Estimating and explaining the spread of COVID-19 at the county level in the USA

Anthony R. Ives[1 *], Claudio Bozzuto[2]

Affiliations:

[1 *] Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI 53706, USA. arives@wisc.edu. ORCHID 0000-0001-9375-9523

[2] Wildlife Analysis GmbH, Oetlisbergstrasse 38, 8053 Zurich, Switzerland. bozzuto@wildlifeanalysis.ch. ORCHID 0000-0003-0355-8379

Corresponding Author: Anthony R. Ives, Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI 53706, USA. 608-238-3771. arives@wisc.edu

Abstract

The basic reproduction number, R_0, determines the rate of spread of a communicable disease and therefore gives fundamental information needed to plan public health interventions. Estimated R_0 values are only useful, however, if they accurately predict the future potential rate of spread. Using mortality records, we estimated the rate of spread of COVID-19 among 160 counties and county-aggregates in the USA. Most of the high among-county variance in the rate of spread was explained by four factors: the timing of the county-level outbreak (partial $R^2 = 0.093$), population size (partial $R^2 = 0.34$), population density (partial $R^2 = 0.13$), and spatial location (partial $R^2 = 0.42$). Of these, the effect of timing is explained by early steps that people and governments took to reduce transmission, and population size is explained by the sample size of deaths that affects the statistical ability to estimate R_0. For predictions of future spread, population density is important, likely because it scales the average contact rate among people. To generate support for a possible explanation for the importance of spatial location, we show that SARS-CoV-2 strains containing the G614 mutation to the spike gene are associated with higher rates of spread ($P = 0.016$). The high predictability of R_0 based on population density and spatial location allowed us to extend estimates to all 3109 counties in the lower 48 States. The high variation of R_0 among counties argues for public health policies that are enacted at the county level for controlling COVID-19.

keywords: covid-19, disease spread, epidemiology, R0

Introduction

The basic reproduction number, R_0, is the number of secondary infections produced per primary infection of a disease in a susceptible population, and it is a fundamental metric in epidemiology that gauges, among other factors, the initial rate of disease spread during an epidemic [1]. While R_0 depends in part on the biological properties of the pathogen, it also depends on properties of the host population such as the contact rate between individuals [1,2]. Estimates of R_0 are required for designing public health interventions for infectious diseases such as COVID-19: for example, R_0 determines in large part the proportion of a population that must be vaccinated to control a disease [3,4]. Because R_0 at the start of an epidemic measures the spread rate under "normal" conditions without interventions, these initial R_0 values can inform policies to allow life to get "back to normal."

Using R_0 estimates to design public health policies is predicated on the assumption that the R_0 values at the start of the epidemic reflect properties of the infective agent and population, and therefore predict the potential rate of spread of the disease when interventions are implemented or in case of a resurgent outbreak. Estimates of R_0, however, might not predict future risks if (i) they are measured after public and private actions have been taken to reduce spread [5,6], (ii) they are driven by stochastic events, such as super-spreading [7,8], or (iii) they are driven by social or environmental conditions that are likely to change between the time of initial epidemic and the future time for which public health interventions are designed [9,10]. The only way to determine whether the initial R_0 estimates reflect persistent properties of the respective populations is to identify those properties: if they are unlikely to change, then so too is R_0 unlikely to change.

Policies to manage for COVID-19 in the USA are set by a mix of jurisdictions from state to local levels. We estimated R_0 at the county level both to match policymaking and to account for possibly large variation in R_0 among counties. To estimate R_0, we performed the analyses on the number of daily COVID-19 deaths [11]. We used death count rather than infection case reports, because we suspected the proportion of deaths due to COVID-19 that were reported is less likely to change compared to reported cases. Due to the mathematical structure of our estimation procedure, unreported deaths due to COVID-19 will not affect our estimates of R_0, provided the proportion of unreported deaths does not change through time. We analyzed data for counties that had at least 100 reported cumulative deaths, and for other counties we aggregated data within the same state including deaths whose county was unknown. This led to 160 final time series representing counties in 39 states and the District of Columbia, of which 36 were aggregated at the state level. Some states, even after aggregating data from all counties, did not reach the 100 threshold of cumulative deaths, and therefore the spread rate for these states was not estimated.

Results

Estimates of the spread rate

Before estimating R_0, we first estimated the rate of spread of the COVID-19 as the rate of increase of the daily death counts, r_0. Although this approach is not typically used in epidemiological studies, it has the advantage of being statistically robust even when the data (death counts) are low and makes the minimum number of assumptions that could affect the estimates in unexpected ways (see SI: Overview of Statistical Methods). We applied a time-varying autoregressive state-space model to each time series of death counts [12]. In contrast to

other models of COVID-19 epidemics [13,14], we do not incorporate the transmission process and the daily time course of transmission, but instead we estimate the time-varying exponential change in the number of deaths per day, $r(t)$. Detailed simulation analyses (SI: Simulation model) showed that estimates of $r(t)$ generally lagged behind the true values. Therefore, we analyzed the time series in forward and reverse directions, and averaged to get the estimates of r_0 at the start of the time series (Fig. S1); this approach counterbalances the lag in the forward direction with the lag in the backwards direction, therefore reducing the lag effect. The model was fit accounting for greater uncertainty when mortality counts were low, and confidence intervals of the estimates were obtained from parametric bootstrapping, which is the most robust approach when counts are low. Thus, our strategy was to use a parsimonious model to give robust estimates of r_0 even for counties that had experienced relatively few deaths, and then calculate R_0 from r_0 after the fitting process using well-established methods [15].

Our r_0 estimates ranged from close to zero for several counties to 0.33 for New York City (five boroughs); the latter implies that the number of deaths increases by a factor of $e^{0.33} = 1.39$ per day. There were highly statistically significant differences between upper and lower estimates (Fig. 1). Although our time series approach allowed us to estimate r_0 at the start of even small epidemics, we anticipated two factors that could potentially affect our estimates of r_0 that are not likely to be useful in explaining future spread rates. The first factor is the timing of the onset of county-level epidemic: 35% of the local outbreaks started after the declaration of COVID-19 as a pandemic by the WHO on 11 March, 2020 [16], and thus we anticipated estimates of r_0 to decrease with the Julian date of outbreak onset. Change in human behaviors caused by public awareness about COVID-19 at the outbreak onset will not necessarily predict future rates of spread. We used the second factor, the size of the population encompassed by the time series,

to factor out statistical bias from the time series analyses. Simulation studies showed that estimates for time series with low death counts were downward biased (Fig. S2). Because for a given spread rate $r(t)$ the total number of deaths in a time series should be proportional to the population size, we used population size as a covariate to remove bias. In addition to these two factors that we do not think have strong predictive value for the future rate of spread, we also anticipated effects of population density and spatial autocorrelation. Therefore, we regressed r_0 against outbreak onset, population size and population density, and included spatially autocorrelated error terms.

Explaining variation in r_0

The regression analysis showed highly significant effects of all four factors (Table 1), and each factor had a substantial partial R^2_{pred} [17]. The overall R^2_{pred} was 0.69, so most of the county-to-county variance was explained. We calculated corrected r_0 values, factoring out outbreak onset and population size, by standardizing the r_0 values by 11 March, 2020 and the most populous county (for which the estimates of r_0 are likely best). Counties with low to medium population density never had high corrected r_0 values, suggesting that population density sets an upper limit on the rate of spread of COVID-19 (Fig. 2A), in agreement with expectations and published results [1,18]. Nonetheless, despite the unequivocal statistical effect of population density ($P < 10^{-8}$, Table 1), the explanatory power was not great (partial $R^2_{pred} = 0.13$), probably because population density at the scale of counties will be only roughly related to contact rates among people.

Spatial autocorrelation, in turn, had strong power in explaining variation in r_0 among counties (partial $R^2_{pred} = 0.42$, Table 1) and occurred at the scale of hundreds of kilometers (Fig.

126 2B). This spatial autocorrelation might reflect differences in public responses to COVID-19
127 across the USA not captured by the variable in the regression model for outbreak onset. For
128 example, Seattle, WA, reported the first positive case in the USA, on 15 January, 2020, and there
129 was a public response before deaths were recorded [19]. In contrast, the response in New York City
130 was delayed, even though the outbreak occurred later than in Seattle [20]. Spatial autocorrelation
131 could also be caused by movement of infected individuals. However, movement would only lead
132 to autocorrelation in our regression analysis if many of the reported deaths were of people
133 infected outside the county. A further possibility is that spatial variation in the rate of spread of
134 COVID-19 reflects spatial variation in the occurrence of different genetic strains of SARS-CoV-
135 2.

136 To investigate whether spatial autocorrelation could potentially be caused by different
137 strains of SARS-CoV-2 differing in infectivity, we analyzed publicly available information about
138 genomic sequences from the GISAID metadata [21]. Scientific debate has focused on the role of
139 the G614 mutation in the spike protein gene (D614G) to increase the rate of transmission of
140 SARS-CoV-2 [22]. We therefore asked whether the proportion of strains containing the G614
141 mutation was associated with higher rates of COVID-19 spread. Because the genomic samples
142 are only located to the state level, we performed the analysis accordingly, for each state selecting
143 the r_0 from the county or county-aggregate with the highest number of deaths (and hence being
144 most likely represented in the genomic samples). We further restricted genomic samples to those
145 collected within 30 days following the outbreak onset we used to select the data for time-series
146 analyses, and we required at least 5 genomic samples per state. This data handling resulted in 28
147 states available for analysis. We again used our regression model, except that now we included
148 the proportion of strains having the G614 mutation instead of spatial location (Eq. 7). The

proportion of samples containing the G614 mutation had a positive effect on r_0 ($P = 0.016$, Table S2). The low proportion of strains containing the G614 mutation in the Pacific Northwest and the Southeast were associated with lower values of r_0 (Fig. 3). Before analyzing the full GISAID data, we analyzed a subset from Nextstrain [23] naïvely, without engaging the specific hypothesis that the G614 mutation increased transmission. This naïve analysis picked up the same pattern ($P = 0.019$, SI: Analysis of SARS-CoV-2 strains).

Higher transmissibility of strains containing the G614 mutation is also suggested by its increasing prevalence in strains in the USA [24]. Nonetheless, our analyses give no information about the mechanisms explaining differences in spread rates among strains. A consensus on the potential impact of SARS-CoV-2 mutations is still lacking [22]: some studies present evidence for a differential pathogenicity and transmissibility [25,26], while others conclude that mutations might be mostly neutral or even reduce transmissibility [27]. Our analyses call for further investigation to better understand the potential link between viral genomic variation and its impact on transmission and mortality [28].

To check whether there are other factors that might explain variation in our estimates of r_0 among counties, we investigated additional population characteristics [29,30] that might be expected to affect the initial spread rate of COVID-19: (i) median age, (ii) adult obesity, (iii) diabetes, (iv) education, (v) income, (vi) poverty, (vii) economic equality, (viii) race, and (ix) political leaning (Table S4). The first three characteristics likely affect morbidity [31], although it is not clear whether higher morbidity will increase or decrease the spread rate. The remaining characteristics might affect health outcomes and responses to public health interventions; for example, education, income and poverty might all affect the need for individuals to work in jobs that expose them to greater risks of infection. Nonetheless, because we focused on the early

spread of COVID-19, we anticipated that these characteristics would have minimal effects. Despite the potential for all nine characteristics to affect estimates of r_0, we found that none was a statistically significant predictor of r_0 when taking the four main factors into account (all $P > 0.1$). We also repeated all of the analyses on estimates of $r(t)$ after COVID-19 was broadly established in the USA (5 May, 2020, assuming an average time between infection and death of 18 days) (Table S5). The corresponding $R^2_{pred} = 0.38$, largely driven by a large positive effect of the date of outbreak onset. The absence of significant effects of the additional population characteristics on r_0, and the lower explanatory power of the model on $r(t)$ at the end of the time series, underscore the importance of population density and spatial autocorrelation in predicting county-level values of r_0.

Extrapolating R_0 to all counties

In the regression model (Table 1), the standard deviation of the residuals was 1.11 times higher than the average standard error of the estimates of r_0. This implies that the uncertainty of an estimate of r_0 from the regression is only slightly higher than the uncertainty in the estimate of r_0 from the time series itself; the regression model explains 81% ($= 1/1.11^2$) of the explainable variance. Therefore, using estimates from death count time series from other counties will give estimates of r_0 for a focal county (lacking reliable estimates) that are almost as precise as the estimate from the county's time series. In turn, this implies that the regression can also be used to extrapolate estimates of r_0 to counties for which deaths were too sparse for time-series analysis. We used the regression to extrapolate values of R_0, derived from r_0, for all 3109 counties in the conterminous USA (Fig. 4, Table S1). The high predictability of r_0, and hence R_0, from the regression is seen in the comparison between R_0 calculated from the raw estimates of r_0 (Fig.

195 4A) and R_0 calculated from the corrected r_0 values (Fig. 4B). Extrapolation from the regression
196 model makes it possible not only to get refined estimates for the counties that were aggregated in
197 the time-series analyses; it also gives estimates for counties within states with so few deaths that
198 county-aggregates could not be analyzed (Fig. 4C,D). The end product is a map of R_0 for the
199 conterminous USA (Fig. 4E).
200
201 **Discussion**
202 It is widely understood that different states and counties in the USA, and different
203 countries in the world, have experienced COVID-19 epidemics differently. Our analyses have
204 put numbers on these differences in the USA. The large differences argue for public health
205 interventions to be designed at the county level. For example, the vaccination coverage in the
206 most densely populated area, New York City, needed to prevent future outbreaks of COVID-19
207 will be much greater than for sparsely populated counties. Therefore, once vaccines are
208 developed, they should be distributed first to counties with high R_0. Similarly, if vaccines are not
209 developed quickly and non-pharmaceutical public health interventions have to be re-instated
210 during resurgent outbreaks, then counties with higher R_0 values will require stronger
211 interventions. As a final example, county-level R_0 values can be used to assess the practicality of
212 contact-tracing of infections, which become impractical when R_0 is high [32].
213 We present our county-level estimates of R_0 as preliminary guides for policy planning,
214 while recognizing the myriad other epidemiological factors (such as mobility [33-35]) and political
215 factors (such as legal jurisdictions [36]) that must shape public health decisions [3,37-39]. Although we
216 have emphasized the predictability of R_0 among counties in the USA, values of R_0 could change
217 if there are changes in the transmissibility of strains that are present; our analyses suggest strain

differences (Fig. 3), and changes towards strains with higher transmissibility could lead to higher R_0 values. This argues for continued efforts to identify the transmissibility of different strains and where those strains are prevalent.

We recognize the importance of following the day-to-day changes in death and case rates, and short-term projections used to anticipate hospital needs and modify public policies [40-42]. Looking back to the initial spread rates, however, gives a window into the future and what public health policies will be needed when COVID-19 is endemic.

Materials and Methods

1. Data selection and handling

1.1 Death data

For mortality due to COVID-19, we used time series provided by the New York Times [11]. We selected the New York Times dataset because it is rigorously curated. We analyzed separately only counties that had records of 100 or more deaths. The District of Columbia was treated as a county. Also, because the New York Times dataset aggregated the five boroughs of New York City, we treated them as a single county. For counties with fewer than 100 deaths, we aggregated mortality to the state level to create a single time series. For thirteen States (AK, DE, HI, ID, ME, MT, ND, NH, SD, UT, VM, WV, and WY), the aggregated time series did not contain 100 or more deaths and were therefore not analyzed.

1.2 County-level variables

We obtained county-level population size and area (km^2) from the US Census Bureau *(21)*. Other socio-economic variables (Table S4) we obtained from Kirkegaard [29]. We selected socio-economic variables *a priori* in part to represent a broad set of population characteristics.

2. Time series analysis

2.1 Time series model

We used a time-varying autoregressive model [12,43,44] designed explicitly to estimate the rate of increase of a variable using non-Gaussian error terms. We assume in our analyses that the proportion of the population represented by a time series that is susceptible is close to one, and therefore there is no decrease in the infection rate caused by a pool of individuals who were infected, recovered, and were then immune to further infection.

The model is

$$x(t) = r(t-1) + x(t-1) \quad (1a)$$

$$r(t) = r(t-1) + \omega_r(t) \quad (1b)$$

$$x^*(t) = x(t) + \phi(t) \quad (1c)$$

Here, $x(t)$ is the unobserved, log-transformed value of daily deaths at time t, and $x^*(t)$ is the observed count that depends on the observation uncertainty described by the random variable $\phi(t)$. Because a few of the datasets that we analyzed had zeros, we replaced zeros with 0.5 before log-transformation. The model assumes that the death count increases exponentially at rate $r(t)$, where the latent state variable $r(t)$ changes through time as a random walk with $\omega_r(t) \sim N(0, \sigma^2_r)$. We assume that the count data follow a quasi-Poisson distribution. Thus, the expectation of

counts at time t is $\exp(x(t))$, and the variance is proportional to this expectation.

We fit the model using the Kalman filter to compute the maximum likelihood [45,46]. In addition to the parameters σ^2_r, and σ^2_ϕ, we estimated the initial value of $r(t)$ at the start of the time series, r_0, and the initial value of $x(t)$, x_0. The estimation also requires an assumption for the variance in x_0 and r_0, which we assumed were zero and σ^2_r, respectively. In the validation using simulated data, we found that the estimation process tended to absorb σ^2_r to zero too often. To eliminate this absorption to zero, we imposed a minimum of 0.02 on σ^2_r, which eliminated the problem in the simulations.

2.2 Parametric bootstrapping

To generate approximate confidence intervals for the time-varying estimates of $r(t)$, we used a parametric bootstrap designed to simulate datasets with the same characteristics as the real data that are then refit using the autoregressive model. We used bootstrapping to obtain confidence intervals, because an initial simulation study showed that standard methods, such as obtaining the variance of $r(t)$ from the Kalman filter, were too conservative (the confidence intervals too narrow) when the number of counts was small. Furthermore, parametric bootstrapping can reveal bias and other features of a model, such as the lags we found during model fitting (Fig. S1A,B).

Changes in $r(t)$ consist of unbiased day-to-day variation and the biased deviations that lead to longer-term changes in $r(t)$. The bootstrap treats the day-to-day variation as a random variable while preserving the biased deviations that generate longer-term changes in $r(t)$. Specifically, the bootstrap was performed by calculating the differences between successive

284 estimates of $r(t)$, $\Delta r(t) = r(t) - r(t\text{-}1)$, and then standardizing to remove the bias, $\Delta r_s(t) = \Delta r(t) -$
285 $E[\Delta r(t)]$. The sequence $\Delta r_s(t)$ was fit using an autoregressive time-series model with time lag 1,
286 AR(1), to preserve any shorter-term autocorrelation in the data. For the bootstrap a new time
287 series was simulated from this AR(1) model, $\Delta\rho(t)$, and then standardized, $\Delta\rho_s(t) = \Delta\rho(t) -$
288 $E[\Delta\rho(t)]$. The simulated time series for the spread rate was constructed as $\rho(t) = r(t) + \Delta\rho_s(t)/$
289 $2^{1/2}$, where dividing by $2^{1/2}$ accounts for the fact that $\Delta\rho_s(t)$ was calculated from the difference
290 between successive values of $r(t)$. A new time series of count data, $\xi(t)$, was then generated using
291 equation (S1a) with the parameters from fitting the data. Finally, the statistical model was fit to
292 the reconstructed $\xi(t)$. In this refitting, we fixed the variance in $r(t)$, σ^2_r, to the same value as
293 estimated from the data. Therefore, the bootstrap confidence intervals are conditional of the
294 estimate of σ^2_r.

295

296 *2.3. Calculating R_0*

297 We derived estimates of $R(t)$ directly from $r(t)$ using the Dublin-Lotka equation [15] from
298 demography. This equation is derived from a convolution of the distribution of births under the
299 assumption of exponential population growth. In our case, the "birth" of COVID-19 is the
300 secondary infection of susceptible hosts leading to death, and the assumption of exponential
301 population growth is equivalent to assuming that the initial rate of spread of the disease is
302 exponential, as is the case in equation 1. Thus,

303

304 $$R(t) = 1/\sum_\tau e^{-r(t)\tau} p(\tau) \qquad (2)$$

305

where $p(\tau)$ is the distribution of the proportion of secondary infections caused by a primary infection that occurred τ days previously. We used the distribution of $p(\tau)$ from Li et al. [47] that had an average serial interval of T0 = 7.5 days; smaller or larger values of T0, and greater or lesser variance in $p(\tau)$, will decrease or increase $R(t)$ but will not change the pattern in $R(t)$ through time. Note that the uncertainty in the distribution of serial times for COVID-19 is a major reason why we focus on estimating r_0, rather than R_0: the estimates of r_0 are not contingent on time distributions that are poorly known. Computing $R(t)$ from $r(t)$ does not depend on the mean or variance in time between secondary infection and death. We report values of $R(t)$ at dates that are offset by 18 days, the average length of time between initial infection and death given by Zhou et al. [48].

2.4. Initial date of the time series

Many time series consisted of initial periods containing zeros that were uninformative. As the initial date for the time series, we chose the day on which the estimated daily death count exceeded 1. To estimate the daily death count, we fit a Generalized Additive Mixed Model (GAMM) to the death data while accounting for autocorrelation and greater measurement error at low counts using the R package mgcv [49]. We used this procedure, rather than using a threshold of the raw death count, because the raw death count will include variability due to sampling small numbers of deaths. Applying the GAMM to "smooth" over the variation in count data gives a well-justified method for standardizing the initial dates for each time series.

2.5. Validation

We performed extensive simulations to validate the time-series analysis approach (SI

Appendix).

3. Regression analysis for r_0

We applied a Generalized Least Squares (GLS) regression model to explain the variation in estimates of r_0 from the 160 county and county-aggregate time series:

$$r_0 = b_0 + b_1\, start.date + b_2 \log(pop.size) + b_3\, pop.den^{0.25} + \varepsilon \quad (3)$$

$$\varepsilon = N(0, \sigma^2\mathbf{\Sigma})$$

where *start.date* is the Julian date of the start of the time series, log(*pop.size*) and $pop.den^{0.25}$ are the log-transformed population size and 0.25 power-transformed population density of the county or county-aggregate, respectively, and ε is a Gaussian random variable with covariance matrix $\sigma^2\mathbf{\Sigma}$. The transforms (log(*pop.size*) and $pop.den^{0.25}$) were used to account for nonlinear relationships with r_0 and were selected to give the highest maximum likelihood of the overall regression. The covariance matrix contains a spatial correlation matrix of the form $\mathbf{C} = u\mathbf{I} + (1-u)\mathbf{S}(g)$ where u is the nugget and $\mathbf{S}(g)$ contains elements $\exp(-d_{ij}/g)$, where d_{ij} is the distance between spatial locations and g is the range [50]. To incorporate differences in the precision of the estimates of r_0 among time series, we weighted by the vector of their standard errors, **s**, so that $\mathbf{\Sigma}$ = diag(**s**) * **C** * diag(**s**), where * denotes matrix multiplication. With this weighting, the overall scaling term for the variance, σ^2, will equal 1 if the residual variance of the regression model matches the square of the standard errors of the estimates of r_0 from the time series. We fit the regression model with the function gls() in the R package nlme [51].

To make predictions for new values of r_0, we used the well-known relationship

$$\hat{e}_i = \bar{e} + \boldsymbol{v_i} * \mathbf{V}^{-1}(\varepsilon_t - \bar{e}) \quad (4)$$

where ε_t is the GLS residual for data i, $\hat{e}_i$ is the predicted residual, $\bar{e}$ is the mean of the GLS residuals, $\mathbf{V}$ is the covariance matrix for data other than i, and $\boldsymbol{v_i}$ is a row vector row containing the covariances between data i and the other data in the dataset [52]. This equation was used for three purposes. First, we used it to compute R^2_{pred} for the regression model by removing each data point, recomputing $\hat{e}_i$, and using these values to compute the predicted residual variance following [17]. Second, we used it to obtain predicted values of r_0, and subsequently R_0, for the 160 counties and county-aggregates for which r_0 was also from time series. Third, we used equation (4) similarly to obtain predicted values of r_0, and hence predicted R_0, for all other counties. We also calculated the variance of the estimates from [52]

$$\hat{v}_i = \sigma^2 - \boldsymbol{v_i} * \mathbf{V}^{-1} * \boldsymbol{v_i}^{\mathbf{t}} \quad (5)$$

Predicted values of R_0 were mapped using the R package usmap [53].

4. Regression analysis for SARS-CoV-2 effects on r_0

The GISAID metadata [21] for SARS-CoV-2 contains the clade and state-level location for strains in the USA; strains G, GH, and GR contain the G614 mutation. For each state, we limited the SARS-CoV-2 genomes to those collected no more than 30 days following the onset of outbreak that we used as the starting point for the time series from which we estimated r_0; from these genomes (totaling 5290 from all states), we calculated the proportion that had the G614

mutation. Only twenty-eight states had five or more genomes, so we limited the analyses to these states. For each state, we selected the estimates of r_0 from the county or county-aggregate representing the greatest number of deaths. We fit these estimates of r_0 with the weighted Least Squares (LS) model as in equation (3) with additional variables for strain.

Figure 3 was constructed using the R packages usmap [53] and scatterpie [54].

Acknowledgements: We thank Steve R. Carpenter, Volker C. Radeloff, and Monica M. Turner for comments on the manuscript. **Funding:** This work was supported by NASA-AIST-80NSSC20K0282 (A.R.I). **Author contributions**: A.R.I and C.B. designed the study, and A.R.I. led the analyses and writing of the manuscript. **Competing interests**: The authors declare no competing interests. **Data and materials availability:** Data and R code for the analyses are presented in the Supplementary Materials.

References

1 Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T. & Jacobsen, K. H. Complexity of the basic reproduction number (R0). *Emerging Infectious Diseases* **25**, 1-4 (2019).

2 Hilton, J. & Keeling, M. Estimation of country-level basic reproductive ratios for novel Coronavirus (COVID-19) using synthetic contact matrices. *Preprint* (2020).

3 Fine, P., Eames, K. & Heymann, D. L. "Herd immunity": a rough guide. *Clinical Infectious Diseases* **52**, 911-916, doi:10.1093/cid/cir007 (2011).

4 Anderson, R. M. The concept of herd immunity and the design of community-based immunization programmes. *Vaccine* **10**, 928-935, doi:10.1016/0264-410X(92)90327-G (1992).

398 5 Flaxman, S. & et al. Estimating the number of infections and the impact of non-
399 pharmaceutical interventions on COVID-19 in 11 European countries. *Report 13,*
400 *Imperial College London* (2020).
401 6 Scire, J. *et al.* Reproductive number of the COVID-19 epidemic in Switzerland with a
402 focus on the Cantons of Basel-Stadt and Basel-Landschaft. *Swiss Medical Weekly* **150**
403 (2020).
404 7 Adam, D. & et al. Clustering and superspreading potential of severe acute respiratory
405 syndrome coronavirus 2 (SARS-CoV-2) infections in Hong Kong. (2020).
406 8 Paull, S. H. *et al.* From superspreaders to disease hotspots: linking transmission across
407 hosts and space. *Frontiers in Ecology and the Environment* **10**, 75-82,
408 doi:10.1890/110111 (2012).
409 9 Lofgren, E., Fefferman, N. H., Naumov, Y. N., Gorski, J. & Naumova, E. N. Influenza
410 seasonality: underlying causes and modeling theories. *Journal of Virology* **81**, 5429-
411 5436, doi:10.1128/JVI.01680-06 (2007).
412 10 Peña-García, V. H. & Christofferson, R. C. Correlation of the basic reproduction number
413 (R0) and eco-environmental variables in Colombian municipalities with chikungunya
414 outbreaks during 2014-2016. *PLoS Neglected Tropical Diseases* **13**, e0007878 (2019).
415 11 New York Times. Coronavirus (Covid-19) data in the United States. (2020).
416 <https://github.com/nytimes/covid-19-data>.
417 12 Ives, A. R. & Dakos, V. Detecting dynamical changes in nonlinear time series using
418 locally linear state-space models. *Ecosphere* **3**, art58, doi:http://dx.doi.org/10.1890/ES11-
419 00347.1 (2012).

13 Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology* **178**, 1505-1512, doi:10.1093/aje/kwt133 (2013).

14 Flaxman, S. & al., e. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. *Report 13, Imperial College London* (2020).

15 Dublin, L. I. & Lotka, A. J. On the true rate of natural increase. *Journal of the American Statistical Association* **20**, 305–339 (1925).

16 Cucinotta, D. & Vanelli, M. WHO declares COVID-19 a pandemic. *Acta Bio Medica Atenei Parmensis* **91**, 157-160, doi:10.23750/abm.v91i1.9397 (2020).

17 Ives, A. R. R2s for correlated data: phylogenetic models, LMMs, and GLMMs. *Systematic Biology* **68**, 234-251, doi:10.1093/sysbio/syy060 (2019).

18 Rader, B. *et al.* Crowding and the epidemic intensity of COVID-19 transmission. *medRxiv*, 2020.2004.2015.20064980, doi:10.1101/2020.04.15.20064980 (2020).

19 Fink, S. in *The New York Times* 1 (New York, NY, 2020).

20 Anon. in *The Economist* Vol. 435 4 (The Economist Newspaper Limited, London, UK, 2020).

21 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1:33-46**, doi:10.1002/gch2.1018 (2017).

22 Grubaugh, N. D., Hanage, W. P. & Rasmussen, A. L. Making Sense of Mutation: What D614G Means for the COVID-19 Pandemic Remains Unclear. *Cell*, doi:https://doi.org/10.1016/j.cell.2020.06.040 (2020).

23 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123, doi:10.1093/bioinformatics/bty407 (2018).

24 NextstrainTeam. Nextstrain. (2020). <https://nextstrain.org/ncov>.

25 Korber, B. *et al.* Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*, 2020.2004.2029.069054, doi:10.1101/2020.04.29.069054 (2020).

26 Yao, H. *et al.* Patient-derived mutations impact pathogenicity of SARS-CoV-2. *medRxiv*, 2020.2004.2014.20060160, doi:10.1101/2020.04.14.20060160 (2020).

27 Dorp, L. v. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *bioRxiv*, 2020.2005.2021.108506, doi:10.1101/2020.05.21.108506 (2020).

28 Eaaswarkhanth, M., Al Madhoun, A. & Al-Mulla, F. Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? *International Journal of Infectious Diseases* **96**, 459-460, doi:10.1016/j.ijid.2020.05.071 (2020).

29 Kirkegaard, E. O. W. Inequality across US counties: an S factor analysis. *Open Quantitative Sociology and Political Science* (2016).

30 United States Census Bureau. USA Counties. (2011). <https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html>.

31 Centers for Disease Control and Prevention. Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019 — United States, February 12–March 28, 2020. *MMWR. Morbidity and Mortality Weekly Report* **69** (2020). <www.cdc.gov>.

32 Fraser, C., Riley, S., Anderson, R. M. & Ferguson, N. M. Factors that make an infectious disease outbreak controllable. *Proceedings for the National Academy of Sciences* **101**, 6146–6151 (2004).

33 Bichara, D., Kang, Y., Castillo-Chavez, C., Horan, R. & Perrings, C. SIS and SIR Epidemic Models Under Virtual Dispersal. *Bulletin of Mathematical Biology* **77**, 2004-2034, doi:10.1007/s11538-015-0113-5 (2015).

34 Roberts, M. G. & Heesterbeek, J. a. P. A new method for estimating the effort required to control an infectious disease. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 1359-1364, doi:10.1098/rspb.2003.2339 (2003).

35 Gatto, M. *et al.* Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences* **117**, 10484-10491, doi:10.1073/pnas.2004978117 (2020).

36 Gorman, S. & Bernstein, S. Wisconsin Supreme Court invalidates state's COVID-19 stay-at-home order. *Reuters* (2020). <https://www.reuters.com/article/us-health-coronavirus-usa-wisconsin/wisconsin-supreme-court-invalidates-states-covid-19-stay-at-home-order-idUSKBN22Q04H>.

37 Lahariya, C. Vaccine epidemiology: A review. *Journal of Family Medicine and Primary Care* **5**, 7-15, doi:10.4103/2249-4863.184616 (2016).

38 Mallory, M. L., Lindesmith, L. C. & Baric, R. S. Vaccination-induced herd immunity: Successes and challenges. *Journal of Allergy and Clinical Immunology* **142**, 64-66, doi:10.1016/j.jaci.2018.05.007 (2018).

485 39 Ridenhour, B., Kowalik, J. M. & Shay, D. K. Unraveling R0: Considerations for public
486 health applications. *American Journal of Public Health* **104**, e32-e41,
487 doi:10.2105/AJPH.2013.301704 (2013).
488 40 Imperial College London. Covid-19 Scenario Analysis Tool. (2020).
489 <https://covidsim.org>.
490 41 Systrom, K. & Vladeck, T. Rt Covid-19. (2020). <https://rt.live>.
491 42 Swiss National Covid-19 Science Task Force. Situation report. (2020). <https://ncs-
492 tf.ch/en/situation-report>.
493 43 Zeng, Z., Nowierski, R. M., Taper, M. L., Dennis, B. & Kemp, W. P. Complex
494 population dynamics in the real world: Modeling the influence of time-varying
495 parameters and time lags. *Ecology* **79**, 2193-2209 (1998).
496 44 Bozzuto, C. & Ives, A. R. (2020).
497 45 Durbin, J. & Koopman, S. J. *Time Series Analysis by State Space Methods*. 2nd edn,
498 (Oxford University Press, 2012).
499 46 Harvey, A. C. *Forecasting, structural time series models and the Kalman filter*.
500 (Cambridge University Press, 1989).
501 47 Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–
502 Infected Pneumonia. *New England Journal of Medicine* **382**, 1199-1207,
503 doi:10.1056/NEJMoa2001316 (2020).
504 48 Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with
505 COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* **395**, 1054-1062,
506 doi:10.1016/S0140-6736(20)30566-3 (2020).

49 Wood, S. N. *Generalized additive models: an introduction with R*. (CRC Press, Chapman and Hall, 2017).

50 Cressie, N. A. C. *Statistics for spatial data*. (John Wiley & Sons, 1991).

51 Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & Team, R. C. nlme: Linear and lonlinear mixed effects models. R package version 3.1-147. (2020). <https://CRAN.R-project.org/package=nlme>.>.

52 Petersen, K. B. & Pedersen, M. S. (Technical University of Denmark, 2012).

53 Di Lorenzo, P. usmap: US Maps Including Alaska and Hawaii. R package version 0.5.0.9999. (2020). < https://usmap.dev>.

54 Yu, G. scatterpie, R package version 0.1.4. (2019). <https://CRAN.R-project.org/package=scatterpie>.

55 Flaxman, S. & et al. State-level tracking of COVID-19 in the United States. *Report 23, Imperial College London* (2020).

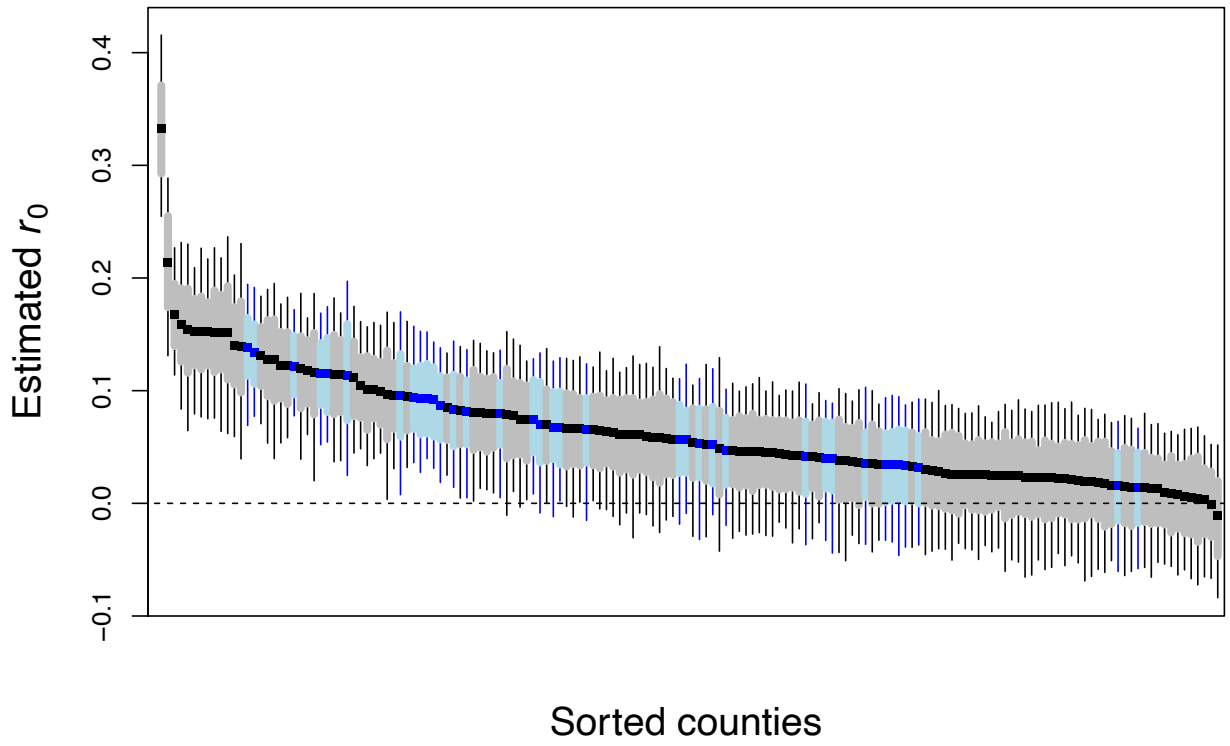
56 Gelman, A. & Hill, J. *Data analysis using regression and multilevel/hierarchical models*. (Cambridge University Press, 2007).

57 Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap*. (Chapman and Hall, 1993).

58 Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936, doi:10.1126/science.abb6936 (2020).

Table 1. For 160 county and county-aggregates, results of the regression of the estimates of the initial spread rate, r_0, against (i) the date of outbreak onset, (ii) total population size and (iii) population density, in which (iv) spatial autocorrelation is incorporated into the residual error. Transforms of population size and density were selected to best-fit the data and satisfy linearity assumptions. The coefficient column contains the estimate of the regression parameters with their associated t-tests; spatial autocorrelation is characterized by a range and nugget for regional and local sources of variation, and their joint significance is given by a likelihood ratio test. For the overall model, $R^2_{pred} = 0.69$, and the residual standard error is 1.11.

	Coefficient	**SE**	**t**	**P**	**partial R^2_{pred}**
onset	-0.0018	0.0004	-4.28	10^{-4}	0.093
log(size)	0.0242	0.0028	8.59	$< 10^{-8}$	0.34
density$^{1/4}$	0.010	0.0017	5.68	$< 10^{-8}$	0.13
space	range = 3.88 nugget = 0.39		$\chi^2_2 = 59$	$< 10^{-8}$	0.42



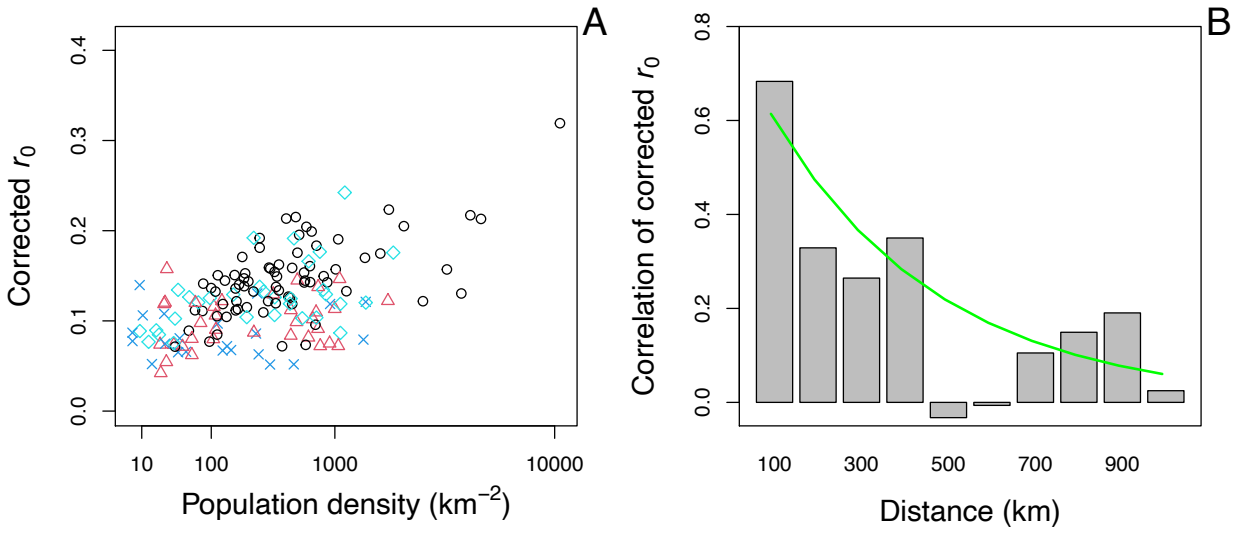
538

539 **Fig. 1.** Estimates of initial spread rate, r_0, for 124 counties (gray) and 36 county-aggregates

540 (blue) with 66% (bars) and 95% (whiskers) bootstrapped confidence intervals.

541

542

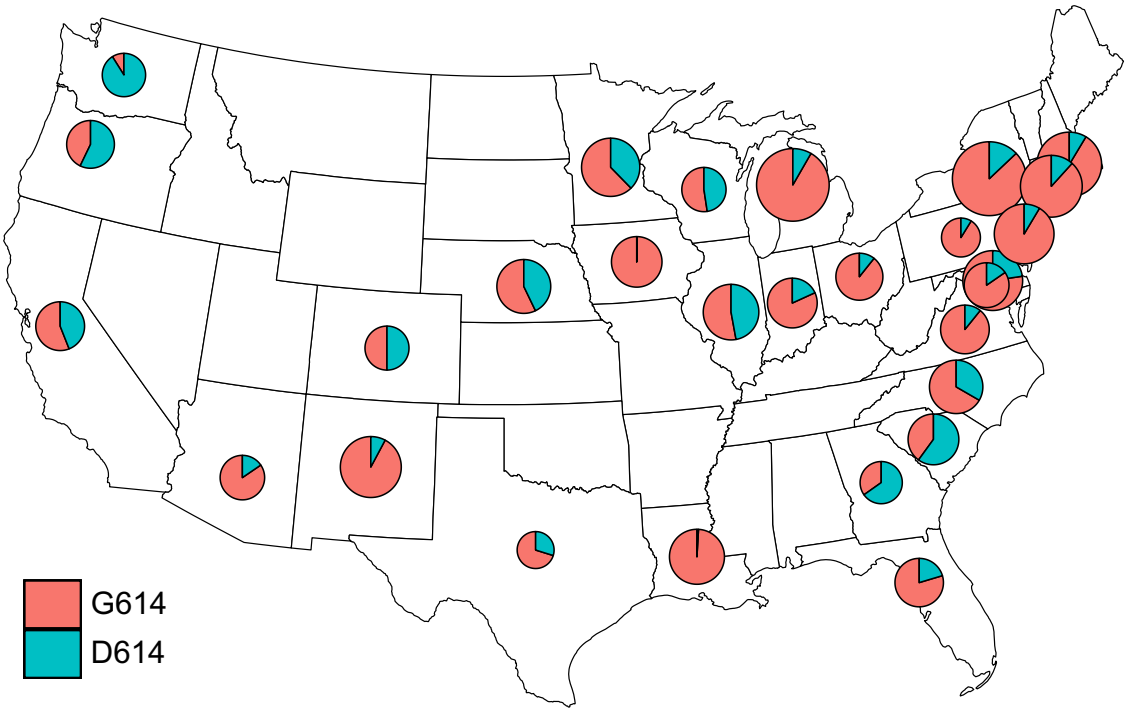


543

544

545 **Fig. 2.** Estimates of initial spread rates, r_0, after correcting for the effects of outbreak onset and
546 the population size. **(A)** Effect of population density: Northeast, black circles; Midwest, cyan
547 diamonds; South, blue x's; West, red triangles. **(B)** Effect of spatial proximity depicted by
548 computing correlations in bins representing 0-100 km, 100-200 km, etc. The line gives the
549 correlation of the residuals from the fitted regression.

550



551

552 **Fig. 3.** Spatial distribution of strains of SARS-CoV-2 having the G614 mutation in the spike
553 gene at the outbreak onset among states. Pie charts give the proportion of samples in states
554 collected within 30 days following the outbreak onset that are in the G clades (blue) [21]. The size
555 of the pie is proportional to the residual values of r_0 after removing the effects of the timing of
556 outbreak onset, population size represented by the time series, and population density. For each
557 state, we used the estimate of r_0 corresponding to the county or county-aggregate that had the
558 greatest number of deaths.

559

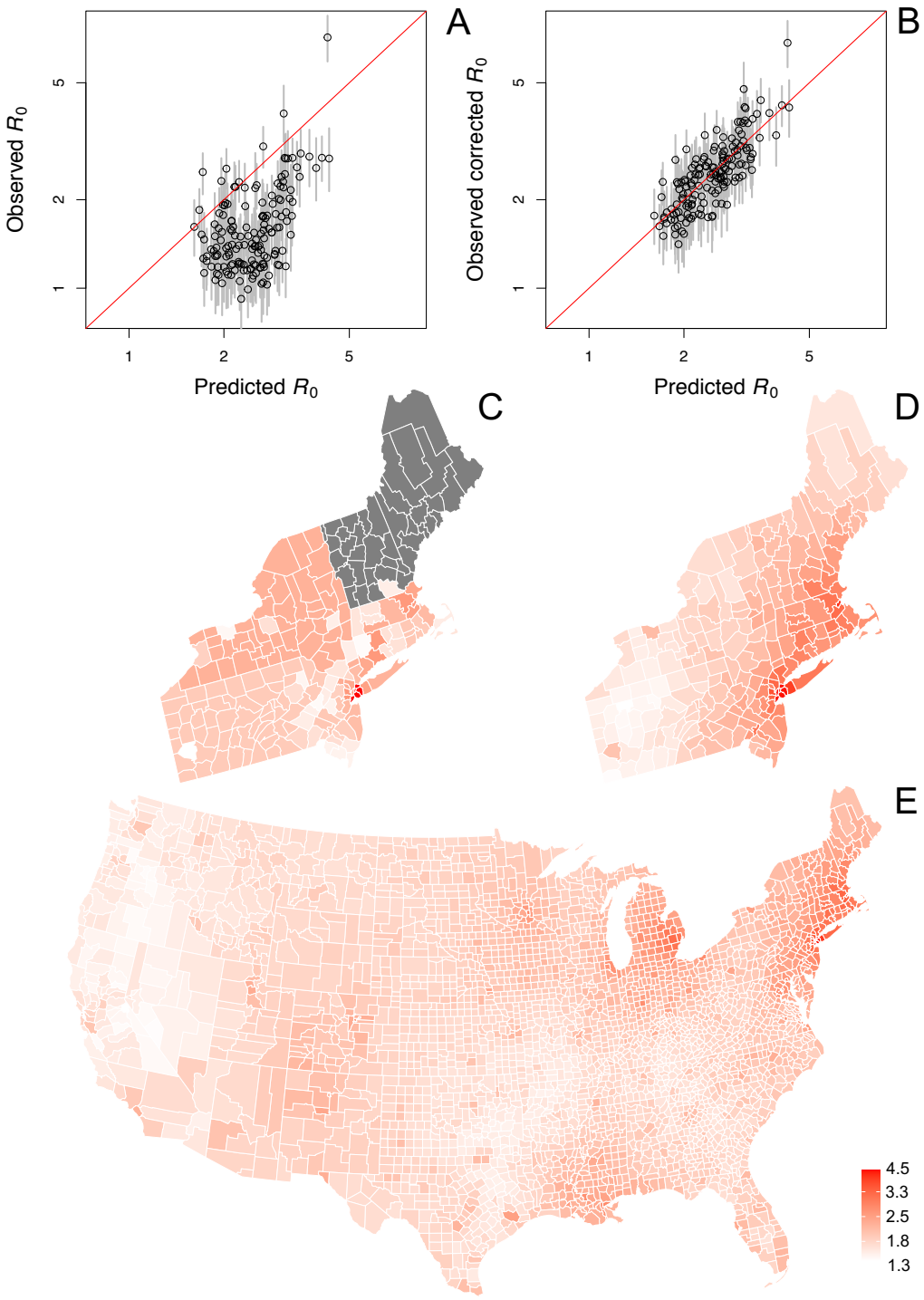


Fig. 4. (A,B) Raw and corrected estimates of R_0 for 160 counties and county-aggregates. The predicted R_0 values are obtained from the regression model, with corrections to standardize values to an outbreak onset of 11 March, 2020, and a population size equal to the most populous

564 county. Comparing the raw estimates of R_0 (A) and the corrected R_0 values (B) shows the
565 predictive power of the regression analysis. We thus used the regression model to predict R_0 for
566 all counties. **(C,D)** To illustrate the prediction process for the northeastern states, the raw
567 estimates (C) are all the same for county-aggregates and could not be made for some states
568 (gray). In contrast, the predictability R_0 in the regression model allows for better estimates (D).
569 This makes it possible to extend estimates of R_0 to all 3109 counties in the conterminous USA
570 **(E)**.
571

Supplementary Information

Estimating and explaining the spread of COVID-19 at the county level in the USA

Anthony R. Ives[1 *], Claudio Bozzuto[2]

Affiliations:

[1 *] Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI 53706, USA. arives@wisc.edu. ORCHID

[2] Wildlife Analysis GmbH, Oetlisbergstrasse 38, 8053 Zurich, Switzerland. bozzuto@wildlifeanalysis.ch. ORCHID

Corresponding Author: Anthony R. Ives, Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI 53706, USA. 608-238-3771. arives@wisc.edu

The SI includes:

Overview of Statistical Methods

The rate of spread of a disease in a population at the early phase of an epidemic, r_0, when the entire population is susceptible depends on the basic reproduction number, R_0, giving the number of secondary infections produced per infected individual, and the distribution of the time between primary and secondary infections. Thus, if the spread rate and distribution of infection times can be estimated, R_0 can then be calculated. Our strategy is to estimate r_0 as the most direct parameter associated with the dynamics of an epidemic, and then subsequently estimate R_0. The advantages of calculating r_0 include: (i) it captures all of the real-life complexities that affect R_0 by simply observing what happened in real life, and (ii) it uses data that are (tragically) becoming more prevalent. The challenges include (i) the changes in $r(t)$ that are to be expected (and hoped for) as people and governments respond to lessen the spread, and (ii) the statistical challenges and uncertainties of determining rates of disease spread when the numbers of deaths are still low.

We developed and tested statistical methods to overcome the two challenges of estimating R_0 from death data. Because the rate of spread of a disease may change rapidly in response to actions that are taken to reduce disease transmission, we used a time-varying autoregressive model that allows for the rate of spread to change through time, $r(t)$. Other models take a related approach [6,55]. The second challenge is that the counts of deaths at the beginning of an epidemic are low. To account for this, the time-series model includes increased uncertainty (measurement error) that depends on the time-varying estimate of the number of deaths. Standard (asymptotic) approaches often have poor statistical properties (type I errors, correctly calculated confidence intervals) when sample sizes are small [56]. Therefore, we use bootstrapping [57] in which simulation time series are reconstructed to share the same pattern as the observed time series; a

large number of simulated time series are then fit using the same statistical model as used to fit the original data. This bootstrapping procedure thus gives estimates and confidence intervals for model fit to the real data. Note that our approach is frequentist, in comparison to the majority of models that use a Bayesian framework.

Our approach focuses on estimating the time-varying rate of spread, $r(t)$, of the number of deaths. Our rationale is that, for statistical fitting, it is better to keep the model as simple as possible, rather than "building in" assumptions about the processes of infection, reporting, and death. Our simple phenomenological model uses the same data as more complicated, process-based models, and therefore both approaches ultimately rely on the same information. The simpler approach, however, does not depend on assumptions about the infection processes. Instead, after estimating r_0, we computed R_0 as $1/\sum_\tau e^{-r(t)\tau}p(\tau)$, where τ is the number days after initial infection, and $p(\tau)$ is the proportion of secondary infections produced per infected individual at τ [15]. This expression assumes that deaths (removal of individuals from the population) occur after all secondary infections have occurred. We used the distribution of $p(\tau)$ that was estimated using contact tracing in Wuhan, China [47].

To validate the statistical method, we constructed a simulation model of the transmission process and spread of infections iterated on a daily time scale. Our simulations considered scenarios in which the transmission rate changed through time either in steps or gradually to capture the extremes of possible changes in real $R(t)$. We varied the initial R_0 and duration of simulations to produce epidemics that qualitatively match the county data we analyzed. Changes in our estimates of $r(t)$ tended to lag behind changes in the true (simulated) value of $r(t)$ (gray line and regions in Fig. S1A,B), and therefore we also estimated $r(t)$ in the reverse direction (blue line and regions in Fig. S1A,B). For the estimate of the initial r_0, we averaged the estimates

641 from the forward and reverse time series. For the scenario of step changes in $R(t)$ (Fig. S1 C), the
642 estimates were unbiased and had accurate confidence intervals, although for the scenario of
643 gradual changes (Fig. S1 D), there was some downwards bias. Nonetheless, the estimates of
644 initial R_0 captured the order of simulations according to the true R_0. In contrast, fitting the same
645 time series with a commonly used Bayesian model that incorporates the transmission process
646 given in the R package EpiEstim [13] gave estimates that poorly reflect the true (simulated) initial
647 R_0 (Fig. S1 E,F).

648 We also used the simulation model to investigate the properties of the statistical method
649 when the number of deaths was low, as occurred in some time series. Reducing the simulated
650 values of R_0 reveals that the estimates of r_0 become biased downwards when the maximum
651 number of reported deaths per day drops below 15 (Fig. S2). This is due to the time series
652 containing too little information about the rate of increase in the number of mortalities for
653 accurate estimates. Because we did not think that our method (or any other) could overcome this
654 challenge, we incorporated population size encompassed by a time series in the subsequent
655 regression analysis. We used population size rather than the maximum number of deaths,
656 because this would introduce a confounding effect: time series with higher r_0 will likely have
657 higher numbers of deaths.

658 In order to extrapolate the estimates of R_0 from 160 time series to the remaining counties
659 in the conterminous USA, we *a priori* selected four predictors. We selected population size
660 encompassed by the time series to account for possible downwards bias in sparse datasets. We
661 selected the Julian date of the outbreak onset to factor out public and private responses to
662 COVID-19. We included population density, because it could potentially affect transmission
663 rates. Population size and density were weakly and negatively correlated among the 160 time

664 series (Pearson correlation between log population size and log density = –0.25), and therefore
665 there were no problems with multicollinearity. Finally, the regression model included spatial
666 autocorrelation based on the latitude and longitude of the midpoint of the counties or county
667 aggregates. Because the regression model had residual variance that was only slightly high than
668 the variance of the estimates of r_0 that the regression predicted, the precision of the estimates
669 from the regression for the counties without time series will be on par with the precision of the
670 counties with time series.
671
672 **Simulation model**
673 To assess the robustness of the statistical model, we built a simulation SIR (susceptible-
674 infected-recovered) model of a hypothetical epidemic. The simulation model was not the same as
675 the statistical model, so the goal was to determine whether the phenomenological statistical
676 model was capable of capturing the rate of infection spread in the process-based simulations.
677 The simulation model tracks the number of infected individuals on day t who were
678 infected τ days previously, $X(t;\tau)$. After 25 days, they are all assumed to be recovered or dead.
679 The probability distribution of the day on which a susceptible is infected, $p(t)$, is given by a
680 Weibull distribution with mean 7.5 days and standard deviation 3.4 *(23)* (Fig. S3 A). For an
681 individual who dies, the day of death, $d(t)$, is given by a Weibull distribution with mean 18.5
682 days and standard deviation 3.4 [47] (Fig. S3 B). Finally, for case data we need to know the time
683 between initial infection and diagnosis, $h(t)$, which we assume is lognormally distributed with
684 mean 5.5 days and standard deviation 2.2 [58] (Fig. S3 C).
685 On day t, the number of new infections produced by individuals who were infected τ days
686 earlier is $b(t)\ p(\tau)$. The term $b(t)$ is closely related to $R(t)$, the number of secondary infections

687 caused per infection. However, because we allow $b(t)$ to fluctuate on a daily basis, here we use a
688 notation that differs from $R(t)$. Note, however, that on average $R(t) = \sum_\tau b(t+\tau)\, p(\tau)$. The total
689 number of new infections on day t is given by a lognormal Poisson distribution in which the
690 mean of the Poisson process is $b(t)\ \alpha(t) \sum_\tau p(\tau) X(t;\tau)$, where the lognormal random variable $\alpha(t)$
691 is included to represent environmental variation.

692 Deaths occur according to a binomial distribution for each infection age category $X(t;\tau)$,
693 so that the probability of death of individuals that had been infected τ days earlier is $(1 - s)\ \beta(t)$
694 $d(\tau)$, where s is the overall survival probability and $\beta(t)$ is a lognormal distribution. We assume
695 that the overall survival probability for COVID-19 is 98%; changes in this assumption had little
696 effect on the simulation study. Once an individual dies, they are removed from the pool of
697 individuals.

698 To illustrate the simulations, we assumed that the expectation of the infection rate, $b(t)$,
699 changes as a step function (Fig. S4 A, black line), while there is also daily variation around this
700 expectation (Fig. S4 A, points). We also calculated $R(t)$ from the asymptotic rate of disease
701 spread (Fig. S4 A, red line). This shows that the expected daily infection rate, $b(t)$, is closely
702 related to the population-level $R(t)$. Over the simulated time series of 60 days, we then recorded
703 the number of deaths (Fig. S4 B) and diagnosed cases (Fig. S4 C). We initiated the simulation
704 with a single cohort of individuals, all infected on day 1 (Fig. S4 C, filled black dot). This gives
705 the "worst-case" situation in which the distribution of time-since-infection is far from the stable
706 age distribution.

707 We fit this simulated dataset using the same procedure as we used for the real data,
708 including the same rules to determine which day to initiate the fitted time series (Fig. S1 A).

We performed a similar exercise while assuming that the expectation of the infection rate, $b(t)$ changes geometrically, producing a linear change in $r(t)$ (Fig. S1 B). In this particular example, the estimated values of $r(t)$ are below the true values in the simulation in the first part of the time series. Because there was a lag in response of the estimates of $r(t)$ relative to $b(t)$, we fit the time series in both the forward and reversed directions, and we averaged these values (and their confidence intervals) for the final estimates. Note that this is possible in our approach, because we estimate $r(t)$ rather than $R(t)$.

We performed 100 simulations with the expectation of $b(t)$ changing as either a step function (Fig. S1 C) or geometrically (Fig. S1 D), to assess the overall robustness of the modeling approach. Simulations were performed by changing the initial value of $b(t)$. Because higher values of $b(t)$ led to much higher numbers of deaths, we shorted the intervals between step changes and increased the decline in geometric changes in $b(t)$ to roughly match the observed time series. Specifically, the simulated time series ranged in length from 55 to 150 days: for the case of step changes, the time series were broken into three equal periods, and for the case of geometric changes, the ending value of $b(t)$ was kept the same. We also estimated $R(t)$ using the R package EpiEstim under default control parameters [13]. EpiEstim has the same general structure of many of the Bayesian models that estimate $R(t)$ directly using information about the transmission process (Fig. S1 E,F). Even though EpiEstim is structurally more complicated than our model, it tended to give values of R_0 that were biased upwards when the true value was low, and biased downward when the true value was high. Finally, we investigated the bias in our estimates of r_0 when the maximum number of deaths in a time series was low by simulating time series for 20 to 70 days, using an initial value of $b(t)$ to correspond to $R_0 = 4$, and changing the timing of step changes or the rate of geometric decline of $b(t)$ to correspond to the length of the

time series. The simulations show that the estimates of r_0 are downward biased when the total numbers of counts are low (Fig. S2).

Analysis of Nextstrain metadata of SARS-CoV-2 strains

In the analyses presented in the main text, we used the GISAID metadata to test the specific assumption that the G614 mutation increases the rate of spread of SARS-CoV-2. Prior to this analysis, however, we analyzed a subset of the genomic data available from Nextstrain [23]. We present this analysis here, because it was a naïve analysis that did not have a specific hypothesis about what strains might lead to higher spread rates. Instead, we asked whether the proportion of different Nextstrain clades (19A, 19B, 20A, 20B, 20C in the USA) within a population were related to r_0 estimates. We used the same statistical approach as we present for the GISAID metadata, except we included the proportion of strains from clades 19A, 19B, 20A, and 20B instead of the proportion in the G clades containing mutation G614; we excluded the largest clade, 20C, because the sums of the proportions must add to one, and therefore all of the information about the distribution of strain 20C among states is contained in the distribution of the other clades. We found that the proportion of samples within clade 19B had a negative effect on r_0 ($P = 0.019$, Table S2). The high proportion of strains from 19B in the Pacific Northwest and the Southeast were associated with lower values of r_0 (Fig. 3).

Supplementary figures and tables

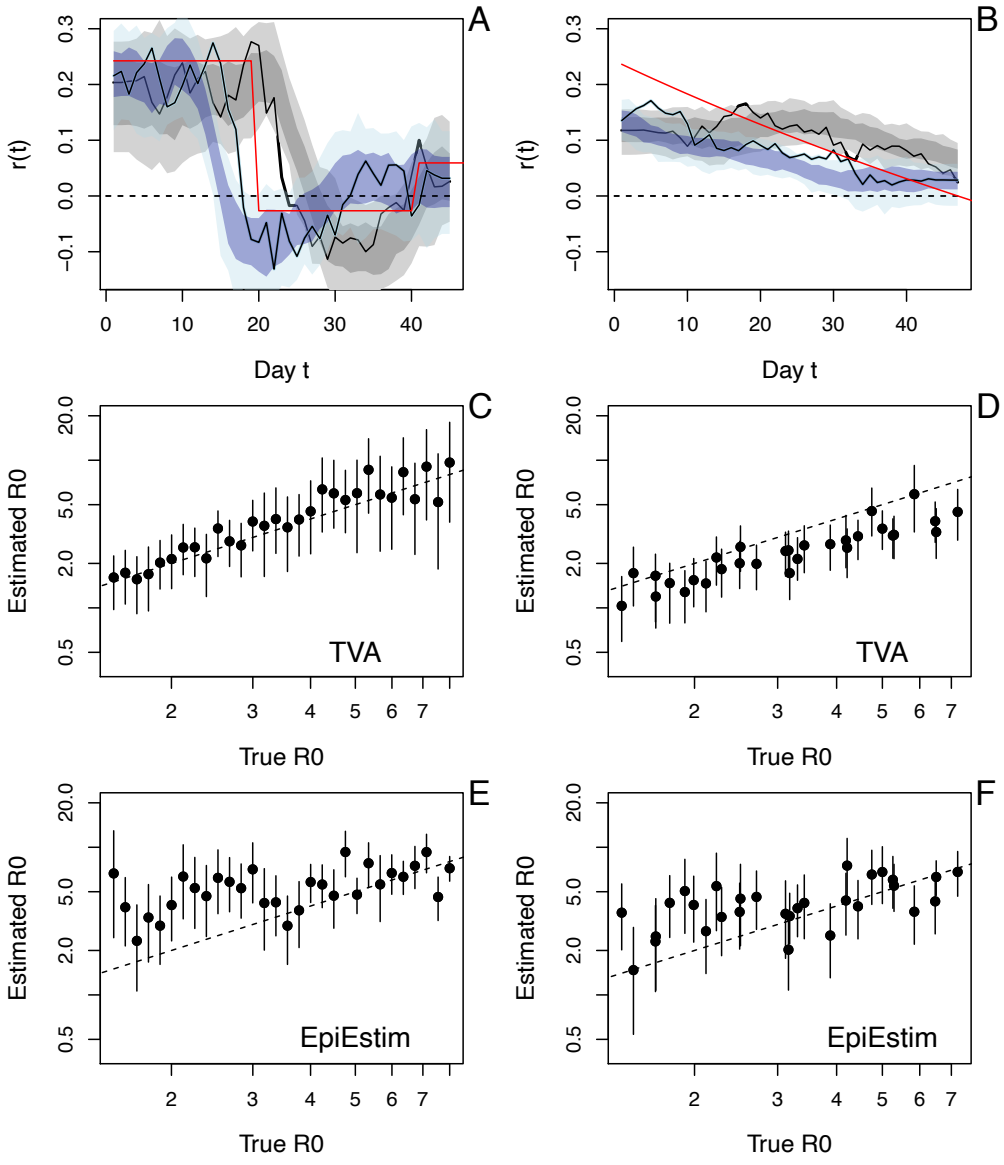


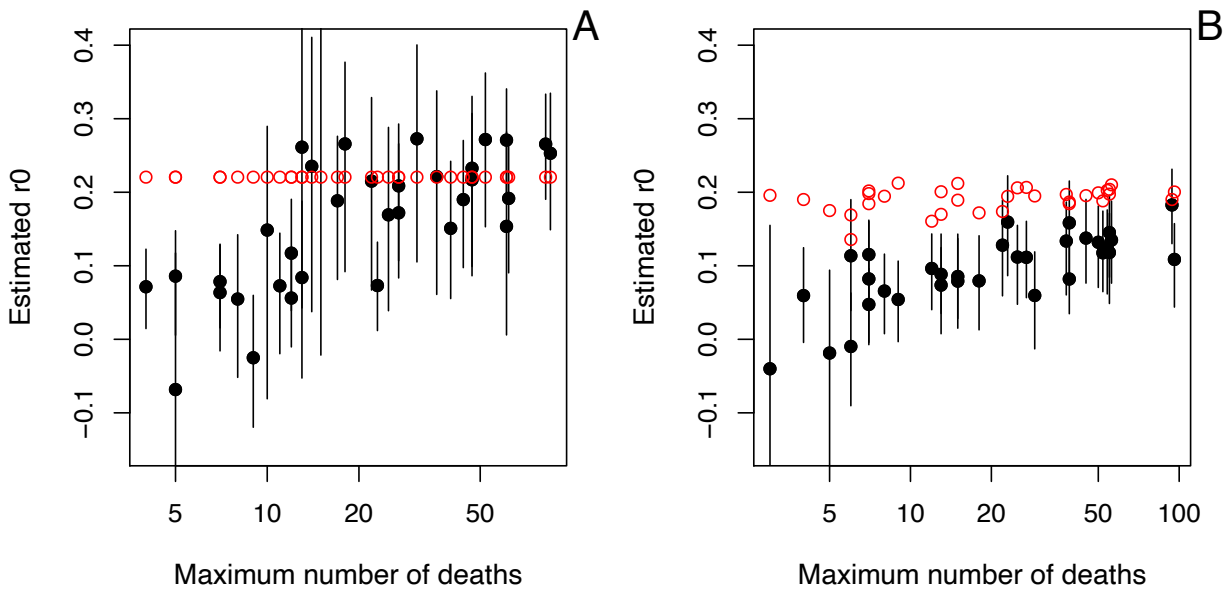
Fig. S1. Simulation study of fitting methods to epidemic death data. Simulations were fit with the time-varying autoregression model (TVA) in the forward (black line with dark and light gray regions giving 66% and 95% approximate confidence intervals) and reverse (blue line and regions) directions when the true value of R(t) (red line) shows either **(A)** a step or **(B)** gradual changes. For each simulation, the forward and reverse estimates were averaged to give an

760 estimate of R0 with 95% confidence intervals, which are plotted against the true values of R0 for

761 step **(C)** and gradual **(D)** changes in R(t). The same simulations with fit using EpiEstim **(E,F)**.

762

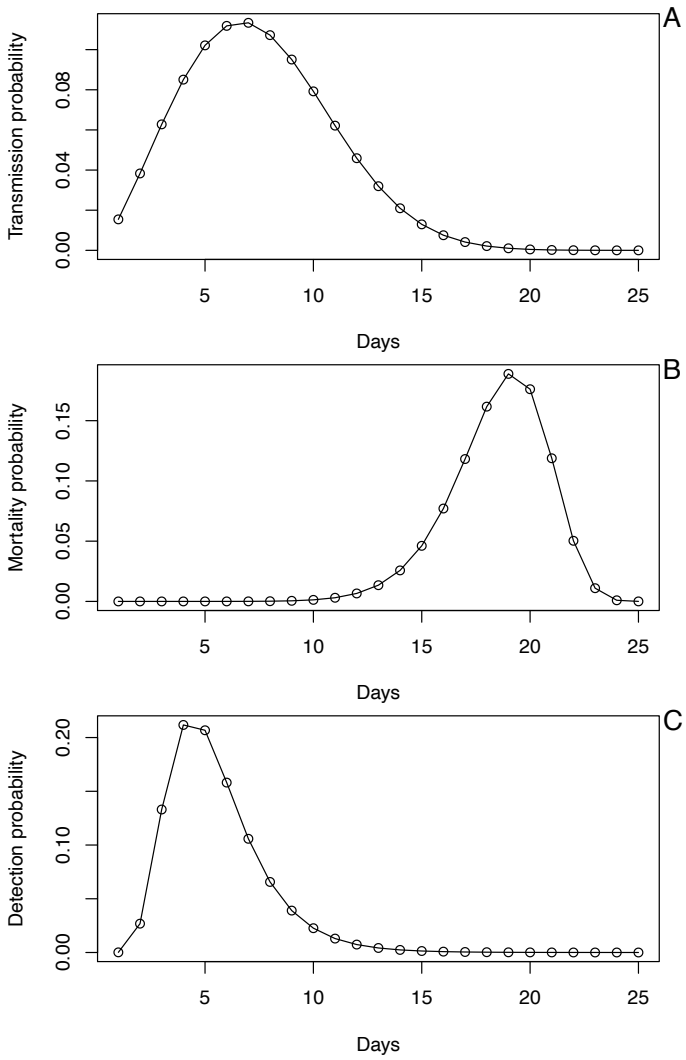
763



764

765

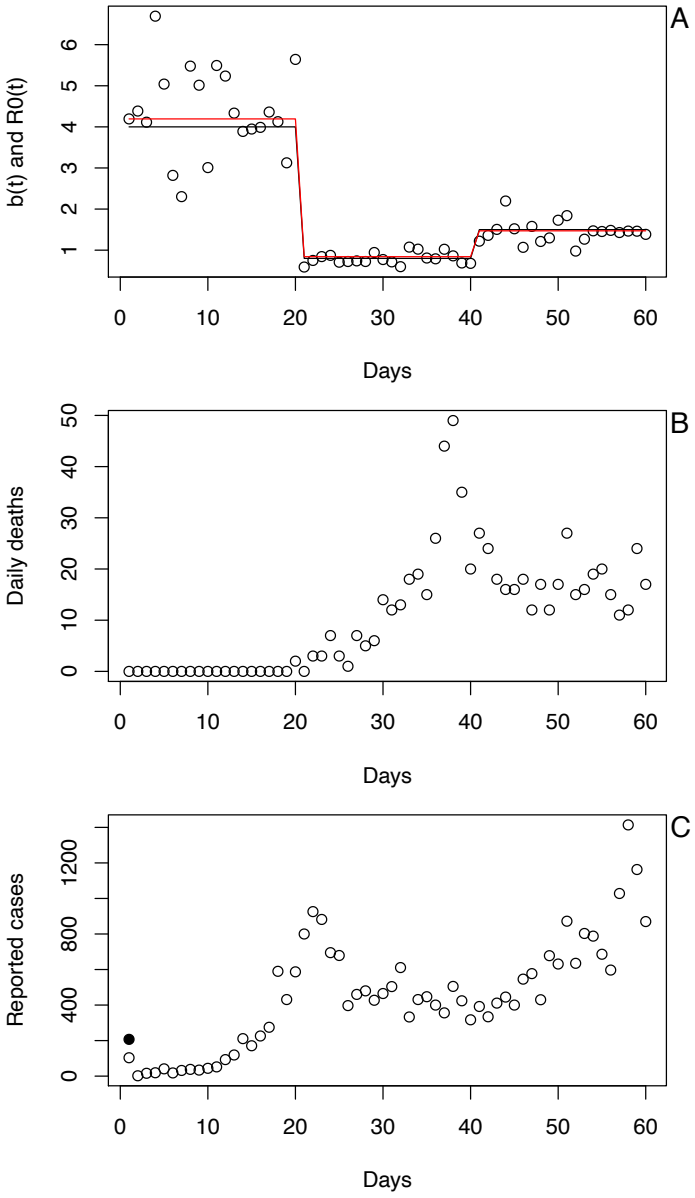
766 **Fig. S2.** Simulation study of the estimation of r_0 from the forward and reverse time-varying
767 autoregressive model for different population sizes. Simulations following those used for Fig. 1
768 were performed assuming $r(t)$ changed either **(A)** in steps or **(B)** gradually. The simulations were
769 performed using the same initial value of r_0, but the length of time of the simulation was varied
770 to change the maximum number of deaths that occurred. Due to the stochastic nature of the
771 simulations, the realized value of r_0 when the analysis was started differed among time series
772 when $r(t)$ changed gradually (red points in B), while they were all 0.22 when $r(t)$ was changed in
773 steps (A). The median in the maximum number of deaths among the real county time series was
774 21.



775

776

777 **Fig. S3.** Probability distributions used in the process-based SIR simulation model used to test
778 methods for robustness. **(A)** The probability distribution of the day on which a susceptible is
779 infected, $p(t)$, which is given by a Weibull distribution with mean 7.5 days and standard
780 deviation 3.4. **(B)** For an individual who dies, the day of death, $d(t)$, which is given by a Weibull
781 distribution with mean 18.5 days and standard deviation 3.4. **(C)** For case data, the time between
782 initial infection and diagnosis, $h(t)$, which is lognormally distributed with mean 5.5 days and
783 standard deviation 2.2.



784

785 **Fig. S4.** Example simulation from the process-based SIR model. **(A)** Changes in the infection

786 rate, $b(t)$, are modeled as a step function (black line) with daily variation (points). R0(t) (red line)

787 tracks changes in $b(t)$. **(B)** and **(C)** The number of deaths (B) and diagnosed cases (C) when the

788 simulation is initiated with a single cohort of individuals, all infected on day 1 (solid black dot).

789

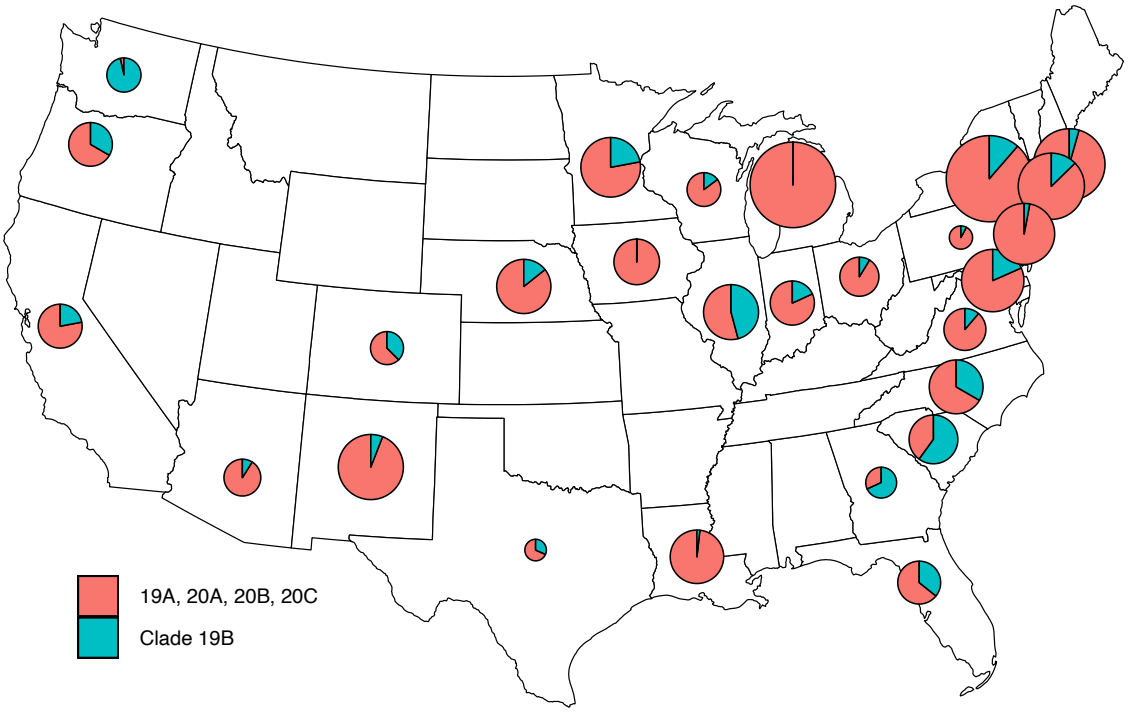


Fig. S5. Spatial distribution of the 19B clade of SARS-CoV-2 at the outbreak onset among states. Pie charts give the proportion of samples in states collected within 30 days following the outbreak onset that are in the 19B clade (blue). The size of the pie is proportional to the residual values of r_0 after removing the effects of the timing of outbreak onset, population size represented by the time series, and population density. For each state, we used the estimate of r_0 corresponding to the county or county-aggregate that had the greatest number of deaths.

799 **Table S1.** Separate spreadsheet giving the following variables for the 3109 counties in the
800 conterminous USA.
801

Variable	Description
ST	two-letter state abbreviation
state_county	state abbreviation with county name
fips	FIPS identifier for counties
lon	longitude
lat	latitude
death.max	maximum number of daily deaths
start.date	state date of the analyzed time series
end.date	end date of the analyzed time series
den	population density
r0.est	estimate of r_0 from time-series analyses
r0.est.se	standard error of the estimate of r_0 from bootstrapping
r0.est.cor	corrected estimate of r_0 removing start.date and the population size
r0.l66.cor	lower 66% confidence interval of the corrected estimate of r_0
r0.u66.cor	upper 66% confidence interval of the corrected estimate of r_0
r0.pred	predicted estimate of r_0 from the regression model
r0.pred.se	standard error of the predicted estimate of r_0
R0.pred	predicted estimate of R_0 from the predicted estimate of r_0
R0.pred.l66	lower 66% confidence interval of the predicted estimate of R_0
R0.pred.u66	upper 66% confidence interval of the predicted estimate of R_0

Table S2. Regression of the initial spread rate, r_0, of COVID-19 against (i) the date of outbreak onset, (ii) total population size, (iii) population density, and (iv) the proportion of samples of SARS-CoV-2 containing the G614 mutation in the spike gene [21]. The estimates of r_0 were for the county or county-aggregate with the greatest number of deaths in the state. All genetic samples were collected within 30 days following the onset of outbreak in a county. Twenty-eight states had five or more genetic samples, and only these states are included in the regression.

	Coefficient	**SE**	**t**	**P**
onset	–0.0027	0.0013	–2.23	0.036
log(size)	0.022	0.009	2.46	0.022
density$^{1/4}$	0.013	0.005	2.85	0.009
G614	0.124	0.048	2.60	0.016

Table S3. Regression of the initial spread rate, r_0, of COVID-19 against (i) the date of outbreak onset, (ii) total population size, (iii) population density, and (iv) the proportion of samples of SARS-CoV-2 in four of the five clades identified in [24]. The estimates of r_0 were for the county or county-aggregate with the greatest number of deaths in the state. All genetic samples were collected within 30 days following the onset of outbreak in a county. Twenty-seven states had five or more genetic samples, and only these states are included in the regression. Transforms of population size and density were selected to best-fit the data and satisfy linearity assumptions.

	Coefficient	**SE**	**t**	**P**
onset	0.0027	0.0014	–1.88	0.076
log(size)	0.023	0.010	2.18	0.042
density$^{1/4}$	0.015	0.005	3.00	0.007
19A	-0.083	0.091	-0.91	0.37
19B	-0.134	0.052	-02.54	0.019
20A	-0.034	0.055	-0.71	0.48
20B	0.008	0.165	-0.05	0.96

819 **Table S4.** Variables giving population characteristics that were including in the regression model
820 (equation S3). No variable was statistically significant. Data from [29,30].

821

Variable	Description
median age	median age 2010
adult obesity	incidence of adult obesity
diabetes	incidence of adult diabetes
education	percent bachelor's degree or higher, 2005-2009
income	median earnings 2010
poverty	percentage people below federal poverty threshold
economic equality	Gini index
race	percent White, non-Latino
political leaning	proportion of votes cast for Donald Trump, 2016

822

823

824 **Table S5.** For 160 county and county-aggregates, regression of spread rate at the end of the time
825 series, corresponding to 5 May, 2020, $r(t_{end})$, against (i) the date of outbreak onset, (ii) total
826 population size and (iii) population density, in which (iv) spatial autocorrelation is incorporated
827 into the residual error. Transforms of population size and density were selected to best-fit the
828 data and satisfy linearity assumptions. The coefficient column contains the estimate of the
829 regression parameters with their associated t-tests; spatial autocorrelation is characterized by a
830 range and nugget for regional and local sources of variation, and their joint significance is given
831 by a likelihood ratio test. For the overall model, $R^2_{pred} = 0.38$.
832

	Coefficient	**SE**	**t**	**P**	**partial R^2_{pred}**
onset	0.0021	0.0003	6.40	$< 10^{-8}$	0.17
log(size)	0.0097	0.0021	4.61	$< 10^{-6}$	0.083
density$^{1/4}$	-0.0008	0.0013	-0.57	0.57	0.003
space	range = 0.29 nugget = 0.18		$\chi^2_2 = 10.3$	0.0056	0.099

833