PLOS MEDICINE

RESEARCH ARTICLE

Investigating associations between COVID-19 mortality and population-level health and socioeconomic indicators in the United States: A modeling study

Sasikiran Kandula*, Jeffrey Shaman

Department of Environmental Health Sciences, Columbia University, New York, New York, United States of America

* sk3542@cumc.columbia.edu

Citation: Kandula S, Shaman J (2021) Investigating associations between COVID-19 mortality and population-level health and socioeconomic indicators in the United States: A modeling study. PLoS Med 18(7): e1003693. https://doi.org/10.1371/journal.pmed.1003693

Academic Editor: Richard Turner, PLOS Medicine, UNITED KINGDOM

Received: February 1, 2021

Accepted: June 12, 2021

Published: July 13, 2021

Data Availability Statement: The data underlying the results presented in the study are available from the original data providers: 1. The New York Times Counties: https://github.com/nytimes/covid-19-data/blob/master/live/us-counties.csv 2. CDC PLACES project https://chronicdata.cdc.gov/500-Cities-Places/PLACES-County-Data-GIS-Friendly-Format-2020-releas/i46a-9kgh 3. CDC Social Vulnerability Rankings, 2018 https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html 4. USAFACTS • Cases: https://static.usafacts.org/public/data/covid-19/covid_

Abstract

Background

With the availability of multiple Coronavirus Disease 2019 (COVID-19) vaccines and the predicted shortages in supply for the near future, it is necessary to allocate vaccines in a manner that minimizes severe outcomes, particularly deaths. To date, vaccination strategies in the United States have focused on individual characteristics such as age and occupation. Here, we assess the utility of population-level health and socioeconomic indicators as additional criteria for geographical allocation of vaccines.

Methods and findings

County-level estimates of 14 indicators associated with COVID-19 mortality were extracted from public data sources. Effect estimates of the individual indicators were calculated with univariate models. Presence of spatial autocorrelation was established using Moran's I statistic. Spatial simultaneous autoregressive (SAR) models that account for spatial autocorrelation in response and predictors were used to assess (i) the proportion of variance in county-level COVID-19 mortality that can explained by identified health/socioeconomic indicators (R^2); and (ii) effect estimates of each predictor.

Adjusting for case rates, the selected indicators individually explain 24%–29% of the variability in mortality. Prevalence of chronic kidney disease and proportion of population residing in nursing homes have the highest R^2. Mortality is estimated to increase by 43 per thousand residents (95% CI: 37–49; $p < 0.001$) with a 1% increase in the prevalence of chronic kidney disease and by 39 deaths per thousand (95% CI: 34–44; $p < 0.001$) with 1% increase in population living in nursing homes. SAR models using multiple health/socioeconomic indicators explain 43% of the variability in COVID-19 mortality in US counties, adjusting for case rates. R^2 was found to be not sensitive to the choice of SAR model form. Study limitations include the use of mortality rates that are not age standardized, a spatial adjacency matrix that does not capture human flows among counties, and insufficient accounting for interaction among predictors.

confirmed_usafacts.csv • Deaths: https://static.usafacts.org/public/data/covid-19/covid_deaths_usafacts.csv 5. County Health Rankings • Data: https://www.countyhealthrankings.org/sites/default/files/media/document/analytic_data2020_0.csv • Documentation etc.: https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation 6. US census Bureau Decennial census 2010" • Table P42 from data.census.gov: https://data.census.gov/cedsci/table?q=P42&g=0100000US.050000&tid=DECENNIALSF12010.P42.

Funding: This work is funded in part by a grant from the National Science Foundation (DMS-2027369) and a gift from the Morris-Singer Foundation to JS . The funders had no role in study design, analysis or decision to publish.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: JS and Columbia University disclose ownership of SK Analytics and JS discloses personal fees from BNI (Business Network International). SK consulted for SK Analytics.

Abbreviations: ACS, American Community Survey; BRFSS, Behavioral Risk Factor Surveillance System; CDC, Centers for Disease Control and Prevention; CHR, County Health Rankings; COPD, chronic obstructive pulmonary disease; COVID-19, Coronavirus Disease 2019; HR, hazard ratio; LM, Lagrange Multiplier; LMICs, low- and middle-income countries; OLS, ordinary least square; OR, odds ratio; PLACES, Population Level Analysis and Community Estimates; RR, relative risk; SAR, simultaneous autoregressive; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; SVI, Social Vulnerability Index.

Conclusions

Significant spatial autocorrelation exists in COVID-19 mortality in the US, and population health/socioeconomic indicators account for a considerable variability in county-level mortality. In the context of vaccine rollout in the US and globally, national and subnational estimates of burden of disease could inform optimal geographical allocation of vaccines.

Author summary

Why was this study done?

- We are interested in evaluating strategies for optimal geographical allocation of Coronavirus Disease 2019 (COVID-19) vaccines.
- We hypothesized that health and socioeconomic indicators of a location can be used to model differential risk of COVID-19 mortality, and, hence, inform vaccine prioritization strategies.

What did the researchers do and find?

- Using spatial simultaneous autoregressive (SAR) models and small-area prevalence estimates for US counties, we found that 43% of the variability in COVID-19 mortality can be explained by the considered health and socioeconomic indicators.
- The prevalence of chronic kidney disease and the proportion of population resident in nursing homes were found to have the largest individual effect estimates.
- Strong spatial autocorrelation in COVID-19 mortality and considerable collinearity in COVID-19–associated health conditions were also detected.

What do these findings mean?

- Our findings reiterate that differential risks of severe outcomes from COVID-19 across populations are dependent on the structures and contexts in which outbreaks occur.
- National and subnational socioeconomic indicators and burden of disease estimates can potentially be leveraged to allocate vaccines optimally and reduce severe outcomes from COVID-19.

Introduction

By the end of 2020, the Coronavirus Disease 2019 (COVID-19) pandemic has resulted in 81.5 million documented cases and 1.8 million deaths globally [1]. The US has contributed nearly a

quarter of these cases and has lost 1 in every 1,000 residents to COVID-19 [2]. The outbreak has affected all states in the US but with considerable differences in the trajectory and severity of individual outbreaks. Besides this inter- and intrastate geographical variability, the likelihood of adverse outcomes among those infected is reported to be associated with individual's age, gender, race/ethnicity, and underlying health conditions [3–6]. An estimated 22% of the global population and 28% of the US have 1 or more of the underlying conditions that pose increased risk of severe outcomes from COVID-19 [7].

Early studies on clinical characteristics of severe outcomes from COVID-19 were reported from China [5,8], after the first large outbreak in Wuhan, and concurring estimates were subsequently published from the United Kingdom, France, US, and elsewhere [3,4,9–12]. Guan and colleagues [8] reported that among 1,100 of the earliest laboratory confirmed cases of COVID-19 in China, the presence of comorbidities such as diabetes, hypertension, and chronic obstructive pulmonary disease (COPD) were more prevalent in those with severe outcomes (admission to ICU, requiring mechanical ventilation, or death), along with a slightly elevated risk among men and by now well-established risk with increasing age. Using a larger data sample of 45,000 cases, Deng and colleagues [5] reported that mortality was associated (relative risk (RR) or hazard ratio (HR)) with cardiovascular disease (RR = 6.75, 95% CI = 5.40 to 8.43), hypertension (HR = 4.48, 95% CI = 3.69 to 5.45), diabetes (RR = 4.43, 95% CI = 3.49 to 5.61), and respiratory disease (RR = 3.43, 95% CI = 2.42 to 4.87, $p < 0.001$). In Italy, Grasselli and colleagues [13] reported associations with COPD (HR = 1.68; 95% CI = 1.28 to 2.19), hypercholesterolemia (HR = 1.25; 95% CI = 1.02 to 1.52), and diabetes (HR = 1.18; 95% CI = 1.01 to 1.39). Relatedly, Palmieri and colleagues reported differences in prevalence of comorbidities between younger (<65 years) and older (65+ years) deceased [14] as well as between the first 2 waves (March to May and June to August of 2020) of the pandemic in Italy [15].

A later, more extensive study [9] from the UK linking 17 million cases to 11,000 deaths also found association between COVID-19 deaths and kidney disease (HR = 2.5, 95% CI = 2.3 to 2.7), diabetes (HR = 1.95, 95% CI = 1.8 to 2.1), extreme obesity (HR = 1.9, 95% CI = 1.7 to 2.1), and several other comorbidities. From a pooled analysis of 75 studies from multiple countries, Popkin and colleagues [11] summarized that individuals with obesity are at increased risk of death (odds ratio [OR] = 1.48; 95% CI = 1.22 to 1.80), hospitalization (OR = 2.13; 95% CI = 1.74 to 2.60), and ICU admission (OR = 1.74; 95% CI = 1.46 to 2.08). Based on these findings and the known prevalence of comorbidities that existed in the population before the emergence of the pandemic, the populations at risk of severe COVID-19 outcomes at county level in the US [16] and in several countries have been estimated [7]. Other studies have examined the associations of socioeconomic characteristics including poverty, income, and race/ethnicity [17–19].

Over the past year, public health attempts to reduce transmission largely centered on nonpharmaceutical interventions such as social distancing, face coverings, and hand hygiene. In the US, these interventions have had limited success, and part of this failure stems from their dependence on collective compliance. The recent availability of high-efficacy vaccines gives individuals an additional tool to protect themselves (vaccine supply permitting), and, importantly, does not require cooperation from collective public.

The availability of vaccines also implies an opportunity to refocus our efforts from reducing infections to reducing severe outcomes by prioritizing vaccination for those at a higher risk of severe outcomes. To date, such strategies have been largely guided by individual characteristics such as age and occupation. We hypothesize that population-level characteristics can also guide the optimal allocation and distribution of vaccines geographically. This points to a

potential 2-layered approach of first identifying high-risk communities within which high-risk individuals can be prioritized.

Here, we assess the feasibility of the first part of such an approach and evaluate the extent to which the geographical variability of mortality in US can be explained by population characteristics that predate the epidemic. Our outcome of interest is COVID-19–associated mortality rates at county resolutions, which we attempt to model as a function of population health and socioeconomic indicators. An initial set of indicators associated with COVID-19 mortality as reported in peer-reviewed studies, and data sources for estimates of these indicators were identified. A smaller subset of the variables were selected based on the correlation between the variables and their independent effects on the response.

Conventional regression models assume that observations are independent of one another, which in the case of spatial data translates to assuming observations in nearby locations are no more closely related than those farther away. Given the transmission dynamics of COVID-19, counties nearby are likely to be have similar case and death rates, and spatial dependence rather than spatial independence is a more appropriate assumption. This spatial dependence also extends to health and socioeconomic indicators and potentially latent and unobservable characteristics that effect mortality.

Spatial simultaneous autoregressive (SAR) models offer a parsimonious way to augment basic regression models with spatial dependence between locations [20] and are an extensively studied family of analytical approaches with applications ranging from econometrics, environmental studies, and health sciences [21–23]. In the current study, we first establish the presence of spatial autocorrelation in the response and explanatory variables, thus motivating the need for spatial models. We apply 3 forms of SAR models, show that they explain a greater proportion of the variability in mortality than linear models, and report effect estimates from each.

Data and methods

County-level indicators of population's health and social status were retrieved from public sources including the US census and large population surveys. In cases where the survey data are not available at county resolutions, data from prior studies on small-area estimates were used. We tried to limit the number of source dependencies, and, when alternative estimates were available from multiple sources, we preferred estimates from the US Centers for Disease Control and Prevention (CDC). See Table 1 for a list of sources and descriptions for each variable; Fig 1 presents summary statistics.

The New York Times

Counts for cumulative cases and deaths through December 31, 2020 were retrieved from The New York Times public repository [24]. These data included both confirmed and probable cases and deaths at the US county level and is based on Times' monitoring and analyses of news conferences, data releases, and communications with public officials. The determination of cases and deaths as either confirmed or probable is made per definitions laid out in the position statement of the Council of State and Territorial Epidemiologists [25]. But as the application can vary across local agencies, here, we treat both confirmed and probable categories identically and use total cases and deaths. Case and death rates as a proportion of residents are based on county population estimates from the American Community Survey (ACS) 2014 to 2018 [26].

County-specific data for the 5 counties in New York City were retrieved from USAFACTS [27] as the Times' data source was found to combine counts for these 5 counties into a single entity. Fig 2 shows maps of reported county case and death rates.

Table 1. Descriptions and sources for variables included in the study.

Variable	Source	Description; primary source
Deaths	The New York Times [24], USAFACTS[27]	Cumulative COVID-19 confirmed and probable deaths through December 31, 2020; per thousand residents
Cases	The New York Times, USAFACTS	Cumulative COVID-19 confirmed and probable cases through December 31, 2020; per 100,000 residents
Obesity	PLACES [28]	Proportion of residents 18+ years of age with calculated BMI $\geq$30 kg/m^2, based on self-reported weight and height; BRFSS [29]
Diabetes	PLACES	Proportion of residents 18+ years of age who report being told by a doctor/nurse/other health professional that they have type 1 or type 2 diabetes; BRFSS
CKD	PLACES	Proportion of residents 18+ years of age who report being told by a doctor/nurse/other health professional that they have kidney disease; BRFSS
CHD	PLACES	Proportion of residents 18+ years of age who report being told by a doctor/nurse/other health professional that they have angina or coronary heart disease; BRFSS
COPD	PLACES	Proportion of residents 18+ years of age who report being told by a doctor/nurse/other health professional that they have COPD, emphysema, or chronic bronchitis; BRFSS
High cholesterol	PLACES	Proportion of residents 18+ years of age who report being told by a doctor/nurse/other health professional that they have high cholesterol; BRFSS
High blood pressure	PLACES	Proportion of residents 18+ years of age who report being told by a doctor/nurse/other health professional that they have high blood pressure; BRFSS
Uninsured	PLACES	Proportion of residents 18–64 years of age who report having no health insurance coverage
Population density	SVI [33]	Number of residents per square mile; Census Cartographic Boundary File—U.S. Tracts 2018 [40]
Income	SVI	Median per capita income (in US$100,000); ACS, 2014–2018 (5 years) [26]
Elderly	SVI	Proportion of residents 65+ years of age; ACS, 2014–2018 (5 years)
Group quarters—nursing	US 2010 Census	Proportion of residents living in nursing/skilled nursing facilities; P042
Inequality	CHR [36]	Ratio of household income at 80th percentile with income at 20th percentile; ACS, 2014–2018 (5-years)
Resident diversity	CHR	Proportion of non-white resident to white residents; ACS, 2014–2018 (5 years)

BRFSS, Behavioral Risk Factor Surveillance System; CHD, chronic heart disease; CHR, County Health Rankings; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; COVID, Coronavirus Disease 2019; PLACES: Population Level Analysis and Community Estimates; SVI: Social Vulnerability Index.

https://doi.org/10.1371/journal.pmed.1003693.t001

Population Level Analysis and Community Estimates (PLACES)

From the PLACES study [29], a collaboration between the CDC and Robert Wood Johnson Foundation, estimates for population-level health and behavioral indicators were retrieved. These small-area estimates of population health outcomes across the US at county resolutions were generated using data collected through the Behavioral Risk Factor Surveillance System (BRFSS) [30], the US decennial 2010 census and the ACS, following a multilevel regression and post-stratification approach [31,32].

Of the 27 indicators available in PLACES, we extracted 5 measures of population-level prevalence of health conditions that are reported to have individual level associations with

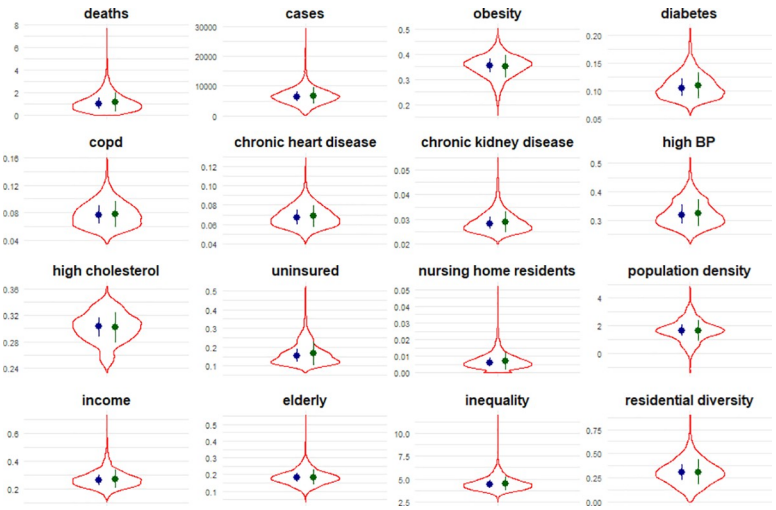

Fig 1. Violin plots of distribution for each variable among counties in the US, along with median (interquartile range) in blue and mean (standard deviation) in green.

https://doi.org/10.1371/journal.pmed.1003693.g001

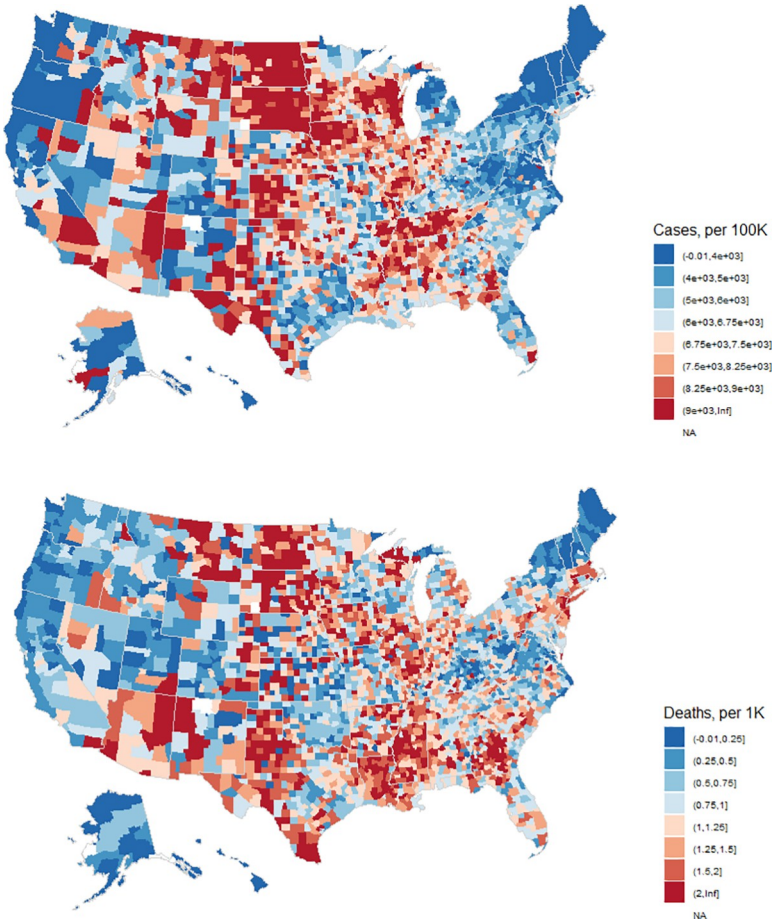

Fig 2. COVID-19 cases (per 100,000 residents) and deaths (per 1,000 residents) in US counties through December 31, 2020. Maps generated with *usmap* R package [28]. COVID-19, Coronavirus Disease 2019.

https://doi.org/10.1371/journal.pmed.1003693.g002

COVID-19 outcomes, namely obesity, diabetes, COPD, and chronic heart and kidney diseases. In addition, 3 related health indicators, the prevalence of high blood pressure and high cholesterol, and proportion of residents uninsured were also included.

Social Vulnerability Index

CDC's Social Vulnerability Index (SVI) is a measure of a county's relative vulnerability to hazardous events [33,34] and is intended to help public officials and planners better prepare for such events. Overall, county ranks are based on 15 socioeconomic indicators collected in the ACS. Three of the factors in the SVI, namely county population density, median per capita income, and proportion of the population that is older than 65 years of age, are hypothesized to be associated with COVID-19 mortality [17,18,35,36]. As association between the other variables in SVI and COVID-19 is uncertain, we limited inclusion to the raw estimates of these 3 variables and ignore the other variables in SVI and the overall index.

County Health Rankings

Two additional variables derived from the ACS 2014 to 2018 and available through the County Health Rankings (CHR) [37] are hypothesized to be measures of socioeconomic disparities in a county and included in this study: ratio of the 80th percentile income to 20th percentile as a measure of income inequality, and the proportion of non-white to white residents as a measure of racial diversity. We observed that estimates for these variables in a small percentage (approximately 1.5%) of counties were missing and used the following 3-step process to impute missing values: (a) the mean of neighboring (defined in later sections) counties that have estimates; (b) if there are no neighbors with estimates, the median of all counties in the state for which estimates are available; and (c) if estimates are missing for all counties in a state, the median across all counties in the US for which estimates are available.

US 2010 Census

It has also been reported that COVID-19 clusters occur in facilities in which people live in group quarters, where the increased vulnerability can result from either the living conditions in such facilities (difficulty to social distance in correctional facilities or on college campuses, for example) or the characteristics of the residents (elderly in nursing homes with underlying health conditions)[38,39]. As mortality from COVID-19 is known to be less likely in younger populations, we focused instead on elderly living in group quarters. An estimate of proportion of the population living in nursing homes or facilities with skilled nursing in each county was included in this analysis (Table 1).

As all data used here are routinely collected aggregate surveillance data, no ethics approval was deemed necessary for this study. A prespecified analysis plan has not been filed; a Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist is available as Supporting information (S1 STROBE Checklist).

Methods overview

We first built linear univariate models for each predictor with county-level COVID-19 mortality as outcome, adjusting for county case rates. These models inform both the individual effects and the proportion of variance in mortality explained by each of these predictors. We followed this with a linear multivariate model, again adjusting for case rates. In both univariate and multivariate models, observational independence is inappropriate because of spatial autocorrelation

in both the response and predictors. We verify this by standard tests on the residual of the multivariate model. We finally build spatial SAR models and report effect estimates.

Spatial weight matrix and spatial autocorrelation

As introduced in an earlier section, a key assumption in standard ordinary least square (OLS) regression models is the independence of observations that does not hold because COVID-19 cases and deaths in a county are related to cases and deaths in other counties (spatial dependence) and often counties adjacent to it (spatial autocorrelation). Models that do not account for spatial dependence and autocorrelation are shown to have inflated type I errors [23,40].

To establish adjacency of counties in the US, we define a simple spatial *n x n* matrix, **W**, using shape files that list county boundaries as an ordered set of geocoded reference points [41]. County adjacency is defined by queen congruity (at least 1 shared boundary point), and the spatial weight matrix is row standardized, i.e., for each county *i*, the weight of link to county *j*, $w_{ij,}$ is the inverse of the number of neighbors of *i*, if *j* is adjacent to *i*, and 0 otherwise; $\sum_j w_{ij} = 1$. A county is assumed to not be a neighbor of itself, i.e., $w_{ij} = 0$ when $i = j$.

Moran's I [42,43], a commonly used measure of global spatial autocorrelation, is calculated as follows:

$$I = \frac{n * \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_i \sum_j w_{ij})(\sum_{i=1}^{n}(x_i - \bar{x})^2)},$$

where *n* is the number of counties, x_i is the variable of interest for county *i*, $\bar{x}$ is the mean across all counties, and w_{ij} is as defined by the spatial weight matrix, **W**. Here, as ***W*** is row standardized, $\sum_i \sum_j w_{ij} = n$ and the above equation can be simplified. The significance of the statistic was tested under the randomization assumption, i.e., x_i are draws from a random distribution and there is no spatial association. A related measure to identify specific regions within the study region that exhibit spatial autocorrelation, the Local Moran's I, was also estimated. Fig 3 shows Moran's I and counties with significant [44] Local Moran's I for each predictor and outcome.

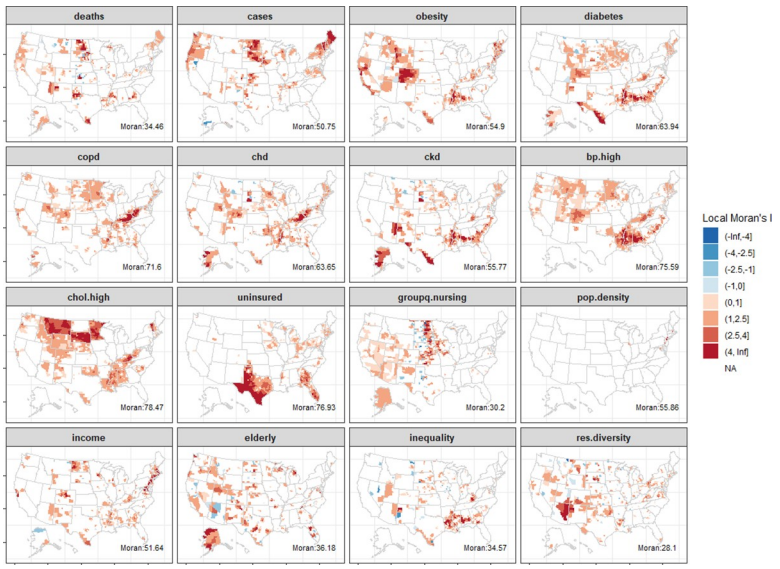

Fig 3. Local Moran's I statistic for spatial autocorrelation for all measures and outcome. Only counties where the statistic is significant ($p < 0.05$) are shown. Significance is tested under Pr[I–E(I)/Var(I)] as given by Anselin [44]. Global Moran's I statistic is denoted by the label in each subpanel and found to be statistically significant for all variables. Maps generated with *usmap* R package [28].

https://doi.org/10.1371/journal.pmed.1003693.g003

We were also interested in determining whether spatial autocorrelation, if present, resided in the response or in the residual, as this also informs the choice of the spatial model. To identify this, we used robust Lagrange Multiplier (LM) tests that can detect possible autocorrelated residuals in the presence of an omitted lagged response and vice versa [45,46]. The statistics reported here are from implementations of these tests in the *spdep* [43,47] R [48] library.

Variable pruning

As the variables selected for inclusion are related, we calculated Spearman correlation between pairs of variables (Fig 4) and found some of the variables to be very highly correlated. Hence, it would not be appropriate to include these pairs together in models. We used the results of the univariate analysis to aid variable selection by only retaining those variables that have a correlation of less than 0.75 with variables of a higher R^2. This led to the elimination of 5 variables—prevalence indicators for diabetes, heart disease, high blood pressure (all highly collinear with kidney disease), high cholesterol (collinear with COPD), and median per capita income. The linear multivariate model and the spatial models were built using this smaller set of predictors ($n = 9$).

Spatial simultaneous autoregressive models

The general form of an autoregressive model in spatial statistics is given by [20,23,49]:

$$y = X\beta + \rho Wy + \lambda Wu + \varepsilon,$$

where y is a $n \times 1$ vector of the response variable, $\mathbf{X}$ is a $n \times k$ matrix of k predictors for n counties, $\mathbf{W}$ is the $n \times n$ spatial weight matrix, ρ is the SAR *lag* coefficient and λ the spatial *error* coefficient, and β, u the coefficient and error vectors, respectively. When $\lambda = 0$, the autoregressive process is assumed to occur in the response only (captured by $\rho\mathbf{W}$) and the model is referred to as a spatial lag model. When $\rho = 0$, the autoregressive process is assumed to occur only in the errors (captured by $\lambda\mathbf{W}$), and the model referred to as spatial error model. Model implementations are per *spatialreg* [47,49,50] library in R.

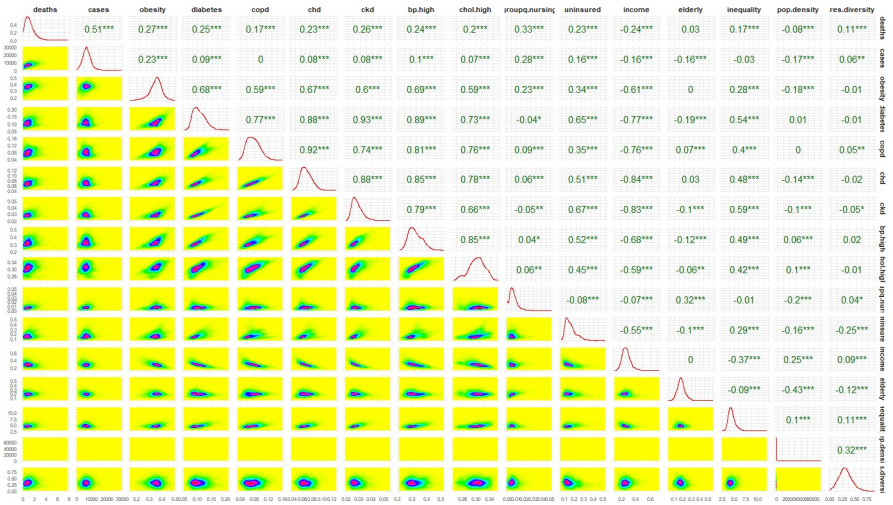

Fig 4. Pairwise surface plots (below diagonal), Spearman correlation (above diagonal), and density (diagonal) of outcome and measures used in the study. * indicates level of statistical significance of the correlation: $p < 0.001$ (***); $0.001 \leq p < 0.01$ (**); $0.01 \leq p < 0.05$ (*).

https://doi.org/10.1371/journal.pmed.1003693.g004

Results

Results from the univariate analysis indicate that the selected variables individually explain 24% to 29% of the variability in mortality, adjusting for case rates. Mortality is estimated to increase by 43 per thousand residents (95% CI: 37 to 49; $p < 0.001$) for every 1% increase in prevalence of chronic kidney disease and by 10.4 (95% CI: 8 to 13; $p < 0.001$) for chronic heart disease, 7.4 (95% CI: 6 to 8; $p < 0.001$) for diabetes, 4.4 (95% CI: 3 to 5.8; $p < 0.001$) for COPD, 3.7 (95% CI: 2.6 to 5.8; $p < 0.001$) for high cholesterol, 2.8 (95% CI: 2.2 to 3.3; $p < 0.001$) for high blood pressure, and 2.6 (95% CI: 2 to 3.2; $p < 0.001$) for obesity prevalence, respectively (Fig 5). These health indicators also explain 28%, 25.5%, 27.5%, 24.6%, 24.6%, 25.9%, and 25.3% of the variability, respectively.

Among socioeconomic indicators, the largest association was seen with the nursing home variable (adjusted R^2: 29%) with an estimated increase of 39 deaths per thousand (95% CI: 34 to 44; $p < 0.001$) for every 1% increase in percent living in nursing homes. Mortality rates are estimated to increase by 2.8 (95% CI: 2.3 to 3.4; $p < 0.001$) and 2.4 (95% CI: 2 to 2.9; $p < 0.001$) for each 1% increase in percentage of the population who are elderly (65+ years) and uninsured 18 to 64 year olds, respectively. In contrast, mortality rate is estimated to decrease by 1.5 (95% CI: 1.05 to 1.87; $p < 0.001$) for every thousand dollar increase in per capita income. On average, the R^2 estimates for socioeconomic indicators are lower than for health indicators.

Following variable pruning to correct for collinearity, the multivariate model explained 38% of the variability in mortality with a few changes in effect estimates. Obesity's association is not statistically significant in the presence of kidney disease, and COPD's association is counterintuitively negative (Table 2).

Moran's I test for spatial autocorrelation in residuals of the above model was found to be statistically significant (18.4, $p < 0.001$). Both robust LM tests were found to be significant indicating possible autocorrelation in both the error (28.7, $p < 0.001$) and response (33.5,

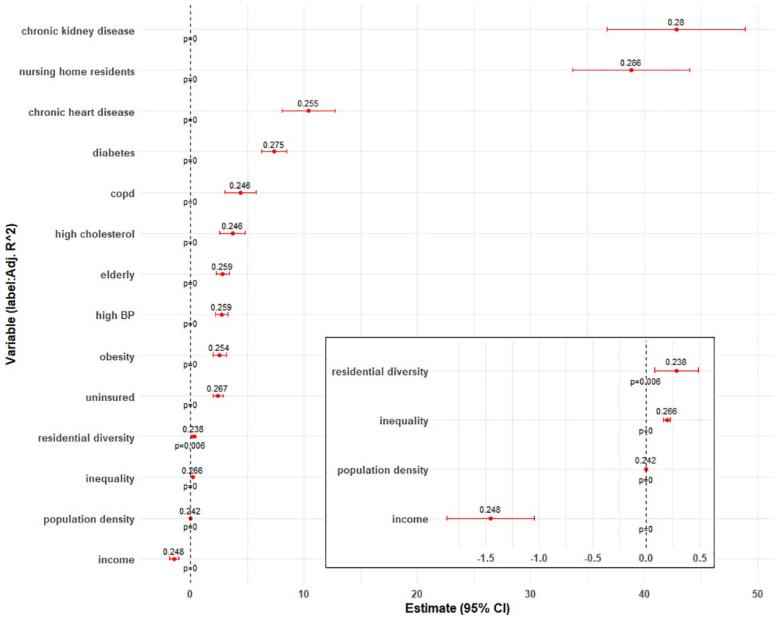

Fig 5. Estimates (95% CI) of health and socioeconomic indicators in a linear univariate model with death rate as outcome and adjusting for COVID-19 case rates. Labels indicate adjusted R^2. Inset magnifies select variables of smaller estimates. COVID, Coronavirus Disease 2019.

https://doi.org/10.1371/journal.pmed.1003693.g005

Table 2. Results of multivariate analysis with linear model, adjusting for case rate.

	Estimate	95% CI	*p*
(Intercept)	−2.007	(−2.27, −1.74)	<0.001
Cases	1.26E-4	(1.16E-4, 1.36E-4)	<0.001
Obesity	−0.654	(−1.43, 0.13)	0.101
COPD	−4.681	(−6.64, −2.72)	<0.001
CKD	53.48	(41.11, 65.84)	<0.001
Nursing home residents	41.66	(36.27, 47.08)	<0.001
Uninsured	1.479	(0.92, 2.03)	<0.001
Elderly	2.449	(1.87, 3.03)	<0.001
Inequality	0.046	(0.005, 0.087)	0.029
Population density	3.4E-5	(1.9E-5, 4.8E-5)	<0.001
Residential diversity	0.557	(0.36, 0.75)	<0.001

Adjusted R^2 = 0.3812; F-statistic = 194 (*p*-value: < 0.001).

CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease.

https://doi.org/10.1371/journal.pmed.1003693.t002

$p < 0.001$). Hence, 3 model forms, the general SAR model, spatial lag, and spatial error models, were attempted.

The proportion of variability explained by the SAR models is about 14% higher than the linear model (Fig 6). The spatial error model had an Nagelkerke R^2 [51] of 43.5% with an estimated autocorrelation error coefficient (λ) of 0.418 (95% CI: 0.37 to 0.46; $p < 0.001$). The spatial lag model and the general model were observed to have an R^2 nearly identical to that of the error model. The autocorrelation coefficient in response (ρ) was found to be 0.347 (95% CI: 0.31 to 0.39; $p < 0.001$) for the spatial model, but when both coefficients were estimated simultaneously in a general model, the lag coefficient was found to be not significant: λ = 0.336 (95% CI: 0.244 to 0.429; $p < 0.001$); ρ = 0.083 (95% CI: −0.007 to 0.174; $p = 0.07$).

Fig 6 demonstrates that the effect estimates of SAR models are generally smaller than the linear model, i.e., accounting for spatial autocorrelation reduces the magnitude of associations.

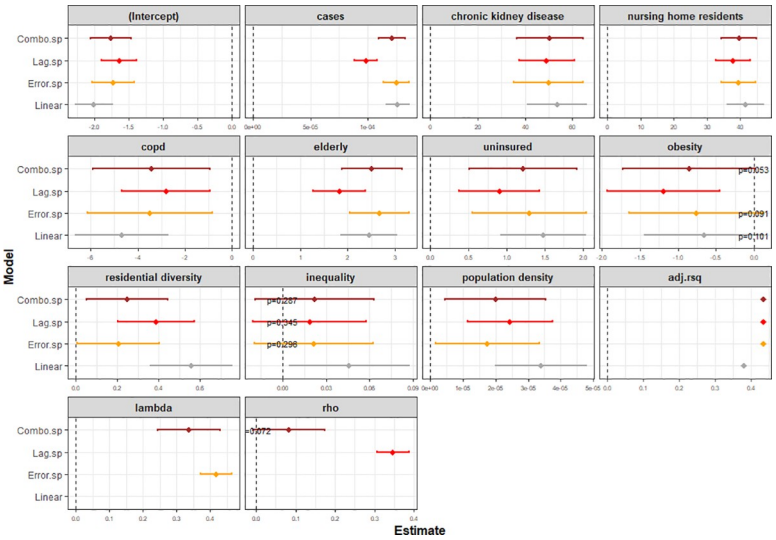

Fig 6. Variables estimates with linear and 3 spatial regression models. *p*-values indicated when $p > 0.05$. *adj.rsq*: adjusted R^2; *lambda* and *rho* denote the spatial error and spatial lag coefficients, respectively.

https://doi.org/10.1371/journal.pmed.1003693.g006

Focusing on the spatial lag model, the most strongly associated health and socioeconomic indicators are the prevalence of chronic kidney disease (49; 95% CI: 37 to 61; $p < 0.001$) and proportion of nursing home residents (38; 95% CI: 33 to 43; $p < 0.001$), respectively, consistent with the univariate analysis. The negative association of COPD seen in the univariate model is also observed with the spatial models. On the other hand, inequality and obesity had significant association in the univariate model, but, after accounting for spatial autocorrelation and in the presence of other indicators, their association tends to be no longer significant ($p > 0.05$).

The Global Moran's test on the residuals of all 3 models found no significant spatial autocorrelation ($p > 0.05$). Fig 7 shows the spatial lag model's fit and residuals. To test for sensitivity of models' R^2 to the variable pruning method, we additionally subset variables using alternative spearman correlation thresholds of 0.5, 0.65, and 0.85 and built linear and spatial models with each. Fig 8 shows that R^2 was not sensitive to the value of threshold, and the spatial models have a consistently higher R^2 than the linear model.

Discussion

We have built models to estimate COVID-19 mortality rates for given case rates and population health and socioeconomic characteristics. Our results indicate that, together, these indicators can explain 43% of the variability in US county mortality rates, when spatial autocorrelation is accounted for. We found that among health indicators considered, the prevalence of chronic kidney disease, and among socioeconomic indicators, the proportion living in nursing homes have the largest associations with mortality.

The choice and timeliness of control strategies in response to an outbreak do affect its progress and caseload. Our findings here show that differential risks of severe outcomes from COVID-19 across populations can be in part estimated from the structures and contexts in which the outbreak occurs, for example, a population's quality of health, its access to healthcare, and the disparities therein. With the availability of vaccines, these population-level indicators can serve as criteria for prioritizing geographical allocation of vaccines.

These findings may also be relevant to low- and middle-income countries (LMICs). It has been reported that almost all of the Pfizer-BioNTech and Moderna vaccine doses to be manufactured through the end of 2021 have been purchased and are reserved for distribution in the US, Canada, UK, and the European Union [52,53]. Of the 42 countries that have rolled out vaccines by early January 2021, only 6 are middle-income countries, and none are low-income countries [54]. The COVAX initiative with participation from governments of several LMIC countries, WHO, and partner nongovernmental organizations aims to achieve equitable and affordable access to vaccines globally through a common vaccine purchase and allocation framework [55]. When allocation decisions need to span multiple countries, national and subnational socioeconomic indicators and burden of disease estimates can potentially be leveraged to reduce overall risk of severe outcomes from COVID-19 as our findings demonstrate.

This study has a few limitations. Case and death counts were retrieved a week after the end of the study period. Given the lags in data reporting, particularly with deaths, events occurring at the end of the study period may not have been recorded, and the rates used are underestimates. Similarly, the outcomes may not yet be known for cases recorded near the end of the study period. Additionally, the case and mortality rates used in this study are crude rates that do not account for differences in age distribution among county populations. County rates standardized to US national age distribution would be more appropriate, but as age-stratified case and deaths counts at county scales are not publicly available, age standardization has not been possible. However, a supplementary analysis (S1 Text and Figs A–D in S1 Text) using

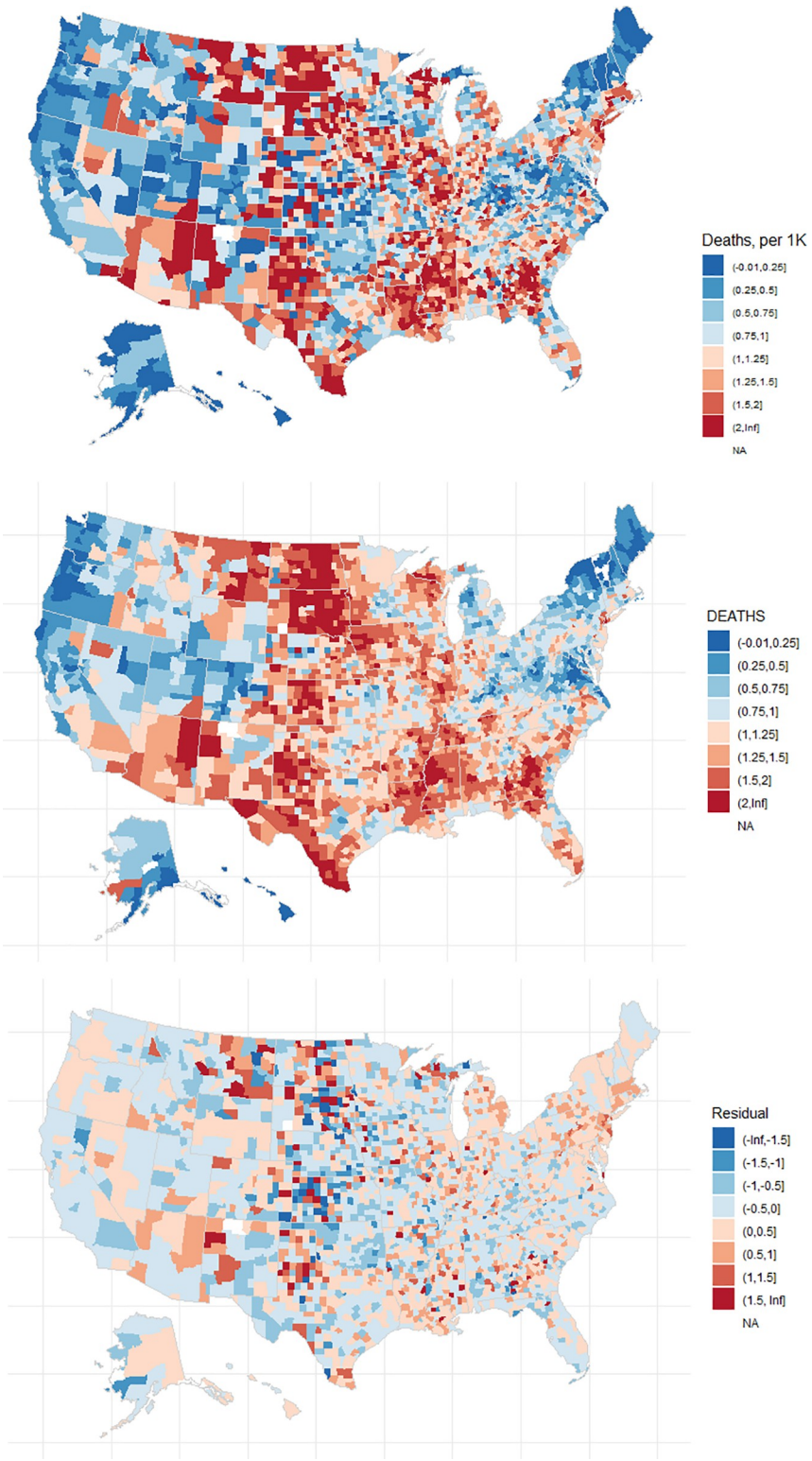

Fig 7. Observed death rate (as in Fig 2), model fit, and residual of the spatial lag model. Maps generated with *usmap* R package [28].

https://doi.org/10.1371/journal.pmed.1003693.g007

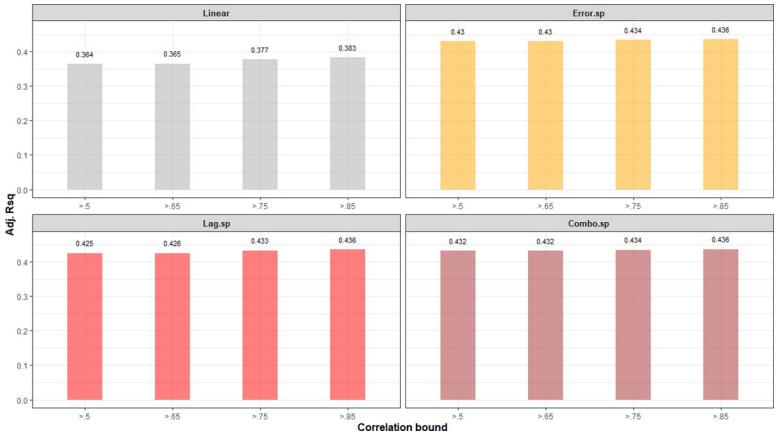

Fig 8. Sensitivity of adjusted R^2 to Spearman correlation threshold used in variable pruning.

https://doi.org/10.1371/journal.pmed.1003693.g008

age-specific rates at the state level as proxy for county rates indicate that the results presented here may be robust except for the 2 indicators directly related to age (proportion of populations in nursing home and proportion 65+ years of age).

Second, the adjacency based spatial weight matrix that was used in this study does not sufficiently capture the spread of COVID-19. Cases that occur in a county are not only correlated with those in counties geographically adjacent to it, but also with counties with which it has strong population mixing; for example, counties with metropolitan centers into which commuters travel from the suburbs or counties with major airports. Spatial weight matrices that capture mobility patterns may be more appropriate and lead to better spatial models. Similarly, methods that can explicitly account for spatial autocorrelation in predictors remain to be explored.

Finally, the model structure presented may not be parsimonious in the number of predictors. Although we dropped a third of the predictors initially considered (to correct observed collinearity), model forms with a smaller subset of independent variables may yield near identical R^2 and need to be explored. This is also belied by the lack of significance of some of the predictors included in the spatial models. One approach could start with a minimal set of predictors and incrementally add predictors, while evaluating goodness of the resulting model in each iteration and terminating when the improvement is below a threshold. Similarly, the variable pruning discussed above is ad hoc; the variables included in the model may be interchangeable with those discarded with only marginal change in model performance. Exploration of interaction between indicators and inclusion of significant interactions in the SAR models could be a potential extension to the analysis presented.

Supporting information

S1 STROBE Checklist. Checklist for STROBE guidelines for observational studies. STROBE, Strengthening the Reporting of Observational Studies in Epidemiology.
(DOC)

S1 Text. Supplementary analysis to check for changes in model effect estimates with approximate age-standardized mortality rates. Fig A: Distribution of crude COVID-19 mortality rates (deaths per thousand residents) for each of 8 age groups in US states. Each data point indicates a US state, and the bounded region indicates the distribution. Note that the y-axis is on $\log_{10}$ scale. **Fig B:** Scatter plot of crude (y-axis) and age-standardized (x-axis)

mortality rates in US states. A state below the diagonal (black dashed line) indicates that the mortality rate increases when standardized. **Fig C:** Effect estimates (95% CI) with a linear univariate linear model using crude mortality rate (red) and age-stratified mortality rate (in green) as response variables. Labels indicate adjusted R^2. Inset magnifies select variables of smaller estimates. All estimates are significant ($p < 0.05$). The slight difference between the effect estimates with crude rates here and Fig 5 of the main text is due to the use of different data sources (The New York Times in the main text and NCHS provisional counts here). **Fig D:** Variables estimates of spatial lag models built using crude (orange) and age-standardized (brown) rates as the response variable. *p*-values indicated when $p > 0.05$. *adj.rsq*: adjusted R^2; *rho*: spatial lag coefficients. COVID, Coronavirus Disease 2019.
(DOCX)

Author Contributions

Conceptualization: Sasikiran Kandula.

Data curation: Sasikiran Kandula.

Formal analysis: Sasikiran Kandula, Jeffrey Shaman.

Funding acquisition: Jeffrey Shaman.

Investigation: Sasikiran Kandula.

Methodology: Sasikiran Kandula, Jeffrey Shaman.

Software: Sasikiran Kandula.

Supervision: Jeffrey Shaman.

Validation: Sasikiran Kandula.

Visualization: Sasikiran Kandula.

Writing – original draft: Sasikiran Kandula.

Writing – review & editing: Sasikiran Kandula, Jeffrey Shaman.

References

1. WHO Coronavirus Disease (COVID-19) Dashboard 2020 [1/11/2021]. Available from: https://covid19.who.int/.
2. CDC Case Task Force. United States COVID-19 Cases and Deaths by State over Time 2021 [01/11/2021]. Available from: https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36.
3. Chow N, Fleming-Dutra K, Gierke R, Hall A, Hughes M, Pilishvili T, et al. Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019—United States, February 12–March 28, 2020. 2020.
4. CDC. Coronavirus disease 2019 (COVID-19). Evidence used to update the list of underlying medical conditions that increase a person's risk of severe illness from COVID-19. 2020.
5. Deng G, Yin M, Chen X, Zeng F. Clinical determinants for fatality of 44,672 patients with COVID-19. Crit Care. 2020; 24(1):1–3. https://doi.org/10.1186/s13054-019-2683-3 PMID: 31898531
6. Stokes EK, Zambrano LD, Anderson KN, Marder EP, Raz KM, Felix SEB, et al. Coronavirus disease 2019 case surveillance—United States, January 22–May 30, 2020. Morb Mortal Wkly Rep. 2020; 69 (24):759. https://doi.org/10.15585/mmwr.mm6924e2 PMID: 32555134
7. Clark A, Jit M, Warren-Gash C, Guthrie B, Wang HH, Mercer SW, et al. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. Lancet Glob Health. 2020; 8(8):e1003–e17. https://doi.org/10.1016/S2214-109X(20)30264-3 PMID: 32553130

8. Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020; 382(18):1708–20. https://doi.org/10.1056/NEJMoa2002032 PMID: 32109013

9. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020; 584(7821):430–6. https://doi.org/10.1038/s41586-020-2521-4 PMID: 32640463

10. Simonnet A, Chetboun M, Poissy J, Raverdy V, Noulette J, Duhamel A, et al. High prevalence of obesity in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) requiring invasive mechanical ventilation. Obesity. 2020; 28(7):1195–9. https://doi.org/10.1002/oby.22831 PMID: 32271993

11. Popkin BM, Du S, Green WD, Beck MA, Algaith T, Herbst CH, et al. Individuals with obesity and COVID-19: A global perspective on the epidemiology and biological relationships. Obes Rev. 2020; 21 (11):e13128. https://doi.org/10.1111/obr.13128 PMID: 32845580

12. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. BMJ. 2020; 369:m1985. https://doi.org/10.1136/bmj.m1985 PMID: 32444460

13. Grasselli G, Greco M, Zanella A, Albano G, Antonelli M, Bellani G, et al. Risk factors associated with mortality among patients with COVID-19 in intensive care units in Lombardy, Italy. JAMA Intern Med. 2020; 180(10):1345–55. https://doi.org/10.1001/jamainternmed.2020.3539 PMID: 32667669

14. Palmieri L, Vanacore N, Donfrancesco C, Lo Noce C, Canevelli M, Punzo O, et al. Clinical characteristics of hospitalized individuals dying with COVID-19 by age group in Italy. J Gerontol A Biol Sci Med Sci. 2020; 75(9):1796–800. https://doi.org/10.1093/gerona/glaa146 PMID: 32506122

15. Palmieri L, Palmer K, Noce CL, Meli P, Giuliano M, Floridia M, et al. Differences in the clinical characteristics of COVID-19 patients who died in hospital during different phases of the pandemic: national data from Italy. Aging Clin Exp Res. 2020:1–7. https://doi.org/10.1007/s40520-020-01764-0 PMID: 33345291

16. Razzaghi H, Wang Y, Lu H, Marshall KE, Dowling NF, Paz-Bailey G, et al. Estimated county-level prevalence of selected underlying medical conditions associated with increased risk for severe COVID-19 illness—United States, 2018. Morb Mortal Wkly Rep. 2020; 69(29):945. https://doi.org/10.15585/mmwr.mm6929a1 PMID: 32701937

17. Adhikari S, Pantaleo NP, Feldman JM, Ogedegbe O, Thorpe L, Troxel AB. Assessment of community-level disparities in coronavirus disease 2019 (COVID-19) infections and deaths in large US metropolitan areas. JAMA Netw Open. 2020; 3(7):e2016938–e. https://doi.org/10.1001/jamanetworkopen.2020.16938 PMID: 32721027

18. Baena-Díez JM, Barroso M, Cordeiro-Coelho SI, Díaz JL, Grau M. Impact of COVID-19 outbreak by income: hitting hardest the most deprived. J Public Health. 2020; 42(4):698–703. https://doi.org/10.1093/pubmed/fdaa136 PMID: 32776102

19. Townsend MJ, Kyle TK, Stanford FC. Outcomes of COVID-19: disparities in obesity and by ethnicity/race. Nat Publ Group. 2020:1807–9.

20. LeSage JP. An introduction to spatial econometrics. 2008(123):19–44.

21. Fischer MM, Getis A. Handbook of applied spatial analysis: software tools, methods and applications: Springer Science & Business Media; 2009.

22. LeSage JP, Pace RK. Spatial econometric models. Handbook of applied spatial analysis: Springer; 2010. p. 355–76.

23. Dormann CF, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography. 2007; 30 (5):609–28.

24. Coronavirus (Covid-19) Data in the United States [Internet]. 2020 [cited 1/11/2021]. Available from: https://github.com/nytimes/covid-19-data.

25. Council of State and Territorial Epidemiologists. Coronavirus Disease 2019 (COVID-19) 2020 Interim Case Definition [1/11/2021]. Available from: https://wwwn.cdc.gov/nndss/conditions/coronavirus-disease-2019-covid-19/case-definition/2020/08/05/.

26. U.S. Census Bureau. 2014–2018 American Community Survey 5-year Public Use Microdata Samples 2019 [1/7/2021]. Available from: https://www.census.gov/programs-surveys/acs/microdata/access.2018.html.

27. USAFACTS. Coronavirus data resource [1/12/2021]. Available from: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

28. Di Lorenzo P. usmap: US maps including Alaska and Hawaii. 2018.

29. Centers for Disease Control and Prevention. PLACES Project [1/11/2021]. Available from: https://www.cdc.gov/places.

30. Remington PL, Smith MY, Williamson DF, Anda RF, Gentry EM, Hogelin GC. Design, characteristics, and usefulness of state-based behavioral risk factor surveillance: 1981–87. Public Health Rep. 1988; 103(4):366. PMID: 2841712

31. Zhang X, Holt JB, Lu H, Wheaton AG, Ford ES, Greenlund KJ, et al. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. Am J Epidemiol. 2014; 179(8):1025–33. https://doi.org/10.1093/aje/kwu018 PMID: 24598867

32. Zhang X, Holt JB, Yun S, Lu H, Greenlund KJ, Croft JB. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. Am J Epidemiol. 2015; 182(2):127–37. https://doi.org/10.1093/aje/kwv002 PMID: 25957312

33. Flanagan BE, Gregory EW, Hallisey EJ, Heitgerd JL, Lewis BA. A social vulnerability index for disaster management. J Homel Security Emerg Manage. 2011; 8(1).

34. Centers for Disease Control and Prevention. Social Vulnerability Index [1/11/2021]. Available from: https://www.atsdr.cdc.gov/placeandhealth/svi/index.html.

35. Lamb MR, Kandula S, Shaman J. Differential COVID-19 case positivity in New York City neighborhoods: Socioeconomic factors and mobility. Influenza Other Respir Viruses. 2021; 15(2):209–17. https://doi.org/10.1111/irv.12816 PMID: 33280263

36. Sannigrahi S, Pilla F, Basu B, Basu AS, Molter A. Examining the association between socio-demographic composition and COVID-19 fatalities in the European region using spatial regression approach. Sustain Cities Soc. 2020; 62:102418. https://doi.org/10.1016/j.scs.2020.102418 PMID: 32834939

37. Remington PL, Catlin BB, Gennuso KP. The county health rankings: rationale and methods. Popul Health Metrics. 2015; 13(1):11. https://doi.org/10.1186/s12963-015-0044-2 PMID: 25931988

38. Abrams HR, Loomer L, Gandhi A, Grabowski DC. Characteristics of US Nursing Homes with COVID-19 Cases. J Am Geriatr Soc. 2020; 68(8):653–6.

39. Barnett ML, Grabowski DC. Nursing homes are ground zero for COVID-19 pandemic. JAMA Health Forum. 2020. https://doi.org/10.1001/jamahealthforum.2020.0961 PMID: 33728420

40. Anselin L. Under the hood issues in the specification and interpretation of spatial regression models. Agric Econ. 2002; 27(3):247–67.

41. U.S. Census Bureau. Catrographic Boundary Files—Shapefile 2018 [1/11/2021]. Available from: https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html.

42. Cliff AD, Ord JK. Spatial processes: models & applications: Taylor & Francis; 1981.

43. Bivand RS, Wong DW. Comparing implementations of global and local indicators of spatial association. TEST. 2018; 27(3):716–48.

44. Anselin L. Local indicators of spatial association—LISA. Geogr Anal. 1995; 27(2):93–115.

45. Anselin L. Spatial econometrics: methods and models: Springer Science & Business Media; 2013.

46. Anselin L, Bera AK, Florax R, Yoon MJ. Simple diagnostic tests for spatial dependence. Regional science urban economics. 1996; 26(1):77–104.

47. Bivand RS, Pebesma EJ, Gómez-Rubio V, Pebesma EJ. Applied spatial data analysis with R: Springer; 2008. https://doi.org/10.1016/j.trstmh.2008.11.012 PMID: 19117584

48. R Core Team. A language environment for statistical computing. R Foundation for Statistical Computing: version. 2018; 5(0):3.

49. Bivand R, Piras G. Comparing implementations of estimation methods for spatial econometrics. J Stat Softw. 2015; 63(18).

50. Bivand R, Hauke J, Kossowski T. Computing the J acobian in G aussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods. Geogr Anal. 2013; 45(2):150–79.

51. Nagelkerke NJ. A note on a general definition of the coefficient of determination. Biometrika. 1991; 78 (3):691–2.

52. LaFraniere S, Thomas K, Weiland N. Trump administration officials passed when Pfizer offered months ago to sell the U.S. more vaccine doses. The New York Times. 2020. 12/07/2020.

53. Mullard A. How COVID vaccines are being divvied up around the world. 2020 11/30/2020.

54. Director-General's opening remarks at the media briefing on COVID-19–8 January 2021 [Internet]. 1/8/2021. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-8-january-2021.

55. COVAX. GAVI Alliance [1/15/2021]. Available from: https://www.gavi.org/covax-facility.