Drivers of COVID-19 in U.S. counties: A wave-level analysis

Christopher F Baum[*] Andrés García-Suaza[†] Miguel Henry[‡] Jesús Otero [§]

2024-04-13

Abstract

Since the initial outbreak of COVID-19 in the United States, researchers from a variety of scientific disciplines have sought to understand the factors influencing the evolution of cases and fatalities. This paper proposes a two-stage econometric modeling approach to analyze a range of socioeconomic, demographic, health, epidemiological, climate, pollution, and political factors as potential drivers of the spread of COVID-19 across waves and counties in the United States. The two-step modeling strategy allows us to (i) accommodate the observed heterogeneity across waves and counties in the transmissibility of the virus, and (ii) assess the relative importance of the cross-sectional measures. We leverage the availability of daily data on confirmed cases and deaths of COVID-19 in counties across the 48 contiguous states and the District of Columbia, spanning a two-year period from March 2020 to March 2022. We find that socioeconomic and demographic factors generally had the greatest influence on the transmissibility of the virus and the associated mortality risk, with health and climate factors playing a lesser role.

Keywords: COVID-19, coronavirus, geographic heterogeneity, covariate selection

JEL Classification: C13, C21, R15, R23

[*] Corresponding author. Boston College, Department of Economics and School of Social Work, Chestnut Hill, MA, USA; baum@bc.edu.

[†] Facultad de Economía, Universidad del Rosario, Bogotá, Colombia; andres.garcia@urosar io.edu.co

[‡] QuantEcon Research, Boston, MA, USA; mhenry@quanteconresearch.com

[§] Facultad de Economía, Universidad del Rosario, Bogotá, Colombia; jesus.otero@urosario.edu.co

1 Introduction

"The spread of COVID-19 across the USA confirms that not all Americans are equally at risk of infection, severe disease, or mortality." (Chin et al., 2020, p.1)

Since the first COVID-19 deaths were officially reported in February 2020 in Washington[1] by the Centers for Disease Control and Prevention (CDC) and in California,[2] COVID-19 became the third leading cause of mortality in 2020 and 2021 (Ahmad et al., 2022). It surpassed heart disease and cancer during this period. Additionally, COVID-19 was the number one cause of death for people ages 45–84 in January 2022 (Ortaliza et al., 2022). During 2022, COVID-19 was ranked as the fourth underlying cause of death in the U.S. (Ahmad et al., 2023). The course of the novel coronavirus pandemic in the country varied dramatically over time and geographically.[3] By mid-March 2020, COVID-19 transmission had become widespread, initially clustering in certain urban areas of high economic activity like New York City, New Orleans, Albany, and Georgia (Mukherji, 2022). This accelerated the spread with case counts rising more than 1,000-fold within weeks. This rapid nationwide transmission deepened the consequences of inequality.[4] As the virus spread to afflict more communities over time - including urban, rural, suburban, and exurban areas - various actions were taken across the country to slow and contain the coronavirus pandemic. These included non-pharmaceutical public health interventions,[5] as well as vaccination campaigns for the U.S. public. As a result of these efforts, the geography of COVID-19 infection and death rates shifted notably over the course of the public health emergency.

Between March 2020 and March 2022, the U.S. experienced six distinct COVID-19 waves, with each wave exhibiting unique complexities and patterns. These episodes are documented by Jones (2022). To analyze each of these waves individually, we adopted the following time ranges: March 15–June 30, 2020 for the first wave; July 1–

[1] The first confirmed case of COVID-19 in the United States was reported in January 2020 in the state of Washington (Holshue et al., 2020).

[2] See https://time.com/5825320/california-coronavirus-february-first-death/.

[3] "The Changing Geography of COVID-19 in the U.S.", Pew Research Center, at https://www.pewresearch.org/politics/2020/12/08/the-changing-geography-of-covid-19-in-the-u-s/.

[4] "As Coronavirus Deepens Inequality, Inequality Worsens Its Spread" at https://www.nytimes.com/2020/03/15/world/europe/coronavirus-inequality.html?searchResultPosition=1.

[5] Raifman et al. (2020) provide a comprehensive list of the interventions implemented in each U.S. state along with their effective dates.

September 30, 2020 for the second wave; October 1, 2020–March 31, 2021 for the third wave; April 1–July 31, 2021 for the fourth wave; August 1–November 30, 2021 for the fifth wave; and December 1, 2021–March 19, 2022 for the sixth wave.[6] As of March 17, 2022, the CDC reported a total cumulative number of 79,723,281 confirmed COVID-19 cases and 971,072 deaths nationwide.[7] However, the number of excess deaths (the number of people who died in a given period compared to the number that would be expected to die in the same span of time in the past) was even higher (Wang et al., 2022), totaling 1,105,736 between March 7, 2020 and March 5, 2022 according to CDC estimates.[8] Additionally, provisional COVID-19 deaths from death certificates reached 992,691 by March 19, 2022 according to the National Vital Statistics System.[9]

From the early stages of the COVID-19 pandemic to the present day, numerous researchers across a wide range of scientific fields, such as Allcott et al. (2020), Knittel and Ozaltun (2020), Oronce et al. (2020), Wu et al. (2020), Andersen et al. (2021), Chernozhukov et al. (2021), Liao and De Maio (2021), Papageorge et al. (2021), Baum and Henry (2022), Carozzi et al. (2022), Desmet and Wacziarg (2022), Welsch (2022), Bollyky et al. (2023), Haimerl and Hartl (2023) and Ho et al. (2023), have employed different empirical strategies, data sources, and assumptions to model U.S. COVID-19 cases and deaths. They aimed to uncover the drivers influencing the progression of both cases and deaths. Unlike most existing literature on this topic, we employ a two-stage panel plus cross-section sequential modeling approach to identify the drivers contributing to the progression of COVID-19 cases and deaths in the U.S. over time. Another important contribution of our paper is the use of daily panel data from March 15, 2020 through March 19, 2022 (a 739-day period). This spans all six identified pandemic waves for confirmed cases and deaths attributed to COVID-19 across 3,014 U.S. counties in the 48 contiguous U.S. states and Washington D.C. Ours is the first study using such extensive county-level confirmed cases and mortality data covering all six COVID-19 waves. We employ the one-covariate-at-a-time (OCMT) variable selection algorithm proposed by Chudik et al. (2018) to guide the choice of drivers that are in-

[6] Due to irregular data reporting from some states and discontinued updates from Nebraska and Missouri as of Spring 2022, our study data ends in March 2022.

[7] Aggregate case and death count data reported to the CDC's COVID Data Tracker by states, territories, and other jurisdictions (https://covid.cdc.gov/covid-data-tracker/#datatracker-home).

[8] More details on the CDC's methodology for estimating excess deaths is available at https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm.

[9] See https://covid.cdc.gov/covid-data-tracker/#trends_totaldeaths_select_00.

cluded in the cross-sectional analysis of our second-stage model.

Given the considerable variation in the severity of the COVID-19 pandemic across the U.S., influenced by evolutionary factors such as emerging viral variants, widespread vaccinations, and advancements in treatments[10] such as Paxlovid,[11] our analysis recognizes the need for a more nuanced modeling approach. Specifically, we employ a two-stage sequential econometric modeling approach that accounts for the temporal stability of the estimated models and considers the distinct waves of the COVID-19 pandemic.[12]

In contrast to most cited studies, analyzing two full years of daily COVID-19 data requires evaluation of the temporal stability of an estimated model.[13] This is due to evolutionary factors like emerging viral mutations that caused wide swings in cases and deaths. A single model applied to the entire pandemic history cannot adequately capture such variations. Therefore, we offer a two-stage sequential econometric modeling approach to examine a number of socioeconomic, demographic, health, epidemiological, climate, pollution, and political drivers of COVID-19 spread across U.S. counties and pandemic waves. Our econometric modeling approach is conceptually similar to that studied by Saxonhouse (1976) and Hornstein and Greene (2012), and has been used by Giulietti et al. (2014) in studying long-term relationships between pairs of crude oil prices. We adopt the six distinct pandemic wave definitions from the Pew Research Center in Jones (2022) and Figure 1 shows the trajectory of deaths across these waves.

The structure of the paper is as follows. Section 2 provides a brief discussion and comparison to related economic literature examining the importance of demographic and socioeconomic factors on the evolution of COVID-19 cases and deaths. Section 3

[10] The last two factors —widespread vaccinations and advancements in treatments such as Paxlovid—The are particularly important, as they have reduced the likelihood of mortality for those infected across most segments of the population.

[11] The first oral outpatient antiviral for the treatment of mild-to-moderate COVID-19 in adults and pediatric patients (12 years of age and older weighing at least 40 kilograms or about 88 pounds) with positive results of direct SARS-CoV-2 testing, and who are at high risk for progression to severe COVID-19, including hospitalization or death. The antiviral was given an FDA emergency use authorization (EUA) on December 22, 2021.

[12] Martinez-Beneito et al. (2023) examined the first six waves of COVID-19 cases in Barcelona, Spain from March 2020 to March 2022 using a Bayesian multilevel logistic modeling framework. Their framework allowed for heterogeneous point estimates by wave in a single model.

[13] We did not formally test for the coefficient stability across the first stage models. However, Figures 2(a), 2(b), 3(a), and 3(b) provide optical evidence that some of the key coefficients exhibited considerable variation across waves.

describes the data and the variables used in our study. Section 4 presents our empirical modeling strategy, while Section 5 discusses and presents the main results. We conclude by summarizing our key findings in Section 6. Replication files and programs are available upon request from the authors.

2 Related economic literature

There is a limited but important body of economic literature on COVID-19 using a two-stage sequential modeling approach, including Brown and Ravallion (2020), Aron and Muellbauer (2022), and Mukherji (2022). However, there are several key differences between those prior works and our study, including our motivation for adopting a two-stage modeling approach.

Brown and Ravallion (2020) focused on county-level socioeconomic factors like median household income, race, income inequality, and poverty, as well as epidemiological and health characteristics. They estimated exponential conditional mean models for COVID-19 infections and deaths across U.S. counties in the first half of 2020. To address endogeneity from including cumulative case counts as a covariate in the second-stage death model, they added a residual-based control function derived from the first-stage case equation.

Aron and Muellbauer (2022) took a different approach to handling endogeneity in a two-stage framework. In the first stage, they modeled the time of arrival of a significant level of COVID-19 infection (timing of onset) for each state from week 9 of 2020 to week 8 of 2021. Using the fitted onset timing as an instrument, they estimated a second-stage death model, over the 52-week pandemic period, linking the derived instrument, 2020 spring temperature, and pre-pandemic socioeconomic, demographic, and political factors to state-level weekly deaths (cumulative per capita *excess* ('all-causes') deaths, the ratio of the cumulative *excess* ('all-causes') deaths relative to cumulative normal deaths, and cumulative per capita COVID-19 deaths).

In contrast, Mukherji (2022) used residuals from a first-stage epidemiological panel data model of early pandemic COVID-19 infections (March 30 to April 19, 2020) to compute county-level fixed effects and a social vulnerability index for 770 U.S. counties. In the second stage, the author evaluated the importance of county-level socioeconomic (income, unemployment rate, income inequality, access to housing), demographic (population size, density), spatial (weighted international air passengers served by the top international airports in the U.S.), and health factors (e.g., percentage of the population that receives the flu vaccine) for this index. The same factors except the spatial measure, along with a 14-day lagged case covariate, were also used separately to examine reported deaths in a pooled regression model over the 20-day period.

Our study updates and complements this COVID-19 economic literature through a two-stage, county-level panel and cross-section sequential modeling approach. We exploit the availability of two years of daily confirmed case and death data across 3,014 U.S. counties and all six pandemic waves. Our models allow for county-level random effects, lagged cases, and vaccination rates of people who were fully vaccinated against COVID-19. None of the aforementioned studies include COVID-19 vaccination measures.

3 Data

This section describes the datasets used to assemble the final dataset utilized for our study. Detailed information on the data sources, construction, and transformations is provided in Appendix A.

To construct a consistent final sample of data, we retained 2,215,290 daily observations for cases, deaths and vaccinations across 3,014 counties with no missing data, representing 97% of the total 3,108 counties. Table 1 provides descriptive statistics for the various factors included in our econometric models. Table 2 reports the average of confirmed cases, deaths, and vaccination rates by wave as well as for low and high population density counties separately.

3.1 Cases and mortality data

We constructed a county-level panel dataset using daily cumulative counts of confirmed COVID-19 cases and deaths from USAFacts. The data covers counties in the 48 contiguous U.S. states and Washington D.C. from March 15, 2020 through March 19, 2022. The resulting longitudinal data include the state abbreviation, the county name, the Federal Information Processing System (FIPS) codes for states and counties, the daily calendar date when cases and deaths are reported, and the daily cumulative confirmed case and death counts attributed to COVID-19.

During the first two waves of the pandemic (March—September 2020), confirmed cases were, on average, more prevalent in counties with high population density (fourth quartile). However, from October 2020 onwards (waves 3–6), counties with lower population density (the first three quartiles) experienced a more rapid increase in cases, as detailed in Table 2. Similarly, although death rates in high-density counties were initially on average twice as high those in lower-density areas during the first wave, this pattern shifted by the third wave. By the sixth wave (beginning December 2021), death rates in low-density populated counties were 30% higher than those in more densely populated areas. The early stages of the pandemic were characterized by shortages of treatment facilities in a number of hard-hit urban areas, so the rates of cases and deaths in high-density areas in waves 1 and 2 (March-September 2020) is not surprising. Nevertheless, the progression of the pandemic in subsequent waves highlighted disparities in state-level policies concerning lockdowns, quarantines, and treatment protocols that were largely political in nature. Lower-density 'red' counties generally imposed fewer restrictions on personal liberty compared to 'blue' urban counties. Given these political dynamics, it is not surprising that the high-density counties exhibited lower case and death rates from October 2020 onward.

Appendix B contains county-level *bivariate maps* from Naqvi (2022),[14] showing the interplay of population-adjusted COVID-19 case and death rates across the six waves with a two-dimensional color legend. Using common color scheme properties, these maps reveal geographic shifts over time. In the first wave, high cumulative rates concentrated along the Atlantic and Gulf coasts. The second wave saw concentration in the South and Southwest. Coastal states had lower risk of both cumulative cases and deaths in waves three and four, with a swath along the Mississippi becoming evident. Despite growing vaccine availability, the fifth wave still saw severe outcomes in some Mountain states. Finally, the sixth wave displayed relatively low cumulative case rates but high death rates in states like Georgia and Texas. Overall, the variations in cumulative case and death rates across time and space are evident in these bivariate maps.

[14] To produce these maps, we used spatial data (the spatial unit identifiers, geographic coordinates, and geographic entity codes (GEOIDs) (i.e, the FIPS identifier) for each county and the state boundaries, from the U.S. Census Bureau's TIGER geographic database. For more details on the construction of the spatial data see Appendix A.

3.2 Vaccination data

We obtained county-level data on the daily cumulative vaccination rates of residents fully vaccinated against COVID-19, with a second dose of a two-dose vaccine or one dose of a single-dose vaccine, from the CDC Immunization Information Systems. Appendix C contains *bivariate maps* visualizing the interplay between population-adjusted COVID-19 death rates and vaccination rates across the pandemic waves. In wave 3, when vaccines became available, Southern states clearly showed low vaccination rates combined with high cumulative death rates. This pattern persisted into wave 4, while Atlantic coast states hard hit early in the pandemic now had high vaccination and death rates. Through waves 5 and 6, states in the Mississippi valley and other Southern states continued facing low vaccination rates and high cumulative deaths. By the end of wave 6, several states in New England, Florida, and the Pacific Northwest recorded high vaccination rates and relatively low death rates. These differences in vaccination rates are also reflected in the comparison of low-density and high-density counties in Table 2. Specifically, starting with wave 4 (April 2021) and continuing thereafter, counties with higher population densities (categorized in the fourth quartile) exhibit, on average, higher vaccination rates. As previously mentioned, these discrepancies may have a political basis, with the importance of widespread vaccination being more prominently emphasized in 'blue' states.

3.3 Socioeconomic and demographic data

Using 2020 county-level resident population estimates from the U.S. Census Bureau we gathered demographic information on sex, race, and ethnicity. In addition, we collected socioeconomic data at the county level, including median household income, education levels, and poverty rates. Specifically, median income and education data from the 2022 County Health Rankings. Poverty rates were drawn from the 2020 U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) Program. Data on the percentage of owner-occupied housing in 2018 came from Wu et al. (2020).

Other socioeconomic factors included in our analysis are the Social Vulnerability Index (SVI), population density, and county-level Democratic vote share in the 2020 presidential election. The 2018 SVI data came from the CDC\ATSDR, providing the latest available county-level percentile ranking of social vulnerability to disasters like pandemics. According to ATSDR, *"it [SVI] refers to the potential negative effects on communities caused by external stresses on human health. Such stresses include natural or human-caused disasters or disease outbreaks."* (Flanagan et al., 2011). We obtained the Democratic vote shares from the MIT Election Data and Science Lab (2018) county-level dataset. As described in Appendix A, we constructed the population density measure for each county.

3.4 Health data

The health factors included county-level data on: the percentage of residents under 65 without health insurance, age-adjusted percentages of adults who currently smoke, age-adjusted percentages of adults aged 20 and above with diagnosed diabetes, and the percentage of the population with adequate access to locations for physical activity. This data came from the 2022 County Health Rankings database. We also added a 2020 county-level Severe COVID-19 Health Risk from PolicyMap (2020), along with mean county-level estimates of the number of years from birth a person can expect to live (life expectancy) from Dwyer-Lindgren et al. (2022). Intensive care unit (ICU) bed data by county came from Kaiser Family Foundation (KFF)'s Kaiser Health News Program. Finally, we incorporated state-level Medicaid expansion status as of March 2022, also from KFF.

3.5 Climate and pollution data

For climate and pollution data, we used county-level average seasonal temperature, relative humidity, and PM2.5 values for 2000–2016. Specifically, we relied on seasonal estimates produced by Wu et al. (2020) for summer (June to September) and winter (December to February). The PM2.5 data reflects fine inhalable particulate matter in the air (aerodynamic diameter ≤ 2.5 micrometers) in units of micrograms per cubic meter.

4 Econometric modeling strategy

Our econometric strategy employs a two-stage (panel plus cross-section), sequential modeling approach to examine the drivers described above across COVID-19 waves and U.S. counties. We start our analysis by estimating separate linear mixed models for each wave. These models yield autoregressive fixed and random coefficients for each of the 3,014 counties. These coefficients then serve as the dependent variables in the second stage of our cross-sectional analysis, taking into account the uncertainty associated with the estimation of the first-stage coefficients. In the second stage, we evaluate cross-sectional measures, including socioeconomic, demographic, health, epidemiological, climate, pollution, and political characteristics, for each county, after employing the novel OCMT variable selection procedure to identify the most relevant factors. The second-stage models are estimated by wave, using a weighted seemingly unrelated regression (SUR) system estimator, allowing for the importance of explanatory factors to vary during the course of the pandemic.

4.1 First-stage county-level daily panel analysis

In the first stage, we estimated two linear mixed models utilizing county-level daily panel data on confirmed cases, deaths, and vaccination rates from March 2020 through March 2022. These data are generated according to the methodology outlined in Appendix A. Following Rabe-Hesketh and Skrondal (2022), we used random coefficient models[15] to accommodate the observed heterogeneity across waves and counties. We estimated the models separately by wave to allow for variations over those episodes in the transmissibility of the virus, the impact of vaccinations, and improvements in treatment regimes. The models employed an unstructured covariance matrix, providing flexibility for the county-level random effects.

The confirmed cases and deaths series were differenced based on the unweighted average of the seven prior days' cumulative count series adjusted for population (see Appendix A). The first model explains the county-level confirmed case rate as a function of the same variable lagged $j = 14$ days. The second model explains the county-level death rate as a function of the confirmed case rate lagged $j = 14$ days. Thus, both models are autoregressive with a single regressor: the county-level confirmed cases 14 days prior, capturing transmissibility of the disease.[16] When vaccination rates of

[15] We assessed the appropriateness of the mixed models using a conservative likelihood-ratio test. We tested whether the random intercepts and random slopes are needed against a pooled OLS alternative under the null. The null hypothesis was soundly rejected at the 99% confidence level (p-value of zero to four decimal places) in all cases. Thus, adding random slopes into the mixed models brings significant improvement. For a thorough review of random coefficient models see Hsiao and Pesaran (2008).

[16] Our choice of a 14-day lag ($j = 14$) in both mixed models follows the literature using a biweekly data frequency. Earlier pandemic research, when medical resources were extremely stressed and treatment options limited, proposed an 8-day infection-to-death delay Jin (2021). Later studies, under less resource strain and with expanded treatments, suggested a delay of 17-21 days IHME (2023). To select a consistent lag across

people who were fully vaccinated against COVID-19 become available in waves 3–6, we also included the 14-day lagged county vaccination rate in the models.

Letting i and t denote the county and time subscripts, respectively, the econometric mixed model for the county-level confirmed case rate, allowing both intercepts and slopes for cases and vaccines rates lagged 14 days to vary by county, is

$$c_{it} = \alpha_0 + \alpha_i + \beta_0 c_{i,t-14} + \beta_i c_{i,t-14} + \gamma_0 v_{i,t-14} + \gamma_i v_{i,t-14} + \epsilon_{it}, \tag{1}$$

where c_{it} and $c_{i,t-14}$ denote the confirmed cases in county i at time t and $t-14$, respectively. In turn, $v_{i,t-14}$ indicates the percentage of county residents that are fully vaccinated with a second dose of a two-dose vaccine or a dose of a single-dose vaccine; α_0, β_0 and γ_0 denote unknown fixed parameters; α_i, β_i and γ_i denote county-level random effects; and ϵ_{it} is the disturbance term. Notice that in equation (1), the county-specific random coefficients relax the homogeneity assumption in the fixed parameters, allowing the impact of the lagged case rate and vaccination rate to vary over both time and counties.

In order to model the county-level deaths rate in county i at time t, d_{it}, the equivalent mixed model, capturing the mortality risk for those infected, is

$$d_{it} = \delta_0 + \delta_i + \kappa_0 c_{i,t-14} + \kappa_i c_{i,t-14} + \lambda_0 v_{i,t-14} + \lambda_i v_{i,t-14} + \varepsilon_{it}. \tag{2}$$

Here, δ_0, κ_0 and λ_0 are unknown fixed parameters; δ_i, κ_i and λ_i denote county-level random effects; and ε_{it} is the disturbance term. Notice that in equation (2), deaths are associated to the county-level confirmed cases 14 days prior.

Using the wave-specific point estimates of β_0 and κ_0, we computed county-level composite estimates to take into account the uncertainty associated with the estimation of the first-stage autoregressive coefficients. This calculation is accomplished by adding the point estimates to the estimated best linear unbiased predictions (BLUPs) of the county-level random effects. In waves 3–6, county-level composite estimates are also computed for vaccinations by using the corresponding point estimates of γ_0 and λ_0 and the BLUPs of the county-level random effects. The generated county-level and wave-specific composite estimates

$$\hat{\beta}_0 + \hat{\beta}_i \tag{3}$$

all six waves, we chose an intermediate 14-day delay for infections and deaths. We validated our findings using a 21-day lag; these robustness results are available upon request.

$$\hat{\kappa}_0 + \hat{\kappa}_i \tag{4}$$

$$\hat{\gamma}_0 + \hat{\gamma}_i \tag{5}$$

$$\hat{\lambda}_0 + \hat{\lambda}_i \tag{6}$$

and their associated precision estimates[17] are then stored and merged as a cross-sectional dataset on 3,014 counties along with the set of time-invariant socioeconomic, demographic, health, climate, pollution, and political variables shown in Table 1.

The importance of allowing for variations in the two mixed models (1) and (2), over the six waves, is illustrated in Figures 2(a), 2(b), 3(a), and 3(b). Considerable variation is evident across waves and between counties with differing levels of population density. Regarding Figure 2(a), we note that the magnitude of the transmissability coefficients is always higher in the high population density counties in the fourth quartile of that empirical distribution. For waves 3–6, we consider the same comparison for the county-level vaccination coefficients, illustrated in Figure 2(b). As states and counties had very different stances and policies toward vaccinations, there is considerable variability in these measures. The variations in the county-level lagged cases coefficients in the cumulative deaths equation can be visualized in Figure 3(a). Here, too, the coefficients gauging mortality rate for those infected are systematically higher in those counties with high population density, although the differences are less extreme. Likewise, Figure 3(b) illustrates the variability in the county-level vaccination coefficients over waves 3–6, as well as geographic differences related to vaccination efforts.

4.2 Second-stage system cross-section analysis

Using the cross-sectional dataset from the first stage, we assessed the relative importance of the cross-sectional measures (socioeconomic, demographic, health, epidemiological, climate, pollution, and political characteristics) in the second stage. The county-level and wave-specific estimates from eqs. (3) – (6) were used as dependent variables. To account for the uncertainty associated with the estimation of the first-stage estimates, each wave's equation was estimated by weighted least squares (WLS). The inverse variance weights were applied to the equations of each wave individually, rather than using a fixed set of weights for the entire estimation.

[17] The composite precision estimates for (3)–(6) were computed by adding the standard errors of the fixed effect point estimates to the standard errors of the BLUPs of the random effect standard errors. This assumes the variance-covariance matrix (VCE) of the error process is separable between fixed and random effects.

Before estimating the cross-sectional models, we applied the OCMT iterative model selection procedure[18] proposed by Chudik et al. (2018). Both dependent and independent variables (including the constant) were inversely weighted, with dependent variables scaled by 100 for interpretation. We used analytic weights proportional to the square root of the inverse of the variance of the county-level and wave-specific composite precision estimates. Using OCMT allowed for the identification of the most relevant predictors in each wave through a more parsimonious model specification in which only the selected factors are included in the second-stage models.

As many aspects of both infection and mortality changed during the several waves, there is no reason to assume that these factors had constant effects on the transmissibility and mortality risk of the virus. For example, when vaccinations were first available in the U.S., they were given to the most vulnerable populations: older citizens and those who were immunocompromised, as well as critical personnel in health care and public safety roles. We take the county-level age structure as fixed, but recognize that it may have played different roles before and after vaccinations were available.

4.2.1 SUR for systems estimation

OCMT selected a different set of factors for each wave and cross-sectional dependent variables from eqs. (3) – (6) for the cases and death models, producing varying coefficients. These cross-section estimates can be considered as a system of six wave equations with differing specifications and time-varying coefficients. The system of equations was estimated with a novel application of the Zellner (1962) seemingly unrelated regression (SUR) technique, using the iterative SUR estimator (Oberhofer and Kmenta, 1974) and WLS. Both dependent and independent variables, including the constant, were weighted by using the inverse of the county-level and wave-specific composite precision estimates from the first stage as weights. In our WLS iterated SUR application, the cross-sectional wave equations correspond to different time periods. Standard errors for each point estimate were calculated using nonparametric bootstrapping.[19]

[18] See Appendix D and Núñez and Otero (2021) for more details.

[19] We have also calculated cluster-bootstrap standard errors (clustered by state) based on the argument that all counties in a state would be subject to the same state-level COVID-19 policies, making it likely that county-level errors within a state would be correlated. Nevertheless, this approach gave us estimates that were very imprecisely estimated possibly due to the fact that the clusters are extremely heterogeneous in

The usual rationale for using SUR as a system estimator is the degree to which each equation's error process might be contemporaneously correlated with other units' errors at each point in time. If those correlations are sizable, SUR can yield efficiency gains relative to the single-equation estimation of each equation. In our case, the error correlations that can be exploited are those for each county across the six pandemic waves. We expect sizable correlations as they reflect unobservable factors at the county level that have not been captured by the time-invariant regressors selected for each wave. Correlations of the errors across waves might occur due to omitted common effects, spatial effects, or as a result of interactions within socioeconomic networks (Chudik and Pesaran, 2015).

The degree to which the correlations increase the precision of the estimates is evaluated by using the Breusch and Pagan (1980) Lagrange multiplier test for error independence across equations, with the null hypothesis that the 6×6 residual correlation matrix for cases and deaths is diagonal; that is, the errors are independent. For vaccinations the residual correlation matrix is 4×4. Under the null, the test statistic follows a χ^2 distribution with $(m(m-1)/2)$ degrees of freedom, where m is the number of equations. For each of the outcomes from eqns. (3)–(6), the null hypothesis that the errors are independent across equations is strongly rejected at the 1% level. There are statistically significant correlations between the errors in the wave equations, implying efficiency gains in the SUR estimator relative to estimating each equation separately. The residual correlations improved the precision of the estimates reported in Tables 3–6.

5 Main Results

This section presents and discusses the main results from the two-stage econometric cases and death models examining the impacts of socioeconomic, demographic, health, epidemiological, climate, pollution, and political drivers on the evolution of COVID-19 across U.S. counties and waves. We report the second stage SUR cross-sectional model results by wave and outcome variable based on the OCMT variable selection procedure.

each wave. MacKinnon et al. (2023) discusses the consequences of having severely heterogeneous clusters.

5.1 Cases and Vaccinations

Table 3 reports weighted SUR estimates and the statistical significance for the factors identified by the OCMT procedure as key determinants in the progression of county-level confirmed case rates across waves in the case model. Some factors consistently play a crucial role throughout the six waves from March 2020 through March 2022. For example, the proportion of Black residents in each county significantly influences transmissibility in all waves except for the sixth. Population density also emerges as a significant positive factor in every wave except the fourth, with very sizable effects in waves 1, 2, and 5. The Democratic vote share, used as a proxy for 'blue' states, influences transmissibility in all waves except the second, generally increasing it except in the third wave. The poverty rate, a measure of income inequality after controlling for median income as discussed by Brown and Ravallion (2020), exhibits a significant positive impact on cases rates in waves 2 and 3. This observation suggests that U.S. counties with greater income inequality experienced higher infection rates during these particular waves. The implementation of Medicaid expansion at the state level, which is also associated with 'blue' states, shows sizable significant positive effects in waves 1 and 4, switching to significantly negative in waves 5 and 6. These variations may reflect differences in state and local policies such as lockdown measures and vaccination campaigns. Finally, the average winter temperature in each county plays a role in affecting transmissibility in all waves, with the impact differing according to the season.

Table 4 offers insights into the factors that OCMT has identified as crucial in determining how county-level lagged vaccination rates influence confirmed case rates during waves 3 to 6. Notably, the poverty rate is found to significantly decrease the impact of vaccination on case numbers in waves 3 and 5, while increasing that effect in waves 4 and 6. This may reflect the variations in vaccination initiatives across urban and rural areas. Other significant determinants include the percentage of uninsured residents under 65, the percentage of adults aged 20 and older diagnosed with diabetes, and the percentage of adults who are current smokers within each county. Each of these factors showed significant effects in two or more waves when vaccinations were available. Additionally, both summer and winter average temperatures are generally found to have a significant impact, as is the level of $PM_{2.5}$ air pollution. A particularly striking observation is made in wave 5, which was marked by the prevalence of the Delta variant. Here, the age distribution, especially for individuals younger than 80 years, is highly significant. This trend suggests that older individuals were less likely to be vaccinated during this wave, underlining the importance of age as a factor in vaccination strategies. The racial/ethnic composition of Black and Hispanic residents is also significant.

5.2 Deaths and Vaccinations

Motivated by the observed patterns in the previous section, results for the deaths model are presented in Table 5. The OCMT selection procedure, employing the same tuning parameters $\delta = 1$ and $\delta^{\star} = 2$, and a significance level of 0.01 as in the case model, identified several factors influencing mortality risk from COVID-19 infection during waves 1 through 3, spanning from the onset through the year 2020. These factors include the age, race/ethnicity distributions, life expectancy, access to exercise, median income, as well as weather-related effects. In wave 4, occurring in Spring 2021 when vaccinations became widely available, many fewer factors were identified as influential; however, Medicaid expansion and the health risk index both have positive and highly significant effects in wave 4. In wave 5, from late summer 2021 during the surge of the Delta variant, several factors become relevant. Conversely, wave 6, spanning from December 2021 to March 2022, revealed an important reduction in the number of influential factors, likely due to the availability of improved treatment options during this timeframe. Interestingly, the proportion of males in the population was negatively correlated with mortality in waves 1, 5, and 6, suggesting a complex interplay of demographics and health outcomes throughout different phases of the pandemic. The poverty rate exhibits a significant positive impact on case- rates in waves 3 and 5, suggesting that U.S. counties with greater income inequality experienced higher mortality rates during these two waves.

Table 6 presents the factors chosen by OCMT as important determinants of the impact of vaccinations on deaths for waves 3–6. Black and Hispanic population fractions increased the vaccination coefficient in several instances. The percentages of county residents uninsured, suffering from diabetes or smokers all had significant effects during waves 4-6. Medicaid expansion had significant effects: negative in wage 4 but positive in waves 3, 5 and 6. There is no detectable impact of poverty rate on the impact of vaccination rates.

6 Concluding remarks

This study analyzed two years of U.S. county daily data on COVID-19 cases, deaths, and vaccinations across six distinct waves from March 2020 to March 2022. The two-stage modeling approach allowed unobservable factors to affect both virus transmissibility and mortality risk in each wave. The cross-sectional coefficients for cases, deaths, and vaccinations produced in the first stage and used as dependent variables in the second stage helped identify key socioeconomic, demographic, health, and climate drivers of the outcomes. This flexible approach with wave-specific models provided insights into how factors influencing the pandemic's severity and treatment evolved over time and space as the virus mutated. The variation in the estimated coefficients across waves likely reflects increased transmissibility but lower deadliness of new variants, along with improvements in treatment. By modeling each wave separately, this study illustrated which factors played significant roles in COVID-19 spread, mortality, and vaccination rates across U.S. counties.

A key finding from this study is that the socioeconomic and demographic factors that we studied generally had a more significant influence on COVID-19 outcomes compared to health and climate factors, though those effects weakened at certain times. In particular, after wave 3 when COVID-19 vaccines were first administered and in the final wave, the progression from cases to deaths diminished. This suggests vaccines and improved treatments moderated the effects of risk factors on mortality as the pandemic progressed. However, socioeconomic and demographic variables remained more influential drivers overall compared to health and climate throughout the six distinct waves of the sample period.

Acknowledgments

We gratefully acknowledge the support of Alianza EFI-Colombia Científica grant with code 60185 and FP44842-220-2018. Errors, opinions and omissions are our own and do not represent our institutions. We also thank participants in the 2022 Swiss, 2023 German, 2023 Northern European and 2023 UK Stata Conferences for their helpful comments and suggestions on earlier drafts. The standard disclaimer applies.

References

Ahmad, F. B., J. A. Cisewski, and R. N. Anderson (2022). Provisional Mortality Data – United States, 2021. *Morbidity and Mortality Weekly Report 71*(17), 598–600. https://www.cdc.gov/mmwr/volumes/71/wr/mm7117e1.htm.

Ahmad, F. B., J. A. Cisewski, J. Xu, and R. N. Anderson (2023). Provisional Mortality Data - United States, 2022. *Morbidity and Mortality Weekly Report 72*(18), 488–492. https://www.cdc.gov/mmwr/volumes/72/wr/mm7218a3.htm.

Ahmed, R. and M. H. Pesaran (2022). Regional heterogeneity and U.S. presidential elections: Real-time 2020 forecasts and evaluation. *International Journal of Forecasting 38*(2), 662–687. https://doi.org/10.1016/j.ijforecast.2021.06.007.

Allcott, H., L. Boxell, J. Conway, B. Ferguson, M. Gentzkow, and B. Goldman (2020). What explains temporal and geographic variation in the early us coronavirus pandemic? NBER Working Papers 27965, National Bureau of Economic Research, Inc. http://www.nber.org/papers/w27965.

Andersen, L. M., S. R. Harden, M. M. Sugg, J. D. Runkle, and T. E. Lundquist (2021). Analyzing the spatial determinants of local Covid-19 transmission in the United States. *Science of the Total Environment 754*(142396), 1–10. https://doi.org/10.1016/j.scitotenv.2020.142396.

Aron, J. and J. Muellbauer (2022). Excess Mortality Versus COVID-19 Death Rates: A Spatial Analysis of Socioeconomic Disparities and Political Allegiance Across U.S. States. *The Review of Income and Wealth 68*(2), 348–392. https://doi.org/10.1111/roiw.12570.

Baum, C. and M. Henry (2022). Socio-economic and demographic factors influencing the spatial spread of COVID-19 in the USA. *International Journal Computational Economics and Econometrics 12*, 366–380. https://doi.org/10.1504/IJCEE.2022.126313.

Bollyky, T. J., E. Castro, A. Y. Aravkin, K. Bhangdia, J. Dalos, E. N. Hulland, S. Kiernan, A. Lastuka, T. A. McHugh, S. M. Ostroff, P. Zheng, H. T. Chaudhry, E. Ruggiero, I. Turilli, C. Adolph, J. O. Amlag, B. Bang-Jensen, R. M. Barber, A. Carter, C. Chang, R. M. Cogen, J. K. Collins, X. Dai, W. J. Dangel, C. Dapper, A. Deen, A. Eastus, M. Erickson, T. Fedosseeva, A. D. Flaxman, N. Fullman, J. R. Giles, G. Guo, S. I. Hay, J. He, M. Helak, B. M. Huntley, V. C. Iannucci, K. E. Kinzel, K. E. LeGrand, B. Magistro, A. H. Mokdad, H. Nassereldine, Y. Ozten, M. Pasovic, D. M. Pigott, R. C. R. Jr, G. Reinke, A. E. Schumacher, E. Serieux, E. E. Spurlock, C. E. Troeger, A. T. Vo, T. Vos, R. Walcott, S. Yazdani, C. J. L. Murray, and J. L. Dieleman (2023). Assessing COVID-19 pandemic policies and behaviours and their economic and educational trade-offs across US states from jan 1, 2020, to july 31, 2022: an observational analysis. *The Lancet April*(401), 1341–1360. https://doi.org/10.1016/S0140-6736(23)00461-0.

Breusch, T. S. and A. R. Pagan (1980). The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics. *The Review of Economic Studies 47*(1), 239–253. https://doi.org/10.2307/2297111.

Brown, C. S. and M. Ravallion (2020). Inequality and the Coronavirus: Socioeconomic Covariates of Behavioral Responses and Viral Outcomes Across US Counties. NBER Working Papers 27549, National Bureau of Economic Research, Inc. http://www.nber.org/papers/w27549.

Carozzi, F., S. Provenzano, and S. Roth (2022). Urban density and COVID-19: understanding the US experience. *The Annals of Regional Science*. https://doi.org/10.1007/s00168-022-01193-z.

Chen, L., J. J. Dolado, and J. Gonzalo (2021). Quantile Factor Models. *Econometrica 89*, 875–910. https://doi.org/10.3982/ECTA15746.

Chernozhukov, V., H. Kasahara, and P. Schrimpf (2021). Causal impact of masks, policies, behavior on early covid-19 pandemic in the U.S. *Journal of Econometrics 220*, 23–62. https://doi.org/10.1016/j.jeconom.2020.09.003.

Chin, T., R. Kahn, R. Li, J. T. Chen, N. Krieger, C. O. Buckee, S. Balsari, and M. V. Kiang (2020). US-county level variation in intersecting individual, household and community characteristics relevant to COVID-19 and planning an equitable response: a cross-sectional analysis. *BMJ Open 10*(9). https://doi.org/10.1136/bmjopen-2020-039886.

Chudik, A., G. Kapetanios, and M. H. Pesaran (2018). A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models. *Econometrica 86*(4), 1479–1512. https://doi.org/10.3982/ECTA14176.

Chudik, A. and M. H. Pesaran (2015). Large panel data models with cross-sectional dependence a survey. In B. H. Baltagi (Ed.), *The Oxford Handbook of Panel Data*. New York: Oxford University Press.

Chudik, A., M. H. Pesaran, and M. Sharifvaghefi (2023). Variable Selection in High Dimensional Linear Regressions with Parameter Instability. Working papers. https://arxiv.org/pdf/2312.15494.pdf.

Desboulets, L. (2018). A Review on Variable Selection in Regression Analysis. *Econometrics 6*, 1–27. https://doi.org/10.3390/econometrics6040045.

Desmet, K. and R. Wacziarg (2022). Understanding spatial variation in COVID-19 across the United States. *Journal of Urban Economics 127*(103332), 1–10. https://doi.org/10.1016/j.jue.2021.103332.

Dwyer-Lindgren, L., P. Kendrick, Y. O. Kelly, D. O. Sylte, C. Schmidt, B. F. Blacker, F. Daoud, A. A. Abdi, M. Baumann, F. Mouhanna, E. Kahn, S. I. Hay, G. A. Mensah, A. M. Nápoles, E. J. Pérez-Stable, M. Shiels, N. Freedman, E. Arias, S. A. George, D. M. Murray, J. W. R. Phillips, M. L. Spittel, C. J. L. Murray, and A. H. Mokdad (2022). Life expectancy by county, race, and ethnicity in the USA, 2000–19: a systematic analysis of health disparities. *The Lancet 400*, 25–38. https://doi.org/10.1016/S0140-6736(22)00876-5.

Flanagan, B. E., E. W. Gregory, E. J. Hallisey, J. L. Heitgerd, and B. Lewis (2011). A Social Vulnerability Index for Disaster Management. *Journal of Homeland Security and Emergency Management 8*(1), 1–22. https://doi.org/10.2202/1547-7355.1792.

Flanagan, B. E., E. J. Hallisey, E. Adams, and A. Lavery (2018). Measuring Community Vulnerability to Natural and Anthropogenic Hazards: The Centers for Disease Control and Prevention's Social Vulnerability Index. *Journal of Environmental Health 80*(10), 34–36. https://pubmed.ncbi.nlm.nih.gov/32327766.

Genusso, K., M. Stevenson, C. Muganda, S. Johnson, and M. Givens (2022). County health rankings national findings 2022. Technical report, University of Wisconsin Population Health Institute. Available at https://www.countyhealthrankings.org/2022-measure.

Giulietti, M., A. Iregui, and J. Otero (2014). Crude oil price differentials, product heterogeneity and institutional arrangements. *Energy Economics 46*, S28–S32. https://doi.org/10.1016/j.eneco.2014.10.006.

Haimerl, P. and T. Hartl (2023). Modeling COVID-19 Infection Rates by Regime-Switching Unobserved Components Models. *Econometrics 11*(2), 1–15. https://doi.org/10.3390/econometrics11020010.

Ho, P., T. Lubik, and C. Matthes (2023). How to go viral: A COVID-19 model with endogenously time-varying parameters. *Journal of Econometrics 232*, 70–86. https://doi.org/10.1016/j.jeconom.2021.01.001.

Holshue, M. L., C. DeBolt, S. Lindquist, K. H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural, G. Diaz, A. Cohn, L. Fox, A. Patel, S. I. Gerber, L. Kim, S. Tong, X. Lu, S. Lindstrom, M. A. Pallansch, W. C. Weldon, H. M. Biggs, T. M. Uyeki, and S. K. Pillai (2020). First Case of 2019 Novel Coronavirus in the United States. *New England Journal of Medicine 382*(10), 929–936. http://hdl.handle.net/10.1056/NEJMoa2001191.

Hornstein, A. S. and W. H. Greene (2012). Usage of an estimated coefficient as a dependent variable. *Economics Letters 116*(3), 316–318.

Hsiao, C. and M. H. Pesaran (2008). Random Coefficient Models. In L. Mátyás and P. Sevestre (Eds.), *The Econometrics of Panel Data Fundamentals and Recent Developments in Theory and Practice* (Third ed.), Volume 33, pp. 185–213. Springer.

IHME (2023). COVID-19 Projections. Technical report, Institute for Health Metrics and Evaluation. Available at https://covid19.healthdata.org/united-states-of-america?view=mask-use&tab=trend.

Iregui, A. M., H. M. Núñez, and J. Otero (2021). Testing the efficiency of inflation and exchange rate forecast revisions in a changing economic environment. *Journal of Economic Behavior and Organization 187*(C), 290–314. https://doi.org/10.1016/j.jebo.2021.04.037.

Jin, R. (2021). The Lag between Daily Reported Covid-19 Cases and Deaths and Its Relationship to Age. *Journal of Public Health Research 10*(3), 2049. https://doi.org/10.4081%2Fjphr.2021.2049.

Jones, B. (2022). The changing political geography of covid-19 over the last two years. Technical report, Pew Research Center. Available at https://www.pewresearch.org/politics/2022/03/03/the-changing-political-geography-of-covid-19-over-the-last-two-years/.

Knittel, C. R. and B. Ozaltun (2020). What Does and Does Not Correlate with COVID-19 Death Rates. NBER Working Papers 27391, National Bureau of Economic Research, Inc. https://www.nber.org/papers/w27391.

Liao, T. F. and F. De Maio (2021). Association of Social and Economic Inequality With Coronavirus Disease 2019 Incidence and Mortality Across US Counties. *JAMA Network Open 4*(1), 1–10. http://jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2020.34578.

MacKinnon, J. G., M. Ørregaard Nielsen, and M. D. Webb (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics 232*(2), 272–299. https://doi.org/10.1016/j.jeconom.2022.04.001.

Martinez-Beneito, M. A., M. Marí-Dell'Olmo, N. Sánchez-Valdivia, M. Rodríguez-Sanz, G. Pérez, M. I. Pasarín, C. Rius, L. Artazcoz, R. Prieto, K. Pérez, and C. Borrell (2023). Socioeconomic inequalities in COVID-19 incidence during the first six waves in Barcelona. *International Journal of Epidemiology 68*. https://doi.org/10.1093/ije/dyad105.

MIT Election Data and Science Lab (2018). County presidential election returns 2000-2020. https://doi.org/10.7910/DVN/VOQCHQ.

Mukherji, N. (2022). The Social and Economic Factors Underlying the Incidence of COVID-19 Cases and Deaths in US Counties During the Initial Outbreak Phase. *The Review of Regional Studies 52*(52), 127–150. https://doi.org/10.52324/001c.35255.

Naqvi, A. (2022). BIMAP: Stata module to produce bivariate maps. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s459063.html.

Núñez, H. M. and J. Otero (2020). OCMT: Stata module to perform multiple testing approach in high-dimensional linear regression. Statistical Software Components, Boston College Department of Economics.

Núñez, H. M. and J. Otero (2021). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models: A replication in a narrow sense. *Journal of Applied Econometrics 36*(6), 833–841. https://doi.org/10.1002/jae.2850.

Oberhofer, W. and J. Kmenta (1974). A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models. *Econometrica 42*(3), 579–590. https://doi.org/10.2307/1911792.

Oronce, C. I., C. A. Scannell, I. Kawachi, and Y. Tsugawa (2020). Association Between State-Level Income Inequality and COVID-19 Cases and Mortality in the USA. *Journal of General Internal Medicine 35*(9), 2791–2793. https://doi.org/10.1007%2Fs11606-020-05971-3.

Ortaliza, J., K. Amin, and C. Cox (2022). COVID-19 leading cause of death ranking. Technical report, Peterson-KFF Health System Tracker. https://www.kff.org/coronavirus-covid-19/issue-brief/covid-19-leading-cause-of-death-ranking/.

Papageorge, N., M. Zahn, M. Belot, E. van den Broek-Altenburg, S. Choi, J. Jamison, and E. Tripodi (2021). Socio-demographic factors associated with self-protecting behavior during the covid-19 pandemic. *Journal of Population Economics 34*, 691–738. https://doi.org/10.1007/s00148-020-00818-x.

Picard, R. and M. Stepner (2015). GEO2XY: Stata module to convert latitude and longitude to xy using map projections. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s457990.html.

PolicyMap (2020). Severe COVID-19 health risk index. Technical report, PolicyMap [accessed on May 2020]. Available at https://PolicyMap.com.

Rabe-Hesketh, S. and A. Skrondal (2022). *Multilevel and Longitudinal Modeling Using Stata* (Fourth ed.), Volume I: Continuous Responses. College Station, Texas: Stata Press.

Raifman, J., K. Nocka, D. Jones, J. Bor, S. Lipson, J. Jay, M. Cole, P., K. N., P. Chan, and S. Galea (2020). COVID-19 US state policy database. Technical report, Inter-University Consortium for Political and Social Research. https://statepolicies.com/.

Saxonhouse, G. R. (1976). Estimated parameters as dependent variables. *American Economic Review 66*(1), 178–183.

Snyder, J. P. (1984). *Map Projections used by the U.S. Geological Survey. Geological Survey Bulletin 1532* (Second ed.). Washington D.C.: U.S. Department of the Interior.

Tibshirani, R. J. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Wang, H., K. R. Paulson, S. A. Pease, S. Watson, H. Comfort, P. Zheng, A. Y. Aravkin, C. Bisignano, R. M. Barber, T. Alam, J. E. Fuller, E. A. May, D. P. Jones, M. E. Frisch, C. Abbafati, C. Adolph, A. Allorant, J. O. Amlag, B. Bang-Jensen, G. J. Bertolacci, S. S. Bloom, A. Carter, E. Castro, S. Chakrabarti, J. Chattopadhyay, R. M. Cogen, J. K. Collins, K. Cooperrider, X. Dai, W. J. Dangel, F. Daoud, C. Dapper, A. Deen, B. B. Duncan, M. Erickson, S. B. Ewald, T. Fedosseeva, A. J. Ferrari, J. J. Frostad, N. Fullman, J. Gallagher, A. Gamkrelidze, G. Guo, J. He, M. Helak, N. J. Henry, E. N. Hulland, B. M. Huntley, M. Kereselidze, A. Lazzar-Atwood, K. E. LeGrand, A. Lindstrom, E. Linebarger, P. A. Lotufo, R. Lozano, B. Magistro,

D. C. Malta, J. Månsson, A. M. M. Herrera, F. Marinho, A. H. Mirkuzie, A. T. Misganaw, L. Monasta, P. Naik, S. Nomura, E. G. O'Brien, J. K. O'Halloran, L. T. Olana, S. M. Ostroff, L. Penberthy, R. C. Reiner Jr, G. Reinke, A. L. P. Ribeiro, D. F. Santomauro, M. I. Schmidt, D. H. Shaw, B. S. Sheena, A. Sholokhov, N. Skhvitaridze, R. J. D. Sorensen, E. E. Spurlock, R. Syailendrawati, R. Topor-Madry, C. E. Troeger, R. Walcott, A. Walker, C. S. Wiysonge, N. A. Worku, B. Zigler, D. M. Pigott, M. Naghavi, A. H. Mokdad, S. S. Lim, S. I. Hay, E. Gakidou, and C. J. L. Murray (2022). Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21. *The Lancet March*(10), 1–24. https://doi.org/10.1016/S0140-6736(21)02796-3.

Welsch, D. (2022). The Impact of Mask Usage on COVID-19 Deaths: Evidence from US Counties Using a Quasi-Experimental Approach. *The B.E. Journal of Economic Analysis & Policy 22*, 1–28. https://doi.org/10.1515/bejeap-2021-0157.

Wu, X., R. C. Nethery, M. B. Sabath, D. Braun, and F. Dominici (2020). Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Science Advances 6*(45), 1–6. https://doi.org/10.1126/sciadv.abd4049.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias. *Journal of the American Statistical Association 57*, 500–509. https://doi.org/10.1080/01621459.1962.10480664.

Table 1: Descriptive statistics of county-specific drivers (N = 3,014)

	Mean	SD	Min	Max
Socioeconomic variables				
Age 1-19 yrs (%)	24.24	3.54	5.86	45.31
Age 20-39 yrs (%)	24.00	4.07	10.87	54.35
Age 40-59 yrs (%)	24.48	2.09	10.67	34.27
Age 60-79 yrs (%)	22.44	4.65	5.48	54.64
Black (%)	10.49	14.63	0.05	86.94
Hispanic (%)	9.78	13.60	0.61	96.33
Male (%)	50.04	2.15	43.18	72.96
Median income (log)	10.93	0.23	10.17	11.98
HS completion (%)	0.88	0.06	0.49	0.99
Some college (%)	0.59	0.12	0.19	0.92
Poverty (%)	13.70	5.38	3.00	43.90
Owner-occupied housing (%)	71.52	8.08	19.61	92.40
Democratic share 2020 (%)	33.55	15.80	5.04	92.15
Population density (log)	2.95	1.69	-2.26	10.22
Social vulnerability index	0.50	0.29	0.00	1.00
Health variables				
Uninsured (%)	13.88	6.07	2.78	43.45
Diabetes (%)	10.77	2.29	5.50	21.00
Smoking (%)	20.36	4.14	6.50	38.20
Life expectancy (years)	77.38	2.59	64.50	91.72
Health risk index	0.01	1.00	-3.65	3.27
Access to exercise (%)	55.33	23.48	0.00	100.00
ICU (beds per capita)	12.99	23.92	0.00	757.36
Medicaid expansion	0.64	0.48	0.00	1.00
Climate and pollution variables				
Summer avg. temperature (°C)	30.94	3.15	19.40	41.72
Summer rel. humidity (%)	89.03	9.69	31.64	99.78
Winter avg. temperature (°C)	8.21	6.58	-7.46	26.19
Winter rel. humidity (%)	87.47	4.81	58.16	97.67
$PM_{2.5}$ (µg per cubic meter)	6.09	1.39	2.49	12.37

Notes: All variables in Table 1 are at the county level, except for the Medicaid expansion indicator which is at the state level. The sample size N represents the number of U.S. counties (FIPS) included. SD gives the standard deviation. HS stands for high school completion. ICU beds per capita are expressed per 100,000 population.

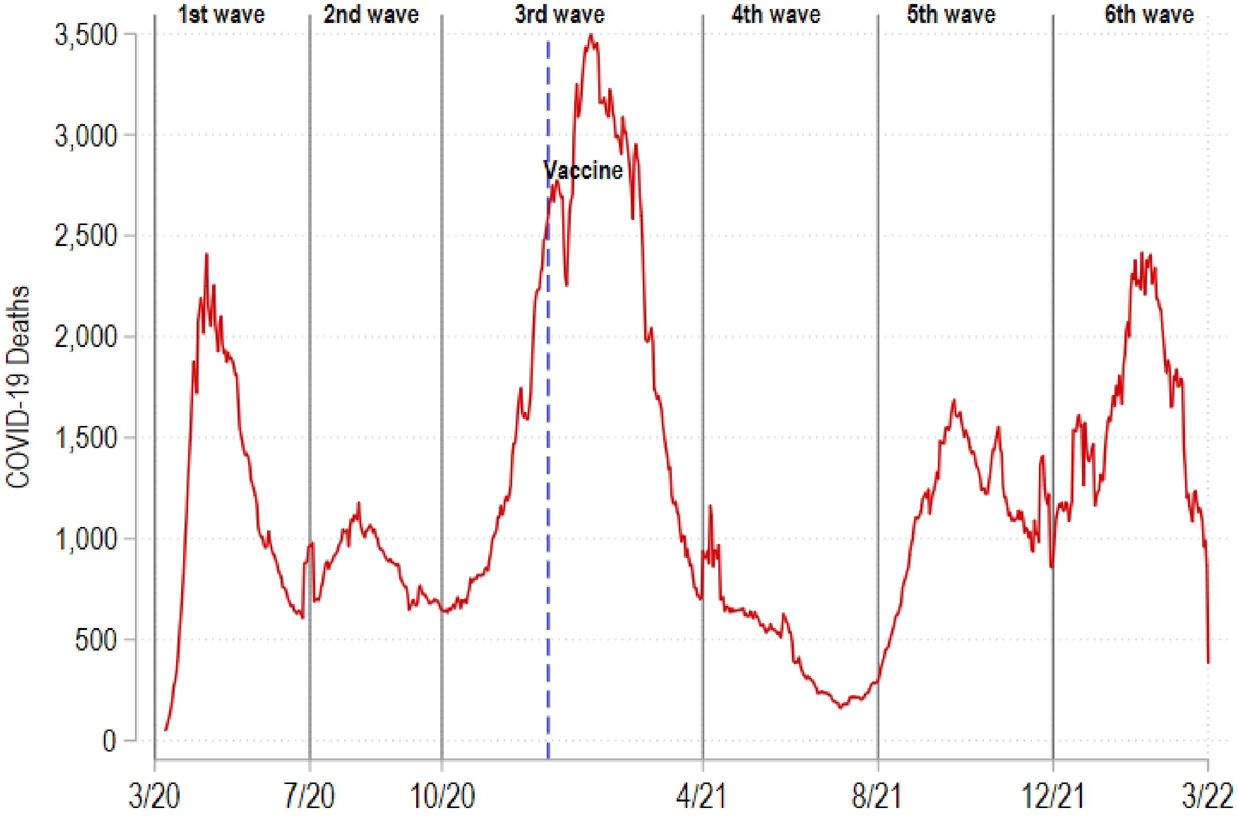
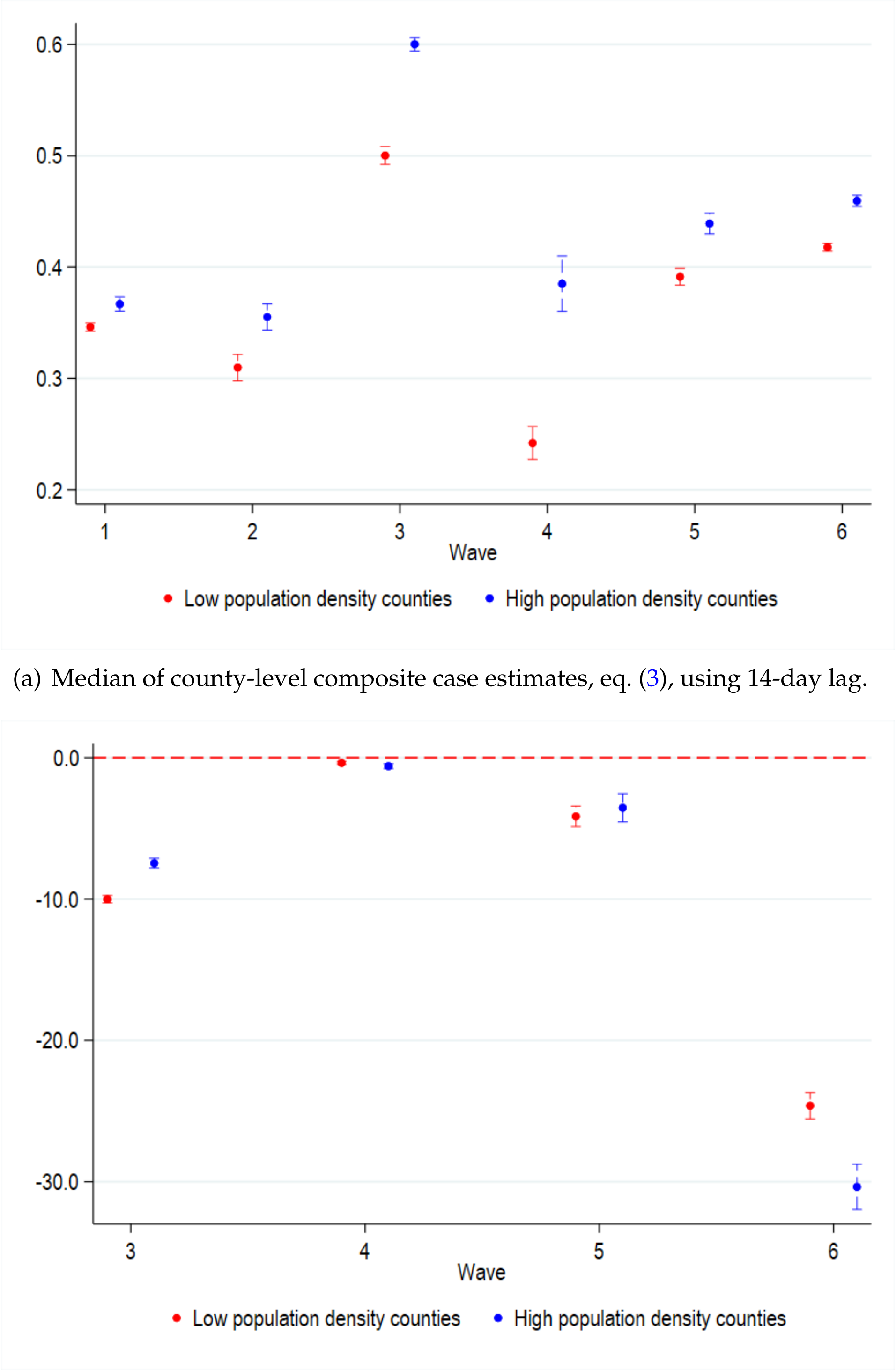


Figure 1: **Daily moving average of COVID-19 deaths in the United States** (excluding U.S. territories and freely associated states) **from March 15,2020 through March 19,2022**. The six waves depicted are defined in Jones (2022), with the first wave starting March 15, 2020 and the last ending March 19, 2022. The solid vertical lines denote the beginning of each new wave. The x-axis shows the month and year. The blue vertical dashed line indicates December 14, 2020, when COVID-19 vaccine were first administrated to the U.S. Public. Wave 3 partially captures the initial vaccine rollout in mid-December 2020, wave 5 reflects the Delta variant surge, and wave 6 shows the Omicron variant dominating. The moving averages were generated using an uncentered 7-term moving average filter. Data Source: USAFacts and U.S. Department of Health and Human Services.

Table 2: Average of confirmed cases, deaths, and vaccination rates by wave

Wave starting: (Month/Year)		3/20	7/20	10/20	4/21	8/21	12/21
Cases	Total	522.71	1,962.47	9,394.12	10,595.04	15,728.08	23,992.87
	Low	479.73	1,959.10	9,573.60	10,731.38	16,117.79	24,172.88
	High	651.76	1,972.59	8,855.23	10,185.63	14,557.91	23,452.38
Deaths	Total	17.65	43.95	189.10	211.17	284.94	357.31
	Low	13.83	41.52	199.08	222.26	303.84	379.28
	High	29.12	51.26	159.13	177.85	228.16	291.35
Vaccinations	Total	0.00	0.00	13.66	32.65	45.54	50.90
	Low	0.00	0.00	13.71	30.89	43.42	48.47
	High	0.00	0.00	13.49	37.93	51.91	58.20

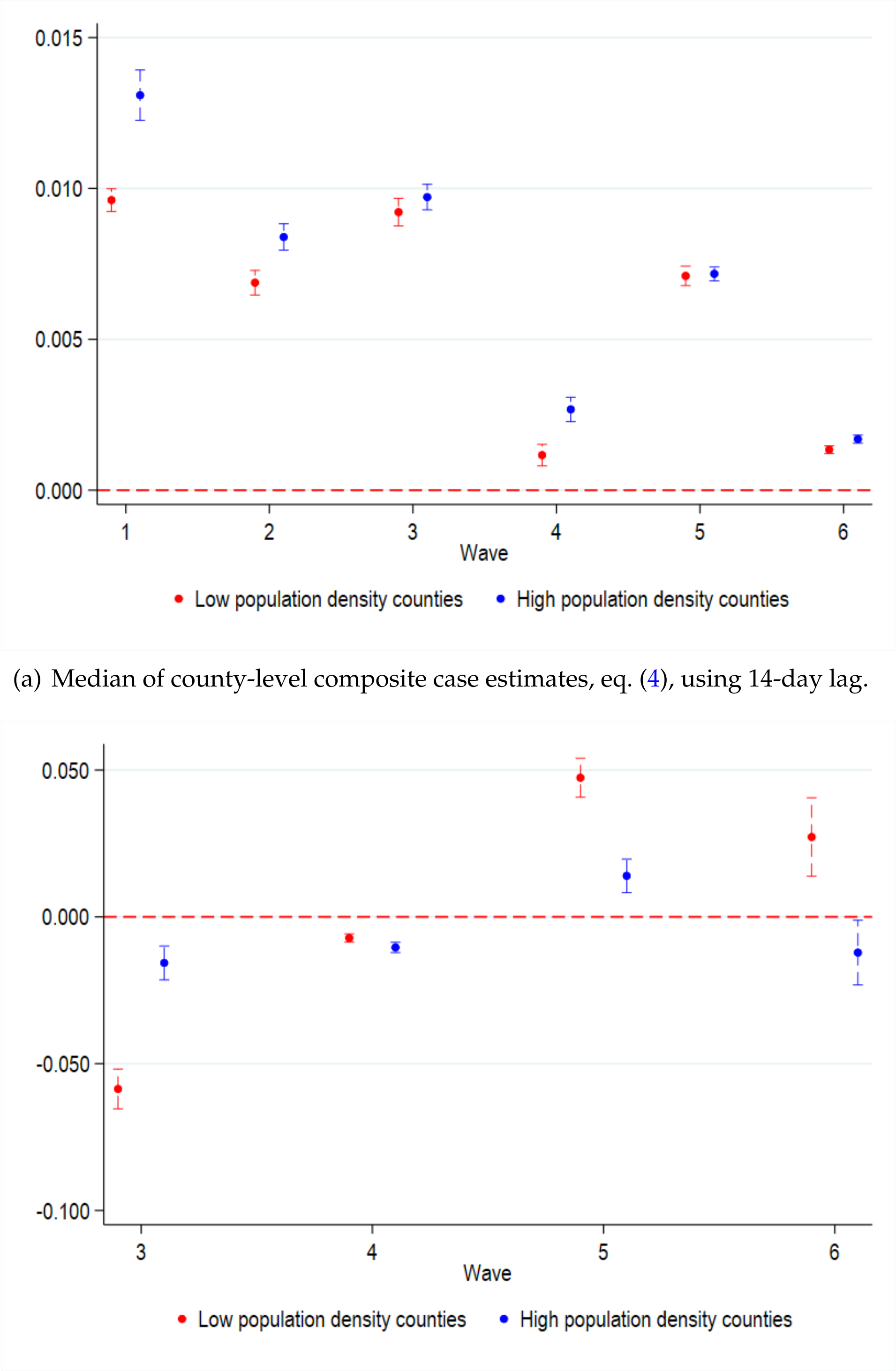
Notes: Waves are defined in Jones (2022) with the end date for the last wave on March 19, 2022. The reported calculations in this table are based on the cumulative cases, cumulative deaths, and cumulative vaccination rates recorded at the end of each wave within each county. Cases and deaths reflect variations in population-adjusted cumulative rates by county. Vaccinations are county-level rates of people who were fully vaccinated against COVID-19. Low and high cases are the average of cases for counties with population density in the first three quartiles (2,261 counties) and in the fourth quartile (753 counties), respectively. Similar calculations apply to deaths and vaccinations.



(a) Median of county-level composite case estimates, eq. (3), using 14-day lag.

(b) Median of county-level composite vaccination estimates, eq. (5), using 14-day lag.

Figure 2: Median of county-level composite estimates from the mixed model results for the confirmed case rate, equation (1), with 95% confidence intervals by wave for low and high population density counties (N = 3,014). High density counties are those with population density in the fourth quartile (753 counties). Low density counties fall in the first three quartiles (2,261 counties).



(a) Median of county-level composite case estimates, eq. (4), using 14-day lag.

(b) Median of county-level composite vaccination estimates, eq. (6), using 14-day lag.

Figure 3: Median of county-level composite estimates from the mixed model results for the confirmed death rate, equation (2), with 95% confidence intervals by wave for low and high population density counties (N = 3,014). High density counties are those with population density in the fourth quartile (753 counties). Low density counties fall in the first three quartiles (2,261 counties).

Table 3: Drivers of 14-day lagged case rates in the case model ($\hat{\beta}_0 + \hat{\beta}_i$ in eq. 1)

Wave starting (Month/Year):	3/20	7/20	10/20	4/21	8/21	12/21
Socioeconomic demographic variables						
Age 1-19 years (%)	2.3747***	1.3922***	0.0651	1.8694***	-0.3896	0.0346
Age 20-39 years (%)	-1.2581***	-0.9886**	-0.9949		0.1420	
Age 40-59 years (%)	-1.2873*	-0.6613	0.1099	0.0345	0.1492	0.1128
Age 60-79 years (%)			-0.4904	-0.0714	-0.2938	
Black (%)	-0.3897***	0.4109***	0.1163*	0.4941**	-0.3967***	-0.0230
Hispanic (%)	-0.1779	0.4199**	0.5044***	0.3999**		0.0805
Male (%)	0.2800	0.0870	0.1165	0.6203	-0.4741	-0.2602*
Median income (log)	-7.2108	-1.9213	6.0835	-44.5141***	3.6967	
HS completion (%)	21.4376	19.0636	38.9028*	45.5196	42.0529***	3.3322
Some college (%)	-16.7773	-0.7530	8.1486		-15.1328*	4.3463
Poverty (%)	0.6316	0.6147**	0.5957*	-0.3337	-0.1166	
Owner-occupied housing (%)	-0.3257		-0.3216***	0.6845***	0.1910**	
Democratic share 2020 (%)	0.4818***		-0.1497*	0.4675***	0.1663**	0.1490**
Population density (log)	2.8756**	4.4504***	2.3250***	0.8094	3.0533***	1.6812***
Social vulnerability index		-6.0135	-0.2703	22.1052***	8.8651***	
Health variables						
Uninsured (%)		-0.2729	-0.7198***			-0.4958***
Diabetes (%)		-4.8183***	-0.0906	-13.2329***	-0.5232	-1.4475***
Smoking (%)	-0.0270	1.4753*		4.4344***	1.8721***	0.9387***
Life expectancy (years)	1.6445**	0.5559	-0.6823*	2.4125***	-0.4739	
Health risk index	-0.9438	0.2177	0.4593	-2.8733	1.6706**	2.0782***
Access to exercise (%)	-0.0376	-0.0495	-0.0159			
ICU (beds per capita per 100K)	-0.0167	0.0400	0.0007	0.0696		0.0213*
Medicaid expansion	5.6995**	-1.9527	-2.2676	26.3241***	-9.0255***	-3.3048***
Climate and pollution variables						
Summer avg. temperature (°C)		1.0407	0.7719*	2.4289***	-1.3772***	
Summer rel. humidity (%)	0.2646	-0.3324**		0.6984**	-0.4562***	-0.1581***
Winter avg. temperature (°C)	1.3312***	-0.9743***	-1.4947***	3.5748***	0.6113***	0.6807***
Winter rel. humidity (%)	-0.9555***		-0.0657		0.4745***	0.2360***
$PM_{2.5}$ (µg per cubic meter)	-1.3700	-0.2569	-0.2336	-2.6660*	-0.0760	0.9296***
Constant	72.9941***	44.2813***	454.0419***	140.3864***	239.7158***	281.0088***
N	3,014	3,014	3,014	3,014	3,014	3,014

Notes: The dependent variable (3) was scaled by 100 in estimation to facilitate the interpretation of the model coefficients. Standard errors for each point estimate were computed by performing nonparametric bootstrap estimation over 1,000 replications, but these are not reported for brevity. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Drivers of 14-day lagged vaccination rates in the case model ($\hat{\gamma}_0 + \hat{\gamma}_i$ in eq. 1)

Wave starting: (Month/Year)	10/20	4/21	8/21	12/21
Socioeconomic demographic variables				
Age 1-19 years (%)	0.0278	0.1479***	-0.7588***	-0.2742
Age 20-39 years (%)		-0.0082	-0.8684***	0.0074
Age 40-59 years (%)			-0.8927***	
Age 60-79 years (%)	0.2404***	-0.0068	-0.6310***	-0.2790
Black (%)		0.0128**	0.0007	-0.2415***
Hispanic (%)		-0.0472***	0.1098***	-0.1627**
Male (%)	-0.0529		0.4443***	
Median income (log)	-2.2927*	-0.2054	-2.5085	5.2935***
HS completion (%)	9.9755	-3.3032	28.5046***	-6.7053
Some college (%)	3.6313	-2.1258***	4.8738*	-6.9593
Poverty rate (%)	-0.2929***	0.0709***	-0.1430*	0.4635***
Owner-occupied housing (%)		-0.0231		-0.0426
Democratic share 2020 (%)	0.0767***		0.0360*	0.0636
Population density (log)	0.7317***	-0.4196***	0.2774	-0.6847
Social vulnerability index	1.5716	-0.9975*	-0.7633	-3.0629
Health variables				
Uninsured (%)	0.0648*	0.0236	0.2313***	0.3110***
Diabetes (%)	0.5451**	-0.3225**	0.8442***	0.3513
Smoking (%)	0.0649		-0.8478***	-0.4052*
Life expectancy (years)	0.0638	0.1126	0.0715	-0.0565
Health risk index	-1.0438***	-0.1650	-0.3710	2.1537***
Access to exercise (%)	-0.0194**	0.0000	-0.0164**	0.0034
ICU (beds per capita per 100K)	-0.0056			
Medicaid expansion	1.2497***	0.2041	3.0729***	
Climate and pollution variables				
Summer avg. temperature (°C)	-0.5874***	0.0067	1.4884***	-0.4456**
Summer rel. humidity (%)			0.1139***	-0.1337***
Winter avg. temperature (°C)	0.2510***	0.2361***	-1.3917***	0.0809
Winter rel. humidity (%)	0.0869***	-0.0387***	0.1352***	0.1063
$PM_{2.5}$ (µg per cubic meter)		0.2375***	-1.0964***	-1.9314***
Constant	-1.2002***	-0.7157**	-1.3948***	-2.0162***
N	3014	3014	3014	3014

Notes: Standard errors for each point estimate were computed by performing non-parametric bootstrap estimation over 1,000 replications, but these are not reported for brevity. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Drivers of 14-day lagged case rates in the death model ($\hat{\kappa}_0 + \hat{\kappa}_i$ in eq. 2)

Wave starting: (Month/Year)	3/20	7/20	10/20	4/21	8/21	12/21
Socioeconomic demographic variables						
Age 1-19 years (%)		0.1066***	-0.0384		-0.0100	
Age 20-39 years (%)	0.0559	0.0538**	-0.0615**			
Age 40-59 years (%)			-0.0778**		0.0084	
Age 60-79 years (%)	0.1242**	0.1187***	-0.0118			
Black (%)	0.0036		0.0050*		0.0088***	0.0020***
Hispanic (%)			0.0170***		0.0007	
Male (%)	-0.1411***	-0.0247	-0.0092		-0.0283***	-0.0037***
Median income (log)	1.0144*		0.6942**	0.0006	0.4579***	
HS completion (%)	-6.3305*		0.3281		1.7128**	
Some college (%)	3.0123**	0.2246	0.1432		0.0109	
Poverty rate (%)		0.0066	0.0463***		0.0168**	
Owner-occupied housing (%)	-0.0020	0.0150				0.0024***
Democratic share 2020 (%)	0.0133				-0.0110***	
Population density (log)	0.4787***		0.0530		-0.0210	
Social vulnerability index			0.1781		0.3161***	
Health variables						
Uninsured (%)	-0.0298		-0.0157**		0.0202***	
Diabetes (%)			-0.0627		0.0045	
Smoking (%)	0.0503	-0.0400**		0.0020	-0.0089	
Life expectancy (years)	0.0033	-0.0520**	-0.0435**		-0.0445***	
Health risk index		0.1312		0.1843***	0.0520	
Access to exercise (%)	-0.0146**		-0.0047**			
ICU (beds per capita per 100K)	-0.0109*					0.0009
Medicaid expansion			-0.1022	0.3100***	0.0087	
Climate and pollution variables						
Summer avg. temperature (°C)			0.0421**		-0.0589***	
Summer rel. humidity (%)						
Winter avg. temperature (°C)	0.0393**		-0.0586***		-0.0022	
Winter rel. humidity (%)	-0.0378**	-0.0095	0.0029		-0.0016	
$PM_{2.5}$ (µg per cubic meter)	-0.2080*				0.0393**	
Constant	-15.5367	11.8505	142.4560***	36.7336***	160.7601***	109.3487***
N	3014	3014	3014	3014	3014	3014

Note: The dependent variable (4) was scaled by 100 in estimation to facilitate the interpretation of the model coefficients. Standard errors for each point estimate were computed by performing nonparametric bootstrap estimation over 1,000 replications, but these are not reported for brevity. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Drivers of 14-day lagged vaccination rates in the death model ($\hat{\lambda}_0 + \hat{\lambda}_i$ in eq. 2)

Wave starting: (Month/Year)	10/20	4/21	8/21	12/21
Socioeconomic demographic variables				
Age 1-19 years (%)		0.0201	-0.4221*	0.0524
Age 20-39 years (%)			-0.7254***	
Age 40-59 years (%)	0.5771***		-0.2453	
Age 60-79 years (%)			-0.7566***	-0.4118**
Black (%)		0.0754*		0.1707***
Hispanic (%)		0.2498***	0.1172***	
Male (%)	-0.1792			-0.2492
Median income (log)	2.0827	-1.6051	2.4298	
HS completion (%)		-17.4913**		4.3661
Some college (%)	9.0750*	12.7743***		2.2008
Poverty rate (%)		-0.0025		0.1880
Owner-occupied housing (%)	-0.1113**		-0.1728***	-0.0631
Democratic share 2020 (%)	0.0067	-0.0412		
Population density (log)	0.6765**			
Social vulnerability index		-0.0698		-2.3029
Health variables				
Uninsured (%)	-0.0134	-0.9247***	0.1194*	0.8489***
Diabetes (%)		-0.0687	-1.4792***	-2.3155***
Smoking (%)		0.4853***	1.7480***	
Life expectancy (years)		0.6465**	0.2584	0.2238
Health risk index		0.4912	-0.6527	3.9141***
Access to exercise (%)	-0.0218	-0.0296**	-0.0000	
ICU (beds per capita per 100K)			-0.0253*	
Medicaid expansion	6.3810***	-4.2273***	6.5334***	6.1602***
Climate and pollution variables				
Summer avg. temperature (°C)	-1.0521***	-0.5873***	0.0758	0.2286
Summer rel. humidity (%)				0.1823***
Winter avg. temperature (°C)	0.7509***	-0.0900	0.2214***	0.8311***
Winter rel. humidity (%)	-0.0723		-0.1807***	-0.2679**
$PM_{2.5}$ (μg per cubic meter)				0.1191
Constant	-23.3391*	-108.6129***	94.8942***	-0.3167
N	3014	3014	3014	3014

Note: The dependent variable (6) was scaled by 100 in estimation to facilitate the interpretation of the model coefficients. Standard errors for each point estimate were computed by performing nonparametric bootstrap estimation over 1,000 replications, but these are not reported for brevity. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Appendix A Data characteristics and their configuration

All the data below were merged for 3,014 counties (FIPS), out of 3,108 total counties. We constructed datasets for counties in the 48 U.S. contiguous states and Washington D.C. with no missing data. The data covers March 15, 2020 through March 19, 2022. Alaska and Hawaii were excluded on the grounds of severe limitations for the climate and pollution data at the county level. Additionally, one could argue that the mechanisms that may underpin the propagation of COVID-19 across states within the United States may not operate in these two states because they are not geographically contiguous. U.S. territories and associated states were also excluded from the analysis.

Confirmed cases, mortality, and vaccination data

For confirmed cases and deaths, we used USAFacts data. Unlike other similar data sources (e.g., Johns Hopkins Center for Systems Science and Engineering (CSSE) and the New York Times data), USAFacts distinguishes cases and deaths in the five New York City boroughs rather than reporting them as a single entry. Per USAFacts, daily cases are assigned to the county where the person was diagnosed with COVID-19. Deaths are counted as COVID-19–related according to the person's place of residency and if the virus played a direct role in causing death. This implies that coronavirus can be one of several causes of death.[20]

Using the daily cumulative confirmed case and death counts attributed to COVID-19, we constructed a panel data set for U.S. counties.[21] As these data series are not monotonically increasing in all U.S. counties (FIPS), we applied a backward correction using the counts to ensure the cumulative series for both cases and deaths increase monotonically over time in each county. Given sizable differences in population across counties, the number of confirmed COVID-19 cases and deaths are then expressed as population-adjusted rate measures in each county by dividing the confirmed counts by the total population of each county in 2020: the latest county resident population estimates available from the U.S. Census Bureau, Population Division[22] at the time of

[20] For more details on USAFacts' COVID-19 data collection process, refer to their methodology and sources.

[21] A very few daily death counts (0.001%) were replaced with zero deaths when cases were zero and death counts were equal to 1.

[22] Annual Resident Population Estimates for States and Counties [accessed on May 17, 2022], corresponding to POPESTIMATE2020, were used in this paper.

writing this article. Next, we applied a one-sided seven-day lagged moving average filter on the population-adjusted rate measures for both cases and deaths to reduce the variability of the cumulative count series adjusted for population by expressing each value as a deviation from the unweighted average of the seven prior days' values. Finally, we took the difference between the smoothed series and the cumulative count series adjusted for population in order to capture the changes in the cumulative series, using a 'long difference' approach.

Regarding vaccinations, while the first administration of the COVID-19 vaccine in the U.S. was officially reported on December 14, 2020,[23] the first fully vaccinated individuals in the 48 contiguous U.S. states and the District of Columbia, with vaccination rates[24] of 0.1%, occurred on December 24, 2020, according to our COVID-19 vaccination CDC data accessed on May 17, 2022. As a few daily cumulative data series are not monotonically increasing in some counties, we applied the same backward correction approach for the daily cumulative vaccination rates to ensure these series increase monotonically within each of the U.S. counties (FIPS). Missing vaccinations rates were replaced with zeros.

Socioeconomic and demographic data

Using the 2020 U.S. Census Bureau county-level resident population estimates,[25] we gathered information on sex (total male and female population), race (Black or African American alone or in combination male and female population, and White alone or in combination male and female population), and ethnicity (Hispanic male and female population) at the county level across five aggregated age groups.[26] These data are included in the cross-sectional, second-stage analysis in percentage units. The variables White, Black, and Hispanic include both males and females but do not add up to 100%, as Hispanic ethnicity may be combined with any or several of the U.S. Census Bureau racial categories (White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or other Pacific Islander).

[23] See https://www.wsj.com/articles/covid-19-vaccinations-in-the-u-s-slated-to-begin-monday-11607941806.

[24] Percent of people who have completed a primary series (have a second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction and county where vaccine recipient lives.

[25] Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin, corresponding to YEAR 13, were used in this paper [accessed on May 17, 2022].

[26] Below 20 years, between 20 and 39 years, between 40 and 59 years, between 60 and 79 years, and 80 years or more.

In addition to the demographic information, we gathered socioeconomic county-level data on median household income and education (high school completion[27] and some college[28]) from the 2022 County Health Rankings, University of Wisconsin Population Health Institute, 2020 poverty rates[29] from the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) Program using the official national poverty line, and the percent of owner-occupied housing in 2018. Median household income is from the 2020 U.S. Census Bureau SAIPE, education from the 2016-2020 ACS, State-specific sources and EDFacts, and the owner-occupied housing data from Wu et al. (2020).[30]

Other factors that we included are the Social Vulnerability Index (SVI), population density and the 2020 presidential election Democratic vote share. The county-level percentile-ranked metric of social vulnerability to natural disasters (i.e., the SVI) ranges from 0 to 1, with 1 being the highest level of social vulnerability for a county. The SVI[31] used in this study are for 2018, the latest county-level estimates available from the CDC\ATSDR. The county population density was constructed by forming the ratio of the U.S. Census Bureau 2020 county population and its 2020 land area in square kilometer from the U.S. Census Bureau Gazetteer Files., while the county-level Democratic share was derived by dividing the number of votes obtained by Joseph R. Biden Jr. (Democratic Party) in the 2020 presidential election in county i by the total number of votes recorded in county i in the 2020 presidential election. These data were

[27] Percentage of adults ages 25 and over with a high school diploma or equivalent.

[28] Percentage of adults ages 25-44 with some post-secondary education.

[29] Estimated percent of people of all ages in poverty. The SAIPE program poverty county estimates are based on the official measure of poverty as defined by the federal government, and derived based on combining direct American Community Survey (ACS) estimates and Empirical Bayes (or "shrinkage") techniques. For the 2020 SAIPE estimates, the official Census Bureau poverty threshold for a family of four containing two related children under age 18 was $26,246. Hence, a family, and all individuals from the family, are considered in poverty if their total money income (pre-tax) is less than the poverty threshold for their family size and age composition.

[30] See https://github.com/wxwx1993/PM_COVID [accessed on May 17, 2022] and the Supplementary Materials. A housing unit is owner-occupied if the owner or co-owner lives in the unit, even if it is mortgaged or not fully paid for. Mobile homes occupied by owners with installment loan balances also are included in this category. The homeownership rate is computed by dividing the number of owner-occupied housing units by the number of occupied housing units or households. For details, see https://www2.census.gov/programs-surveys/acs/tech_docs/subject_definitions/2020_ACSSubjectDefinitions.pdf.

[31] This index is based on 15 U.S. Census variables, categorized into one of four themes: socioeconomic status (poverty, unemployment, income, no high school diploma), household composition and disability (aged 65 or older, aged 17 or younger, disability, single-parent households), minority status and language (racial or ethnic minority group, speak English "less than well"), and housing type and transportation (multi-unit housing, mobile homes, crowded housing, vehicle access, group quarters) Flanagan et al. (2018).

gathered from the MIT Election Data and Science Lab (2018).[32]

Health data

The health factors are taken from a variety of data sources. Using the same 2022 County Health Rankings database, we gathered county-level data on the percent of residents under age 65 lacking health insurance, the percentage of adults who are current smokers (age-adjusted), the percentage of adults aged 20 and above with diagnosed diabetes (age-adjusted), and the percentage of population with adequate access to locations for physical activity (exercise opportunities).[33] The health uninsured measure comes from the U.S. Census Bureau's 2019 Small Area Health Insurance Estimates Program, the smoking and diabetes prevalence data are taken from the 2019 CDC's Behavioral Risk Factor Surveillance System (BRFSS), and the exercise opportunities information comes from the 2010 & 2021 ArcGIS Business Analyst, ESRI, YMCA & US Census Tigerline Files. We also added a county-level index of Severe COVID-19 Health Risk from PolicyMap (2020). This index,[34] presented in z-score form,[35] incorporates the prevalence of five health conditions which have been considered as risk factors for COVID-19 infections[36] by the CDC's BRFSS at the county level. We also included mean county-level estimates of the number of years from birth a person can expect to live (life expectancy), produced by Dwyer-Lindgren et al. (2022).[37] Finally, we added the number of Intensive Care Unit (ICU) beds[38] by county from the Kaiser

[32] See https://electionlab.mit.edu/data [accessed on August 20, 2021]. This dataset contains county-level returns for presidential elections from 2000 to 2020.

[33] *"Counties are assigned a missing value when no locations for access to exercise have been identified in either the Business Analyst or ESRI Parks datasets. In contrast, counties are assigned a 0% when they have a location for access to exercise but the county population does not live within the defined buffers of that location."* Genusso et al. (2022)

[34] The underlying information used to calculate this 2020 index comes from the CDC's 2018 BRFSS survey, the CDC's 2016 Diabetes Atlas, the 2010 U.S. Census Bureau, and the 2014-2018 American Community Survey.

[35] *"A county's z score shows how many standard deviations above or below the average a county's risk level falls. A score of 0.6, for example, would mean that the county has a higher risk than average, but is still within one standard deviation of the average and is therefore not unusually high."* PolicyMap (2020)

[36] Obesity, diabetes, high blood pressure, heart disease, and chronic obstructive pulmonary disease. Asthma is not included in the health risk index due to data inconsistency on asthma risk.

[37] Life expectancy estimates in 31 counties, where the average annual population is less than 1000 and in cases where the width of the uncertainty interval is greater than 10 years, are missing [accessed on July 1, 2022].

[38] ICU Beds reported by Medicare-certified institutional providers (hospitals) in the Fiscal Years 2018 and 2019 to the Centers for Medicare & Medicaid Services (Healthcare Cost Report Information System; Hospitals - 2010), including the categories "intensive care unit," "coronary care unit," "burn intensive care unit" and "surgical intensive care unit." It includes coronary, trauma, surgical, burn, and general ICU beds in community and non-Federal hospitals. Hospitals for veterans run by the Department of Defense are not

Family Foundation (KFF)'s Kaiser Health News Program and the status of state action on the Medicaid expansion decision as of March 2022, from KFF.

We converted the counts of ICU beds into per capita ICU beds by dividing the number of ICU beds in a county over the total population of that county in 2018.

Climate and pollution data

For the climate data, we relied upon 2000–2016 county-level average seasonal temperature and relative humidity panel data as well as measures of PM2.5 air pollution, produced by Wu et al. (2020)[39] for summer (June to September) and winter (December to February). These aggregate measures are a result of averaging maximum daily temperature and relative humidity information on 4 km × 4 km gridded rasters from gridMET via Google Earth Engine over the period 2000–2016 and across grid cells in each county. Using this data, we calculated separate average estimates for summer and winter for temperature and relative humidity. Finally, because the raw temperature data is in Kelvin degrees, we converted them into degrees Celsius by subtracting 272.15 from the calculated average temperatures.

The pollution data refers to fine inhalable particular matter in the air, in micrograms per cubic meter, with an aerodynamic diameter of 2.5 micrometers and smaller (PM2.5). The PM2.5 air pollution estimates used in this study for 2018 are publicly available for each county and produced by Wu et al. (2020).[40]

Spatial data used for the construction of bivariate maps

Using the geographic identifiers (GEOIDs) (i.e, the FIPS identifier) numeric codes for each U.S. county from the 2020 U.S. Census Bureau's TIGER geographic database (cb_2020_us_county_500k.zip) and the 2020 Cartographic Boundary File with geographic coordinates for each county, both accessed on June 15, 2022, and after removing the state FIPS identifiers for Alaska, Hawaii, Puerto Rico, and other noncontiguous territories, we adjusted the geographic U.S. Census Bureau latitude and longitude of each spatial unit (counties) to Cartesian coordinates (x,y) using the Albers equal-area conic map projection (See Picard and Stepner, 2015), which is one

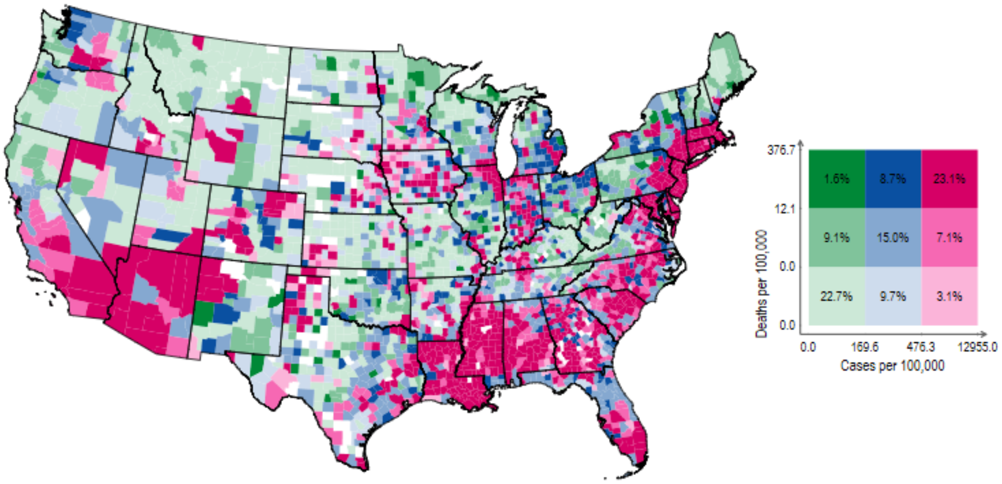
included in the data. Data accessed on June 11, 2021.

[39] See https://github.com/wxwx1993/PM_COVID [accessed on May 17, 2022].

[40] See https://github.com/wxwx1993/PM_COVID [accessed on May 17, 2022]. For more details about the construction of this measure, see the Supplementary Materials.

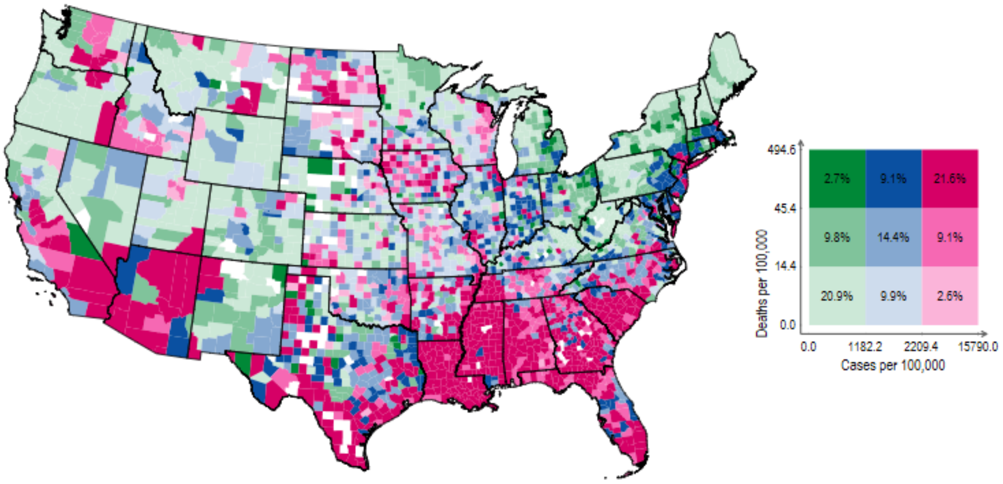
of the most commonly used projections for maps of the conterminous United States (Snyder, 1984, p.93). The same configuration applies to states, but using the 2020 U.S. Census Bureau's TIGER geographic database cb_2020_us_state_500k.zip datafile and the 2020 Cartographic Boundary File for states. To translate the geographic shape files contained in the zip files to Stata format, we used the Stata routine `spshape2dta`.

Appendix B Bivariate maps for population-adjusted COVID-19 case and death rates



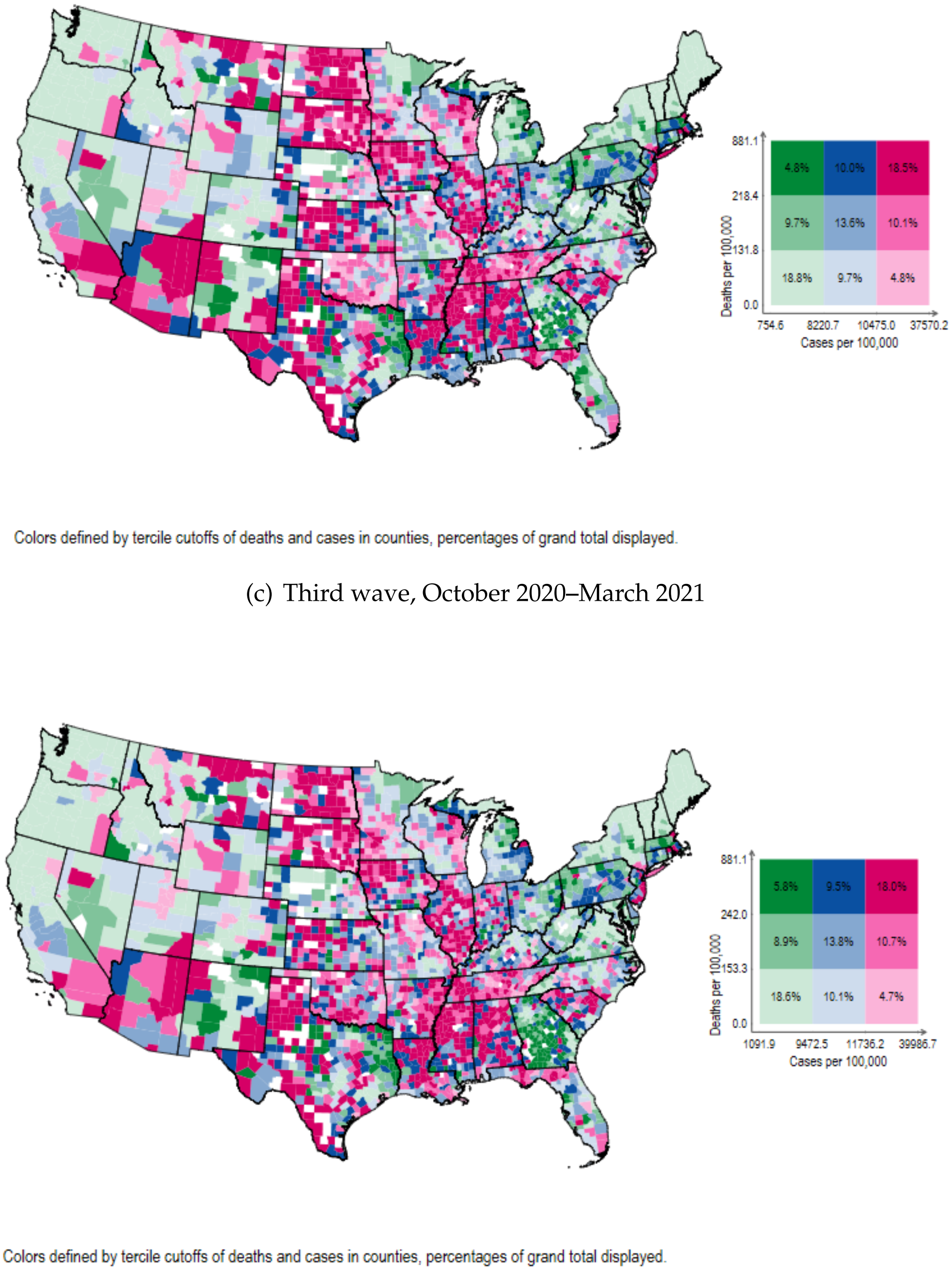
Colors defined by tercile cutoffs of deaths and cases in counties, percentages of grand total displayed.

(a) First wave, March–June 2020



Colors defined by tercile cutoffs of deaths and cases in counties, percentages of grand total displayed.

(b) Second wave, July–September 2020

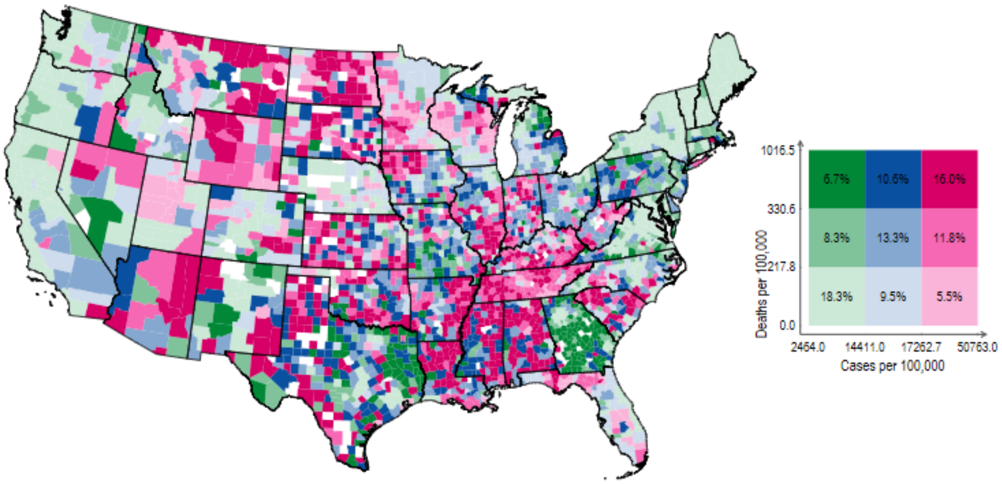


Colors defined by tercile cutoffs of deaths and cases in counties, percentages of grand total displayed.

(c) Third wave, October 2020–March 2021

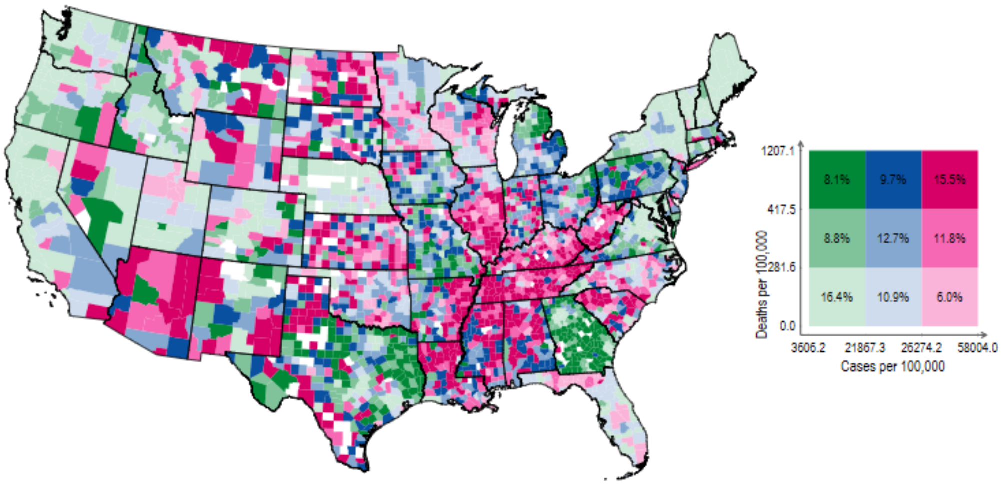
Colors defined by tercile cutoffs of deaths and cases in counties, percentages of grand total displayed.

(d) Fourth wave, April–July 2021



Colors defined by tercile cutoffs of deaths and cases in counties, percentages of grand total displayed.

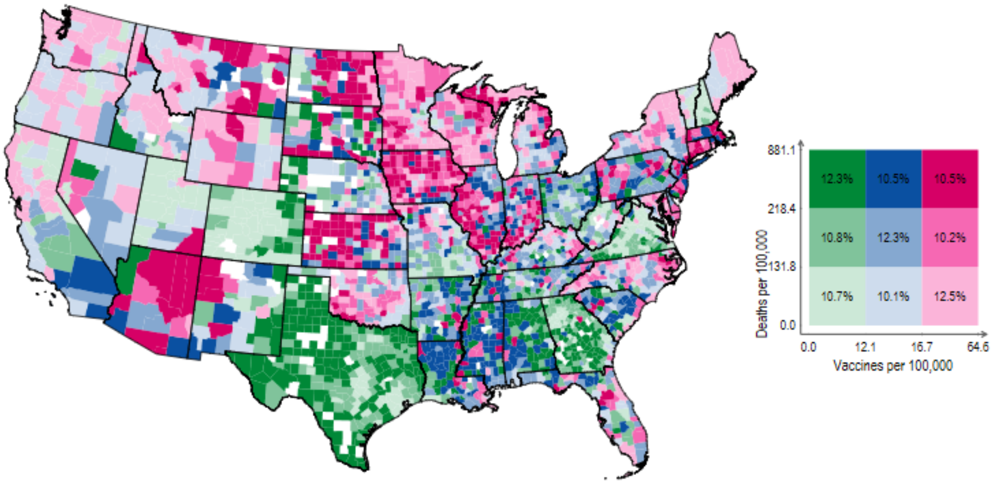
(e) Fifth wave, August–November 2021



Colors defined by tercile cutoffs of deaths and cases in counties, percentages of grand total displayed.

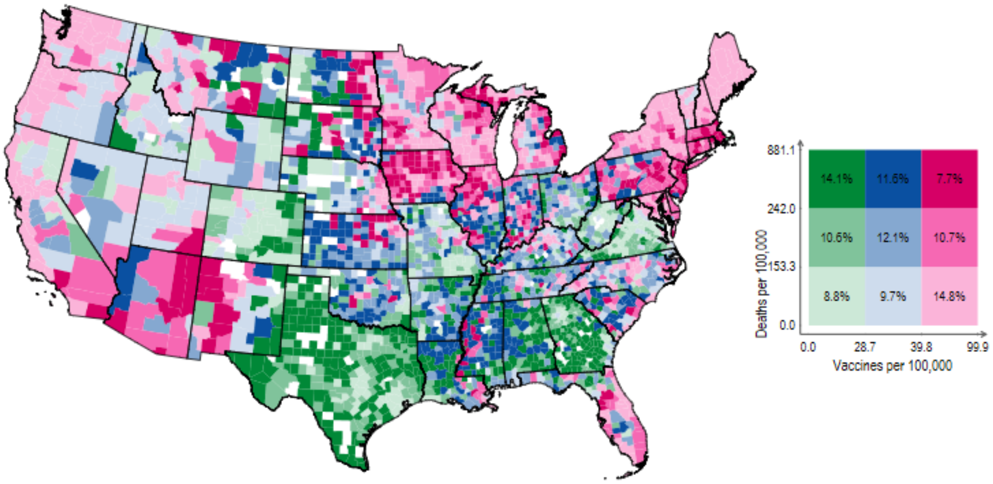
(f) Sixth wave, December 2021–March 2022

Appendix C Bivariate maps for population-adjusted COVID-19 death and vaccination rates



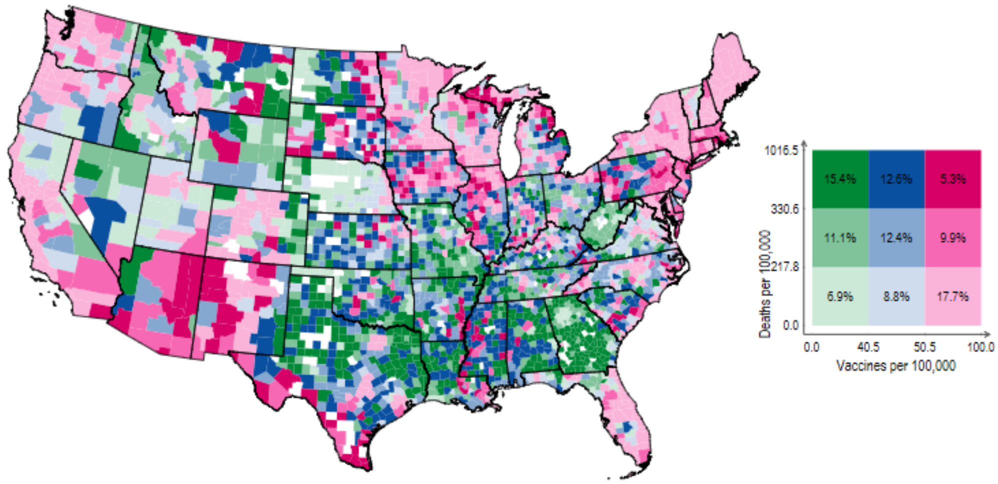
Colors defined by tercile cutoffs of deaths and vaccines in counties, percentages of grand total displayed.

(a) Third wave, September 2020–March 2021



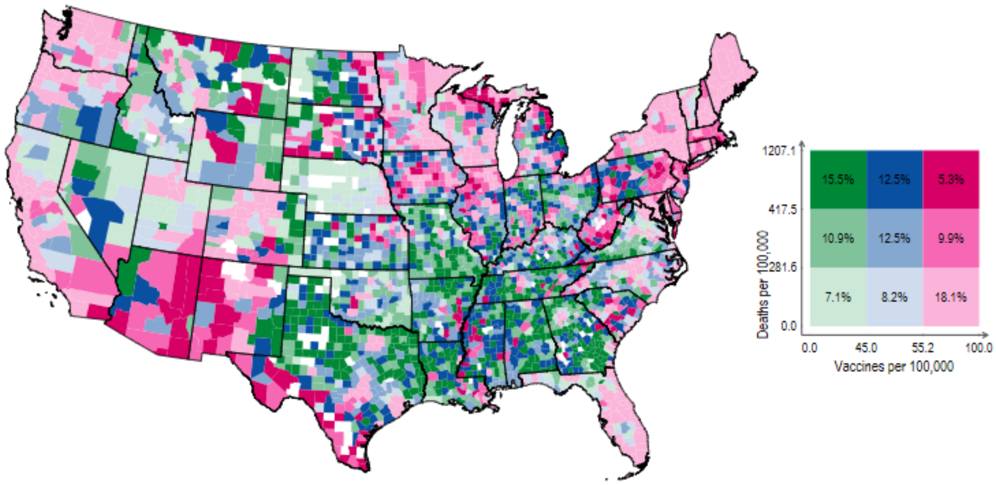
Colors defined by tercile cutoffs of deaths and vaccines in counties, percentages of grand total displayed.

(b) Fourth wave, April–July 2021



Colors defined by tercile cutoffs of deaths and vaccines in counties, percentages of grand total displayed.

(c) Fifth wave, August–November 2021



Colors defined by tercile cutoffs of deaths and vaccines in counties, percentages of grand total displayed.

(d) Sixth wave, December 2021–March 2022

Appendix D The OCMT variable selection procedure

OCMT is an alternative approach to penalized regression methods (e.g., the Least Absolute Shrinkage and Selection Operator developed by Tibshirani (1996); LASSO) and other widely used procedures (e.g., stepwise regressions)[41] for variable selection in high-dimensional linear regression models.[42] It selects variables based on multiple-testing corrected statistical significance. The objective of OCMT is to find a set of predictors that is sufficient to approximate the true data generating process underlying the variable of interest. Among the several advantages of OCMT over penalized regression methods, Chudik et al. (2018) highlight its ease of interpretation, its relation to classical statistical analysis, computational speed, and good performance in small samples. Moreover, the variable selection under OCMT is separated from the forecasting stage, while in the penalized regression methods the variable selection and estimation are performed simultaneously (Chudik et al., 2023). Unlike OCMT, LASSO "introduces a penalty term in the minimand used for estimation and calibrates the extent of penalization by cross-validation (typically 10-fold cross-validation). The use of cross-validation is supported by Monte Carlo evidence for standard models with homoskedastic and cross-sectionally independent errors, but both of these assumptions are likely to be violated in the case of the panel regressions" on U.S. county-level data (Ahmed and Pesaran, 2022). For the OCMT implementation, we used the community-contributed routine `ocmt` developed for the Stata environment, available from the Statistical Software Components (SSC) Archive (Núñez and Otero, 2020).

As the name implies, OCMT tests the statistical significance of all covariates one at a time and selects those whose t-statistics are in absolute value greater than a given critical value threshold. The critical value is computed using the critical value function $c_p(K,\delta) = \Phi^{-1}\Big(1 - \frac{\mathtt{p}}{2f(K,\delta)}\Big)$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function; $f(K,\delta) = cK^{\delta}$ for some positive constant $c = 1$ and δ, the critical value exponent; $0 < \mathtt{p} < 1$ is the nominal size of the individual test statistics; and K is the number of covariates in the regression model of interest. All of the covariates that satisfy the stated condition are selected jointly to form the initial specification of the

[41] Although stepwise regression has been extensively used among practitioners, it does not ensure consistent selection in several situations (Desboulets, 2018).

[42] Chen et al. (2021), Iregui et al. (2021), Ahmed and Pesaran (2022) and Chudik et al. (2023) constitute some applications of OCMT.

model.

In a second stage of the algorithm, OCMT uses this initial specification and once again tests the statistical significance of the covariates not selected before one at a time. The procedure continues until there are no more statistically significant covariates. Chudik et al. (2018) point out that OCMT is fast because the number of covariates bounds the number of stages required for convergence. To account for the multiple testing nature of the problem, the critical value function in the second and subsequent stages of OCMT is given by $c_p(K, \delta^\star) = \Phi^{-1}\Big(1 - \frac{\mathtt{p}}{2f(K,\delta^\star)}\Big)$, where it is required that $\delta^\star > \delta$. In their Monte Carlo simulations and empirical illustration, Chudik et al. (2018) set the value of $\delta = 1$, equivalent to applying the well-known Bonferroni adjustment to the critical value from the standard normal distribution, for a given significance level $\mathtt{p}$. We follow Chudik et al. (2018) for δ and $\delta^\star$ and set them equal to 1 and 2, respectively. It proves helpful to think of the positive constants δ and $\delta^\star$ as tuning parameters that play the role of adjusting the critical values used for inference. To assess the robustness of our findings, we also set $\delta^\star = 1.5$, which yielded qualitatively similar results to those based on $\delta^\star = 2$. For the statistical significance, we choose $\mathtt{p} = \mathtt{0.01}$, a significance level that is tighter than the recommended value of $\mathtt{p} = \mathtt{0.05}$, as we aim to be more conservative when selecting covariates. This higher level of significance ensures that only the most robust and reliable relationships between variables are considered for inclusion in the analysis.