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

Ting Tian[a,1], Jingwen Zhang[a,1], Liyuan Hu[a,1], Yukang Jiang[a,1], Congyuan Duan[a], Zhongfei Li[d], Xueqin Wang[b,2], and Heping Zhang[c,2]

[a]*School of Mathematics, Sun Yat-sen University*

[b]*School of Management,University of Science and Technology of China*

[c]*School of Public Health, Yale University*

[d]*School of Management,Sun Yat-sen University*

Abstract

Background

The number of cumulative confirmed cases of COVID-19 in the United States has risen sharply since March 2020. 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.

Methods

To find out the risk factors associated with county-level mortality of COVID-19 with various levels of prevalence, a negative binomial design was applied to the county-level mortality counts of COVID-19 as of April 15, 2020 in the United States. In this design, the infected counties were categorized into three levels of infections using

[1]Ting Tian, Jingwen Zhang,Liyuan Hu and Yukang Jiang contributed equally to this article

[2]Xueqin Wang and Heping Zhang are corresponding authors

clustering analysis based on time-varying cumulative confirmed cases from March 1 to April 15, 2020. COVID-19 patients were not analyzed individually but were aggregated at the county-level, where the county-level deaths of COVID-19 confirmed by the local health agencies.

Findings

2692 infected counties were assigned into three classes corresponding to low, median, and high prevalence levels of infection. Several risk factors were significantly associated with the mortality counts of COVID-19, where elder (0.221, P=0.001) individuals were more vulnerable and higher level of air pollution (0.186, P=0.005) increased the mortality in the metropolis areas. The segregation between non-Whites and Whites had higher likelihood of risk of the deaths in all infected counties.

Interpretation

The mortality of COVID-19 depended on sex, race/ethnicity, and outdoor environment. The increasing awareness of the impact of these significant factors may lead to the reduction in the mortality of COVID-19.

1. INTRODUCTION

COVID-19 is an infectious disease caused by a novel coronavirus with an estimated average incubation period of 5.1 days (Lauer et al., 2020). It spreads through person-to-person transmission, and has now infected 210 countries and regions with over 2 million total confirmed cases as of April 15, 2020 (National Health Commission of the People's Republic of China, 2020). The United States had 652,474 confirmed cases on April 15, 2020, the highest in the world, but there were only 69 confirmed cases on March 1, 2020 (Centers for Disease Control and Prevention, 2020).

The United States has been suffering from a severe epidemic, with COVID-19 related deaths 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 was infected (Centers for Disease Control and Prevention, 2020). Therefore, it is of great interest to find out the risk factors that influence the number of deaths of COVID-19. It is known that infectious diseases are affected by factors other than medical treatments (Hadler et al., 2016; Noppert et al., 2017). For example, influenza A is associated with obesity (Maier et al., 2018), and the spread of SARS depends on seasonal temperature changes (Lin et al., 2006).

The County Health Rankings and Roadmaps program was launched by both the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute (A Robert Wood Johnson Foundation program, 2020). This program has been providing annual sustainable source data including health outcomes, health behaviors, clinical care, social and economic factors, physical environment and demographics since 2010. We explored putative risk factors that may affect the mortality of COVID-19 in different areas of the United States in order to increase awareness of the disparity and aid the development of risk reduction strategies.

2. METHODS

2.1 Data sources

We collected the number of cumulative confirmed cases and deaths from March 1 to April 15, 2020, for counties in the United States from the New York Times (New York Times, 2020). The COVID-19 confirmed cases and deaths were identified by the laboratory RNA test and specific criteria for symptoms and exposures from health departments and U.S. Centers for Disease Control and Prevention (CDC). The county health rankings reports from year 2020 were compiled from the County Health Rankings and Roadmaps program official website (A Robert Wood Johnson Foundation program, 2020). There were 77 measures in each of 3142 counties, including the health outcome, health behaviors, clinical care, social and economic factors, physical environment, and demographics. We refer to the official website of the County Health Rankings and Roadmaps program (A Robert Wood Johnson Foundation program, 2020) for detailed information.

2.2 Study areas

As of April 15, 2020, a total of 2,692 counties reported confirmed cases in the United States, leaving 450 counties without confirmed cases of COVID-19 which were excluded from this study. The total number of deaths as of April 15, 2020 was considered as the outcome of this study.

2.3 Assessment of covariates in health factors

We divided the putative risk factors (A Robert Wood Johnson Foundation program, 2020) into 5 categories: health behaviors (e.g., access to exercise opportunities, insufficient sleep), clinical care (e.g. primary care physicians ratio), social and economic factors (e.g., racial segregation index), physical environment (e.g., transit problems and air quality), and demographics (age, sex, rural, and race/ethnicity). For example, there were previous studies which identified the air pollution may relate to high levels of COVID-19 (Conticini et al., 2020) and elder population had the high risk in the COVID-19 (Onder et al., 2020). Besides

these identified risk factors, we were interested in the adverse health factors may link to the mortality of COVID-19. Table 1 presented descriptive definition, sources and literature of 12 risk factors. All deaths resulted from complications of COVID-19.

Table 1: The definitions and sources of the selected five risk factors, and the literature supporting these factors.

Category	**Variable**	**Description**	**Sources**	**Related topics**	**Previous studies**
HEALTH BEHAV-IORS	% With Access to Exercise Opportunities	Percentage of population with access to locations for physical activity	Business Analyst, Delorme map data, ESRI, & US Census Tigerline Files	Social Distanc-ing;Outdoor and indoor exercise	Novel
	% Insufficient Sleep	Percentage of adults who report fewer than 7 hours of sleep on average	Behavioral Risk Factor Surveillance System	Immune system	Novel
CLINICAL CARE	Primary Care Physicians Ratio	Ratio of population to primary care physicians	Area Health Resource File/American Medical Association	Medicare services	Novel
SOCIAL & ECO-NOMIC FACTORS	Segregation index-non-Whites/Whites	Index of dissimilarity where higher values indicate greater residential segregation between non-White and White county residents	American Community Survey, 5-year estimates	Poor health status	Novel

Continued on next page

Table 1 – continued from previous page

Category	Variable	Description	Sources	Related topics	Previous studies
	% Long Commute - Drives Alone	Among workers who commute in their car alone, the percentage that commute more than 30 minutes	American Community Survey, 5-year estimates	Psychological risks	Novel
PHYSICAL ENVIRONMENT	% Drive Alone to Work	Percentage of the workforce that drives alone to work	American Community Survey, 5-year estimates	Psychological risks	Novel
	Average Daily $PM_{2.5}$	Average daily density of fine particulate matter in micrograms per cubic meter ($PM_{2.5}$)	Environmental Public Health Tracking Network	Air quality	Conticini, Frediani and Caro(2020), Contini and Costabile(2020), Martelletti and Martelletti(2020)
	Population	Resident population	Census Population Estimates	Population density	Novel
	%Rural	Percentage of population living in a rural area	Census Population Estimates	Demographical science	Novel
DEMOGRAPHICS	%Hispanic	Percentage of population that is Hispanic	Census Population Estimates	Minorities	Novel
	%Female	Percentage of population that is female	Census Population Estimates	Sex	Wenham, et.al(2020)

Continued on next page

Table 1 – continued from previous page

Category	Variable	Description	Sources	Related topics	Previous studies
	% 65 and over	Percentage of population ages 65 and older	Census Population Estimates	Age structure	Onder, et.al(2020), Abdulamir and Hafidh(2020)

2.4 Statistical analysis

The trend of the cumulative confirmed cases varied greatly in counties of the United States. We used the partitioning around medoids (PAM) clustering algorithm (Zhang et al., 2012; Lei et al., 2012) to assign counties with similar trends into a homogenous class after standardizing the time series of cumulative confirmed cases from March 1 to April 15, 2020. Based on the clustering results, we used the Kruskal-Wallis test (Brunner et al., 2018) to detect whether there were significant differences in the distributions of 12 risk factors across different classes of counties. The 12 risk factors were used to build a negative binomial model (Hilbe, 2011; Zeileis et al., 2008) for every class of the counties. The analysis was conducted in R version 3.6.1.

2.5 Validation analysis

We randomly divided counties (samples) into training (70% of the counties) and testing (30% of the counties) in each class. The model obtained from the training data was employed to predict the death counts of COVID-19 in the testing data, and the accuracy was assessed by the root mean square error (RMSE) of the mortality ratio (the number of deaths divided by the number of cumulative confirmed cases).

3. RESULTS

3.1 Three classes of county-level infection in the United States

The clustering analysis grouped the 2,692 counties were assigned into 3 classes. There were 2,523 counties in the first class with the lowest overall cumulative confirmed cases. Its medoid was Austin County in Texas. There were 141 counties in the second class with a median level of overall cumulative confirmed cases. Its medoid was Monroe County in Pennsylvania. There were 28 counties in the third class with the highest overall cumulative confirmed cases. Its medoid was Fairfield County in Connecticut. Here, the PAM algorithm selected the county with most representative data as the medoid in a class (Hilbe, 2011; Zeileis et al., 2008). The geographical distribution of the counties by class was shown in Figure 1, where the size of a circle indicated the cumulative confirmed cases on April 15, 2020. The distribution of deaths on April 15, 2020, which clearly differed among the three classes, was also presented in Figure 1. 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 (New York Times, 2020).

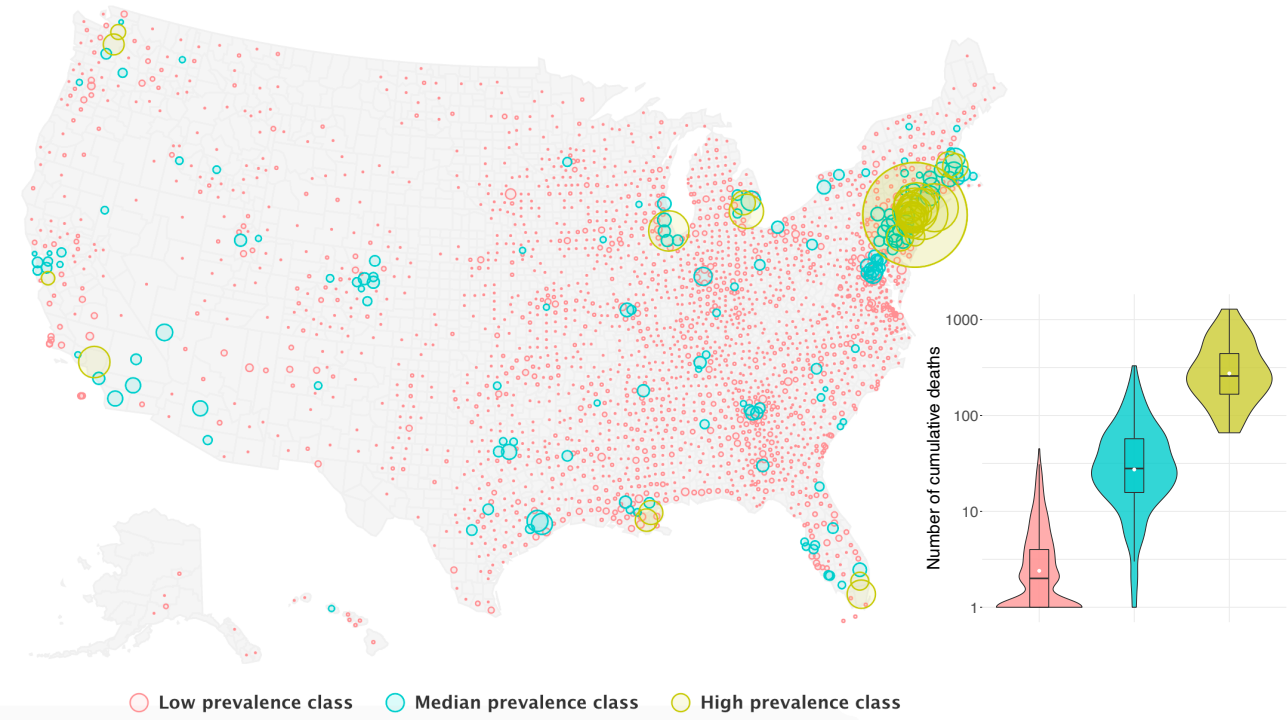

Figure 1: **The geographical distribution of three classes of counties.** The clustering was based on time-variant cumulative confirmed cases from March 1 to April 15, 2020. The size of circle represented the total confirmed cases on April 15, 2020. The distributions of deaths on April 15, 2020 in the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

3.2 Distributions of 12 selected risk factors in the three classes of counties

Figure 2 showed the distributions of the 12 selected risk factors by the class of counties. The distributions were significant different (P <0.001) for most of the 12 risk factors. For example, the average population in the low prevalence class was 63,438, which was 8% and 4% of the average populations in the median and high prevalence classes, respectively. The average proportion of rural residents in the low prevalence class was 57.58%, versus 2.5% in the high prevalence class. The segregation index of non-Whites versus Whites was the largest in the high prevalence class, but the smallest in the low prevalence class.

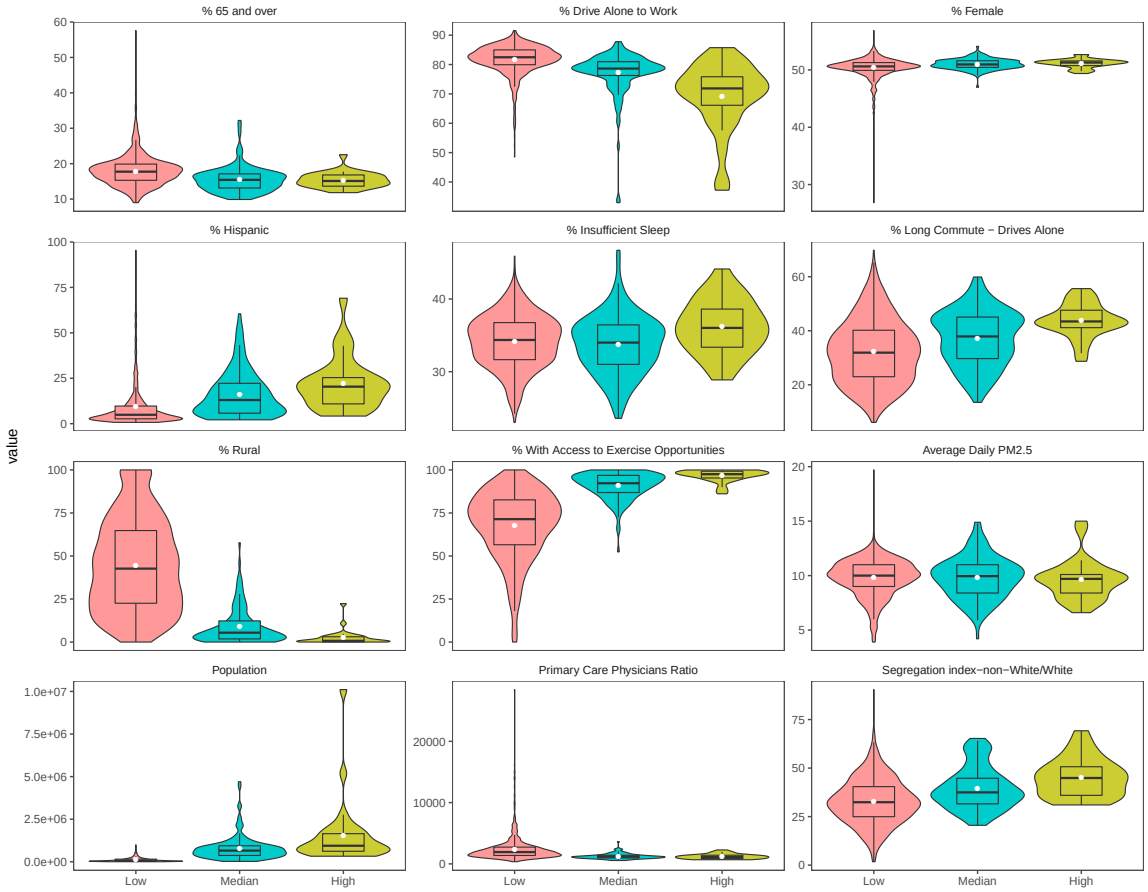

Figure 2: **Violin diagram and boxplot of the distributions of the 12 selected risk factors in the three classes of counties.** The low, median and high prevalence classes of counties were represented by red, blue and green colors.

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

There was one common factor, namely residential segregation between non-Whites and Whites, which had a statistically significant (P <0.05) effect on mortality in all classes. The negative binomial model was used to understand the within-class effects of this factor on mortality of COVID-19 as shown in Figure 3. Note that the higher value of residential segregation between non-Whites and Whites the higher mortality of COVID-19. In the high prevalence class, an increase in the residential segregation between non-Whites and Whites resulted in more deaths than other two classes of counties.

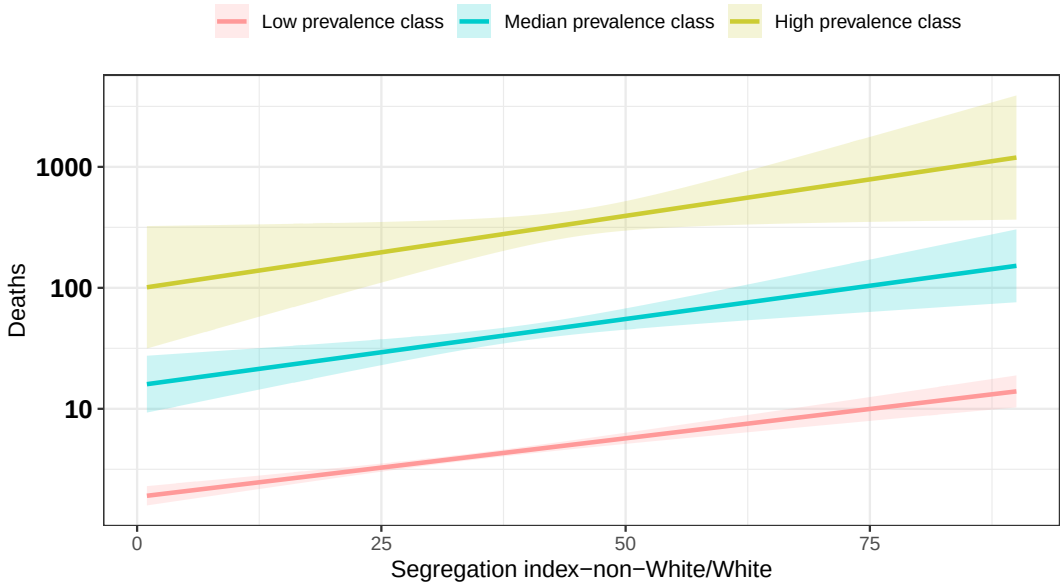

Figure 3: **Segregation index non-Whites versus Whites as the risk factor common in the three classes of counties.** Direct curves were generated using the negative binomial model with the segregation index between non-Whites and Whites as the single covariate.

Table 2 presented the significant factors specific to each class based on the training data. Specifically, in the low prevalence class, six variables were significantly associated with the mortality of COVID-19. Higher values in the resident population (P <0.001), segregation index (0.015, P <0.001), the percentage of workforce that had more than 30 minutes commute driving alone (0.013, P <0.001), and the percentage of adults who reported less than average 7 hours sleeping (0.026, P=0.017) significantly increased the number of deaths, while more people living in rural areas (-0.014, P <0.001) and the percentage of Hispanic population (-0.011, P=0.001) decreased the number of deaths of COVID-19.

In the median prevalence class, three variables were significantly associated with the deaths of COVID-19. Higher values in the percentage of workforce driving alone to work (0.058, P=0.006), segregation index (0.033, P=0.004), and the percentage of workforce that had more than 30 minutes commute driving alone (0.031, P=0.002) led to an increase in deaths.

In the high prevalence class, four variables were significantly associated with mortality. Higher values in the average daily density of $PM_{2.5}$ (0.186, P=0.005), segregation index (0.032, P=0.023), the percentage of adults who reported less than average 7 hours sleeping (0.081, P=0.021), and the percentage of population aged over 65 (0.221, P=0.001) caused more deaths.

For each class of counties, the model obtained from the training data was employed to predict the deaths of COVID-19 on April 15, 2020 using the testing data. The corresponding RMSE values for the mortality ratio were 0.09, 0.07 and 0.03, respectively, in the low, median, and high prevalence classes.

Table 2: Variables significantly related to the mortality rate of COVID-19 in the three classes of counties.

	Variable	**Estimate**	**Pr($>|t|$)**
	% Insufficient Sleep	0.026	0.017
	Segregation index non-White/white	0.015	< 0.001
Low	%Long Commute-Drives Alone	0.013	< 0.001
Prevalence	Population	1.68E-06	< 0.001
Class	%Hispanic	-0.011	0.001
	%Rural	-0.014	< 0.001
Median	Segregation index non-Whites/Whites	0.033	0.004
Prevalence	% Drive Alone to Work	0.058	0.006
Class	% Long Commute - Drives Alone	0.031	0.002
	% Insufficient Sleep	0.081	0.021
High	Segregation index non-White/White	0.032	0.023
Prevalence	Average Daily $PM_{2.5}$	0.186	0.005
Class	% 65 and over	0.221	0.001

4. DISCUSSION

Using the time trends of the cumulative confirmed cases in 2,692 counties in the United States, we categorized those counties into three levels of infection. The low prevalence class counted for 93.7% of the 2,692 counties. Their resident population was remarkably smaller than the other two classes of counties. Thus, the resident population density appeared to be a significant contributor to the mortality of COVID-19. A higher population density may increase more contacts in social distancing (Dowd et al., 2020; Greenstone and Nigam, 2020), leading to a higher risk in mortality of COVID-19. On the contrary, a higher percentage of residents living in rural areas in the median prevalence class of counties may reduce the mortality. The segregation index between non-Whites and Whites revealed the racial disparity in health, leading to differences in health status not only at the individual level but also at the community level (Williams and Collins, 2012). A higher values in the segregation index indicated the poor health status, which may increase the mortality of COVID-19 (Dowd et al., 2020). This health inequality increased the mortality rates of COVID-19 in all classes of counties.

For the low prevalence class of counties, a higher percentage of long-distance commuting workforce was linked to a high level of anxiety for commuters (Van Rooy, 2006). Sleeping time was reported to be associated with the health system (Besedovsky et al., 2019), the higher number of people who have inadequate sleeping time, the adverse effects of sleep on immunity were identified (Irwin, 2002). These two factors together may increase psychological distress and subsequently make people feel vulnerable to COVID-19 (Mazza et al., 2020; Qiu et al., 2020; Wang et al., 2020). Disparities in race and ethnicity were found in the infected populations. For example, Blacks were reported to be prone to COVID-19 (Hooper et al., 2020; Laurencin and McClinton, 2020). However, Hispanic populations in more rural areas may be more protective to COVID-19.

For the median prevalence class of counties, more workforce driving alone to work and commuting long-distance may increase the levels of anxiety (Van Rooy, 2006), leading to the high mortality in COVID-19.

For the high prevalence class of counties, there was an age trend in the mortality rate of COVID-19. In those counties, there was a higher percentage of elderly, indicating a larger population of individuals aged over 65, which increased the mortality rate of COVID-19 (Onder et al., 2020). The air quality also was associated with the mortality rate of COVID-19 (Conticini et al., 2020; Wu et al., 2020; Contini and Costabile, 2020).

This study identified several significant risk factors associated with the mortality of COVID-19, and our findings are highly valuable and timely for the decision-makers to develop strategies in reducing the mortality of COVID-19. The study relied on mortality data on April 15, 2020. The counties were randomly divided into the training and testing data once. However, we offered the epidemiological picture to facilitate the identification of important factors influencing the mortality of COVID-19 across different levels of infected counties in the United States. Regardless of the regions, the factors linked to the poor health status contributed to higher mortality of COVID-19. Improving the clinical care and eliminating the racial health inequality, combined with improving physical environment were expected to significantly decrease the mortality rate of COVID-19. Thus, we recommended that local governments should reduce physical and psychological risks in residential environments.

ACKNOWLEDGMENTS

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

REFERENCES

A Robert Wood Johnson Foundation program (2020), "County Health Rankings Reports,". **URL:** *https://www.countyhealthrankings.org/reports/county-health-rankings-reports*

Abdulamir, A. S., and Hafidh, R. R. (2020), "The Possible Immunological Pathways for the Variable Immunopathogenesis of COVID–19 Infections among Healthy Adults, Elderly and Children.," *Electronic Journal of General Medicine*, 17(4).

Besedovsky, L., Lange, T., and Haack, M. (2019), "The sleep-immune crosstalk in health and disease," *Physiological reviews*, 99(3), 1325–1380.

Brunner, E., Konietschke, F., Bathke, A. C., and Pauly, M. (2018), "Ranks and Pseudo-Ranks-Paradoxical Results of Rank Tests," *arXiv preprint arXiv:1802.05650*, .

Centers for Disease Control and Prevention (2020), "CCoronavirus Disease 2019 (COVID-19),". **URL:** *https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html*

Conticini, E., Frediani, B., and Caro, D. (2020), "Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy?," *Environmental pollution*, p. 114465.

Contini, D., and Costabile, F. (2020), "Does Air Pollution Influence COVID-19 Outbreaks?,".

Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., Liu, Y., and Mills, M. C. (2020), "Demographic science aids in understanding the spread and fatality rates of COVID-19," *Proceedings of the National Academy of Sciences*, 117(18), 9696–9698.

Greenstone, M., and Nigam, V. (2020), "Does social distancing matter?," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2020-26).

Hadler, J. L., Yousey-Hindes, K., Pérez, A., Anderson, E. J., Bargsten, M., Bohm, S. R., Hill, M., Hogan, B., Laidler, M., Lindegren, M. L. et al. (2016), "Influenza-related hospi-

talizations and poverty levelsUnited States, 2010–2012," *Morbidity and Mortality Weekly Report*, 65(5), 101–105.

Hilbe, J. M. (2011), *Negative binomial regression* Cambridge University Press.

Hooper, M. W., Nápoles, A. M., and Pérez-Stable, E. J. (2020), "COVID-19 and racial/ethnic disparities," *Jama*, .

Irwin, M. (2002), "Effects of sleep and sleep loss on immunity and cytokines," *Brain, behavior, and immunity*, 16(5), 503–512.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020), "The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application," *Annals of internal medicine*, 172(9), 577–582.

Laurencin, C. T., and McClinton, A. (2020), "The COVID-19 pandemic: a call to action to identify and address racial and ethnic disparities," *Journal of Racial and Ethnic Health Disparities*, pp. 1–5.

Lei, D., Chen, J., Lin, H., and Yang, P. (2012), "Automatic PAM Clustering Algorithm for Outlier," *Journal of software*, 7(5), 1045.

Lin, K., Fong, D. Y.-T., Zhu, B., and Karlberg, J. (2006), "Environmental factors on the SARS epidemic: air temperature, passage of time and multiplicative effect of hospital infection," *Epidemiology & Infection*, 134(2), 223–230.

Maier, H. E., Lopez, R., Sanchez, N., Ng, S., Gresh, L., Ojeda, S., Burger-Calderon, R., Kuan, G., Harris, E., Balmaseda, A. et al. (2018), "Obesity increases the duration of influenza a virus shedding in adults," *The Journal of infectious diseases*, 218(9), 1378–1382.

Martelletti, L., and Martelletti, P. (2020), "Air pollution and the novel Covid-19 disease: a putative disease risk factor," *SN Comprehensive Clinical Medicine*, pp. 1–5.

Mazza, C., Ricci, E., Biondi, S., Colasanti, M., Ferracuti, S., Napoli, C., and Roma, P. (2020), "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*, 17(9), 3165.

National Health Commission of the People's Republic of China (2020), "Distribution of COVID-19 cases in the world. [accessed 2020 April 15],".
URL: *http://2019ncov.chinacdc.cn/2019-nCoV/global.html*

New York Times (2020), "Coronavirus in the US: Latest Map and Case Count,".
URL: *https://github.com/nytimes/covid-19-data*

Noppert, G. A., Yang, Z., Clarke, P., Ye, W., Davidson, P., and Wilson, M. L. (2017), "Individual-and neighborhood-level contextual factors are associated with Mycobacterium tuberculosis transmission: genotypic clustering of cases in Michigan, 2004–2012," *Annals of epidemiology*, 27(6), 371–376.

Onder, G., Rezza, G., and Brusaferro, S. (2020), "Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy," *Jama*, 323(18), 1775–1776.

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

Van Rooy, D. L. (2006), "Effects of automobile commute characteristics on affect and job candidate evaluations: A field experiment," *Environment and Behavior*, 38(5), 626–655.

Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., Ho, C. S., and Ho, R. C. (2020), "Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China," *International journal of environmental research and public health*, 17(5), 1729.

Wenham, C., Smith, J., and Morgan, R. (2020), "COVID-19: the gendered impacts of the outbreak," *The Lancet*, 395(10227), 846–848.

Williams, D. R., and Collins, C. (2012), "Racial residential segregation," *Race, Ethnicity, and Health: A Public Health Reader*, 26, 331.

Wu, X., Nethery, R. C., Sabath, B. M., Braun, D., and Dominici, F. (2020), "Exposure to air pollution and COVID-19 mortality in the United States," *medRxiv*, .

Zeileis, A., Kleiber, C., and Jackman, S. (2008), "Regression models for count data in R," *Journal of statistical software*, 27(8), 1–25.

Zhang, L. S., Yang, M. J., and Lei, D. J. (2012), An improved PAM clustering algorithm based on initial clustering centers,, in *Applied Mechanics and Materials*, Vol. 135, Trans Tech Publ, pp. 244–249.