed novel features that included offset and normalised case growth as well as dynamic time window features derived from state health department data from July 2020 to January 2021. Time independent enrichment data, including both quantitative and qualitative SDOH with demographic information, were joined into this feature set to generate county-specific data.[4] Data were aggregated

from multiple sources to minimize the impact of any implicit bias and missing values removed or ignored depending on model type. The outcome for predictive modelling was defined as the future relative case growth normalised to the population in Tennessee and Georgia counties. We performed a grid search of generalised linear and tree-based machine learning models, training and testing each model with 4–6 weeks of historical COVID-19 case data to generate predictions using the most recent data available. From the ~50 regression models that we built for each time point, models were chosen using cross-validation model metrics (eg, R^2, Tweedie deviance) and prediction accuracy for COVID-19 case growth.[5] We identified the top third of Tennessee and Georgia counties

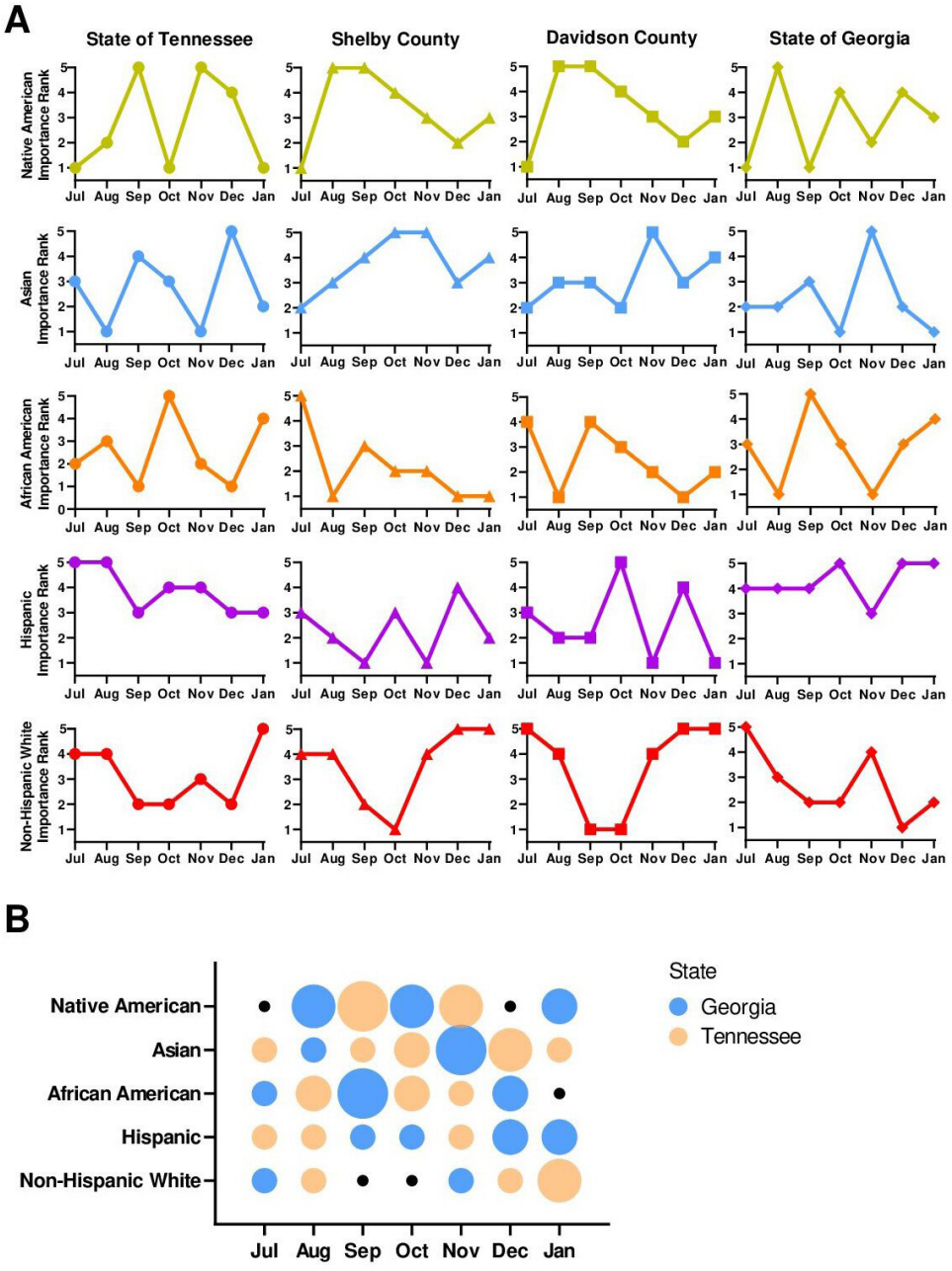

Figure 1 Influence of racial demographic features linked to COVID-19 case growth exhibit dynamic shifts over time in Tennessee and Georgia. (A) Relative rank of demographic feature importance across top predictive models is reported for the entire state of Tennessee (●) and the two most populous counties in Tennessee, Shelby County (▲) and Davidson County (■) as well as the state of Georgia (◆). A score of 5 on the importance rank indicates the most important demographic feature relative to the other four demographic features. Groups include Native American (●), Asian (●), African American (●), Hispanic (●) and non-Hispanic white (●). (B) Differences in the rank of demographic feature importance in Tennessee and Georgia over time. The colour of the bubble (Tennessee (●); Georgia (●)) indicates the state that exhibited a higher importance rank of the specific demographic feature for predicting COVID-19 case growth. Black dots (●) designate months where the two states displayed the same importance rank for an individual demographic feature. The size of the bubbles shows the difference in importance of each demographic feature between the two states. Larger bubbles connote greater difference in importance.

at highest risk for case growth and assessed our prediction accuracy versus actual case growth over time. Finally, we analysed each data feature's impact at the state and county level to track the rank order of demographic data features that drove COVID-19 case growth at each time point.

RESULTS

Candidate models for Tennessee and Georgia achieved excellent metrics across all time points including a mean R^2 value of 0.998 (Tennessee and Georgia), mean Tweedie deviance of 0.003 (Tennessee) and 0.002 (Georgia) as well as a mean absolute error of 0.357 (Tennessee) and 0.337 (Georgia) (online supplemental figure 1A). Prediction accuracy was >90% in all models across both states when compared with actual case growth (online supplemental figure 1B).

Racial demographics produced variable trends at both the state and county levels. The two most populous counties in Tennessee, Shelby and Davidson, revealed an identical pattern of importance for Native American racial demographics in determining future case growth while exhibiting differences among Asians. Shelby County displayed a gradual increase in importance in the Asian demographic, while Davidson County saw a more pronounced spike between October and November. Comparing racial demographic importance at the Tennessee state level versus individual counties yields similar patterns (non-Hispanic white) as well as contrasting trends (African American). Further, the Hispanic ethnicity risk factor trends across Tennessee differed from the individual Tennessee counties' more acute fluctuation of Hispanic risk factor importance (figure 1A).

Additionally, similarities and differences in racial demographic trends extend across state borders. While Hispanic ethnicity displayed the most meaningful importance in Tennessee during July and August, Georgia saw a similar increase in importance starting in September. Comparison of the two states' top racial demographic drivers showed a potential macropattern in which the most important driver for one state often preceded its rise to top importance in the other (figure 1A,B).

DISCUSSION

Our study developed highly accurate modeling to predict COVID-19 case growth and discover associations between state and county SDOH characteristics connected to risk for future spread of infection. Analysis of the most influential racial and ethnic demographic data at each time point discovered localized, evolving patterns of risk that correlate with state-level and county-level SARS-CoV-2 case growth. These patterns can shift dramatically month to month, increasing or decreasing over time and vary by geography, even among similarly sized counties within a state or between two neighbouring states. The state-specific and county-specific modelling results we describe for Tennessee and Georgia may bias or limit the validity of extrapolating our specific modelling results to other localities. However, the approach is extensible to all US states and counties.

Early identification of the specific SDOH risk drivers tied to disease outcomes in a pandemic could help decision-makers promote health equity and deliver targeted interventions to mitigate disease risk in vulnerable populations. Closing the loop to address certain SDOH risk factors also enhances community resilience to future viral respiratory infections.[6]

Applications of this approach extend beyond acute respiratory infection to chronic disease outcomes. A growing percentage (approximately 10%) of patients infected with SARS-CoV-2 develop long COVID.[7] These patients experience prolonged, debilitating symptoms months after infection and emergence or exacerbation of chronic illness. Targeted approaches to mitigate spread of disease can lessen future acute and chronic disease burden.

Twitter Jamieson D Gray @jamiesongray, Coleman R Harris @colemanrharris and Charles F Spurlock, III @cfspurlock

Acknowledgements The authors would like to acknowledge Cody Heiser for providing a constructive review of the manuscript.

Contributors CFS had full access to all of the data in the study, devised the concept and study design and takes responsibility for the integrity of the data and the accuracy of the data analysis. 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, Inc., IQuity Labs, Inc., and grants from the National Institutes of Health (AI124766, AI129147 and AI145505).

Competing interests JDG, LSW and CFS are shareholders in IQuity Labs, Inc. (Nashville, Tennessee) and Decode Health, Inc. (Nashville, Tennessee). IQuity Labs, Inc. develops blood-based RNA tools to aid in the diagnosis and treatment of human disease. Decode Health, Inc. develops artificial intelligence approaches to predict chronic and infectious disease risk in patient populations.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

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

REFERENCES

1 The COVID tracking project. racial data Dashboard. Available: https://covidtracking.com/race/dashboard [Accessed 25 Nov 2020].
2 Johns Hopkins University Coronavirus Resource Center. COVID-19 United States cases by County. Available: https://coronavirus.jhu.edu/us-map [Accessed 25 Nov 2020].

3 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.
4 Kolak M, Bhatt J, Park YH, *et al*. Quantification of neighborhood-level social determinants of health in the continental United States. *JAMA Netw Open* 2020;3:e1919928.
5 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.
6 Graves E, Weiss A, Rickles M, *et al*. Community resiliency to COVID-19 in a subset of US communities. Available: https://about.sharecare.com/wp-content/uploads/2020/08/National_COVID-whitepaper_PROOF_08.28.20.pdf [Accessed 25 Nov 2020].
7 Greenhalgh T, Knight M, A'Court C, *et al*. Management of post-acute covid-19 in primary care. *BMJ* 2020;370:m3026.