# Risk factors associated with mortality of COVID-19 in 2692 counties of the United States

Ting Tian[1#], Jingwen Zhang[1#], Liyuan Hu[1#], Yukang Jiang[1#], Congyuan Duan[1], Xueqin, Wang[2,1]*, Heping Zhang[3]*

**Affiliations:**

[1]School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong, China

135 Xingang Xi Road, Guangzhou, Guangdong, 510275, China

[2]School of Management, University of Science and Technology of China, Hefei, Anhui, China

96 Jinzhai Road, Hefei, Anhui, 230026, China

[3]School of Public Health, Yale University, New Haven, CT, USA

60 College Street, New Haven, CT, 06520-8034, USA

*Corresponding authors: wangxq20@ustc.edu.cn; +8613600006896

heping.zhang@yale.edu; +12037855185

[#]Dr. T. Tian, J. Zhang, L. Hu and Y. Jiang contributed equally to this article

*Dr. X. Wang and Dr. H. Zhang are corresponding authors

Word counts: 2360

## Abstract

**Background:** The number of cumulative confirmed cases of COVID-19 in the United States has risen sharply since March. A county health ranking and roadmaps program has been established to identify factors associated with disparity in mobility and mortality of COVID-19 in all counties in the United States.

**Objective:** To find out the risk factors associated with mortality of COVID-19 with various levels of prevalence.

**Design:** A negative binomial design was applied to the county-level mortality counts of COVID-19 on April 15, 2020 in the United States. In this design, the infected counties were categorized into three levels of infections using clustering analysis based on time-variant cumulative confirmed cases from March 1 to April 15, 2020.

**Setting**: United States

**Participants:** COVID-19 patients in various counties of the United States from March 1 to April 15, 2020.

**Measurements:** The county-level cumulative confirmed cases and mortality of COVID-19.

**Results.** 2692 infected counties were assigned into three classes where the mild, moderate, and severe prevalence of infections were identified, respectively. Several risk factors are significantly associated with the mortality of COVID-19, where Hispanic (0.024, $P$=0.002), female (0.253, $P$=0.027), elder (0.218, $P$=0.017) and Native Hawaiian or other Pacific islander (2.032, $P$=0.027) individuals are more vulnerable to the mortality of COVID-19. More locations open to exercise (0.030,

*P*=0.004), higher levels of air pollution (0.184, *P*=0.044) and segregation between non-White and White increased the mortality rate.

**Limitation:** The study relied on mortality data on April 15, 2020.

**Conclusion.** The mortality of COVID-19 depends on sex, ethnicity, and outdoor environment. The increasing awareness of these significant factors may lead to the reduction in the mortality rate of COVID-19.

**Funding Source:** The National Key Research and Development Program of China, the National Natural Science Foundation of China (NSFC), the Key Research and Development Program of Guangdong, China and Fundamental Research Funds for the Central Universities.

## Introduction

COVID-19 is an infectious disease caused by a novel coronavirus with an estimated average incubation period of 5.1 days(1). It is spread through person-to-person transmission, and has now spread to 210 countries and regions with over 2 million total confirmed cases as of April 15(2). The United States has the highest number of infections, taking up approximately one-third of the total confirmed cases in the world. Its cumulative confirmed cases were 652,474 on April 15, 2020, an increase of 9,456 times compared with 69 confirmed cases on March 1, 2020(3).

Currently, the entire United States is suffering from a rapidly increasing epidemic situation, with deaths resulted from COVID-19 occurring all over the country. For instance, New York City had the largest number of total deaths, accounting for the vast majority of deaths in the country, while no one in Madison county, North Carolina is infected(3). Therefore, it is of great interest to find out the risk factors that influence the mortality of COVID-19. It is known that infectious diseases are affected by factors other than medical treatments(4, 5). For example, influenza A is associated with obesity(6), and the spread of SARS depended on seasonal temperature changes(7).

The county health ranking and roadmaps program was launched by both the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute(8). This program has provided annual sustainable source data including health outcomes, health behaviors, clinical care, social and economic factors, physical environment and demographics since 2010, which incorporates a total of 64 factors

possibly influencing health across all counties in 50 states. The details about those factors are available on the official website of county health ranking and roadmaps program(8). This paper aims to explore putative risk factors that may affect the mortality of COVID-19 (excluding deaths caused by other causes rather than COVID-19) in different areas of the United States and to increase awareness of the disparity and to form risk reduction strategies.

## Methods

### Data sources

We collected the number of cumulative confirmed cases and total deaths from March 1 to April 15, 2020, for counties in the United States from the New York Times(9). The county health rankings reports from the year 2020 were compiled from the County Health Rankings and Roadmaps program official website(8). There are 77 measures in each of 3142 counties, including the health outcome, health behaviors, clinical care, social and economic factors, physical environment, and demographics.

### Study areas

There were 2692 counties which reported confirmed cases until April 15, 2020. Also, 450 counties had no confirmed cases of COVID-19 and were not considered in this study. To find out the relationship between the risk factors and the mortality of COVID-19, we considered the total number of deaths on April 15, 2020 as the outcome.

**Assessment of covariates in health factors**

Putative risk factors(8) were categorized as 5 types of measures: health behaviors, clinical care, social and economic factors, physical environment and demographics. For health behaviors, there were tobacco, alcohol and drug use, diet and exercise, sexual activity and insufficient sleep. Clinical care data included access to and quality of care were considered. Social and economic factors included education, employment, income, family and social support and community safety. For the physical environment, air and water quality and housing and transit were considered. Overall, there were 56 possible risk variables included in the study. All deaths occurred as a result of COVID-19.

**Statistical analyses**

The trend of the total number of confirmed cases varied greatly in various areas of the United States. We used the PAM clustering algorithm(10, 11) to ensure that the similar trends were assigned to a homogenous class by standardizing the time-series of total confirmed cases from March 1 to April 15, 2020. Based on the clustering results, we used the Kruskal-Wallis test(12) and Chi-square test(13) to detect significant risk factors across different classes of counties. The most important risk factors were identified using random forest(14) in each class, based on which the top 15 factors were selected to build a negative binomial model(15, 16) in each class of the counties. All analysis was conducted in R version 3.6.1.

**Role of the Funding Source**

The funder of this study had no role in study design, data collection, data analysis, data interpretation, and writing of the report. The corresponding authors had full access to study data and final responsibility for the decision to submit for publication.

## Results

### Three classes of epidemic development in the United States

According to the clustering, 2692 counties were assigned into 3 classes. There were 2523 counties in the first class with the lowest overall cumulative confirmed cases. It is referred to as the mild class of infections. Its medoid is Austin county in Texas. There were 141 counties in the second class with overall relatively moderate cumulative confirmed cases. We call it the moderate class of infections. Its medoid is Monroe county in Pennsylvania. There were 28 counties in the third class with the highest overall cumulative confirmed cases, which we named as the severe class of infections. Its medoid is Fairfield county in Connecticut. The geographical distribution of the counties in different classes is shown in Figure 1, where the size of a circle indicated the total confirmed cases on April 15, 2020. Note that the east and west coasts were the most severely hit areas by COVID-19. Most counties in New York and New Jersey belonged to the third class of counties(9).

**Demographical distribution in three classes of counties**

Table 1 shows the significant difference in demographical distribution between the three classes of counties ($P < 0.001$). The average population in the mild class was 63,438, which is 8% and 4% of the average populations in the moderate class and severe class, respectively. The average proportion of rural residents in the mild class was 57.58%, vs 2.5% in the severe class. The average proportion of Black in the mild class was 9.75%, as opposed to 16.52% in the severe class. There were differences in the race, ethnicity and geographical location in the three classes of counties.

**Distribution of significant risk factors in three classes of counties**

Figure 2 shows the distribution of five risk factors in the three classes, and the *P*-values of these five variables are much smaller than 0.05. We found that the percentage of the population of adults with obesity in the severe class was relatively small, and comparatively high in the mild class. There are similar patterns in the percentage of youth in poverty, the number of prevention hospitalization stays, the percentage of adult smokers and the percentage of individuals under age 65 without health insurance. There were also significant differences in the distributions of other risk factors among the three classes of counties (Supplement Figure 1, and Supplement Table 1)

**Factors influencing mortality of COVID-19 in the three classes**

The importance scores of the risk factors were obtained by random forest, and one common factor, namely residential segregation between non-White and White, was identified in each of the three classes of counties. The negative binomial model for this single covariate was used to explore its association with mortality of COVID-19. Residential segregation between non-White and White in was the significant factor associated with the mortality of COVID-19 across the three classes of counties as shown in Figure 3. Note that the higher value of residential segregation between non-White and White the higher mortality of COVID-19. In the severe class of counties, an increase in the residential segregation between non-White and White resulted in more deaths than other two classes of counties.

In the mild class, there were five variables significantly associated with the mortality of COVID-19. Higher values in the resident population ($P$<0.001), segregation index (0.014, $P$<0.001), violent crime rate (0.001, $P$=0.032), and the percentage of workforce that had more than 30 minutes commute driving alone (0.016, $P$<0.001) significantly increased the number of deaths, while more people living in rural areas (-0.011, $P$=0.0004) decreased the number of deaths of COVID-19.

In the moderate class, there were six variables significantly associated with the deaths of COVID-19. Higher values in the percentage of workforce driving alone to work (0.022, $P$=0.033), the percentage of Hispanic population (0.024, $P$=0.002), the percentage of population with adequate access to locations for physical activity (0.030, $P$=0.004), the number of hospital stays for ambulatory-care sensitive conditions per

100,000 Medicare enrollees (0.001, *P*=0.005), the percentage of female population (0.253, *P*=0.027) and resident population (*P*=0.049) led to an increase in deaths.

In the severe class, there were six variables significantly associated with the deaths. Higher values in the average daily density of PM 2.5 (0.184, *P*=0.044), the percentage of adults who reported less than average 7 hours sleeping (0.085, *P*=0.028), the percentage of population aged over 65 (0.218, *P*=0.017), segregation index (0.050, *P*=0.008) and the percentage of native Hawaiian or other Pacific Islander population (2.032, *P*=0.027) caused more deaths, while more primary care physicians (-0.001, *P*=0.006) decreased the deaths of COVID-19 (Table 2).

## Discussion

Using the time trends of cumulative confirmed cases in 2692 counties in the United States, we categorized those counties into three levels of infection. The mild class counted for 93.7% of all counties. Their resident population was remarkably smaller than other two classes of counties. Thus, the resident population appeared to be a significant contributor to the mortality of COVID-19. The higher population may increase more contacts in the social distancing (17), leading to a higher risk in the deaths of COVID-19. On the contrary, higher percentage of residents living in rural areas in the mild infections class of counties may reduce the mortality. The segregation index between non-White and White revealed the disparity in health between non-White and White, leading to differences in health status not only at the individuals level but also at the community level (18). Higher values in the

segregation index indicated the poor health status, which may increase the mortality of COVID-19 (17). This health inequality had a positive association with the mortality of COVID-19 in the mild and the severe class of counties.

For the mild class of counties, higher number of reported violent crime offenses per 100,000 population indicated the poor community safety, and a higher percentage of long-distance commuting workforce was linked to the high level of anxiety for commuters(19). These two factors combined may increase psychological distress and subsequently make people feel vulnerable to COVID-19(20-22).

For the moderate class of counties, gender and racial differences in the population were linked to the persons contracting a new disease, where being female and Hispanic populations may be less protective to the COVID-19. Also, more people reported to stay in ambulatory-care sensitive conditions and more workforce driving alone to work indicated poor health status, leading to the high mortality in COVID-19. However, there was a higher percentage of people reported to have access to exercise in locations, resulting in the increase of mortality of COVID-19. This may be caused by gathering and crowding in the exercise locations (23, 24). Thus, reducing the access to exercise in those locations may improve the safety of the public.

For the severe class of counties, there is an age structure difference in the mortality of COVID-19. There is remarkably large resident population in severe class of counties, a higher percentage of elder indicated the larger population of individuals aged over 65, which increased the deaths of COVID-19(25). Sleep time was reported to be associated with the health system(26). We found a higher percentage of people

reported to have insufficient sleep time among the deaths from COVID-19. The air quality also had a positive association with an increase in the mortality of COVID-19(27). COVID - 19 deaths were significantly associated with the ratio of population to primary care physicians, namely the adequacy of medical resources. If enough medical staff were available, the infected population could be treated promptly, thereby could reduce the deaths.

## Conclusions

There are different key factors influencing the mortality of COVID-19 in across different counties in the United States. Regardless of the regions, the factors linked to the poor health status contributed to increasing mortality of COVID-19. Improving the health system and eliminating the racial disparity to enhance health equity, combining with reducing outdoor physical activities in the metropolitan areas could significantly decrease the mortality of COVID-19.

Therefore, it is recommended that governments in different regions should reduce physical and psychological risks in residential environments to reduce the mortality of COVID-19.

## References

1. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Annals of Internal

Medicine. 2020.

2. National Health Commission of the People's Republic of China 2020;Pageshttp://2019ncov.chinacdc.cn/2019-nCoV/global.html on April 26, 2020.

3. 2020;Pageshttps://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html.

4. Hadler JL, Yousey-Hindes K, Pérez A, Anderson EJ, Bargsten M, Bohm SR, et al. Influenza-related hospitalizations and poverty levels—United States, 2010–2012. 2016;65(5):101-5.

5. Noppert GA, Yang Z, Clarke P, Ye W, Davidson P, Wilson MLJAoe. Individual-and neighborhood-level contextual factors are associated with Mycobacterium tuberculosis transmission: genotypic clustering of cases in Michigan, 2004–2012. 2017;27(6):371-6. e5.

6. Maier HE, Lopez R, Sanchez N, Ng S, Gresh L, Ojeda S, et al. Obesity increases the duration of influenza a virus shedding in adults. The Journal of infectious diseases. 2018;218(9):1378-82.

7. Lin K, Fong DY-T, Zhu B, Karlberg J. Environmental factors on the SARS epidemic: air temperature, passage of time and multiplicative effect of hospital infection. Epidemiology & Infection. 2006;134(2):223-30.

8. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Statistical methods in medical research. 2007;16(3):259-75.

9. Times TNY 2020;Pageshttps://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html.

10. Zhang LS, Yang MJ, Lei DJ. An improved PAM clustering algorithm based on initial clustering centers. Applied Mechanics and Materials, 2012. Trans Tech Publ: 244-9.

11. Lei D, Zhu Q, Chen J, Lin H. Automatic PAM Clustering Algorithm for Outlier Detection. Journal of Software. 2012.

12. Brunner E, Konietschke F, Bathke AC, Pauly M. Ranks and Pseudo-Ranks-Paradoxical Results of Rank Tests. arXiv preprint arXiv:1802.05650. 2018.

13. McHugh ML. The chi-square test of independence. Biochemia medica. 2013;23(2):143-9.

14. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.

15. Hilbe JM. Negative binomial regression: Cambridge University Press; 2011.

16. Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. Journal of statistical software. 2008;27(8):1-25.

17. Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, et al. Demographic science aids in understanding the spread and fatality rates of COVID-19. 2020.

18. Williams DR, Collins CJR, Ethnicity,, Reader HAPH. Racial residential segregation. 2012;26:331.

19. Van Rooy DLJE, Behavior. Effects of automobile commute characteristics on affect and job candidate evaluations: A field experiment. 2006;38(5):626-55.

20. Mazza C, Ricci E, Biondi S, Colasanti M, Ferracuti S, Napoli C, et al. A Nationwide Survey of Psychological Distress among Italian People during the COVID-19 Pandemic: Immediate Psychological Responses and Associated Factors. International Journal of Environmental Research and Public Health. 2020;17(9):3165.

21. Qiu J, Shen B, Zhao M, Wang Z, Xie B, Xu Y. A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. General psychiatry. 2020;33(2):e100213-e.

22. Wang C, Pan R, Wan X, Tan Y, Xu L, Ho CS, et al. Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China. 2020;17(5):1729.

23. McCloskey B, Zumla A, Ippolito G, Blumberg L, Arbon P, Cicero A, et al. Mass gathering events and reducing further global spread of COVID-19: a political and public health dilemma. 2020;395(10230):1096-9.

24. Weber J 2020;Pageshttps://www.dw.com/en/coronavirus-are-outdoor-sports-healthy-exercise-or-a-dangerous-risk/a-52971973.

25. Onder G, Rezza G, Brusaferro S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. JAMA. 2020.

26. Besedovsky L, Lange T, Haack M**.** The sleep-immune crosstalk in health and disease. Physiological reviews. 2019;99(3):1325-80.

27. Wu X, Nethery RC, Sabath BM, Braun D, Dominici FJm**.** Exposure to air pollution and COVID-19 mortality in the United States. 2020.

**Acknowledgments**

We would like to thank all individuals who are collecting epidemiological data of the COVID-19 outbreak, and people collecting health ranking county data in the county health ranking and roadmaps program.

**Funding:** X. Wang was partly supported by the National Key Research and Development Program of China (Grant No. 2018YFC1315400), the National Natural Science Foundation of China (Grant No. 11771462) and the Key Research and Development Program of Guangdong, China (Grant No. 2019B020228001). T. Tian was supported by Fundamental Research Funds for the Central Universities (Grant No. 19lgpy236).

**Author contributions:** T. Tian, X. Wang and H. Zhang developed the idea and research. T. Tian, J. Zhang, L. Hu and Y. Jiang wrote the first draft of the manuscript and all other authors discussed results and edited the manuscript. J. Zhang, L. Hu and Y. Jiang collected and validated epidemiological data, C. Duan collected county health ranking data.

**Conflict of Interest:** All authors declare no conflict.

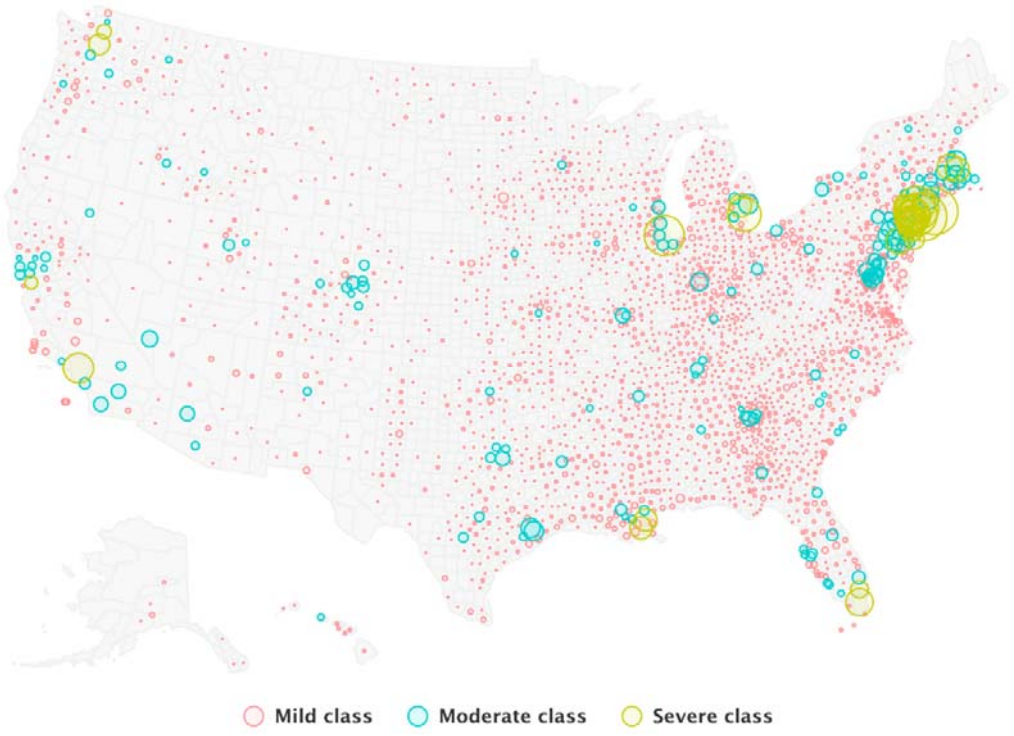

**Figure 1. The geographical distribution of three classes of counties.** The clustering was based on time-variant cumulative confirmed cases from March 1 to April 15, 2020. The size of circle represented the total confirmed cases on April 15, 2020.

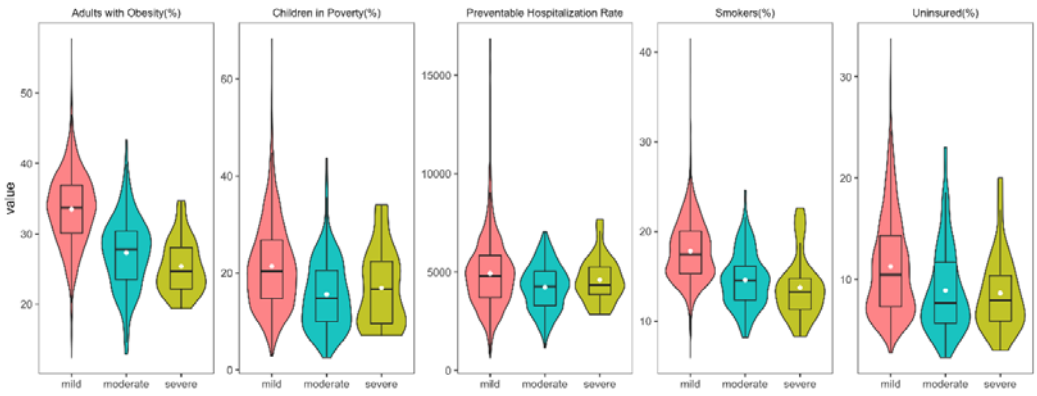


**Figure 2. Violin diagram and boxplot of the distribution of five factors in three classes of counties.** The mild, the moderate and the severe classes of counties are represented by red, blue and green colors.

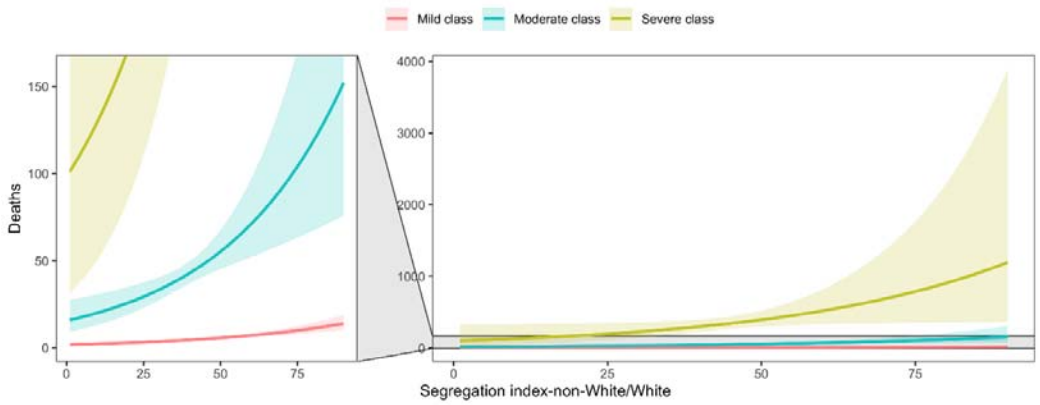


**Figure 3. The common factor of Segregation index non-White/White in three classes of counties.** Direct curves were generated using the Negative Binomial model with single covariate segregation index between non-White and White.

**Table 1. The differences in demographical distribution of three classes of counties.**

| | | Mild class | Moderate class | Severe class |
|---|---|---|---|---|
| DEMOGRAPHICS | P value | Mean±Sd | Mean±Sd | Mean±Sd |
| Resident population | <0.001 | 63438±94810 | 793232±728742 | 1564780±1944636 |
| % 65 and over | <0.001 | 19.007±4.403 | 15.495±3.615 | 15.362±2.178 |
| % Black | <0.001 | 9.745±14.927 | 14.844±14.714 | 16.519±13.735 |
| % Asian | <0.001 | 1.293±1.862 | 6.443±6.658 | 9.496±8.043 |
| % Native Hawaiian/ Other Pacific Islander | 0.011 | 0.123±0.420 | 0.27±0.845 | 0.173±0.194 |
| % Hispanic | <0.001 | 8.819±12.880 | 16.056±12.361 | 22.209±14.558 |
| % Non-Hispanic | <0.001 | 76.656±19.351 | 59.943±17.629 | 49.839±17.614 |

| | | | | |
|---|---|---|---|---|
| White | | | | |
| % Not Proficient in English | <0.001 | 1.517±2.477 | 3.792±2.635 | 7.014±4.523 |
| % Female | <0.001 | 49.966±2.191 | 50.985±0.927 | 51.137±0.792 |
| % Rural | <0.001 | 57.582±28.668 | 9.053±10.552 | 2.501±4.533 |

**Table 2. Variables significantly related to the deaths of COVID-19 in three classes.**

| | Variable | Variable description | Estimate | Pr(>\|t\|) | RF ranking |
|---|---|---|---|---|---|
| Mild Class | Population | Resident population | $1.40 \times 10^{-6}$ | <0.001 | 1 |
| | Segregation index-non-White/White | Index of dissimilarity where higher values indicate greater residential segregation between non-White and White county residents | 0.014 | <0.001 | 4 |
| | Violent Crime Rate | Number of reported violent crime offenses per 100,000 population | 0.001 | 0.032 | 7 |
| | % Long Commute - Drives Alone | Among workers who commute in their car alone, the percentage that commute more than 30 minutes | 0.052 | <0.001 | 9 |
| | % Rural | Percentage of population living in a rural area. | -0.011 | 0.001 | 10 |
| Moderate Class | % Drive Alone to Work | Percentage of the workforce that drives alone to work. | 0.022 | 0.033 | 1 |
| | % Hispanic | Percentage of population that is Hispanic | 0.024 | 0.002 | 4 |

| | | | | | |
|---|---|---|---|---|---|
| | % With Access to Exercise Opportunities | Percentage of population with adequate access to locations for physical activity. | 0.030 | 0.004 | 7 |
| | Preventable Hospitalization Rate | Rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees | 0.001 | 0.005 | 8 |
| | % Female | Percentage of population that is Female | 0.253 | 0.027 | 10 |
| | Population | Resident population | $2.29 \times 10^{-7}$ | 0.049 | 13 |
| Severe Class | Average Daily PM2.5 | Average daily density of fine particulate matter in micrograms per cubic meter (PM2.5) | 0.184 | 0.044 | 3 |
| | % Insufficient Sleep | Percentage of adults who report fewer than 7 hours of sleep on average | 0.085 | 0.028 | 6 |
| | % 65 and over | Percentage of population ages 65 and older. | 0.218 | 0.017 | 8 |
| | Segregation index-non-White/White | Index of dissimilarity where higher values indicate greater residential segregation between non-White and White county residents. | 0.050 | 0.008 | 11 |

| | | | | |
|---|---|---|---|---|
| Primary Care Physicians Ratio | Ratio of population to primary care physicians. | -0.001 | 0.006 | 12 |
| % Native Hawaiian/Other Pacific Islander | Percentage of population that is Native Hawaiian or Other Pacific Islander. | 2.032 | 0.027 | 13 |