J Urban Health (2021) 98:205–210
https://doi.org/10.1007/s11524-021-00514-5

Predictors of COVID-19-Confirmed Cases and Fatalities in 883 US Counties with a Population of 50,000 or More: Estimated Effect of Initial Prevention Policies

Leon S. Robertson

Accepted: 6 January 2021 / Published online: 25 January 2021

Abstract Control of diseases transmitted from person to person may be more effectively and less economically damaging if preventive and ameliorative efforts are focused on the more vulnerable local areas rather than entire countries, provinces, or states. The spread of the COVID-19 virus is highly concentrated in urban US counties. Sixteen factors known or thought to be related to spread of the COVID-19 virus were studied by Poisson regression analysis of confirmed cases and deaths in 883 US counties with a population of 50,000 or more as of May 31, 2020. Evidence of crowding in homes, workplaces, religious gatherings, preexisting health conditions in the population, and local economic and demographic conditions, with one exception, was predictive of incidence and mortality. Based on the correlation of cases and deaths to length of stay-at-home orders, the orders were associated with about 52% reduced cases and about 55% reduced deaths from those expected without the orders.

Keywords COVID-19 · corona virus · infectious disease · social factors,economic factors, · demographic factors

Introduction

Various mathematical models using different methods initially produced quite different predictions of COVID-19 cases and deaths in the USA because of the variance in assumptions about how the virus and people would behave [1]. As more data became available, the projections converged somewhat but still varied substantially [2]. The early predictions did serve the purpose of motivating most US state governments to adopt policies to slow the spread of the virus. In the USA, the initiation and timing of stay-at-home (shutdown) orders varied among the states. The government and business operations and other gatherings prohibited during the shutdown also varied among the states but all left plenty of leeway for the virus to continue to spread. Several state governors did not issue such orders and many announced partial or complete termination of the orders in late April and early May, 2020. In Wisconsin and Oregon, the order was voided by judges. Post shutdown, some state and local governments issued standards for physical distancing in businesses and other organization as well as wearing face masks. Nevertheless, several state governors prohibited local governments from requiring distancing and masks.

In response to warnings and shutdown orders, behavior in the USA varied from substantial recommended risk avoidance behavior (e.g., reduced travel, reduced physical proximity to other people, frequent hand washing, wearing face masks) to mockery of those who did so and protests against requirements to do so [3]. Photographs and videos of street protesters against shutdowns showed many people in close proximity to one another with no face masks [4]. Testing kits provided by the Centers for Disease Control and Prevention were found inaccurate resulting in delays in testing [5]. The virus was well established in many communities

L. S. Robertson (✉)
Yale University, New Haven, CT, USA

before testing, tracking those exposed and quarantine was initiated.

Estimating the transmission rate of the COVID-19 virus using traditional models of epidemics is problematic because of its behavior. These models require information on infection and recovery rates as well as immunity which involves testing vast numbers of people representative of the population [6]. Many people who are infected do not experience significant or any symptoms but shed virus that infects others in physical proximity or in contact with surfaces where it dwells for a time. For example, tests of residents in Westchester County, New York, the first “hot spot” in that state, indicated that 16.7% of people had antibodies to COVID-19 [7]. If the sample is representative of the population, 161,673 people in Westchester County had been exposed to enough of the virus to produce antibodies (0.167 × 967,506 people in the population) but there were only 31,294 “confirmed cases” reported by the County Health Department as of May 10, 2020. Some 81% of those who may be positive for antibodies but not reported as cases either experienced mild enough symptoms to avoid seeking help or no symptoms at all. In New York City, 21% of the tested had antibodies, but in Bronx County, only about 13.8% of those had turned up in the “confirmed cases” count by May 10; about 86 percent did not. The difference between Westchester and Bronx counties could be a result of fluctuations in sampling or they could represent differences in help-seeking among people with relatively high (Westchester) or low (Bronx) incomes or other differences among the populations.

The authors of one study of the effect of warnings and shutdowns claimed that the shutdowns prevented 60 million COVID-19 cases in the USA based on models of early exponential growth in time among US states [8]. Using state data ignores the wide variation in growth rates among counties within states. A comparison of counties in Iowa with no shutdown and Illinois before and during the Illinois shutdown indicated 30 percent excess cases in Iowa counties through April 20, 2020 [9]. Travel data based on tracing cell phone movements indicate that travel decreased substantially prior to the adoption of stay-at-home orders in many metropolitan areas in the USA but increased in time later [10] suggesting that risk avoidance began before the orders and deteriorated thereafter.

Governments and other organizations in the USA collect data on health, education, and other characteristics of the populations as well as indicators of crowding in housing, businesses, and religious institutions specific to US counties that are thought to increase or decrease the risk of human transmission of pathogens or the severity of the illnesses they cause. The purpose of this study is to estimate the effect of state shutdowns corrected for available social, economic, and demographic data from a variety of sources that are predictive of the counts of COVID-19-confirmed cases and deaths in US counties with populations of 50,000 or more as of May 31, 2020.

Materials and Methods

This is a cross-sectional ecological study of predictors of the accumulation of COVID-19-confirmed cases and deaths among urban US counties as of May 31, 2020. Sixteen potential predictors of COVID-19 spread and severity were included in the study. In addition to the number of days of shutdown orders, data from US counties were found on six factors that likely increase the probability of human interaction that would facilitate spread of a contagious virus: population density per square kilometer, average number of persons per household, average employees per business, average religious adherents per congregation, and average number of social acquaintances per population. Four factors that are known to be related to the severity of the disease were included separately: percent of the population with obesity, diabetes, and elderly cardiovascular hospitalizations and persons 65 years and older. Social and economic factors that are often related to health status were also included: percent of adults with at least a high school education, percent unemployment, median family income, income inequality, and percent African American and Hispanic ethnicity.

Daily numbers of accumulated confirmed cases and deaths in each county through May 31, 2020, were downloaded from usafacts.org (https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/). Dates of shutdown orders were obtained from Littler.com (https://www.littler.com/publication-press/publication/stay-top-stay-home-list-statewide). Population density was obtained from the US Census Bureau (https://www.census.gov/quickfacts/fact/note/US/LND110210). Estimated 2019 population, percent unemployed, and median household income prior to the pandemic for each county were downloaded from the US Department of

Agriculture website (https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/) based on estimates from the US Census Bureau and Bureau of Labor statistics. Persons per household, social acquaintances, high school graduates, economic inequality, percent 65 years or older, percent with diabetes, and percent obese were downloaded from files accumulated from various sources by the Robert Wood Johnson Foundation (https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation). Medicare hospital discharges for cardiovascular diseases were obtained from CDC Wonder (https://nccd.cdc.gov/DHDSPAtlas/Reports.aspx). Numbers of religious adherents and congregations were obtained from therda.com (http://www.thearda.com/Archive/Files/Downloads/RCMSCY10_DL.asp). Numbers of businesses and employees were downloaded from the Bureau of Labor Statistics (https://data.bls.gov/cew/apps/table_maker/v4/table_maker.htm#type=1&year=2019&qtr=3&own=5&ind=10&supp=0). Percent African American and Hispanic were obtained from randaalolson.com (http://www.randalolson.com/2014/04/29/u-s-racial-diversity-by-county/).

Poisson regression models of cases and deaths in US counties with 50,000 or more population were applied to the data separately for confirmed cases and deaths corrected for population size. The form of the regression equation is:

Accumulated number of cases (or separately, deaths) as of May 31, 2020 =

b1 (days from the shutdown order, if any, until May 31, 2020) +
b2 (log(population density, estimated 2019 residents per square kilometer)) +
b3 (average number of persons per household) +
b4 (log(average employees per business enterprise)) +
b5 (log(average religious adherents per congregation)) +
b6 (log(average number of social acquaintances reported per person)) +
b7 (percent of the population that is obese) +
b8 (percent of the population with diabetes) +
b9 (Medicare cardiovascular hospitalization discharges 2015-2017) +
b10 (log (percent of the population 65 years or older))+
b11 (percent of adults who finished high school) +
b12 (log(median family income before the pandemic)) +
b13 (income inequality before the pandemic) +
b14 (percent unemployed before the pandemic) +
b15 (√Percent African American) +
b16 (√Percent Hispanic)

Log(population) was included as an offset variable to correct for differences in population size among the counties. The logarithmic transformations or square roots on selected variables were used because the frequency distributions of those variables were skewed. The study was limited to counties with 50,000 or more population to avoid random variation in small numbers. To estimate the number of cases and deaths likely prevented by warnings and shutdowns, the regression equation was used to predict the number of cases and deaths in each county expected when shutdown days was set to zero. The numbers for each county were added and the totals compared to the actual number of aggregated cases and deaths. The regression analysis was confined to cases and deaths as of May 31, 2020, by which time most of the shutdowns were totally or partially abandoned.

Results

The variation in distribution of COVID-19 cases among counties is illustrated in Fig. 1 showing the 200 counties with the most confirmed cases. Of the 3141 counties and county equivalents nationally, as of May 31, 2020, 79% of cases and 85% of deaths occurred in the 200 counties with the most cases. These 200 counties were highly urbanized, containing 50% of the estimated 2019 US population.

Data on all variables included in the Poisson regressions were available for 883 counties and county equivalents with more than 50,000 populations. Although only 28% of the 3141 counties and county equivalents in the USA, the total population of these counties (263,489,065) was about 80% of the estimated 2019 US population. As of May 31, 2020, confirmed COVID-19 cases in the studied counties (1,515,780) were about 85% of cases of the all-US total. Deaths attributed to COVID-19 in the studied counties as of that date were 92,136, about 90% of the total for all counties.

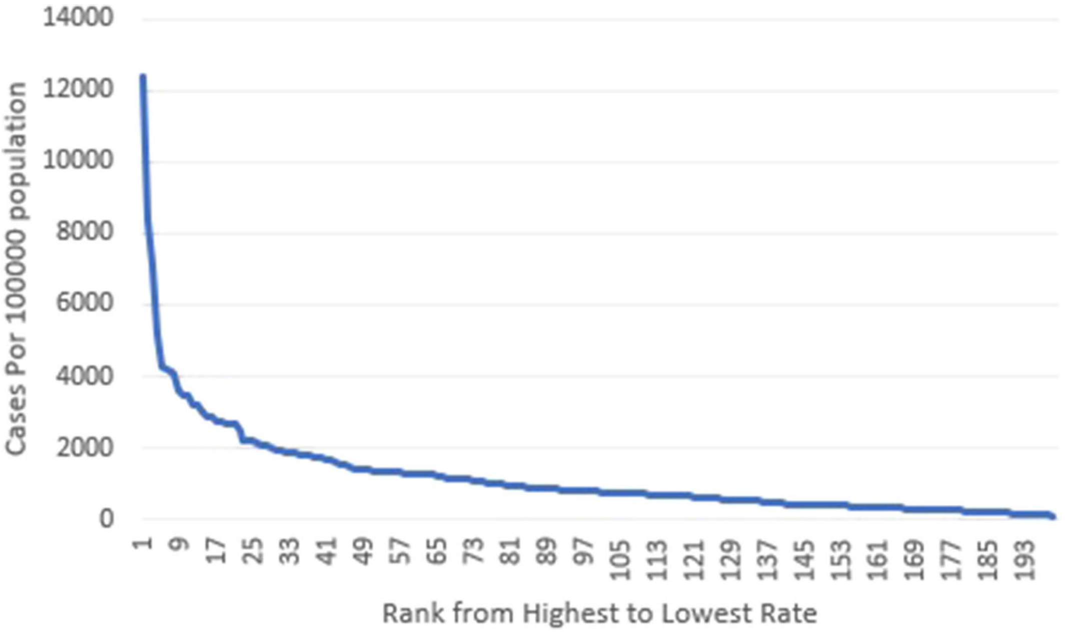

Fig. 1 COVID-19 cases per 100,000 population in the 200 counties with the most cases, USA, May 31, 2020

The parameters and 95% confidence intervals of the regression estimates are shown in Table 1 separately for confirmed cases and deaths attributed to COVID-19. Most of the variables are predictive of cases and deaths with remarkably narrow confidence intervals. Counties with greater population density, crowding in housing (cases but not deaths), workplaces, and religious congregations as well as self-reported social contacts per person had more cases and deaths. Greater numbers of cases and deaths occurred in counties where greater proportions of populations with known higher risk conditions (diabetes, heart disease, and obesity) and where more elderly people live. If a lower percent of the population finished high school, cases and deaths were higher. Cases and deaths were associated with higher median incomes and higher pre-pandemic unemployment. Income inequality was associated with cases but reversed in the death model. Cases and deaths occurred more frequently in counties with a larger proportion of African Americans in the population but were less frequent in counties with a larger proportion of Hispanics. The squared correlation between predicted and actual cases was 0.78. The R^2 between predicted and actual deaths was also 0.78. These indicate reasonably good fits of the models.

A cautionary note: the differences among the coefficients on particular variables do not indicate greater or less importance. The variables are measured on different scales and each coefficient refers only to covariation with increments of the scale it modifies. Coefficients on specific variables are not necessarily indicative of causation but are included for their predictive value.

With the shutdown coefficient set to zero, the regression model predicted 3,150,826 cases and 203,190 deaths in the studied counties through May 31, 2020. Compared to the actual 1,515,780 cases and 92,136 deaths, the results suggest that the shutdowns and accompanying policies reduced confirmed cases by about 52% and deaths by about 55%.

Discussion

The spread of the COVID-19 virus was mainly concentrated in urban areas and was substantially predicted in the first 3 months of the pandemic in the US by numerous factors—crowding in homes, workplaces, religious gatherings, preexisting health conditions in the population, and local economic and demographic conditions. The large variation in cases and deaths among counties raises the issue of the best strategies to reduce the spread of COVID-19 and other pathogens that are easily spread by human contact. A vigorous program of testing, tracing and quarantine, required masks, and physical distancing in public places or resort to shut down in counties that are likely to have higher numbers of cases

Table 1 Poisson regression coefficients of factors thought to contribute to incidence and severity of COVID-19 infections, 883 US counties as of May 31, 2020

	COVID-19 cases (95 % C.I.)	COVID-19 deaths (95% C.I.)
Days from shutdown	− 0.014 (− 0.0134, − 0.0138)	− 0.015 (− 0.016, − 0.014)
Log (population/square kilometer)	0.278 (0.276, 0.280)	0.494 (0.482, 0.506)
Average persons per household	1.110 (1.094, 1.126)	− 0.050 (− 0.067, 0.018)
Log(average employees per business)	0.107 (0.102, 0.112)	0.248 (0.228, 0.268)
Log(average religious per number of congregations)	1.028 (1.021, 1.035)	1.328 (1.298, 1.358)
Log(claimed social acquaintances)	0.420 (0.413, 0.427)	0.389 (0.315, 0.421)
Percent obese in the population	0.008 (0.007, 0.009)	− 0.005 (− 0.008, − 0.002)
Percent diabetic in the population	0.036 (0.034, 0.038)	0.058 (0.050, 0.066)
Cardiovascular hospital discharge rate	0.654 (0.640, 0.668)	1.143 (1.082, 1.204)
Percent aged 65 and older	0.896 (0.883, 0.909)	1.478 (1.424, 1.532)
Percent adults finished high school	− 0.013 (− 0.014, − 0.012)	− 0.016 (− 0.017, − 0.015)
Log(median family income)	1.098 (1.086, 1.110)	0.925 (0.874, 0.976)
Income inequality	0.076 (0.073, 0.079)	− 0.074 (− 0.086, − 0.062)
Percent unemployed before COVID-19	0.421 (0.410, 0.432)	1.283 (1.228, 1.338)
Percent African American	0.078 (00.076, 0.080)	0.065 (0.058, 0.072)
Percent Hispanic	− 0.231 (− 0.234, − 0.228)	− 0.054 (− 0.066, − 0.042)
Intercept	− 35.375	− 40.507
Predicted vs. actual R^2	0.78	0.78

as indicated by the data in this study may be more effective, efficient, and less economically damaging than statewide shutdowns.

The validity and reliability of measurement of each of the predictor variables is difficult to establish. The point of the study is to assess the degree to which the data available, whether flawed or not, can be used to predict the spread of an easily transmitted pathogen. The cause of the disease is the virus. The cause of transmission is human living conditions and behavior the correlates of which may or may not be good enough to be used for prediction. With the exception of the number of shutdown days, inclusion of a given variable does not imply that it is hypothesized as a causal variable, only that it may have predictive value and should be controlled in an assessment of the effect of preventive measures.

Most of the predictor coefficients are in the expected direction but the lower numbers of cases in counties with a larger percentage of Hispanics in the population is contrary to claims that Hispanics work disproportionately in occupations where the virus spread rapidly. Ecological correlations occasionally misrepresent the correlation of individual characteristics and behavior with outcomes and collinearity can also distort regression coefficients. In this study, percent Hispanics was not correlated with average employees per business but was correlated to population density ($R^2 = 0.36$) and congregants per religious facility ($R^2 = 0.32$). The zero-order correlations between percent Hispanics and cases ($R^2 = 0.15$) and deaths ($R^2 = 0.13$) were positive so apparently the collinearity with other factors distorted the coefficients on percent Hispanics. The vast majority of the correlations among the predictor variables were substantially less than those mentioned for percent Hispanics.

When predictor variables are highly correlated, the sign and values of individual coefficients may be distorted but the overall prediction valid [11]. The substantial correlation between predicted and actual cases and deaths among counties indicates that the equation has value for allocation of preventive measures and preparation of the health care system for the onslaught of the aggregate number of cases expected in a county.

The maximum R^2 of shutdown days with each of the other variables was 0.02 so it would not have been distorted and could be used to estimate the effect of shutdowns. The coefficients on shutdown days without the other variables entered in the analysis were − 0.012

for both cases and deaths, only slightly less than the coefficients adjusted for the other variables.

The data suggest that warnings and state shutdowns in late March and April prevented about 1.6 million cases and 111,000 deaths. The study that estimated the warnings and shutdowns prevented 60 million COVID-19 cases in the USA [8] appears to be grossly inflated by a factor of 42. Without the shutdowns, apparently the COVID-19 virus would likely have killed about twice as many people as had died through the end of May and caused more than enough severe illnesses to overwhelm the medical care system sooner in many urban counties in the USA. While this estimate does not account for the variations in effectiveness of specific preventive measures (warnings, physical distancing, wearing masks, and personal hygiene), it does provide an indication of the potential net effect of shutdowns if applied to local areas with growing case numbers. Attempts have been made to evaluate the effects of various countermeasures individually such as school closings at the state level [12], but with many imposed at or near the same time, the estimates of the effect of each are quite problematic.

This study is limited by the lack of data on other factors that are likely predictive of spread of COVID-19. For example, a search for numbers of bars by US counties produced no results. Too many counties fail to report data on use of mass transit that may also be a factor in spread of the virus. Data is available on passenger departures at airports but the counties in which the passengers reside are unknown.

By midsummer 2020, the numbers of cases and positive tests were so numerous in many southern and southwestern cities, and the results of tests were delayed, that there were not enough people assigned to tracing to keep up with the spread of the virus. In the Fall of 2020, the virus spread to the point that many communities had all or most of their intensive care bed occupied. Nevertheless, as of December 24, 2020, 57% of 18.4 million accumulated cases and 55% of 365,657 accumulated deaths had occurred in the 200 counties with the most cases.

Acknowledgments There was no grant to support the research

References

1. Holmdahl SM, Buckee C. Wrong but useful — what Covid-19 epidemiologic models can and cannot tell us. *NEJM*. 2020;383:303–5.
2. Bui Q, Katz J, Parlapiano A, Sanger-Katz M. *Coronavirus models are nearing consensus, but reopening could throw them off again*: The New York Times; 2020. https://www.nytimes.com/interactive/2020/05/12/upshot/coronavirus-models.html. Accessed 13 May 2020.
3. Jones T. Trump supporters harassed Arizona reporters for wearing masks. *Poynter*. 2020. https://www.poynter.org/reporting-editing/2020/trump-supporters-harassed-arizona-reporters-for-wearing-masks/. Accessed 6 May 2020.
4. Fernandez M. In photos: Protesters storm Michigan Capitol over coronavirus restrictions. *Axios*. 2020. https://www.axios.com/michigan-protestors-state-capitol-coronavirus-ebcb34b6-5a48-47fb-b305-360c38baa16a.html. Accessed 1 May 2020.
5. Cohen J. The United States badly bungled coronavirus testing—but things may soon improve. *Science*. 2020. https://www.sciencemag.org/news/2020/02/united-states-badly-bungled-coronavirus-testing-things-may-soon-improve. Accessed 1 March 2020.
6. Perez L, Dragicevic S. An agent-based approach for modeling dynamics of contagious disease spread. *Int J Heal Geogr*. 2009;8:50. https://doi.org/10.1186/1476-072X-8-50.
7. Saplakoglu Y. 1 in 5 people tested in New York City had antibodies for the coronavirus. Livescience.com; 2020. https://www.livescience.com/covid-antibody-test-results-new-york-test.html. Accessed 4 April 2020.
8. Hsiang S, Allen D, Annan-Phan S, Bell K, Bollinger I, Chong T, et al. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nat Online*. 2020, https://www.nature.com/articles/s41586-020-2404-8_reference.pdf. Accessed 10 June 2020.
9. Lyu W, Wehby GL. Comparison of estimated rates of coronavirus disease 2019 (COVID-19) in border counties in Iowa without a stay-at-home order and border counties in Illinois with a stay-at-home order. *JAMA Netw Open*. 2020;3(5):e2011102. https://doi.org/10.1001/jamanetworkopen.2020.11102. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2766229. Accessed 10 June 2020.
10. Ghader S, Zhao J, Lee M, Zhou W, Zhao G, Zhang L. Observed mobility behavior data reveal "social distancing" inertia. Arvix.org; undated. https://arxiv.org/ftp/arxiv/papers/2004/2004.14748.pdf. Accessed 10 June 2020.
11. Mason CH, Perreault WH. Collinearity, power, and interpretation of multiple regression analysis. *J Market Res*. 1991;28:268–80.
12. Auger KA, et al. Association between statewide school closure and COVID-19 incidence and mortality in the US. *JAMA*. 2020; https://jamanetwork.com/journals/jama/fullarticle/2769034. Accessed 30 July 2020.