2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA)

Social Determinates of Health and COVID-19 Mortality Rates at the County level

Sophia Lam, Elizabeth Leeds Hohman, Viveca Pavon-Harr, Jesse Patsolic, Collin Schwantes, Marjorie Willner, Katherine Schulz, Trevor Kent, Kevin Kiernan, Ian McCulloh
Applied Intelligence
Accenture
Washington DC, United States
Email: {sophia.s.lam, e.leeds.hohman, viveca.pavon-harr, jesse.l.patsolic, collin.j.schwantes, marjorie.willner, katherine.schulz, trevor.g.kent, kevin.kiernan, ian.mcculloh}@accenturefederal.com

Abstract*—Understanding how underlying health conditions and social determinants of health affect the severity of COVID-19 is critical for community response planning. Literature reports that groups at higher risk from COVID-19 include those 65 and older, living in nursing homes and long-term care facilities, and with severe obesity, diabetes, chronic lung disease, or asthma. In addition, other studies have shown that the disease disproportionately affects individuals with lower socio-economic status. Our research seeks to validate these findings and observe the effects of health measures and social determinants of health on COVID-19 mortality at the county-level. In addition to COVID-19 research from hospital population samples, public health officials can leverage county-level factors for novel disease mitigation. We use the Johns Hopkins University COVID-19 reports of confirmed cases and deaths to measure disease mortality for each county in the United States. Then, we compare mortality to multiple county social determinants of health such as age, obesity, diabetes, and smoking in hypothesis testing. We fit multivariate linear models as well as non-linear models to predict mortality as a function of these county measures. The analysis shows that there is little evidence of a relationship between the county health measures of obesity, diabetes, or smoking and COVID-19 mortality as of the date of this publication. However, the analysis does reveal a positive relationship between the percent of a county population that is 65 or older and COVID-19 mortality. Other factors such as overcrowding, the percent uninsured, and the length of time since the virus has been detected in the county are also correlated with county COVID-19 mortality. Potential reasons for these findings, including data quality, are discussed. We also emphasize the advantage of collecting high quality, detailed health data at the county-level and explain how such data could be used to understand factors affecting the outcomes from novel diseases in real-time, as a disease is progressing.

***Index Terms*—COVID-19, SARS-CoV2, mortality, death, demographics, social determinants of health, regression**

I. Introduction

The novel coronavirus disease (COVID-19) can present with symptoms ranging from mild to severe, in some cases causing respiratory disease leading to death. The outcomes of the disease are heavily influenced by social determinants of health and preexisting health conditions. The World Health Organization defines social determinants of health as conditions in which people are born, live, and age in and are mostly responsible for dissimilar health effects. The first reports of the disease in China had mortality estimates near 2.3%, but Italy had a mortality rate near 7.2%, likely due to the older age distribution of the population [1]. Early reports indicated that the disease manifested in more severe symptoms and higher mortality rates in older patients as well as those with underlying health conditions, especially respiratory conditions [1], [2], [3]. Sources also indicated that those with obesity and diabetes experienced more severe symptoms [4], [5], [6]. The Center for Disease Control (CDC) states that groups at higher risk to COVID-19 include those 65 years and older, those living in nursing homes and long-term care facilities, those with severe obesity, diabetes, chronic lung disease, or asthma [4]. As with other diseases, socio-economic inequities have led to more severe outcomes for minorities in the United States. Several studies found that COVID-19 disproportionately hurt the African American and immigrant communities [7], [8], [9].

Data limitations in the publications may introduce bias in the results. For example, researchers may resort to using convenience sampling, such as hospital populations, which are biased towards sicker individuals. Therefore, these results may not be representative of the general population. In order to overcome the limitation, we ask the following research questions:

RQ1: Do we observe the effects of smoking, obesity, and diabetes on COVID-19 mortality at the county-level and state-level?

RQ2: Do we observe the effects of age on COVID-19 mortality at the county-level and state-level?

RQ3: Do we observe the effects of social determinants of health, such as socio-economic measures, on COVID-19 mortality at the county-level and state-level?

Understanding how underlying social and health measures affect the disease is important for planning at the hospital and community level. For novel diseases, the ability to infer these contributing risk factors as quickly as possible can lead to more informed preventative and treatment strategies. Decision makers can use previously collected county-level demographics and health measures to develop guidance rather than only relying on new research that emerges as the disease progresses. This paper discusses the importance of the availability of high-quality, detailed health outcomes at the county level and explains how such data can be informative without jeopardizing patient privacy.

II. Data

To test for relationships between county-level demographics and COVID-19 risk, we calculate the risk measure of mortality at the county level as:

$$Mortality = \frac{\text{Number of COVID-19 deaths}}{\text{Number of COVID-19 positives}}, \tag{1}$$

We collect the county totals of confirmed deaths and confirmed positive cases from the Johns Hopkins University Center for Systems Science and Engineering (JHU) on June 10, 2020 [10].

This mortality measure has the advantage of being scaled by county population, which avoids discovering patterns due to variations in county sizes. However, variations in testing capacity across counties leads to variations in the estimates of positive cases. The effects of this are discussed in the Conclusions section.

Counties with less than 100 confirmed positive cases or zero deaths were excluded from the analysis. Excluding counties with low confirmed positives removes counties with overly inflated mortality rates. The resulting data set consists of 1021 counties.

The county-level data are shown in Table I. County-level demographic and health data are aggregated from County Health Rankings, which collects measures from a variety sources to compare health outcomes and access for counties across the United States on an annual basis [11]. The program aggregates measures from the Center for Disease Control and Prevention's Behavioral Risk Factor Surveillance System and United States Diabetes Surveillance System, United States Census's Population Estimates, American Community Survey, and Department of Housing and Urban Development's Comprehensive Housing Affordability Strategy. A major goal of the program is to support local leaders and communities in identifying opportunities for improving public health, in emergencies or crises such as the emergence of a pandemic.

The variable DAYS_SINCE10 is calculated from JHU reporting and is equal to the number of days since the county reported 10 confirmed positive cases. This variable is needed to account for the differences in the amount of time that the virus has been spreading through the community. "Fig. 1"shows that the number of COVID-19 deaths and confirmed positives are exponentially related to the DAYS_SINCE10 variable..

The correlation plot in "Fig. 2" demonstrates that some health and demographic variables are highly correlated. Diabetes, obesity, and smoking are positively correlated with each other in addition to factors associated with socioeconomic status such as PER_CHILD_POV, PER_RURAL, and PER_BLACK. These are well-known correlations between socioeconomic and health factors, however, the first and last rows (and columns) of the figure shows how these factors are correlated with the COVID-19 measures of DAYS_SINCE10 and mortality rate. Although most of these univariate correlations are near zero, we observe a negative relationship between DAYS_SINCE10 and PER_RURAL likely because urban communities were the first to be exposed to the virus.

TABLE I

County-level demographics and health data used in the analysis.

Variable	*Description*	*Data Source*
PER_65PLUS	Percentage of population ages 65 and older.	Census Population Estimates
PER_BLACK	Percentage of population that is Black or African American.	Census Population Estimates
PER_RURAL	Percentage of population living in a rural area.	Census Population Estimates
OVERCROWDING	Measure of severe housing problem. Considered more than 1 person per room.	Comprehensive Housing Affordability Strategy
PER_SMOKE	Percentage of adults who are current smokers.	Behavioral Risk Factor Surveillance System
PER_OBESITY	Percentage of the adult population (age 20 and older) that reports a body mass index (BMI) greater than or equal to 30 kg/m^2.	United States Diabetes Surveillance System
PER_UNINSURED	Percentage of adults under age 65 without health insurance.	Small Area Health Insurance Estimates
PER_CHILD_POV	Percentage of people under age 18 in poverty.	Small Area Income and Poverty Estimates
DAYS_SINCE10	Number of days since the county reported 10 confirmed positive cases	Derived from JHU Reporting
MORT_RATE	Number of COVID-19 deaths divided by the number of COVID-19 positive tests	Derived from JHU Reporting

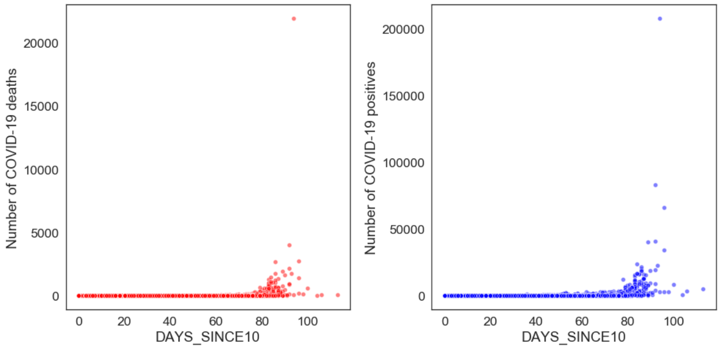

Fig. 1. Exponential relationship between the number of days in a county since ten positive cases were reported and the total number of cases in the county. The county with the highest value is New York, New York.

III. Methods

We tested linear models between each individual health variable and county COVID-19 mortality. Hypothesis testing was also performed to determine whether county health and demographic variables were associated with differing rates of mortality. We then fit a multivariate linear model on all the variables using forward step wise regression for variable selection. A multivariate linear model is explainable and allows us to determine the relative influence of the predictor variables on mortality rate. In addition, we can identify anomalies easily. Variable transformations, such as Box-Cox, were performed

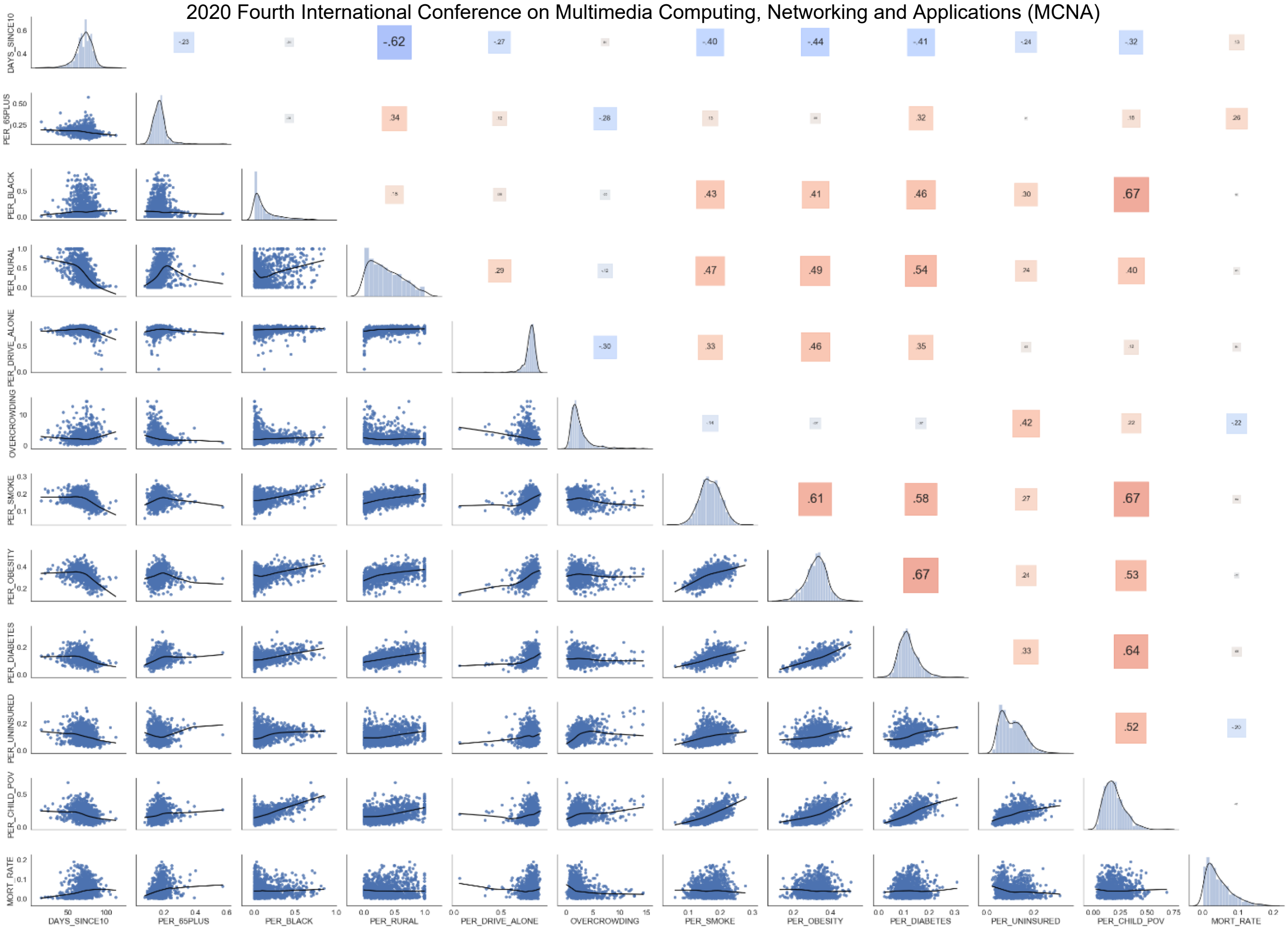

Fig. 2. Correlation plot and paired scatterplots for county variables used in the analysis.

until the linear model assumptions were satisfied. The final model is provided in the Results section.

County-level measures were then rolled up to state level measures to test whether there was a correlation between state-wide health risk measures and COVID-19 mortality. Some relationships may be realized in a larger geographical region.

Finally, a random forest model was built on the county-level data to determine whether nonlinear patterns in the data could be used to predict mortality rate. Variable importance from the random forest was compared to the insights from the linear regression models.

IV. Results

In the individual regressions, only PERCENT_65PLUS, OVERCROWDING, PER_UNINSURED, and DAYS_SINCE10 had statistically significant and linear relationship with the risk measure, mortality rate, in “Fig. 3”. This finding agrees with individual variable hypothesis testing performed to determine whether county health and demographic variables were associated with differing rates of mortality. Based on early COVID-19 research referenced in the Introduction, we hypothesized that counties with older populations, larger Black populations, smaller rural populations, greater household overcrowding, larger smoking populations, larger obese populations, larger diabetic populations, larger uninsured populations, and greater child poverty would have greater mortality. For example, for the variable PER_65PLUS, counties were split into two groups: those with above average 65-plus population and those with below average 65-plus population. Then, a t-test was performed for a difference of means of the mortality rate on the two groups. After correcting for the multiple hypothesis tests, only the variables PER_65PLUS, OVERCROWDING, PER_UNINSURED, and DAYS_SINCE10 separated the counties into groups with statistically significant differences in mean mortality.

We fit a multivariate linear model using forward step wise regression for variable selection. Dependent variable transformation was performed, and the linear model assumptions were satisfied. The final county-level model is shown in Table II.

The model and analysis show none of the expected effects from the health measures of obesity, diabetes, or smoking. This observation may be caused by the correlation between these variables, however, principle components analysis (taking the first principle component of these variables) also failed to result in a significant factor in the linear model. The model also does not show a linear relationship between PER_BLACK and mortality rate. However, the most significant variable in the model is the age demographic of PER_65PLUS. The lack of effect from

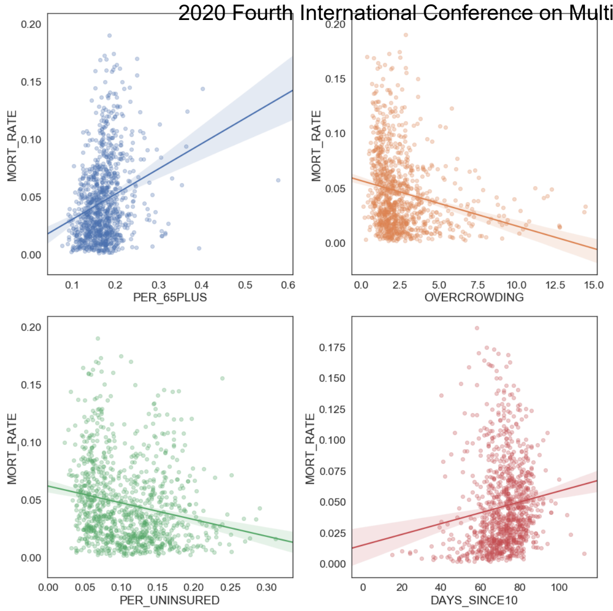

Fig. 3. Linear models between each county variable and mortality rate for statistically significant variables.

TABLE II
Linear model of mortality rate on the subset of variables selected from forward step wise regression.

Linear Model Results			
Variable	***Coefficient***	***P-value***	***Significance***[a]
Intercept	-3.922	$<$ 2e-16	***
PER_65PLUS	2.544	4.22e-12	***
PER_UNINSURED	-1.579	1.28e-6	***
DAYS_SINCE10	0.015	$<$ 2e-16	***
PER_SMOKE	0.445	0.466	
OVERCROWDING	-0.025	0.006	**
PER_RURAL	0.098	0.135	
PER_CHILD_POV	0.633	0.029	*
PER_BLACK	-0.013	0.914	
Residual standard error: 0.3973 on 1012 degrees of freedom Multiple R-squared: 0.2025, Adjusted R-squared: 0.1962 F-statistic: 32.12 on 8 and 1012 DF, p-value: $<$ 2.2e-16			
Linear Assumptions			
Property	***Value***	***P-value***	***Decision***
Global Stat	2.8193	0.5885	Acceptable
Skewness	0.6854	0.4077	Acceptable
Kurtosis	0.3890	0.5328	Acceptable
Link Function	0.5496	0.4585	Acceptable
Heteroscedasticity	1.1953	0.2743	Acceptable

[a] Significance Codes
‘***’ for p-value $<$ 0.001, ‘**’ for p-value $<$ 0.01
’*’ for p-value $<$ 0.05, ‘.’ for p-value $<$ 0.1

the health measures could be due to the limited availability of recorded COVID-19 deaths and confirmed positives at the county level.

To verify whether improved data availability reveals relationships between COVID-19 deaths and health measures, we aggregate the data to the state level. At the state-level, we do not include the variables DAYS_SINCE10 or OVERCROWDING as they lose meaning when the county regions are summed. The individual regressions suggest that PER_BLACK, PER_RURAL, and PER_UNINSURED have a significant linear relationship with mortality shown in “Fig. 4”.

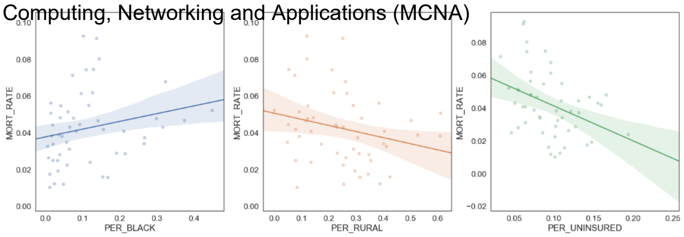

Fig. 4. Linear models between each state variable and mortality rate for statistically significant variables.

However, the state-level multivariate regression, we also fail to find a relationship between health risks and mortality rate. These are shown in “Fig. 5”.

For variable selection, we used forward step wise regression, and we normalized the residuals using Box-Cox transformations. Other variable selection and transformations were explored but did not improve model performance. Hence, these methods are presented for interpretability and explainability. The final model is described in Table III. The variables PER_UNINSURED and PER_BLACK were the only significant variables, with PER_65PLUS having significance at the level of p-value less than 0.1. The state-level results reflect the same conclusions as the county-level regression; that is, health factors are not significant, though the percent of a state’s population that is Black, uninsured, or 65 or older are significant in predicting mortality rate.

TABLE III
Linear model results for county data aggregated to the state level.

Linear Model Results			
Variable	***Coefficient***	***P-value***	***Significance***[a]
Intercept	-1.9159	0.000	***
PER_UNINSURED	-1.1317	0.030	*
PER_BLACK	0.4105	0.029	*
PER_OBESITY	-1.0086	0.193	
PER_65PLUS	1.9282	0.061	.
PER_RURAL	-0.2117	0.202	
PER_SMOKE	1.2626	0.193	
Residual standard error: 0.1159 on 44 degrees of freedom Multiple R-squared: 0.3579, Adjusted R-squared: 0.2703 F-statistic: 4.087 on 6 and 44 DF, p-value: 0.002417			
Linear Assumptions			
Property	***Value***	***P-value***	***Decision***
Global Stat	2.5864	0.6292	Acceptable
Skewness	0.1110	0.7390	Acceptable
Kurtosis	0.4422	0.5061	Acceptable
Link Function	1.8139	0.1780	Acceptable
Heteroscedasticity	0.2194	0.6395	Acceptable

[a] Significance Codes
‘***’ for p-value $<$ 0.001, ‘**’ for p-value $<$ 0.01
‘*’ for p-value $<$ 0.05, ‘.’ for p-value $<$ 0.1

Finally, to explore whether non-linear effects account for mortality risk, we fit a random forest model using all the data shown in Table I. Although model performance was low, we used the variable importance over a sample of models to compare the contribution from each variable. Variable importance is measured as the increase in model Mean Squared Error (MSE) if the variable was omitted from the mode. (The higher this value, the more important the variable.) “Fig. 6” shows

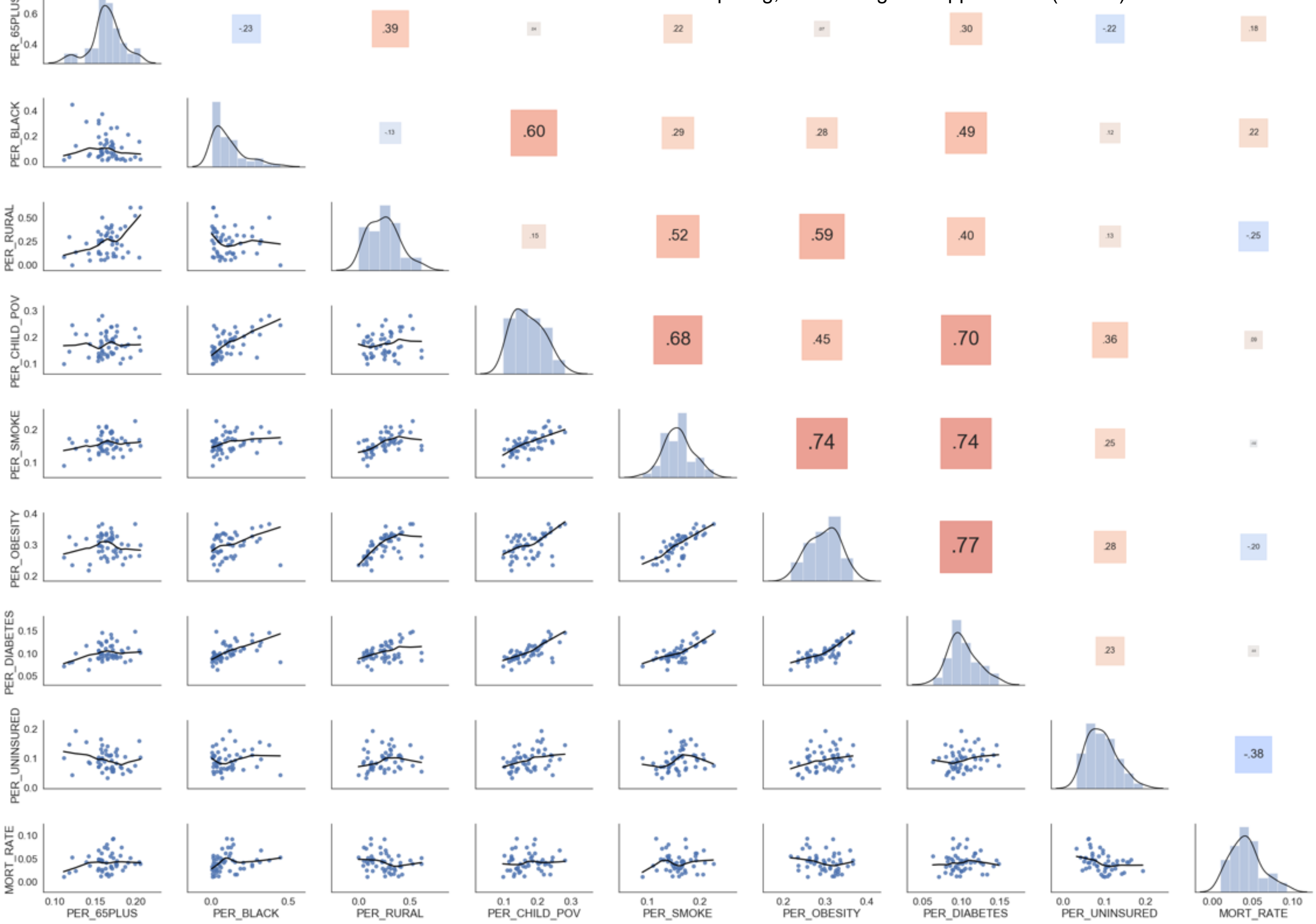

Fig. 5. Correlation plot and paired scatterplots for state variables.

the variable importance over multiple fits of a random forest, with the top four variables colored red. These results agree with the significant variables found in the regression model. That is, obesity, smoking, and diabetes were not as important in predicting mortality and the most important factors were PER_65PLUS, OVERCROWDING, PER_UNINSURED, and DAYS_SINCE10.

V. Conclusions

In RQ1, we found no evidence at the county-level of increased COVID-19 mortality rate in communities with high rates of obesity, diabetes, or smoking. However, we found that the relationship between age and mortality rate is statistically significant with RQ2. The regression model indicates that a 1% increase in the 65-plus population results in a 7.6% increase in mortality rate. In RQ3, we also found that overcrowding and uninsured populations were statistically significant in predicting mortality.

When aggregating data to the state-level, there is still no evidence of a relationship between mortality rate and obesity, diabetes, or smoking as we hypothesized. However, there is evidence of increased risk of mortality for states with higher Black and African American populations supporting RQ3. The

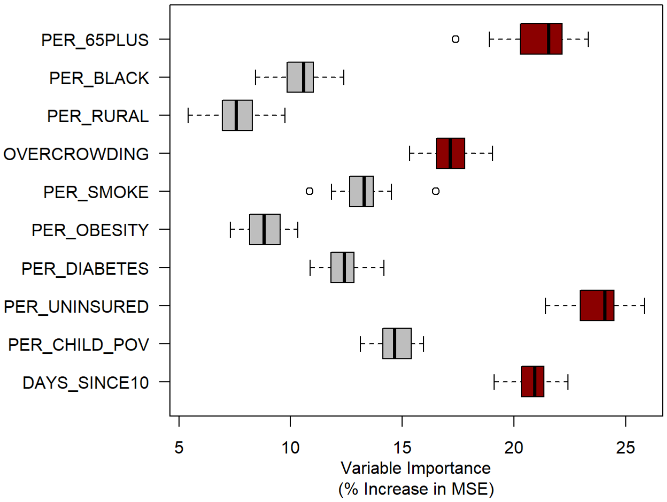

Fig. 6. Results of variable importance over thirty random forest fits. The four most influential variables were PER_65PLUS, OVERCROWDING, PER_UNINSURED, and DAYS_SINCE10.

association between increased risk and increased county age demographics continues to hold at the state level for RQ2.

The lack of an observable relationship between these health factors and mortality from COVID-19 may be the result of limited spread of the virus in smaller communities. Over time, this relationship may appear as more people die from the disease or the disease spreads to more rural communities. However, we do not observe a relationship between these health factors and death when county deaths were aggregated to the state-level. We controlled for the spread of the virus using the variable DAYS_SINCE10. Even when comparing communities with similar emergence of the virus, no relationship was found between these three health measures and COVID-19 risk factors.

We found that indicators of poverty, such as uninsured rates and overcrowding, and race influence COVID-19 mortality outcomes. CDC notes that deaths rates among the Black and African American community were 92.3 people per 100,000 population compared to 45.2 people in white populations [8]. Members of racial minorities more frequently live in multi-generational households where overcrowding may occur and it is difficult to self-isolate [15]. Racial minorities are also over-represented in the essential workforce and have greater exposure to the disease. In addition, ethnic minorities also tend to have lower access to healthcare with higher rates of uninsured individuals. These conditions hinder people's ability to protect themselves and respond to the pandemic, increasing their rate of transmission and death. Research has also shown that implicit bias and racial disparities result in lower quality healthcare for Black populations [7], which may be a contributing factor to higher COVID-19 mortality.

Another consideration is that several studies into risk factors for COVID-19 explore the relationship between severe respiratory disease and obesity or diabetes [5], [6]. That is, outcomes in these studies were not limited to deaths. It could be that although the risk of severe illness is increased due to these heath factors, treatment plans mitigate increased risk of death. The publicly available COVID-19 data only includes confirmed positives and deaths at the county level. Consequently, we are unable to test whether rates of ICU admissions or ventilator use are related to county measures of obesity, diabetes, or smoking.

The current data quality does not allow for robust and informative county-level models predicting COVID-19 severity, slowing down policy making. This work shows the potential benefit of the availability of more specific testing and symptom data at the county-level. If county-level hospitalization numbers, ventilation counts, or ICU counts were available, it may be possible for researchers to quickly discover contributing factors from large amounts of data across the country without requiring sensitive, patient-level data or jeopardizing patient privacy. This could lead to valuable insights about novel diseases that could inform strategy and planning at the county and hospital level.

Acknowledgment

This research was funded by Accenture Applied Intelligence in an effort to better support clients responding to SARS-CoV2 and better estimate future impacts.

References

[1] G. Onder, G. Rezza, and S. Brusaferro. "Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy," *JAMA*, vol. 323, no. 18, pp. 1775–1776, March 2020.

[2] C. Leung. "Risk factors for predicting mortality in elderly patients with COVID-19: A review of clinical data in China," *Mechanisms of Aging and Development*, vol. 188, no. 111255, June 2020.

[3] K. Liu, Y. Chen, R. Lin, and K. Han. "Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients," *Journal of Infection*, vol. 80, no. 6, pp.e14–e18, June 2020.

[4] "People who are at higher risk for severe illness," May 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-higher-risk.html, [Accessed June 5, 2020].

[5] C. Caussy, F. Pattou, F. Wallet, C. Simon, S. Chalopin, C. Telliam, and E. Disse, "Prevalence of obesity among adult inpatients with COVID-19 in France," *The Lancet Diabetes & Endocrinology*, vol. 8, no. 7, pp. 562–564, 2020.

[6] D. A. Kass, P.Duggal, and O. Cingolani, "Obesity could shift severe COVID-19 disease to younger ages," *The Lancet*, vol. 395, no. 10236, pp. 1544–1545, 2020.

[7] W. J. Hall, M. V. Chapman, K. M. Lee, Y. M. Merino, T. W. Thomas, B. K. Payne, E. Eng, S. H. Day, and T. Coyne-Beasley, "Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: A systematic review," *American Journal of Public Health*, vol. 105, no. 12, pp. e60–e76, 2015.

[8] Centers for Disease Control and Prevention, "COVID-19 in racial and ethnic minority groups," 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/racial-ethnic-minorities.html. [Accessed June 10, 2020].

[9] W. Pirtle. "Racial capitalism: A fundamental cause of novel Coronavirus (COVID-19) pandemic inequities in the United States," *Health Educ Behav.*, vol. 47, no. 4, pp.504–508, April 2020.

[10] Johns Hopkins University, "CSSEGISandData/COVID-19," 2020. [Online]. Available: https://github.com/CSSEGISandData/COVID-19. [Accessed: June 10, 2020].

[11] University of Wisconsin Population Health Institute, "Measures & data sources: County health rankings model," 2020. [Online]. Available: https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-health-rankings-model. [Accessed June 10, 2020].

[12] I. McCulloh, K. Kiernan, and T. Kent. "Inferring true COVID19 infection rates from deaths," *Frontiers in Big Data Medicine and Public Health*. September 2020.

[13] H. Lau, T. Khosrawipour, P Kocbach, H. Ichii, J. Bania, and V. Khosrawipour. "Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters," *Pulmonology*, June 2020.

[14] D. Weinberger, J. Chen, and T.Cohen. "Estimation of excess deaths associated with the COVID-19 pandemic in the United States, March to May 2020." *JAMA Intern Med*, vol 180, no.10, pp.1336–1344, July 2020

[15] R. Cholera, O. O. Falusi and J. M. Linton. Sheltering in place in a xenophobic climate: COVID-19 and children in immigrant families. *Pediatrics*, vol 146, no.1, July 2020.