

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

3 Lukasz S. Wylezinski, PhD^{1-2,5}, Coleman R. Harris, BS¹⁻³, Cody N. Heiser, BS^{1-2,4},
4 Jamieson D. Gray, BS¹⁻², and Charles F. Spurlock, III, PhD^{1-2,5-6†}

5 ¹Decode Health, Inc., Nashville, TN, 37203

6 ²iQuity Labs, Inc., Nashville, TN 37203

7 ³Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN,
8 37232

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

11 ⁵Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN,
12 37232

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

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

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

18 **Manuscript Word Count:** 648

19 **Pages:** 8

20 **Abstract**

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

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

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

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

41

42 **Introduction**

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

51

52 **Methods**

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

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

71

72 **Results**

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

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

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

91

92 **Discussion**

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

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

108 **Acknowledgements**

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

115

116 **Conflict of interest statement**

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

122

123

124

125

126

127

128 **References**

- 129 1. Alberti PM, Lantz PM, Wilkins CH. Equitable Pandemic Preparedness and Rapid Response:
130 Lessons from COVID-19 for Pandemic Health Equity. *J Health Polit Policy Law*. 12 2020;45(6):921-935.
131 doi:10.1215/03616878-8641469
- 132 2. Paremoer L, Nandi S, Serag H, Baum F. Covid-19 pandemic and the social determinants of
133 health. *BMJ*. 01 2021;372:n129. doi:10.1136/bmj.n129
- 134 3. Seligman B, Ferranna M, Bloom DE. Social determinants of mortality from COVID-19: A
135 simulation study using NHANES. *PLoS Med*. 01 2021;18(1):e1003490.
136 doi:10.1371/journal.pmed.1003490
- 137 4. Johns Hopkins University Coronavirus Resource Center. COVID-19 United States Cases by
138 County. <https://coronavirus.jhu.edu/us-map>. Accessed February 1, 2021,
- 139 5. Vest JR, Ben-Assuli O. Prediction of emergency department revisits using area-level social
140 determinants of health measures and health information exchange information. *Int J Med Inform*. 09
141 2019;129:205-210. doi:10.1016/j.ijmedinf.2019.06.013
- 142 6. Muhlestein WE, Akagi DS, Chotai S, Chambless LB. The impact of presurgical comorbidities on
143 discharge disposition and length of hospitalization following craniotomy for brain tumor. *Surg Neurol*
144 *Int*. 2017;8:220. doi:10.4103/sni.sni_54_17
- 145 7. Breiman L. *Random Forests*. vol 45. Machine Learning. Kluwer Academic Publishers; 2001.

146
147
148

149 **Figure Legends**

150 **Figure 1. Social determinants of health (SDOH) linked to COVID-19 case growth in**

151 **Tennessee dynamically shift in importance over time.** SDOH include social,

152 physical and environmental factors that impact community health such as age, race,

153 gender, access to transportation, access to primary care and community vaccination

154 rates. Twelve of these SDOH features demonstrated the highest feature importance

155 across all predictive models during the study period. Size and color are used to

156 emphasize SDOH feature importance at each timepoint. Large, red (●) bubbles

157 connote the top ranked SDOH feature while small dark blue (●) bubbles signify least

158 importance of a given feature at each timepoint. Black bubbles (●) represent the least

159 important feature at each time point compared to the other top ranked SDOH data

160 elements.

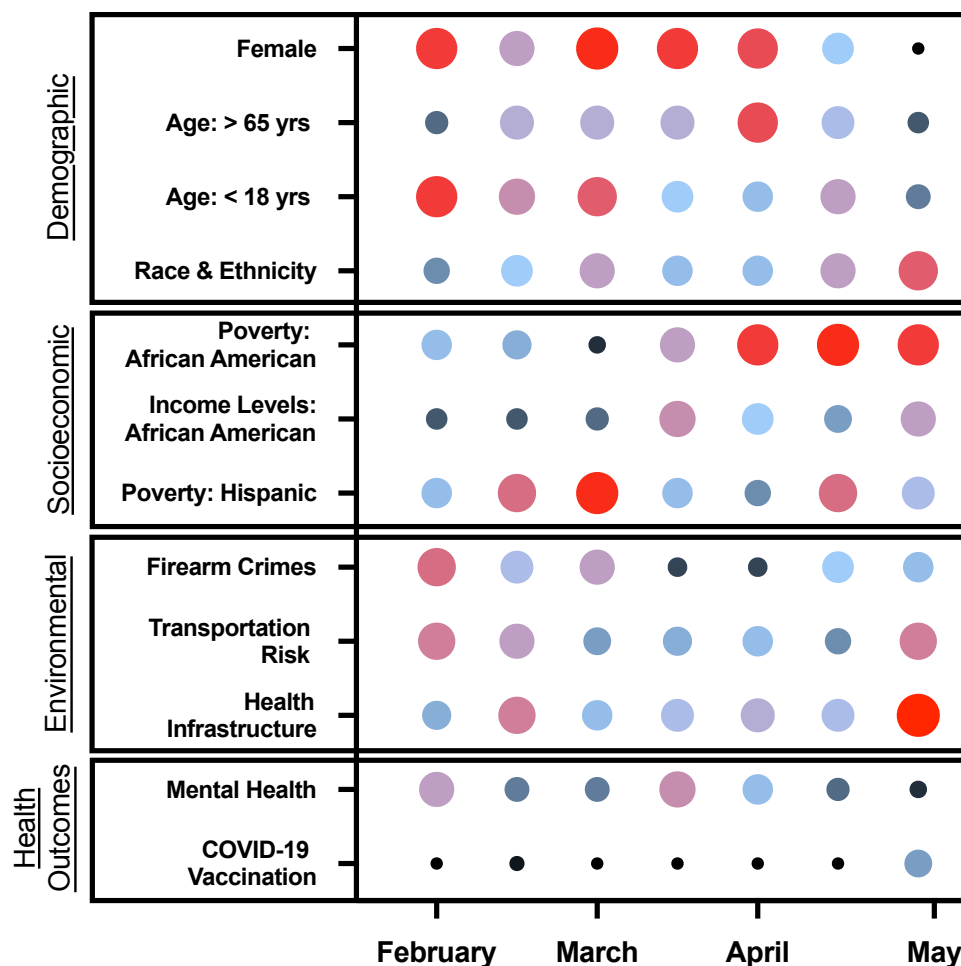


Figure 1: Social determinants of health (SDOH) linked to COVID-19 case growth in Tennessee dynamically shift in importance over time. SDOH include social, physical and environmental factors that impact community health such as age, race, gender, access to transportation, access to primary care and community 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.