Predictable county-level estimates of R_0 for COVID-19 needed for public health planning in the USA

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

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

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

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. Among-county variance in the rate of spread was explained by four factors: the timing of the county-level outbreak, population size, population density, and spatial location. 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, while spatial location can be explained by differences in the transmissibility of SARS-CoV-2 strains in different geographical regions of the USA. 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.

The basic reproduction number, R_0, is the number of secondary infections produced per primary infection of a disease, 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 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 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 morality 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 remains the same. 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.

Estimates of the spread rate, r_0

We applied a time-varying autoregressive state-space model to each time series [12]. In contrast to other models of COVID-19 epidemics [e.g., 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 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 (Supplementary materials). The model was fit accounting for greater uncertainty when mortality counts were low, and confidence intervals of the estimates were obtained from parametric bootstrapping. 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 bias our estimates away from the true value of r_0 in naïve populations before public or private health interventions were taken. 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. 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 few deaths were downward biased (Fig. S2; Supplementary materials). 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 "nuisance" factors, 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 (Supplementary materials).

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 the "nuisance variables" – 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. 2B). This in part reflects the relatively high values of corrected r_0 clustered on the Northeast and Midwest, and the relatively low r_0 clustered in the West and South (Fig. 2B). Nonetheless, these regional differences were not statistically significant when additionally included in the regression analysis. Spatial autocorrelation might also reflect differences in public responses to COVID-19 across the USA not captured by the variable in the regression model for outbreak onset. For example, Seattle, WA, reported the first positive case in the USA, on 15 January, 2020, and there was a public response before deaths were recorded [19]. In contrast, the response in New York City was delayed, even though the outbreak occurred later than in Seattle [20]. Spatial autocorrelation could also be caused by movement of infected individuals. However, movement would only lead to autocorrelation in our regression analysis if many of the reported deaths were of people infected outside the county. A further possibility is that spatial variation in the rate of spread of COVID-19 reflects spatial variation in the occurrence of different genetic strains of SARS-CoV-2.

To investigate whether spatial autocorrelation could potentially be caused by different strains of SARS-CoV-2 differing in infectivity and/or morbidity, we analyzed publicly available genomic sequences [21, 22]. We asked whether the proportion of different clades (19A, 19B, 20A, 20B, 20C in the USA) within a population were related to r_0 estimates. Because many of

the samples are only located at the state level, we performed the analysis accordingly, for each state selecting the r_0 from the county or county-aggregate with the highest number of deaths (and hence being most likely represented in the genomic samples). We further restricted genomic samples to those collected within 30 days following the outbreak onset we used to select the data for time-series analyses. This data handling resulted in 27 states available for analysis. We again used our regression model, except that now we included the proportion of strains from clades 19A, 19B, 20A, and 20B instead of spatial location (Eq. S7, Supporting materials); we excluded the largest clade, 20C, because the sums of the proportions must add to one. 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). These results suggest that the distribution of different SARS-CoV-2 strains affected the local spread of COVID-19 at broad geographic scales.

Lower transmissibility of 19B is also suggested by the increasing prevalence of other strains in the USA [22]. 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: some studies present evidence for a differential pathogenicity and transmissibility [23, 24], while others conclude that mutations might be mostly neutral or even reduce transmissibility [25]. The clade-specific differences we found call for further investigation to better understand the potential link between viral genomic variation and its impact on transmission and mortality [26].

As a check for whether our corrected estimates of r_0 represent values in naïve populations, we investigated additional population characteristics [27, 28] that should not 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 S3). The first three characteristics likely affect morbidity [29] but not transmission rates, and therefore they should not be significant when included with population size (the covariate that we uses to factor out differences in the number of recorded deaths). The remaining characteristics might affect health outcomes and responses to public health interventions, but they should not be significant when included with the date of outbreak onset. Accordingly, none of these characteristics 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 $r(t)$ at the

end of the time series (5 May, 2020, assuming an average time between infection and death of 18 days) (Table S4). 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 predictability of our estimates 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 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. Therefore, using estimates from other time series will give estimates of r_0 for a county 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. 4A) and R_0 calculated from the corrected r_0 values (Fig. 4B). Extrapolation from the regression model makes it possible not only to get refined estimates for the counties that were aggregated in the time-series analyses; it also gives estimates for counties within states with so few deaths that county-aggregates could not be analyzed (Fig. 4C,D). The end product is a map of R_0 for the conterminous USA (Fig. 4E).

County-level policies based on R_0

It is widely understood that different states and counties in the USA, and different countries in the world, have experienced COVID-19 epidemics differently. Our analyses have put numbers on these differences in the USA. The large differences argue for public health interventions to be designed at the county level. For example, the vaccination coverage in the most densely populated area, New York City, needed to prevent future outbreaks of COVID-19 will be much greater than for sparsely populated counties. Therefore, once vaccines are developed, they should be distributed first to counties with high R_0. Similarly, if vaccines are not developed quickly and non-pharmaceutical public health interventions have to be re-instated

during resurgent outbreaks, then counties with higher R_0 values will require stronger interventions. As a final example, county-level R_0 values can be used to assess the practicality of contact-tracing of infections, which become impractical when R_0 is high [30].

We present our county-level estimates of R_0 as preliminary guides for policy planning, while recognizing the myriad other epidemiological factors (such as mobility [31-33]) and political factors (such as legal jurisdictions [34]) that must shape public health decisions [3, 35-37]. Although we have emphasized the predictability of R_0 among counties in the USA, values of R_0 could change 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 [38-40]. 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.

Acknowledgments: 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 PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the basic reproduction number (R0). Emerging Infectious Diseases. 2019;25(1):1-4.

2. Hilton J, Keeling M. Estimation of country-level basic reproductive ratios for novel Coronavirus (COVID-19) using synthetic contact matrices. Preprint [Internet]. 2020 February 27, 2020. doi: 10.1101/2020.02.26.20028167.

3. Fine P, Eames K, Heymann DL. “Herd immunity”: a rough guide. Clinical Infectious Diseases. 2011;52(7):911-6. doi: 10.1093/cid/cir007.

4. Anderson RM. The concept of herd immunity and the design of community-based immunization programmes. Vaccine. 1992;10:928-35. doi: 10.1016/0264-410X(92)90327-G.

5. Flaxman S, et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Report 13, Imperial College London [Internet]. 2020 2020. doi: arXiv:2004.11342

6. Scire J, Nadeau S, Vaughan T, Brupbacher G, Fuchs S, Sommer J, et al. Reproductive number of the COVID-19 epidemic in Switzerland with a focus on the Cantons of Basel-Stadt and Basel-Landschaft. Swiss Medical Weekly [Internet]. 2020; 150(1920). doi: smw.ch/article/doi/smw.2020.20271.

7. Adam D, et al. Clustering and superspreading potential of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections in Hong Kong. 2020 2020-05-21. doi: 10.21203/rs.3.rs-29548/v1.

8. Paull SH, Song S, McClure KM, Sackett LC, Kilpatrick AM, Johnson PT. From superspreaders to disease hotspots: linking transmission across hosts and space. Frontiers in Ecology and the Environment. 2012;10(2):75-82. doi: 10.1890/110111.

9. Lofgren E, Fefferman NH, Naumov YN, Gorski J, Naumova EN. Influenza seasonality: underlying causes and modeling theories. Journal of Virology. 2007;81(11):5429-36. doi: 10.1128/JVI.01680-06.

10. Peña-García VH, Christofferson RC. Correlation of the basic reproduction number (R0) and eco-environmental variables in Colombian municipalities with chikungunya outbreaks during 2014-2016. PLoS Neglected Tropical Diseases [Internet]. 2019 2019-11-7; 13(11):[e0007878 p.]. doi: 10.1371/journal.pntd.0007878.

11. New York Times. Coronavirus (Covid-19) data in the United States. 2020. Available from: https://github.com/nytimes/covid-19-data.

12. Ives AR, Dakos V. Detecting dynamical changes in nonlinear time series using locally linear state-space models. Ecosphere. 2012;3(6):art58. doi: http://dx.doi.org/10.1890/ES11-00347.1.

13. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. American Journal of Epidemiology. 2013;178(9):1505-12. doi: 10.1093/aje/kwt133.

14. Flaxman S, et al. State-level tracking of COVID-19 in the United States. Report 23, Imperial College London [Internet]. 2020 2020. doi: https://doi.org/10.25561/79231.

15. Dublin LI, Lotka AJ. On the true rate of natural increase. Journal of the American Statistical Association. 1925;20:305–39.

16. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Bio Medica Atenei Parmensis. 2020;91(1):157-60. doi: 10.23750/abm.v91i1.9397.

17. Ives AR. R2s for correlated data: phylogenetic models, LMMs, and GLMMs. Systematic Biology. 2019;68(2):234-51. doi: 10.1093/sysbio/syy060. PubMed PMID: WOS:000462830500004.

18. Rader B, Scarpino S, Nande A, Hill A, Reiner R, Pigott D, et al. Crowding and the epidemic intensity of COVID-19 transmission. medRxiv. 2020:2020.04.15.20064980. doi: 10.1101/2020.04.15.20064980.

19. Fink S. Mapping path of virus from first US foothold. The New York Times. 2020 22 April, 2020;Sect. A.

20. Anon. Briefling: Covid-19 in America. The Economist. 2020 30 May, 2020.

21. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3. doi: 10.1093/bioinformatics/bty407.

22. NextstrainTeam. Nextstrain. 2020. Available from: https://nextstrain.org/ncov.

23. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020:2020.04.29.069054. doi: 10.1101/2020.04.29.069054.

24. Yao H, Lu X, Chen Q, Xu K, Chen Y, Cheng L, et al. Patient-derived mutations impact pathogenicity of SARS-CoV-2. medRxiv. 2020:2020.04.14.20060160. doi: 10.1101/2020.04.14.20060160.

25. Dorp Lv, Richard D, Tan CC, Shaw LP, Acman M, Balloux F. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. bioRxiv. 2020:2020.05.21.108506. doi: 10.1101/2020.05.21.108506.

26. 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. 2020;96:459-60. doi: 10.1016/j.ijid.2020.05.071.

27. Kirkegaard EOW. Inequality across US counties: an S factor analysis. Open Quantitative Sociology and Political Science [Internet]. 2016. doi: 10.26775/OQSPS.2016.05.23.

28. United States Census Bureau. USA Counties. 2011. Available from: https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html.

29. 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 [Internet]. 2020 2020; 69. doi: 10.15585/mmwr.mm6913e2.

30. Fraser C, Riley S, Anderson RM, Ferguson NM. Factors that make an infectious disease outbreak controllable. Proceedings for the National Academy of Sciences. 2004;101:6146–51.

31. Bichara D, Kang Y, Castillo-Chavez C, Horan R, Perrings C. SIS and SIR Epidemic Models Under Virtual Dispersal. Bulletin of Mathematical Biology. 2015;77(11):2004-34. doi: 10.1007/s11538-015-0113-5.

32. Roberts MG, Heesterbeek JaP. A new method for estimating the effort required to control an infectious disease. Proceedings of the Royal Society of London Series B: Biological Sciences. 2003;270(1522):1359-64. doi: 10.1098/rspb.2003.2339.

33. Gatto M, Bertuzzo E, Mari L, Miccoli S, Carraro L, Casagrandi R, et al. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. Proceedings of the National Academy of Sciences. 2020;117(19):10484-91. doi: 10.1073/pnas.2004978117.

34. Gorman S, Bernstein S. Wisconsin Supreme Court invalidates state's COVID-19 stay-at-home order. Reuters [Internet]. 2020. Available from: https://www.reuters.com/article/us-health-coronavirus-usa-wisconsin/wisconsin-supreme-court-invalidates-states-covid-19-stay-at-home-order-idUSKBN22Q04H.

35. Lahariya C. Vaccine epidemiology: A review. Journal of Family Medicine and Primary Care. 2016;5(1):7-15. doi: 10.4103/2249-4863.184616.

36. Mallory ML, Lindesmith LC, Baric RS. Vaccination-induced herd immunity: Successes and challenges. Journal of Allergy and Clinical Immunology. 2018;142(1):64-6. doi: 10.1016/j.jaci.2018.05.007.

37. Ridenhour B, Kowalik JM, Shay DK. Unraveling R0: Considerations for public health applications. American Journal of Public Health. 2013;104(2):e32-e41. doi: 10.2105/AJPH.2013.301704.

38. Imperial College London. Covid-19 Scenario Analysis Tool. 2020. Available from: https://covidsim.org.

39. Systrom K, Vladeck T. Rt Covid-19. 2020. Available from: https://rt.live.

40. Swiss National Covid-19 Science Task Force. Situation report. 2020. Available from: https://ncs-tf.ch/en/situation-report.

Table 1. For 160 county and county-aggregates, regression of 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. For the overall model, $R^2_{pred} = 0.69$, and the residual standard error is 1.11.

	Value	**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		$\chi^2_1 = 59$	$< 10^{-8}$	0.42

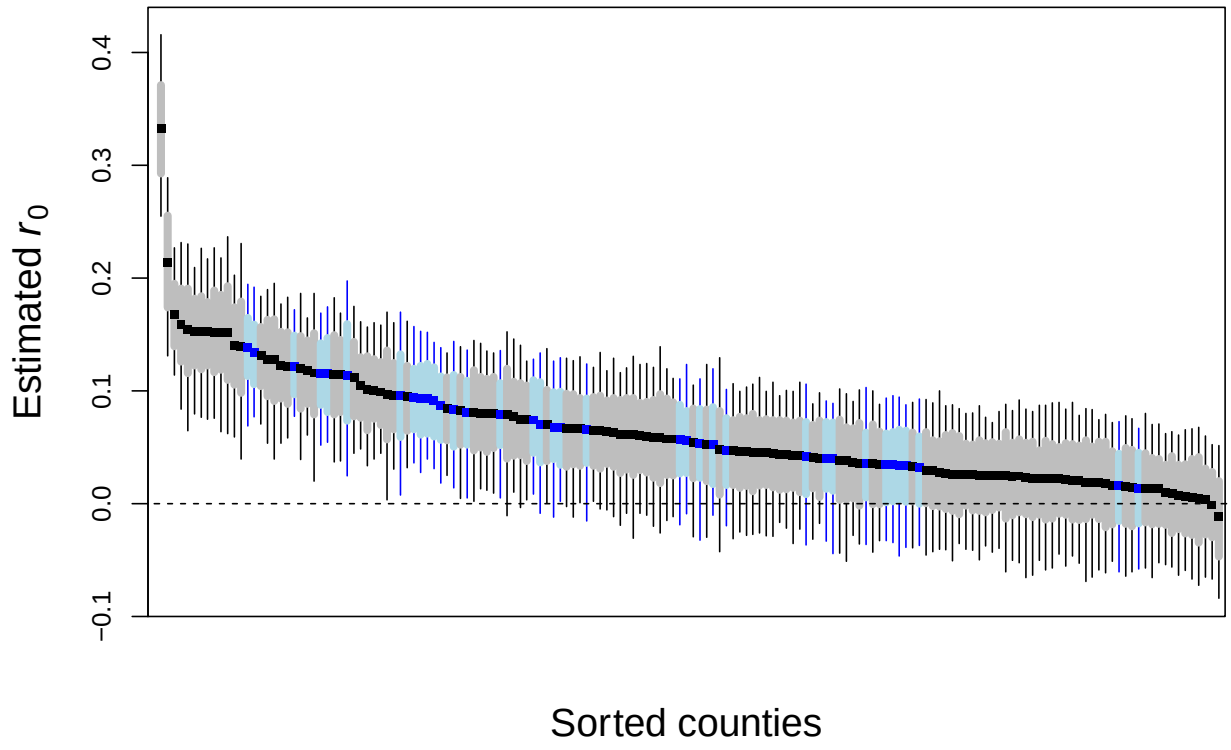

Fig. 1. Estimates of initial spread rate, r_0, for 124 counties (gray) and 36 county-aggregates (blue) with 66% (bars) and 95% (whiskers) bootstrapped confidence intervals.

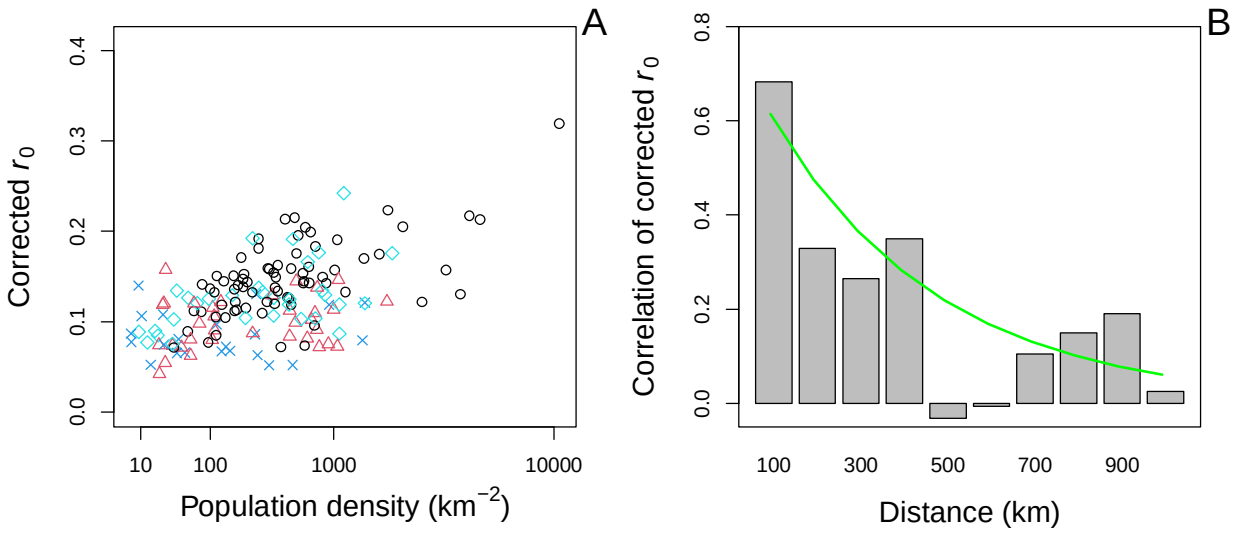

Fig. 2. Estimates of initial spread rates, r_0, after correcting for the effects of outbreak onset and the population size. **(A)** Effect of population density: Northeast, black circles; Midwest, cyan diamonds; South, blue x's; West, red triangles. **(B)** Effect of spatial proximity depicted by computing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

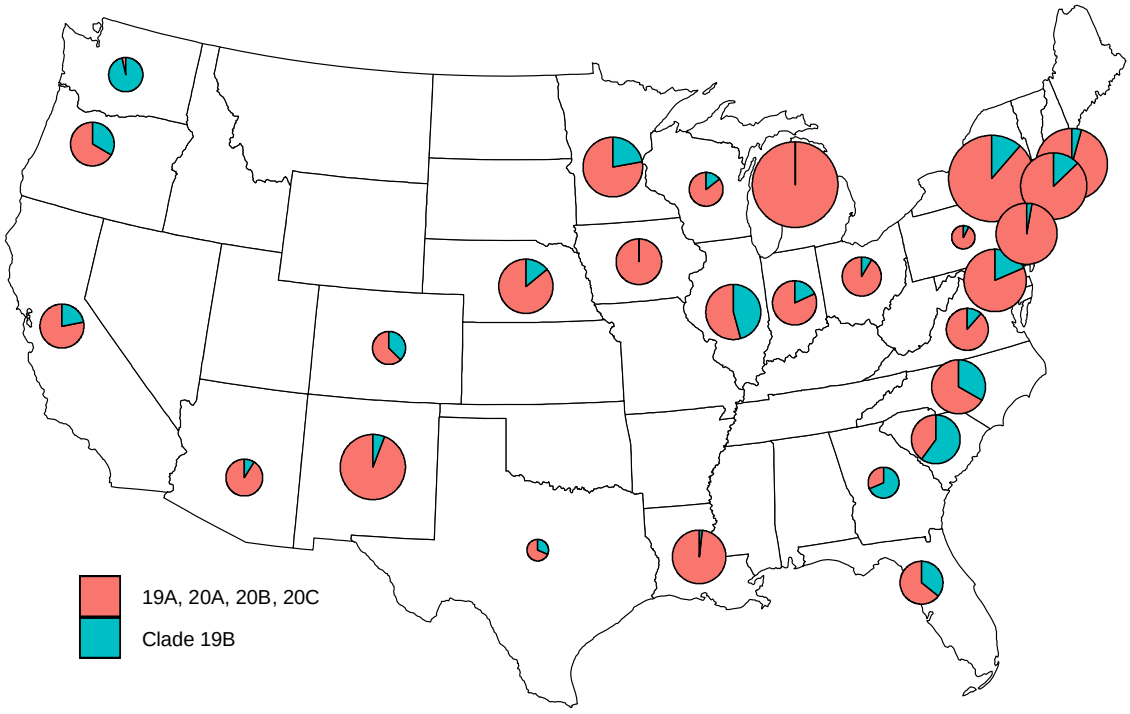

Fig. 3. Spatial distribution of the 19B clade of SARS-CoV-2 at the outbreak onset among states. Pie charts give the proportion of samples in states collected within 30 days following the outbreak onset that are in the 19B clade (blue). The size of the pie is proportional to the residual values of r_0 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.

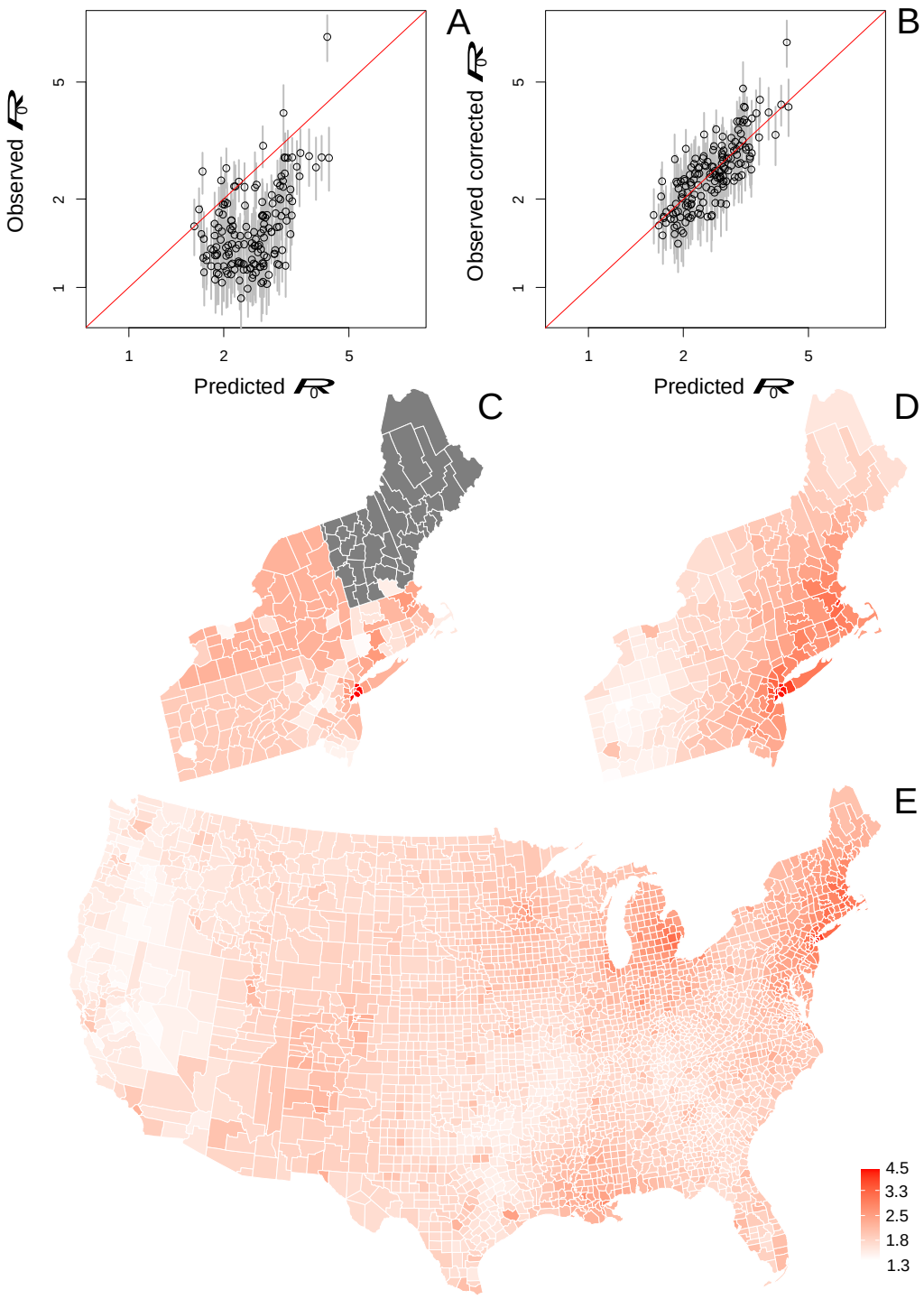

Fig. 4. (A,B) Raw and corrected estimates of R_0 for 160 counties and county-aggregates. The predicted R_0 values are obtained from the regression model, with corrections to standardize values to an outbreak onset of 11 March, 2020, and a population size equal to the most populous county. Comparing the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

all counties. **(C,D)** To illustrate the prediction process for the northeastern states, the raw estimates (C) are all the same for county-aggregates and could not be made for some states (gray). In contrast, the predictability R_0 in the regression model allows for better estimates (D). This makes it possible to extend estimates of R_0 to all 3109 counties in the conterminous USA **(E)**.

Supplementary Materials

Predictable county-level estimates of R_0 for COVID-19 needed for public health planning in the USA

Anthony R. Ives, Claudio Bozzuto.

The Supplementary Materials include:

Materials and Methods,

Figs. S1 to S4,

Tables S1 to S4

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 [1]. 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. For states that contained single counties that satisfied the 100 deaths criterion, while the aggregate of the remainder of counties contained fewer than 100 deaths (DE, NH, NV), the single counties were retained and the aggregated time series dropped in order to maintain the finest-scale spatial resolution in the analyses.

Deciding when to initiate the time series for analysis involves balancing three factors. First, because our goal was to try to estimate R_0 for a "naive" population in which few interventions were taken, pushing the initiation of the time series as early as possible is important. Second, initiating the time series earlier also means that the count data are sparser, increasing uncertainty in the estimates. Third, we wanted the R_0 estimate to reflect conditions within a state and therefore exclude deaths caused by infections contracted elsewhere and brought into the state. To balance these factors, we selected a threshold of one death per day as the starting point of the time series we analyzed. We determined when these thresholds were met using the GAMM in the R package 'mgcv' [2] to smooth the time series (see *3.6. Initial date of the time series*).

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 S3) we obtained from Kirkegaard [3]. We selected socio-economic variables *a priori* in part to represent a broad set of population characteristics,

and in part to represent factors that we anticipated would affect the proportion of the population that would die from COVID-19 without necessarily affecting the transmission rate. Thus, for example, mortality is high for older individuals, and therefore we would anticipate that counties with greater median population age would have higher numbers of deaths. Nonetheless, the increased mortality will not necessarily affect the transmission rate of the disease. Therefore, we used median age as a check on our analyses; if median age were significant, then this would call into question our estimates. Similarly, even though politically right-leaning states might have showed lower responses to COVID-19 in terms of public or private interventions, we would not anticipate differences in very early in the epidemic. Therefore, a check on our estimates is a non-significant effect of the proportion of votes cast for Donald Trump in the 2016 Presidential election.

2. Overview of statistical methods

Here we give an overview of the statistical methods and steps that we took to validate them. In subsequent sections we present the technical details of the methods. Thus, the present section gives a roadmap.

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 [4, 5]. 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 [6]. Therefore, we use bootstrapping [7] 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 [8]. 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 [9].

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 from the forward and reverse time series. For the scenario of step changes in $R(t)$ (Fig. S1 C), the

estimates were unbiased and had accurate confidence intervals, although for the scenario of gradual changes (Fig. S1 D), there was some downwards bias. Nonetheless, the estimates of initial R_0 captured the order of simulations according to the true R_0. In contrast, fitting the same time series with a commonly used Bayesian model that incorporates the transmission process given in the R package EpiEstim [10] gave estimates that poorly reflect the true (simulated) initial R_0 (Fig. S1 E,F).

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

In order to extrapolate the estimates of R_0 from 160 time series to the remaining counties in the conterminous USA, we *a priori* selected four predictors. We selected population size encompassed by the time series to account for possible downwards bias in sparse datasets. We selected the Julian date of the outbreak onset to factor out public and private responses to COVID-19. We treated these two variables as "nuisance" variables that needed to be removed. We included population density, because it could potentially affect transmission rates. Population size and density were weakly and negatively correlated among the 160 time series (Pearson correlation between log population size and log density = –0.25), and therefore there were no problems with multicollinearity. Finally, the regression model included spatial autocorrelation based on the latitude and longitude of the midpoint of the counties or county aggregates. Because the regression model had residual variance that was only slightly high than the variance of the estimates of r_0 that the regression predicted, the precision of the estimates from the regression for the counties without time series will be on par with the precision of the counties with time series.

3. Time series analysis

3.1 Time series model

The time-varying autoregressive model that we applied to the time series is a variant of the TVIRI (time-varying intrinsic rate of increase) model presented in Bozzuto and Ives [11], which is an implementation of time-varying autoregressive models [e.g., 12, 13, 14] that is 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 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. Thus, the variant of the TVIRI model we used here does not include a density-dependent term that would account for decreases in the proportion of susceptibles in the population.

The general specification of the model is

$$x(t) = r(t-1) + x(t-1) \quad \text{(S1a)}$$

$$r(t) = r(t-1) + \omega_r(t) \quad \text{(S1b)}$$

$$x^*(t) = x(t) + \varepsilon(t) \quad \text{(S1c)}$$

Here, $x(t)$ is the unobserved, log-transformed value of deaths at time t, and $x^*(t)$ is the observed count that depends on the observation uncertainty described by the random variable $\varepsilon(t)$. Because a few of the datasets that we analyzed had zeros, we replaced zeros with 0.5 before log-transformation; other ways of treating zeros in the count data gave very similar results. The model assumes that $x(t)$ increases exponentially at rate $r(t)$, where the latent state variable $r(t)$ changes through time as a random walk with $\omega_r(t) \sim N(0, \sigma^2_r)$. This assumption allows $r(t)$ to change through time as dictated by the data, and the estimate of σ^2_r sets the rate at which $r(t)$ can change from one day to the next.

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. On the log-transformed scale of $x^*(t)$, this implies that $\varepsilon(t)$ has mean zero and variance approximately $\sigma^2_\varepsilon + \exp(-x(t))$, where σ^2_ε scales the variance.

We fit the model using the Kalman filter to compute the maximum likelihood [15, 16]. In addition to the parameters σ^2_r, and $\sigma^2_{\square}$, 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; this is a common feature of time-varying autoregressive models. When this occurs, the estimate of r_0 is the average over the entire time series and is thus biased downwards. To eliminate this absorption to zero, we imposed a minimum of 0.02 on σ^2_r, which eliminated the problem in the simulations.

3.2 Applying the model to epidemiological data

The time-varying autoregressive model estimates the rate of spread of the disease, $r(t)$, from the count of deaths observed each day, $x^*(t)$. Any value of $x^*(t)$ reflects the number of people infected over multiple days in the past, and the proportion that is counted as a result of death on day t. If the disease had been spreading exponentially at constant rate for many days, and if the number of infected people was large, then the increase from $x^*(t–1)$ to $x^*(t)$ would approach a constant value; in other words, $r(t)$ would give the exponential rate of spread of the disease. This would be true even if only a small fraction of the infected population died or was diagnosed, provided these fractions did not change through time. However, changes in the infection rate will mean that the disease is not at its "stable infection age distribution", the distribution of time since infection observed in the infected population [17]. While this does not affect the statistical model fitting, it will mean that the observed spread of the disease is not exactly equal to the rate of new infections. Nonetheless, because the distribution of times between infection and counting (deaths) is fairly broad, the assumption that populations are at their "stable infection age distribution" is unlikely to cause a great difference between the observed rate of disease spread and the infection rate. This is addressed in detail in the simulation study (*3.3 Parametric bootstrapping*).

The "true" value of the number of daily deaths in the model, $x(t)$ (S1a), is the probability that a death is counted. After accounting for measurement error (S1c), all of the variation in $x(t)$ is assumed to be given by variation in the spread rate $r(t)$ (S1b). Therefore, the variation $\omega_r(t)$ in

$r(t)$ includes both the day-to-day variation in the spread rate and the longer-term changes in $r(t)$ that results when estimates of $\omega_r(t)$ have a mean different from zero. The assumption that $r(t)$ is a random walk gives it flexibility to track the patterns in the data as the model is fit. We suspect that the true changes in the infection rate do not vary greatly on a day-to-day basis. This might argue for fitting a smoothing curve to $r(t)$ or $x(t)$. Nonetheless, we found that results from curve-fitting models were sensitive to decisions made about the type of curves that were fit. The time-varying autoregressive model was less dependent on *a priori* assumptions, due to few *a priori* assumptions about the data. Further, the bootstrapping method we applied to obtain estimates of the uncertainty of the model fits also acts as a smoothing method.

3.3 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. This procedure answers the question: If it were possible to observe many time series generated by the same process, how variable would be the results of the statistical model fit? This bootstrapping approach requires assumptions about the process underlying the true data. Because the underlying changes in $r(t)$ are of interest, the bootstrap incorporates the time-varying changes in the estimated values of $r(t)$ from the fitted data.

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 estimates of $r(t)$, $\Delta r(t) = r(t) - r(t\text{-}1)$, and then standardizing to remove the bias, $\Delta r_s(t) = \Delta r(t) - E[\Delta r(t)]$. The sequence $\Delta r_s(t)$ was fit using a autoregressive time-series model with time lag 1, AR(1), to preserve any shorter-term autocorrelation in the data. For the bootstrap a new time series was simulated from this AR(1) model, $\Delta\rho(t)$, and then standardized, $\Delta\rho_s(t) = \Delta\rho(t) - E[\Delta\rho(t)]$. The simulated time series for the spread rate was constructed as $\rho(t) = r(t) + \Delta\rho_s(t)/2^{1/2}$, where dividing by $2^{1/2}$ accounts for the fact that $\Delta\rho_s(t)$ was calculated from the difference between successive values of $r(t)$. A new time series of count data, $\xi(t)$, was then generated using equation (S1a) with the parameters from fitting the data. Finally, the statistical model was fit to

the reconstructed $\xi(t)$. In this refitting, we fixed the variance in $r(t)$, σ^2_r, to the same value as estimated from the data. Therefore, the bootstrap confidence intervals are conditional of the estimate of σ^2_r. We imposed this condition, because the estimate of σ^2_r tends to absorb to zero when the change in $r(t)$ is small, with variation in $x(t)$ transferred to the measurement error variance.

3.4. Calculating R_0

We derived estimates of $R(t)$ directly from $r(t)$ as *(15)*

$$R(t) = 1/\sum_{\tau} e^{-r(t)\tau} p(\tau) \qquad \text{(S2)}$$

where $p(\tau)$ is the distribution of the proportion of secondary infections caused when by a primary infection that occurred τ days previously. We used the distribution of $p(\tau)$ from Li et al. [9] 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. We report values of $R(t)$ at dates that are offset by the average length of time between initial infection and death, which is taken as 18 days [18].

3.5. Simulation for validation

To assess the robustness of the statistical model, we built a simulation SIR (susceptible-infected-recovered) model of a hypothetical epidemic. The simulation model was not the same as the statistical model, so the goal was to determine whether the phenomenological statistical model was capable of capturing the rate of infection spread in the process-based simulations.

The simulation model tracks the number of infected individuals on day t who were infected τ days previously, $X(t;\tau)$. After 25 days, they are all assumed to be recovered or dead. The probability distribution of the day on which a susceptible is infected, $p(t)$, is given by a Weibull distribution with mean 7.5 days and standard deviation 3.4 *(23)* (Fig. S3 A). For an individual who dies, the day of death, $d(t)$, is given by a Weibull distribution with mean 18.5 days and standard deviation 3.4 [9] (Fig. S3 B). Finally, for case data we need to know the time between initial infection and diagnosis, $h(t)$, which we assume is lognormally distributed with

mean 5.5 days and standard deviation 2.2 [19] (Fig. S3 C).

On day t, the number of new infections produced by individuals who were infected τ days earlier is $b(t)\, p(\tau)$. The term $b(t)$ is closely related to $R(t)$, the number of secondary infections caused per infection. However, because we allow $b(t)$ to fluctuate on a daily basis, here we use a notation that differs from $R(t)$. Note, however, that on average $R(t) = \sum_\tau b(t + \tau)\, p(\tau)$. The total number of new infections on day t is given by a lognormal Poisson distribution in which the mean of the Poisson process is $b(t)\, \alpha(t) \sum_\tau p(\tau) X(t; \tau)$, where the lognormal random variable $\alpha(t)$ is included to represent environmental variation.

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

Cases are reported according to a binomial distribution for each infection age category $X(t; \tau)$, so that the probability of a person with the infection for τ days being diagnosed is $G\, \chi(t)\, h(\tau)$, where G is the overall probability that a case is reported, and $\chi(t)$ is a logit-normal distribution to represent daily variation in reporting. We assume that the overall reporting probability is $G = 0.5$.

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

We fit this simulated dataset using the same procedure as we used for the real data, 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 [10]. 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).

3.6. 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 daily mortality rate exceeded 1. To estimate the daily mortality rate, 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 [2].

4. Regression analysis for r_0

We fit the estimates of r_0 from the 160 county and county-aggregate time series with the Generalized Least Squares (GLS) model (S3)

$$r_0 = b_0 + b_1\, start.date + b_2 \log(pop.size) + b_3\, pop.den^{0.25} + \varepsilon \qquad \text{(S3)}$$

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

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 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. 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 [20].

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

$$\hat{e}_i = \bar{e} + \boldsymbol{v_i} * \mathbf{V}^{-1}(\varepsilon_t - \bar{e}) \qquad \text{(S4)}$$

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, **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 [21]. 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 [22]. 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. For these predicted values, for each row i of the data matrix, we removed row i and column i from $\mathbf{\Sigma}$ to give $\mathbf{V}$, and we set $\boldsymbol{v_i}$ as row i from $\mathbf{\Sigma}$ to give $\hat{e}_i$. Then the predicted value of r_0, $r_{0[i]}$, is

$$r_{0[i]} = b_0 + b_1\, min.start.date + b_2 \log(max.pop.size) + b_3\, pop.den_{[i]}^{0.25} + \hat{e}_i \qquad \text{(S5)}$$

where *min.start.date* is the earliest start date of any time series, and *max.pop.size* is the maximum population size among counties. Third, we used equation (S4) similarly to obtain predicted values of r_0, and hence predicted R_0, for all other counties. For this, because there is no variance in the predicted values of $r_{0[i]}$ from the time-series analysis, we refit the regression model without the weighting term **s** and applied equation (S4) with **V** as the correlation matrix. For individual counties that were within county-aggregates, we assumed that the spatial distance between individual county and the county-aggregate was the average distance between counties within the aggregate. We also calculated the variance of the estimates from [21]

$$\hat{v}_i = \sigma^2 - \boldsymbol{v_i} * \mathbf{V}^{-1} * \boldsymbol{v_i}^{\mathbf{t}} \qquad \text{(S6)}$$

Predicted values of R_0 were mapped using the R package usmap [23].

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

We downloaded metadata for SARS-CoV-2 from [24] that identified samples to state and gave the time of collection and clade (19A, 19B, 20A, 20B, and 20C). For each state, we selected the estimates of r_0 from the county or county-aggregate representing the greatest number of deaths. We pooled SARS-CoV-2 samples that were collected no more than 30 following the onset of outbreak that we used as the starting point for the time series from which we estimated r_0. Twenty-seven states had five or more samples meeting this criterion. We fit the estimates of r_0 from the 27 states with the weighted Least Squares (LS) model

$$r_0 = b_0 + b_1\, start.date + b_2 \log(pop.size) + b_3\, pop.den^{0.25} + prop.19A + prop.19B + prop.20A + prop.20B + \varepsilon \qquad \text{(S7)}$$

$$\varepsilon = N(0, \sigma^2)$$

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, prop.XXX are the proportions of strains in each of four clades, and ε is a vector of independent Gaussian random variables (Table S2). We also performed the regression with spatial autocorrelation, but this model did not fit the data better than the model without spatial autocorrelation. We also performed the analysis first using a PCA to collapse the genomic clade into axes explaining variation in the proportion of clades among locations, and the results of this analysis were consistent with those from equation S7.

For figure 3, we obtained residuals from the regression model as in equation S7 but without the variables for strains. We then scaled the pie charts according to the residuals. Figure 3 was constructed using the R packages usmap [23] and scatterpie [25] .

Additional figures and tables

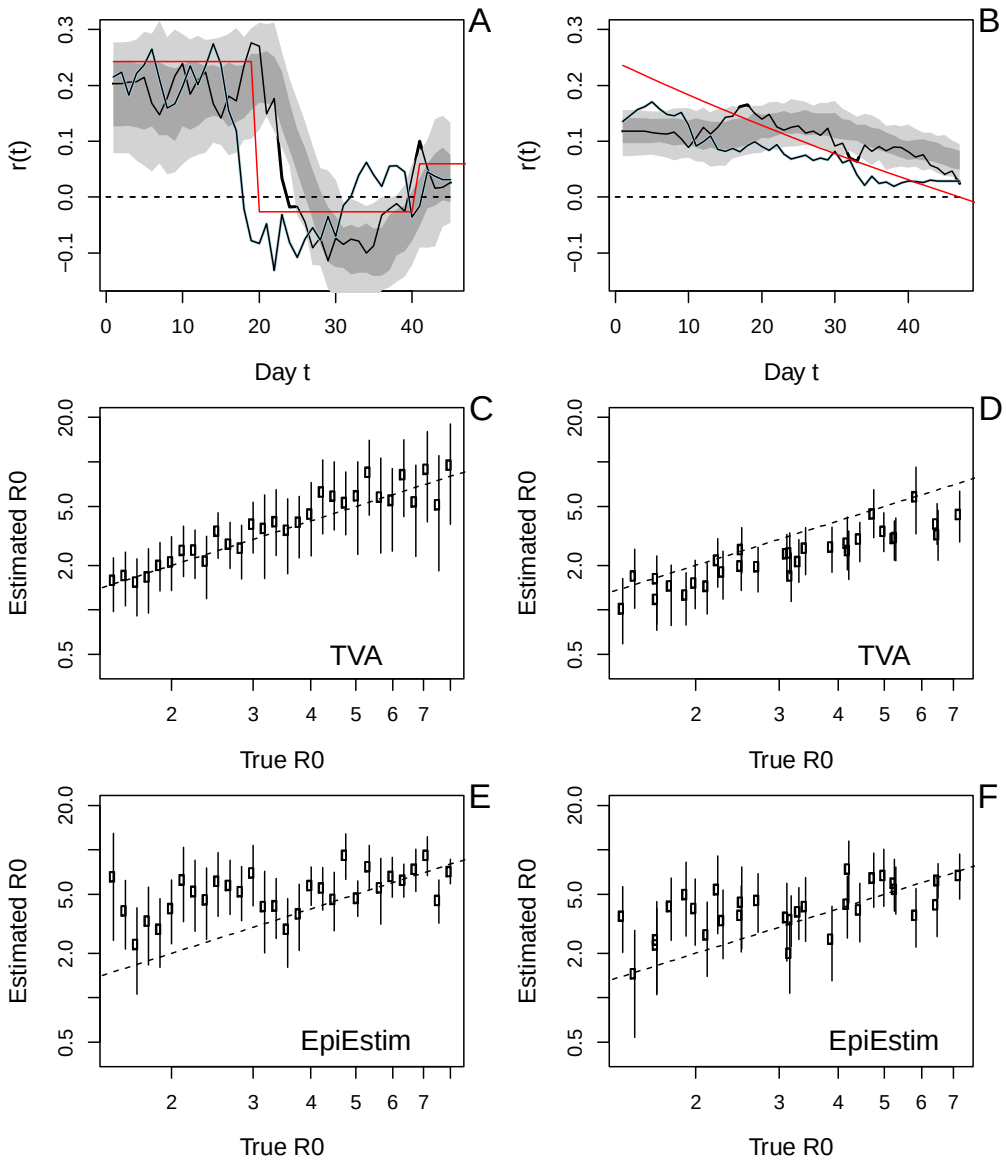

Fig. S1. Simulation study of fitting methods to epidemic death data. Simulations were fit with the time-varying autoregression model (TVA) in the forward (black line with dark and light gray regions giving 66% and 95% approximate confidence intervals) and reverse (blue line and regions) directions when 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 estimate of R0 with 95% confidence intervals, which are plotted against the true values of R0 for step **(C)** and gradual **(D)** changes in R(t). The same simulations with fit using EpiEstim **(E,F)**.

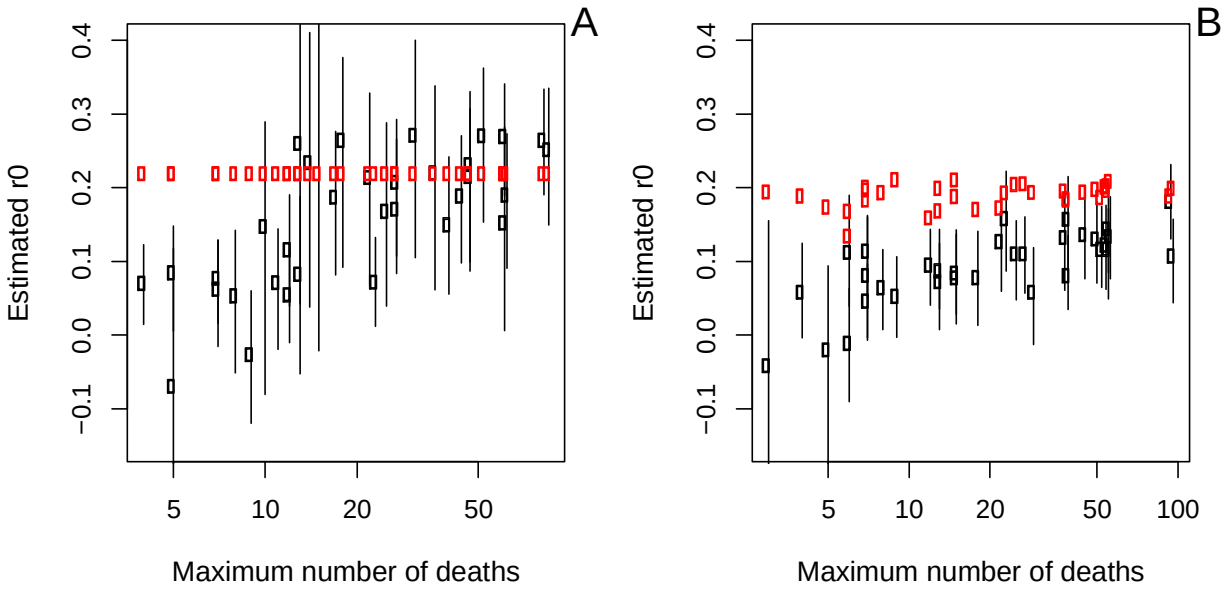

Fig. S2. Simulation study of the estimation of r_0 from the forward and reverse time-varying autoregressive model for different population sizes. Simulations following those used for Fig. 1 were performed assuming $r(t)$ changed either **(A)** in steps or **(B)** gradually. The simulations were performed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

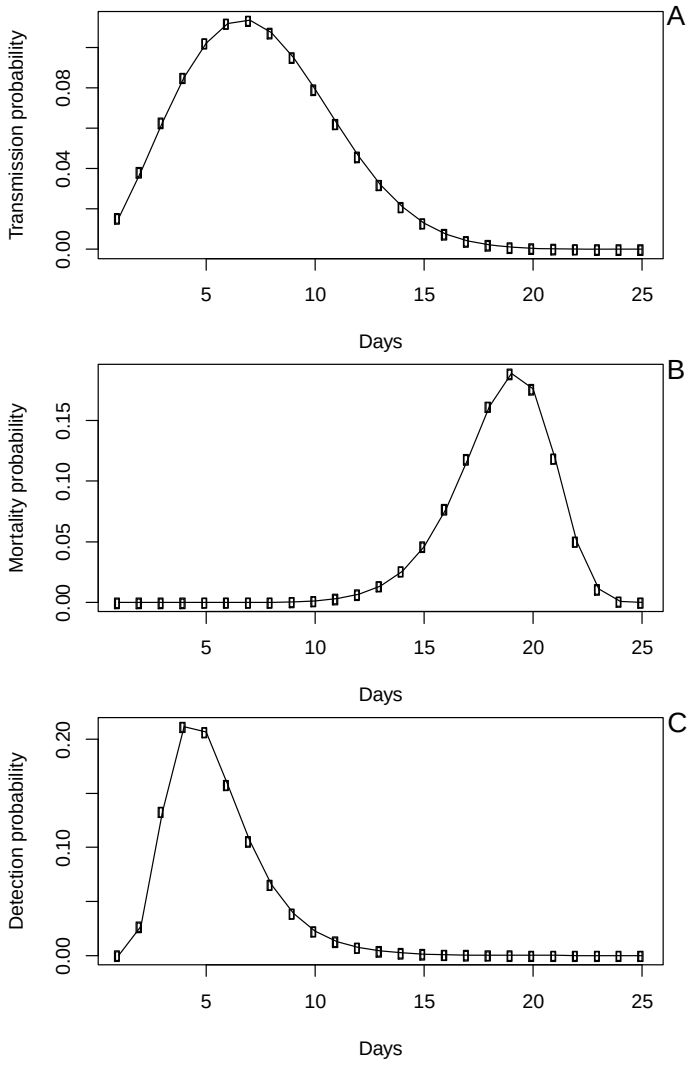

Fig. S3. Probability distributions used in the process-based SIR simulation model used to test methods for robustness. **(A)** The probability distribution of the day on which a susceptible is infected, $p(t)$, which is given by a Weibull distribution with mean 7.5 days and standard deviation 3.4. **(B)**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

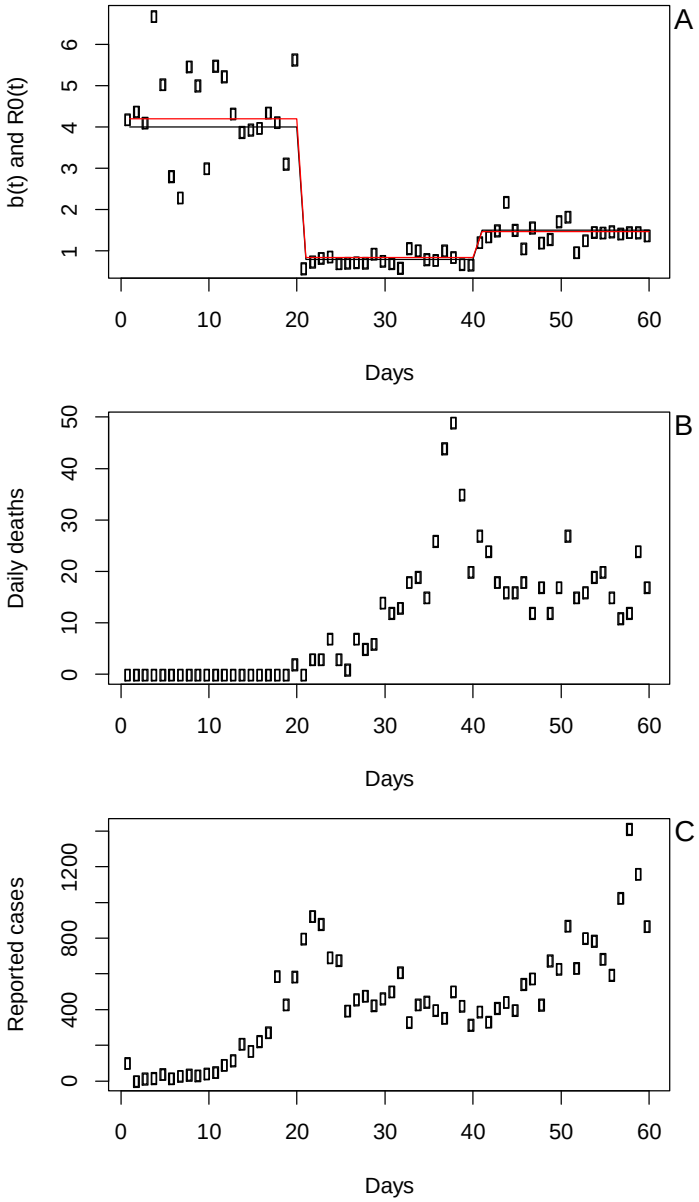

Fig. S4. Example simulation from the process-based SIR model. **(A)** Changes in the infection rate, $b(t)$, are modeled as a step function (black line) with daily variation (points). R0(t) (red line) tracks changes in $b(t)$. **(B)** and **(C)** The number of deaths (B) and diagnosed cases (C) when the simulation is initiated with a single cohort of individuals, all infected on day 1 (solid black dot).

Table S1. Separate spreadsheet giving the following variables for the 3109 counties in the conterminous USA.

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 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.

	Value	**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

Table S3. Variables giving population characteristics that were including in the regression model (equation S3). No variable was statistically significant. Data from [3, 26].

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

Table S4. For 160 county and county-aggregates, regression of spread rate at the end of the time series, corresponding to 5 May, 2020, $r(t_{end})$, 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. For the overall model, $R^2_{pred} = 0.38$.

	Value	**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		$\chi^2_1 = 10.3$	0.0056	0.099

References Supplementary Materials

1. New York Times. Coronavirus (Covid-19) data in the United States. 2020. Available from: https://github.com/nytimes/covid-19-data.

2. Wood SN. Generalized additive models: an introduction with R. Boca Raton: CRC Press, Chapman and Hall; 2017.

3. Kirkegaard EOW. Inequality across US counties: an S factor analysis. Open Quantitative Sociology and Political Science [Internet]. 2016. doi: 10.26775/OQSPS.2016.05.23.

4. Flaxman S, et al. State-level tracking of COVID-19 in the United States. Report 23, Imperial College London [Internet]. 2020 2020. doi: https://doi.org/10.25561/79231.

5. Scire J, Nadeau S, Vaughan T, Brupbacher G, Fuchs S, Sommer J, et al. Reproductive number of the COVID-19 epidemic in Switzerland with a focus on the Cantons of Basel-Stadt and Basel-Landschaft. Swiss Medical Weekly [Internet]. 2020; 150(1920). doi: smw.ch/article/doi/smw.2020.20271.

6. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. New York, NY: Cambridge University Press; 2007. 625 p.

7. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall; 1993.

8. Dublin LI, Lotka AJ. On the true rate of natural increase. Journal of the American Statistical Association. 1925;20:305–39.

9. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine. 2020;382(13):1199-207. doi: 10.1056/NEJMoa2001316.

10. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. American Journal of Epidemiology. 2013;178(9):1505-12. doi: 10.1093/aje/kwt133.

11. Bozzuto C, Ives AR. Inbreeding depression and the detection of changes in the intrinsic rate of increase from time series. 2020. doi: 10.13140/RG.2.2.21603.81447.

12. Ives AR, Dakos V. Detecting dynamical changes in nonlinear time series using locally linear state-space models. Ecosphere. 2012;3(6):art58. doi: http://dx.doi.org/10.1890/ES11-00347.1.

13. Bragina EV, Ives AR, Pidgeon AM, Balčiauskas L, Csányi C, Khoyetskyy P, et al. Wildlife population changes across Eastern Europe after the collapse of socialism. Fontiers in Ecology and Evolution. 2018;16:77-81.

14. Zeng Z, Nowierski RM, Taper ML, Dennis B, Kemp WP. Complex population dynamics in the real world: Modeling the influence of time-varying parameters and time lags. Ecology. 1998;79(6):2193-209.

15. Durbin J, Koopman SJ. Time Series Analysis by State Space Methods. 2nd ed. Oxford, UK: Oxford University Press; 2012.

16. Harvey AC. Forecasting, structural time series models and the Kalman filter. Cambridge, U.K.: Cambridge University Press; 1989.

17. Caswell H. Matrix Population Models. Inc., Sunderland, Massachusetts: Sinauer Associates; 1989.

18. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet. 2020;395(10229):1054-62. doi: 10.1016/S0140-6736(20)30566-3.

19. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science. 2020;368(6491):eabb6936. doi: 10.1126/science.abb6936.

20. Pinheiro J, Bates D, DebRoy S, Sarkar D, Team RC. nlme: Linear and lonlinear mixed effects models. R package version 3.1-147. 2020. Available from: https://CRAN.R-project.org/package=nlme>.

21. Petersen KB, Pedersen MS. The matrix cookbook: Technical University of Denmark; 2012. Available from: http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf.

22. Ives AR. R2s for correlated data: phylogenetic models, LMMs, and GLMMs. Systematic Biology. 2019;68(2):234-51. doi: 10.1093/sysbio/syy060. PubMed PMID: WOS:000462830500004.

23. Di Lorenzo P. usmap: US Maps Including Alaska and Hawaii. R package version 0.5.0.9999. 2020. Available from: https://usmap.dev.

24. NextstrainTeam. Nextstrain. 2020. Available from: https://nextstrain.org/ncov.

25. Yu G. scatterpie, R package version 0.1.4. 2019. Available from: https://CRAN.R-project.org/package=scatterpie.

26. United States Census Bureau. USA Counties. 2011. Available from: https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html.