

## RESEARCH ARTICLE

# Reconstructing the course of the COVID-19 epidemic over 2020 for US states and counties: Results of a Bayesian evidence synthesis model

Melanie H. Chitwood<sup>1</sup>\*, Marcus Russi<sup>1</sup>\*, Kenneth Gunasekera<sup>1</sup>, Joshua Havumaki<sup>1</sup>, Fayette Klaassen<sup>2</sup>, Virginia E. Pitzer<sup>1</sup>, Joshua A. Salomon<sup>3</sup>, Nicole A. Swartwood<sup>2</sup>, Joshua L. Warren<sup>4</sup>, Daniel M. Weinberger<sup>1</sup>, Ted Cohen<sup>1</sup>‡, Nicolas A. Menzies<sup>2</sup>‡

**1** Department of Epidemiology of Microbial Diseases and Public Health Modeling Unit, Yale School of Public Health, Yale University, New Haven, Connecticut United States of America, **2** Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts United States of America, **3** Department of Health Policy, Stanford University, Stanford, California United States of America, **4** Department of Biostatistics and Public Health Modeling Unit, Yale School of Public Health, Yale University, New Haven, Connecticut United States of America

\* These authors contributed equally to this work.

‡ These authors are joint senior authors on this work.

\* [melanie.chitwood@yale.edu](mailto:melanie.chitwood@yale.edu)



## OPEN ACCESS

**Citation:** Chitwood MH, Russi M, Gunasekera K, Havumaki J, Klaassen F, Pitzer VE, et al. (2022) Reconstructing the course of the COVID-19 epidemic over 2020 for US states and counties: Results of a Bayesian evidence synthesis model. *PLoS Comput Biol* 18(8): e1010465. <https://doi.org/10.1371/journal.pcbi.1010465>

**Editor:** Benjamin Althouse, University of Washington, UNITED STATES

**Received:** July 22, 2021

**Accepted:** August 3, 2022

**Published:** August 30, 2022

**Copyright:** © 2022 Chitwood et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All data used in the main analysis are available for use at <https://github.com/covidestim/covidestim-sources>. The covidestim package is available for download at <https://github.com/covidestim/covidestim>.

**Funding:** KG reports grant from National Institutes of General Medical Sciences T32 GM007205 and Fogarty International Center D43 TW010540. VEP, JLW, and DMW report grants from National Institute of Allergy and Infectious Diseases R01

## Abstract

Reported COVID-19 cases and deaths provide a delayed and incomplete picture of SARS-CoV-2 infections in the United States (US). Accurate estimates of both the timing and magnitude of infections are needed to characterize viral transmission dynamics and better understand COVID-19 disease burden. We estimated time trends in SARS-CoV-2 transmission and other COVID-19 outcomes for every county in the US, from the first reported COVID-19 case in January 13, 2020 through January 1, 2021. To do so we employed a Bayesian modeling approach that explicitly accounts for reporting delays and variation in case ascertainment, and generates daily estimates of incident SARS-CoV-2 infections on the basis of reported COVID-19 cases and deaths. The model is freely available as the *covidestim* R package. Nationally, we estimated there had been 49 million symptomatic COVID-19 cases and 404,214 COVID-19 deaths by the end of 2020, and that 28% of the US population had been infected. There was county-level variability in the timing and magnitude of incidence, with local epidemiological trends differing substantially from state or regional averages, leading to large differences in the estimated proportion of the population infected by the end of 2020. Our estimates of true COVID-19 related deaths are consistent with independent estimates of excess mortality, and our estimated trends in cumulative incidence of SARS-CoV-2 infection are consistent with trends in seroprevalence estimates from available antibody testing studies. Reconstructing the underlying incidence of SARS-CoV-2 infections across US counties allows for a more granular understanding of disease trends and the potential impact of epidemiological drivers.

AI137093. TC reports grants from National Institute of Allergy and Infectious Diseases R01 AI112438. NAM reports grants from National Institute of Allergy and Infectious Diseases R01 AI146555-01A1. JAS reports funding from the Centers for Disease Control and Prevention through the Council of State and Territorial Epidemiologists (NU380T000297-02) and the National Institute on Drug Abuse (3R37DA01561217S1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** I have read the journal's policy and the authors of this manuscript have the following competing interests: DMW has received consulting fees from Pfizer, Merck, GSK, and Affinivax for topics unrelated to this manuscript and is Principal Investigator on a research grant from Pfizer on an unrelated topic. VEP has received reimbursement from Merck and Pfizer for travel expenses to Scientific Input Engagements unrelated to the topic of this manuscript. JLV has received consulting fees from Revelar Biotherapeutics for topics unrelated to this manuscript. All other authors have declared that no competing interest exist.

## Author summary

Because many COVID-19 infections go undetected, reported numbers of COVID-19 cases and deaths underestimate the true size of the epidemic. To address this problem, we built a model to estimate the number of new SARS-CoV-2 infections over time in each U.S. state and county. In this paper, we present time trends of infections and other disease outcomes from the first reported case in the U.S. until January 1, 2021, for each state and county. The time series of infection estimates suggest that the US epidemic is best described a series of related epidemics, varying in their timing and intensity. We estimate that over a quarter of the US population was infected with SARS-CoV-2 in 2020 and by the end of 2020 0.12% of the population had died from COVID-19. State-level results were consistent with external measures of disease burden, including estimates of SARS-CoV-2 seroprevalence and excess mortality. Our findings help better understand the epidemic in the pre-vaccine era and demonstrate the feasibility of estimating SARS-CoV-2 infections at local levels using routinely reported case and death data.

## Introduction

The numbers of newly diagnosed cases and confirmed COVID-19 deaths are the most easily observed measures of the health burden associated with COVID-19 and have been widely used to track the trajectory of the epidemic at the national, state, and local level.[1, 2] However, there are at least three limitations of using reported cases and deaths for this purpose. First, testing is primarily organized to identify symptomatic individuals, but a large fraction of SARS-CoV-2 infections are asymptomatic, [3] leading to case counts that are substantially smaller than the true incidence of infection. Second, the degree to which case counts undercount infections is sensitive to the availability and utilization of diagnostic testing, which has varied over time and geography.[4–6] For this reason, it can be difficult to distinguish true trends from changes in testing practices. Third, case and death counts are lagging indicators of the transmission dynamics of the pathogen, as they are affected by delays associated with the incubation period, care-seeking behavior of symptomatic individuals, diagnostic processing times, and reporting practices. Taken together, these limitations present challenges to analyses that rely on these metrics as primary signals of SARS-CoV-2 spread.

A better indicator of changes in local transmission is the effective reproduction number ( $R_t$ ), which represents the average number of secondary infections caused by an individual infected at some time  $t$ . [7]  $R_t$  can signal short-term changes in transmission in response to policy and behavioral changes. However,  $R_t$  is not a directly observable quantity and estimates of  $R_t$  based on raw case reports become biased when reporting delays are incorrectly estimated, [5] weakening their usefulness as a measure of transmission.

Unbiased estimates of COVID-19 cases and the  $R_t$  of SARS-CoV-2 can provide more accurate insight into the size and scope of the United States (US) epidemic and inform current and future COVID-19 control policies. A number of modeling approaches have been developed to reconstruct the time series of infections and deaths over the course of the US epidemic. These approaches typically do not allow for variability in case ascertainment and infection fatality ratios (IFRs) across space and time, nor do they attempt to model SARS-CoV-2 infections or COVID-19 deaths at fine spatial scales, such as at the county level.[8, 9]

Here, we present detailed estimates of viral dynamics for all US states and counties, based on a Bayesian statistical model that combines multiple data sources to estimate SARS-CoV-2

infection patterns from observed case notifications and death reports. We apply our model to publicly available COVID-19 case and death data and report on the trajectory of the epidemic from the first reported case (January 13, 2020) until January 1, 2021. The model is available on GitHub (<https://github.com/covidestim/covidestim/>) as a package for the R programming language (*covidestim*).

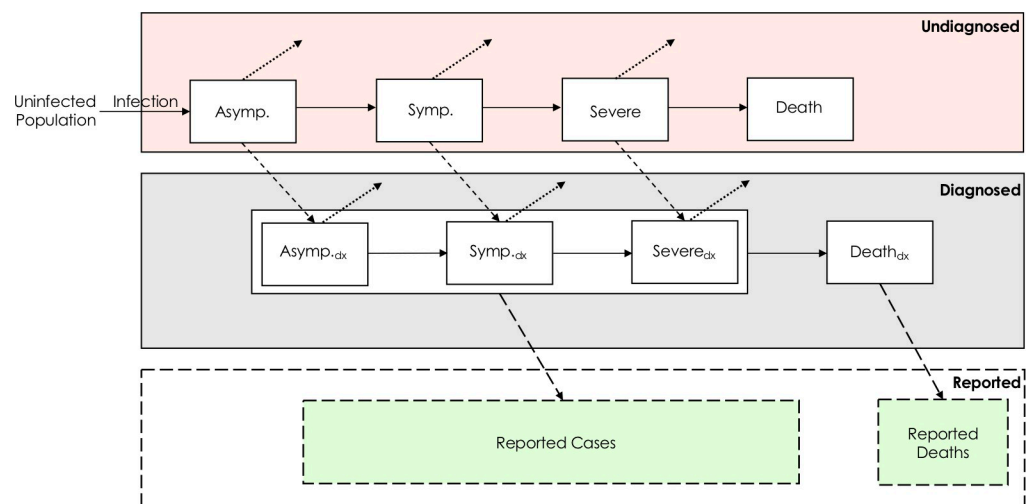
## Results

### Analytic overview

We developed a mechanistic model to back-calculate SARS-CoV-2 infections and subsequent outcomes based on reported COVID-19 cases and deaths. In this model the natural history of COVID-19 is represented using four health states: asymptomatic or pre-symptomatic SARS-CoV-2 infection (*Asymptomatic*), symptomatic but not severe COVID-19 disease (*Symptomatic*), severe COVID-19 disease (*Severe*), and death from COVID-19 (*Death*). In each health state (except *Death*) individuals either recover or transition to a more severe state after some delay. Infected individuals can be diagnosed in the *Asymptomatic*, *Symptomatic*, or *Severe* states, and we assume all diagnosed cases and all deaths among diagnosed individuals are reported after a short delay. Fig 1 shows modeled health states and transitions. The model generates several outcomes of epidemiological importance, including  $R_t$ , total infections, symptomatic cases, total deaths, and case ascertainment; we estimated these outcomes for each US state and county from the start of the epidemic until January 1, 2021.

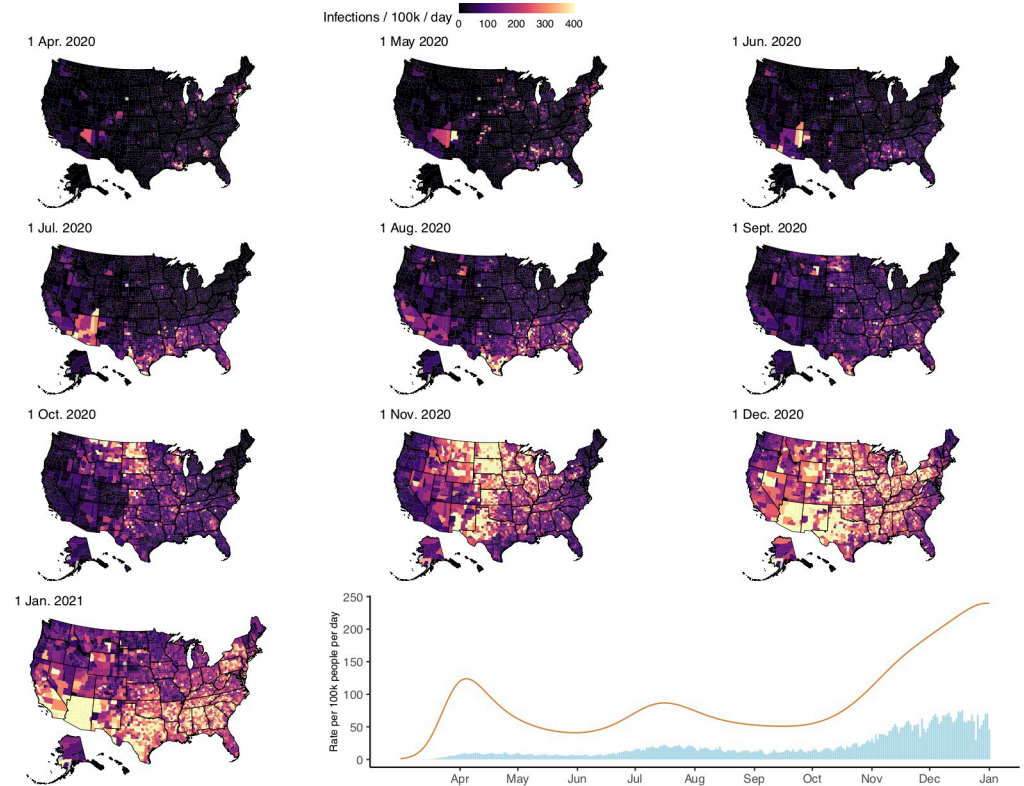
### Main findings

**Incidence and  $R_t$ .** The SARS-CoV-2 epidemic in the US consisted of a series of related outbreaks, which varied greatly in both the intensity of transmission and the extent of geographic spread (Fig 2). The March outbreak in New Jersey was the largest per population in a single state; on April 4, we estimate that New Jersey had 732 (95% credible interval: 464, 1206) infections per 100,000, and 16% (10%, 26%) of all infections in the US on that day. Local surges



**Fig 1. A model schematic of the main health states: *Asymptomatic* (denoted “Asymp.”), *Symptomatic* (denoted “Symp.”), *Severe*, and *Death*.** The subscript “dx” indicates that individuals in that state have received a diagnosis of COVID-19. Each transition (denoted with an arrow) has an associated probability and delay distribution. Solid arrows denote disease progression; dotted arrows denote recovery; short dashed arrows denote diagnosis; long dashed arrows denote reporting. All diagnosed cases and deaths are assumed to be reported after a given delay.

<https://doi.org/10.1371/journal.pcbi.1010465.g001>



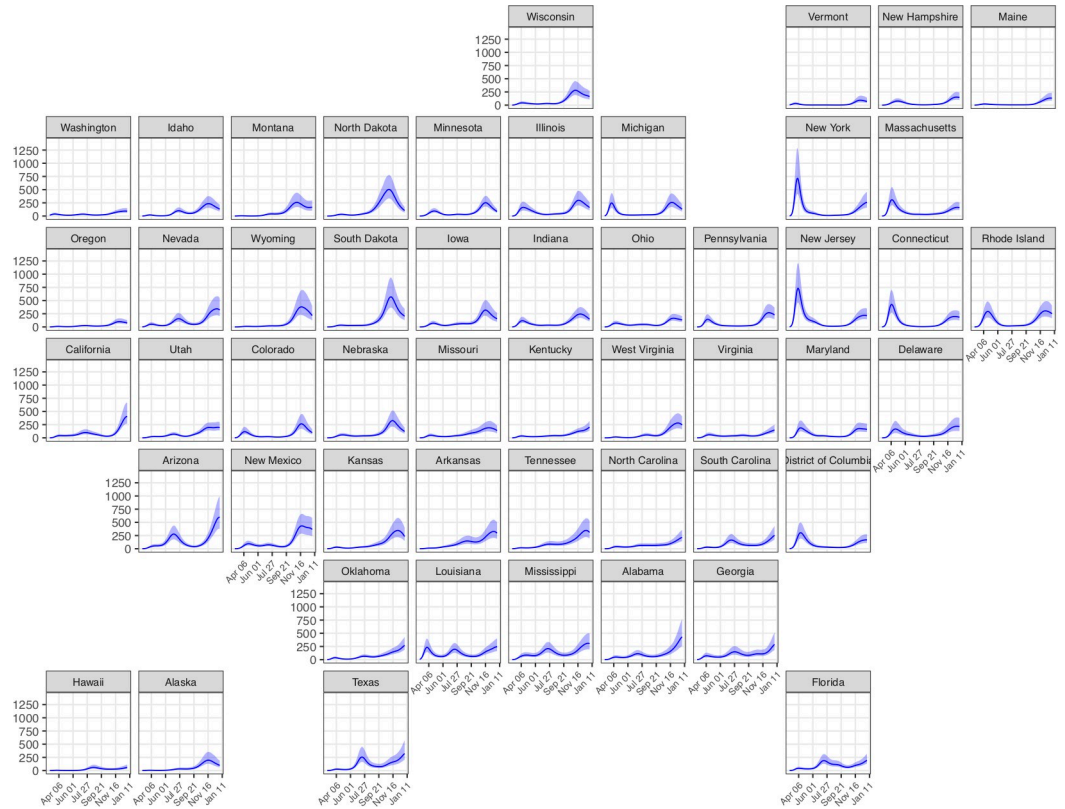
**Fig 2. Panels 1–10: County-level infections per 100,000 population per day at 10 timepoints between April 1, 2020 and January 1, 2021. Panel 11: Time series of national SARS-CoV-2 infection estimates (orange line) and reported COVID-19 diagnoses (blue bars) per 100,000 people per day from March 1, 2020 to January 1, 2021. Maps generated using shapefiles from the *albersusa* package for the R programming language: <https://github.com/hrbrmstr/albersusa>. <https://doi.org/10.1371/journal.pcbi.1010465.g002>**

in infections during the fall and winter of 2020 rivaled New Jersey and New York's spring outbreaks in scale, but occurred in the context of a more generalized US epidemic. South Dakota, for example, had its highest per capita infections of 2020 on November 8 (569 [365, 940] infections per 100,000), but accounted for just 1.2% (0.7%, 1.9%) of all US infections that day. Forty-five states experienced the highest daily infections per capita in November or December (Fig 3). Model fit to data can be found in S1 Fig.

While most states and counties had lower levels of transmission during the summer months, few achieved established thresholds of low levels of community transmission, defined as fewer than 20 confirmed cases per 100,000 per week. [10] We estimate that only four states (Alaska, Montana, Vermont, and West Virginia) had fewer than 20 symptomatic cases per 100,000 inhabitants per week after transmission was established locally. Notably, Vermont remained below this threshold from the week of May 11 until the week of September 28.

Estimates of  $R_t$  at the start of the epidemic varied greatly by state. The median state-level estimate of  $R_t$  on the first day a case was reported in each state was 3.4 (range: 1.7 [1.5–2.0] in Washington to 5.9 [4.3–8.2] in New York). Throughout April,  $R_t$  estimates dropped substantially. Over the period May 1, 2020 to January 1, 2021, state-level estimates of  $R_t$  ranged from 0.7 (0.6, 0.8) to 1.5 (1.3, 1.7) (Fig 4).

**Percent Ever-Infected with SARS-CoV-2.** For each county, we calculated the percentage of the population ever-infected as the sum of all estimated infections divided by county population on January 1, 2021 (Fig 5). This cumulative infection estimate is distinct from reported seroprevalence estimates, as seroprevalence measures may be affected by the lower immune



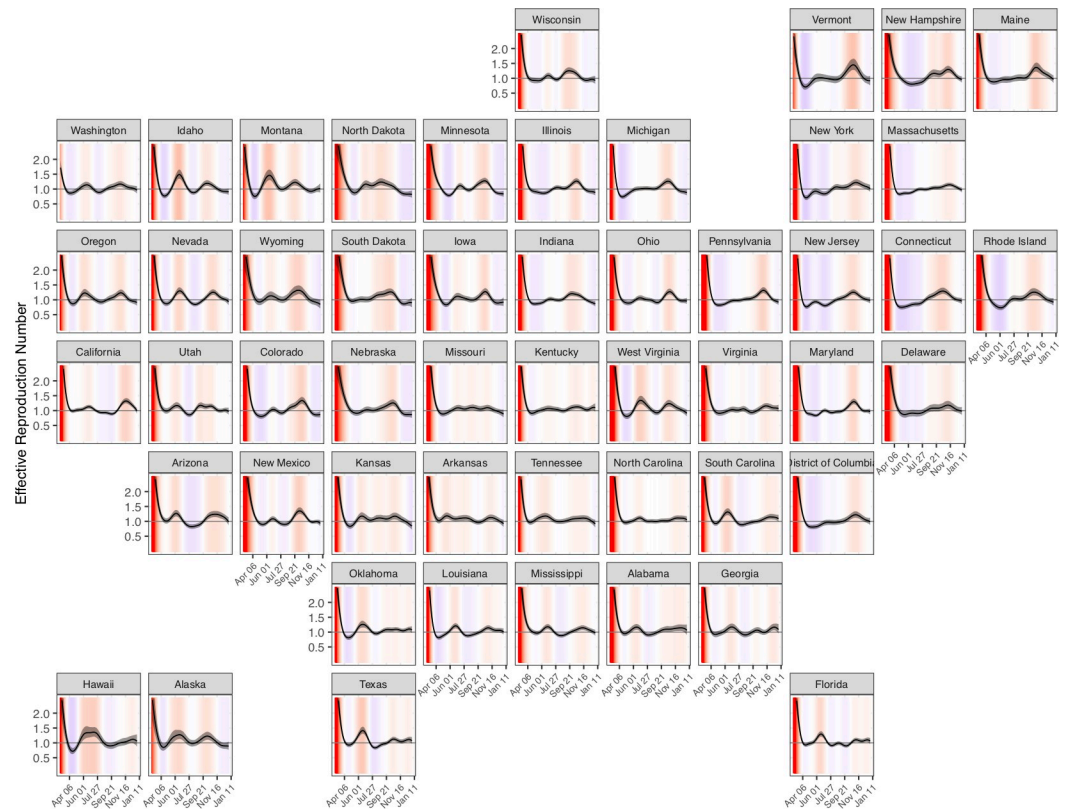
**Fig 3. Incident infections per 100,000 residents per day for each US state from March 1, 2020 to January 1, 2021.** Shaded areas represent 95% credible intervals.

<https://doi.org/10.1371/journal.pcbi.1010465.g003>

response among individuals with mild/asymptomatic infection, possible waning of antibody titers,[11, 12] and non-representativeness of sampled populations.[13] By January 1 2021, we found that the percent of the population ever-infected exceeded 50% in 303 (9.7%) counties and exceeded two-thirds of the population in 42 (1.3%) counties. Conversely, the percent ever-infected was less than 10% in 144 (4.6%) counties and less than 5% in 37 (1.2%) counties. Based on the sum of state estimates (posterior medians), we estimate that 28% of the US population had been infected with SARS-CoV-2 by January 1, 2021. Across states, the percentage ever-infected ranged from 6.5% (4.2%, 11.1%) in Vermont to 45.7% (30.7%, 67.1%) in Arizona (Fig 5).

On January 1, 2021, the US had reported 348,055 cumulative COVID-19 deaths.[14] Based on the sum of state estimates (posterior medians), we estimate there were 404,214 cumulative COVID-19 deaths as of January 1, 2021, 16.1% greater than cumulative reported deaths and approximately 0.12% of the US population on January 1, 2020. Estimates of the size of the infected population were sensitive to assumptions about the IFR, with higher IFR values producing lower estimates of the infected population (S2 Fig). Other major epidemiological outcomes ( $R_t$ , symptomatic cases, severe cases, COVID-19 deaths) had weak relationships with the IFR (absolute rank correlations all <0.2). Alternative assumptions for how county-level IFRs relate to state-level values had a modest impact on cumulative infection estimates (S4 Fig). Cumulative infection estimates and other epidemiological outcomes showed modest changes under different assumptions regarding the time course of COVID-19 disease progression (S5A–S5C Fig) and reporting delays (S5D and S5E Fig).

**Infection ascertainment.** The probability that an infection is diagnosed changed substantially over the course of the U.S. epidemic. Ascertainment was low in the months of March,



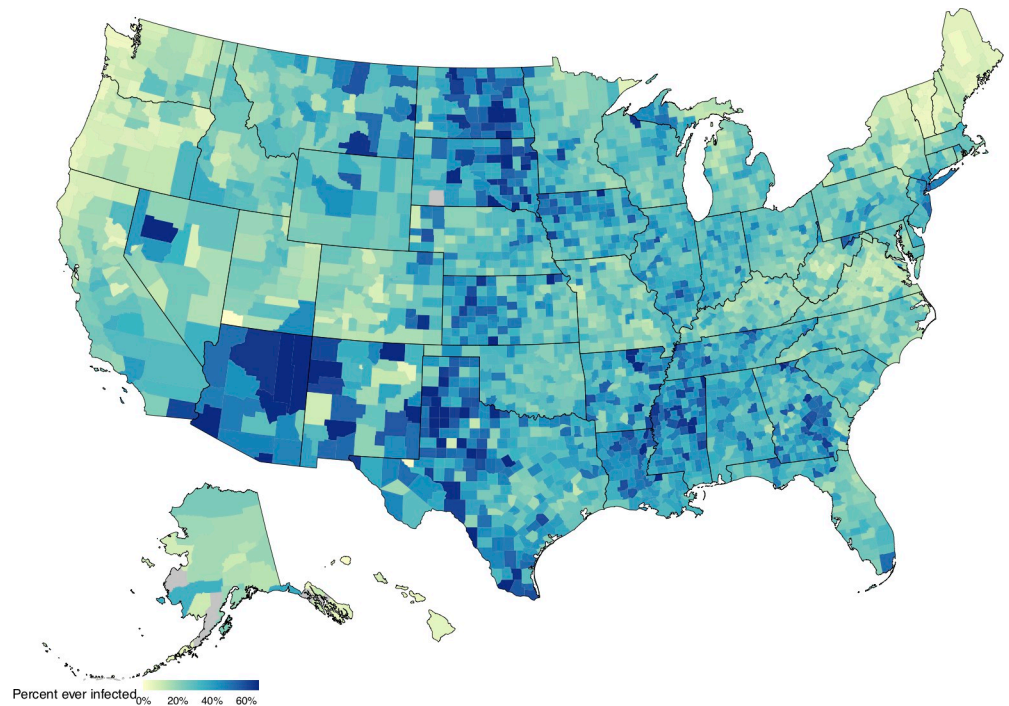
**Fig 4.  $R_t$  estimates for each US state from March 1, 2020 to January 1, 2021.** Background colors indicate whether  $R_t$  is substantially greater than 1 (red), close to 1 (white), or substantially less than 1 (blue). Grey line indicates  $R_t = 1$ . Shaded areas represent 95% credible intervals.

<https://doi.org/10.1371/journal.pcbi.1010465.g004>

April, and May 2020. The national median state-level infection ascertainment (based on state-level posterior medians) in this period was 13.2% (range: 3.2%, 39.7%). Infection ascertainment improved steadily through November 2020, excluding a period of lower ascertainment in July and August; the national mean probability of diagnosis fluctuated between 24% and 35% between September 1, 2020 and January 1, 2021. Infection ascertainment estimates varied significantly across states, and state-level estimates were highly uncertain (Fig 6). Only 3 states achieved greater than 50% ascertainment at any point in 2020 (based on posterior median). State-level model estimates of infection ascertainment each day were negatively correlated with the seven-day moving average fraction of tests that had a positive result[15] (Spearman rank correlation ( $\rho$ ) = -0.36,  $p < 0.001$ ). From the introduction of SARS-CoV-2 in the US until January 1, 2021, we estimate that 22.4% of infections were identified and reported. Infection ascertainment estimates were sensitive to assumptions about the IFR, with higher IFR values producing higher estimates of the fraction of infections identified and reported (S2 Fig). Ascertainment estimates were also sensitive to the natural history delays (S5A–S5C Fig) and reporting delays (S5D and S5E Fig) assumed in the analysis.

### Comparisons to External Covid-19 Burden indicators

We compared our estimates of the percent ever-infected with SARS-CoV-2 to U.S. Centers for Disease Control (CDC) seroprevalence estimates drawn from commercial laboratory data,[16]

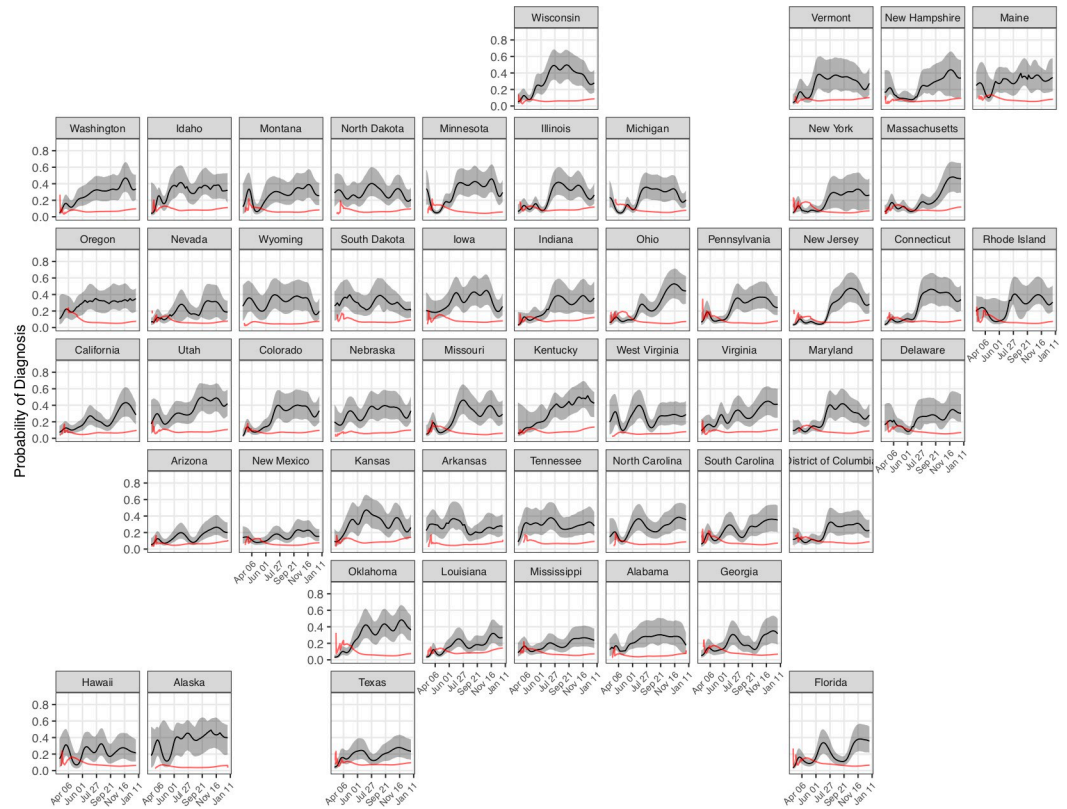


**Fig 5. Percentage of the population ever-infected with SARS-CoV-2 as of January 1, 2021.** Map generated using shapefiles from the *alberusa* package for the R programming language: <https://github.com/hrbrmstr/alberusa>.

<https://doi.org/10.1371/journal.pcbi.1010465.g005>

acknowledging previously noted differences between these outcomes. Derived from a convenience sample of blood specimens collected for reasons unrelated to COVID-19, the seroprevalence estimates provide state-level evidence on SARS-CoV-2 antibody test positivity at multiple time points (Fig 7). However, these estimates are incomplete in some states (e.g. South Dakota), and the series of values declines over time in others (e.g. New York). Comparing these estimates to other reported indicators of cumulative disease burden on December 31, 2020, the modeled estimates of the percent ever-infected were more strongly correlated with cumulative hospitalizations (Spearman rank correlation ( $\rho$ ) = 0.62) and cumulative reported deaths ( $\rho$  = 0.82) than the CDC seroprevalence estimates ( $\rho$  = 0.41 and 0.37 for hospitalizations and deaths respectively).

In addition, we compared model estimates of cumulative COVID-19 deaths (detected and undetected) to state-level estimates of excess all-cause mortality, which reflect both COVID-19 deaths and deviations from expected levels and patterns in non-COVID-19 deaths, [6] (Fig 8) at each weekly timepoint from March 7 to December 19, 2020. On average, modeled estimates of cumulative COVID-19 deaths are less than or approximately equal to estimates of excess all-cause mortality. Notably, three states (Alaska, Hawaii, Maine) have extended periods where the estimated all-cause mortality did not exceed all-cause mortality from previous years (i.e. excess mortality was negative); in periods where all-cause mortality is higher than expected, our estimates of COVID-19 deaths correlate strongly with excess mortality estimates (Spearman rank correlation ( $\rho$ ) = 0.95,  $p < 0.001$ ). Additionally, model estimates of cumulative COVID-19 deaths exceed estimates of excess all-cause mortality in four states (New Jersey, North Dakota, Massachusetts and Rhode Island). Estimates of excess all-cause mortality were not available for Connecticut, North Carolina, or West Virginia.



**Fig 6. The probability that a person infected with SARS-CoV-2 on a given day will be diagnosed for each US state from March 1, 2020 to January 1, 2021.** Shaded areas represent 95% credible intervals. The red line represents a seven-day moving average of the fraction of positive tests.

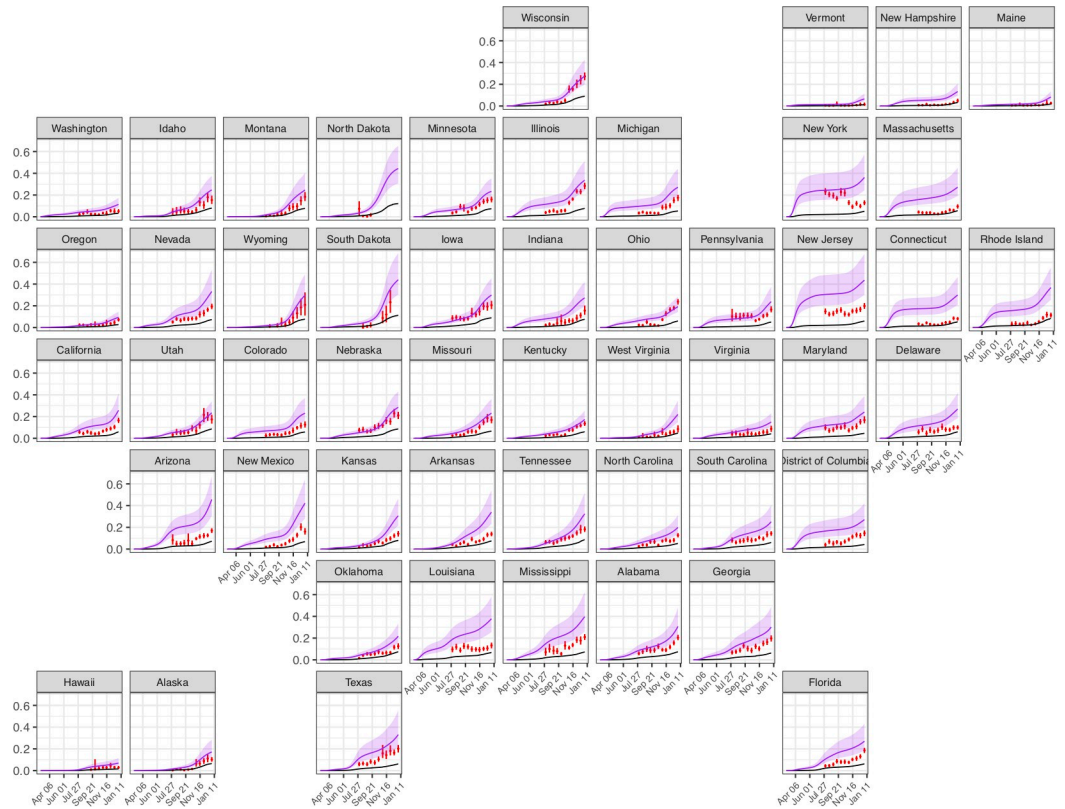
<https://doi.org/10.1371/journal.pcbi.1010465.g006>

## Discussion

We present detailed estimates of the dynamics of SARS-CoV-2 infections in US states and counties through the end of 2020. We found that the viral dynamics are best described as a series of related local and regional epidemics, differing in their timing and magnitude even within individual states. This is evident in the large variation in state- and county-level estimates of percent ever-infected as of January 1, 2021. As case ascertainment has also varied over space and time, these estimates provide insights beyond those that can be inferred from cumulative case counts alone. Ascertainment of infection improved markedly after the first months of the US epidemic, but remained low nationally; we conclude that the reported cumulative case count was approximately one-quarter of the true number of US infections at the end of 2020.

Most notably, we found that model estimates of cumulative infections differ from seroprevalence estimates produce by the CDC. Our estimates of cumulative infections are more strongly correlated with cumulative hospitalizations and deaths across states, potentially reflecting biases in the empirical seroprevalence estimates. Seroprevalence studies have a number of known limitations, including the use of non-representative samples [13] and possible reduced sensitivity associated with waning of antibody titers, as has been reported for some tests. [11, 12] A comparison between model estimates and seroprevalence data therefore suggests that this method provides valuable information about the incidence of infection over time.

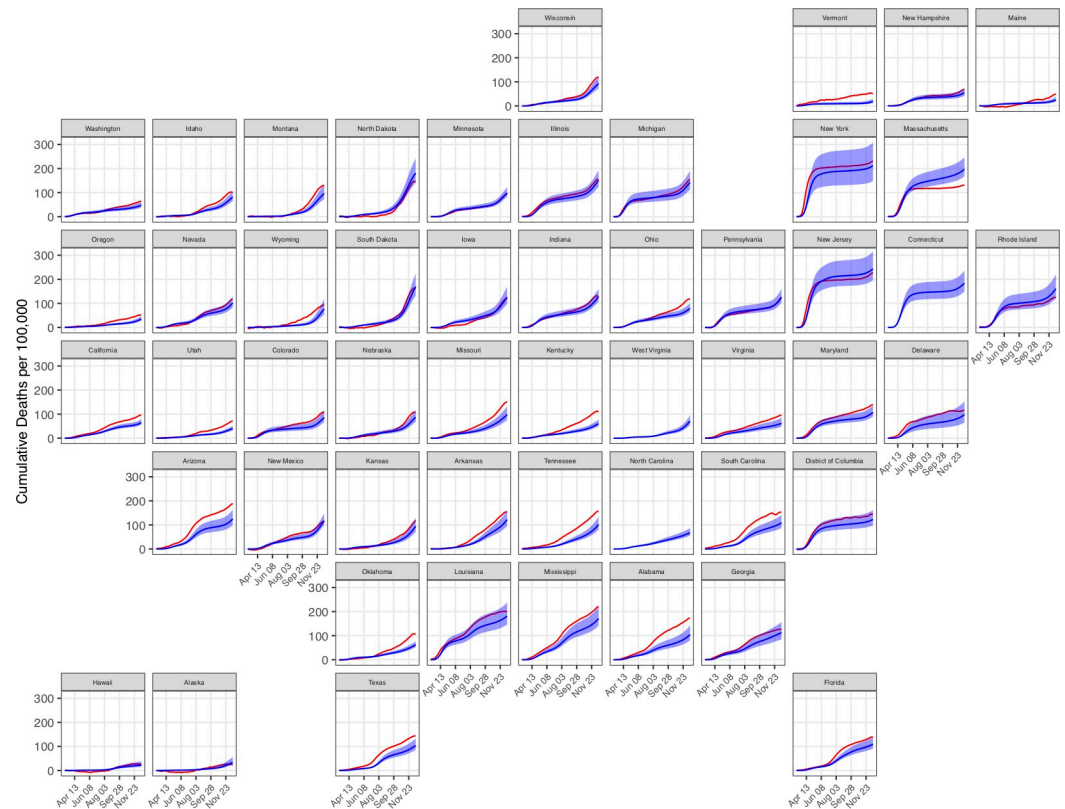




**Fig 7. Comparison of the estimated percent ever-infected with SARS-CoV-2 (purple line, shaded areas represent 95% credible intervals) to CDC seroprevalence estimates from commercial laboratory data (red vertical line) and cumulative reported cases (black line) for each US state from March 1, 2020 to January 1, 2021.**

<https://doi.org/10.1371/journal.pcbi.1010465.g007>

The Bayesian estimation approach used for this analysis makes a number of simplifying assumptions. To reduce model complexity, we rely on fixed distributions to describe delays in disease progression and detection. Because we anchor the analysis on death data (under the assumption that deaths were more consistently reported than cases over the course of the epidemic), model estimates are sensitive to IFR estimates. These IFR estimates are themselves uncertain, being derived from the comparison of death counts to seroprevalence estimates and therefore inheriting the potential biases of these studies. [11–13] While we allow for variation in IFR values at state- and county-level, this variation is based on proxy measures (differences in the age distribution of COVID-19 deaths, differences in reported prevalence of risk factors for severe disease), which may weaken the robustness of this approach. Moreover, it is likely that the IFR has varied over time within each modeled geography due to changes in the age distribution of infections, yet the limited reporting of age-stratified data means that these age-based changes could not be represented in the model. We also allow for modest under-reporting of COVID-19 deaths, consistent with empirical studies of COVID-19 death reporting, [11, 17, 18] but this is an uncertain input to our analysis. For other modeling assumptions—in particular, the delay distributions quantifying the lag between infection and reporting—we assumed consistent values across all locations, due to a lack of data on how these vary across states and counties. Finally, we assume that a previously infected individual cannot be re-infected with SARS-CoV-2. While waning antibody titers suggest that re-infection is possible over time, we do not believe that our assumptions about re-infection meaningfully impact our results. [19, 20]



**Fig 8. Comparison of cumulative COVID-19 deaths (blue) to cumulative excess all-cause mortality (red) for each US state from March 7 to December 19, 2020.** Shaded areas represent 95% credible intervals.

<https://doi.org/10.1371/journal.pcbi.1010465.g008>

In addition, we used data that have been aggregated from state-level reporting mechanisms, which are vulnerable to a number of potential sources of bias. States vary in their reporting criteria (e.g. reporting the number of positive tests as opposed to number of individuals who have tested positive) and the average delay between case detection and reporting. Data are also subject to occasional revisions, often implemented as a single-day change in the cumulative count of cases or deaths. Taken together, these data irregularities lead to additional variance in the reported data and a reduction in the precision of reported estimates. While line-list data would likely improve the precision of model estimate [21], these data are not widely available in the US. Despite these limitations, the method described here may represent an improvement over similar modeling approaches that do not allow for case ascertainment rates and infection fatality ratios that vary over both space and time, [8, 9, 22, 23] or that estimate  $R_t$  using model outputs rather than as part of the modeling framework. [9, 21, 23] Furthermore, our approach uses changes in case and death data to estimate changes in transmission, while others approaches make use of more indirect data on mobility [8, 22] or similar proxies [23] to signal changes in transmission. While mobility has a mechanistic relationship with disease transmission, the association between movement data and viral transmission is complex and variable across time and space, possibly because of changes in mask use and other non-pharmaceutical interventions. [24, 25]

In conclusion, the modeling approach described here provides a coherent framework for simultaneously estimating the trend in SARS-CoV-2 infections and the fraction of the population that has been infected previously, providing key information on the viral dynamics at

county- and state-levels. While the deployment of effective vaccines against the virus represents a great hope for the control of SARS-CoV-2 transmission, vaccine hesitancy and the emergence of more transmissible variants [26] present an ongoing challenge to disease control in the US. Understanding the course of the epidemic in the pre-vaccine era can help guide decision making in a landscape with heterogeneous vaccine coverage. Ongoing, local evidence on trends in  $R_t$  and new and cumulative infections will continue to be important for both governments and individuals.

## Methods

We developed a mechanistic model that uses reported case and death data to back-calculate the natural history cascade of SARS-CoV-2. The model estimates the expected number of cases and deaths reported on a given day as the convolution of the time series of diagnosed cases and deaths (among diagnosed individuals) and fixed reporting delay distributions; the expected number of diagnoses on a given day is estimated with health-state specific and time-varying probabilities of diagnosis. The model represents the natural history of COVID-19 as a series of health state transitions with associated probabilities and delays (Fig 1). The model utilizes delay distributions associated with health state progression, time-invariant probabilities of transitioning from *Asymptomatic* to *Symptomatic* and from *Symptomatic* to *Severe*, and a time-varying probability of transitioning from *Severe* to *Death*. The number of individuals entering *Asymptomatic* is a function of the serial interval, the fraction of the population not yet infected, and  $R_t$ ;  $R_t$  is modeled using a log-transformed cubic b-spline.

## Data

For every state and county in the United States, we extracted daily data on reported COVID-19 cases and deaths from a repository compiled by the Johns Hopkins Center for Systems Science and Engineering (CSSE) [14]. We calculated the time series of new cases and deaths as the difference between cumulative counts reported on consecutive days. In instances in which the reported cumulative count decreased from one day to the next, we assumed that there were zero new cases or deaths on each day until the cumulative count exceeded the previous maximum. In several instances the data reported by CSSE fail to capture the beginning of the epidemic in early 2020, or exhibit irregularities during this period. To reconstruct the time series for this period we used data compiled by the Covid Tracking Project. [15]

## Mathematical model

We constructed a deterministic mathematical model relating reported cases and deaths to unobserved COVID-19 natural history. A flexible function for  $R_t$  determines the number of individuals infected on a given day, and the model then tracks the progression of the infected cohort through health states of increasing disease severity, with modeled quantities— $A_t$  (*Asymptomatic*),  $S_t$  (*Symptomatic*),  $V_t$  (*Severe*), and  $D_t$  (*Death*)—reflecting the number of individuals entering a given health state on day  $t$ . From each health state, an individual can either recover or progress to the next health state, with this transition governed by a defined delay distribution. Ultimately, the model estimates an expected number of reported cases and deaths on each day, which are fit to observed data via negative binomial likelihood functions.

## New infections

We modeled the daily number of newly-infected individuals ( $A_t$ ) entering the *Asymptomatic* state. For each modeled location, we specified a random intercept ( $A_0$ ) 28 days before the first

reported COVID-19 case, and calculated changes in  $A_t$  as a function of the effective reproduction number ( $R_t$ ) and the mean serial interval ( $z$ ), measured in days (derivation shown in supplement).

$$A_{t+1} = A_t R_t^{\frac{1}{z}} \text{ for } t \geq 0 \tag{1}$$

We modeled the time trend in  $R_t$  using a log-transformed cubic b-spline ( $X_{R,t}$ ) with knots every 10 days (S3 Fig), allowing flexibility in the evolution of the epidemic curve over time. Penalties on first and second differences of the spline parameters were used to dampen oscillations not supported by the data. We assumed that individuals can only be infected once and multiplied the spline by the fraction of the population ( $N$ ) uninfected at each timepoint, penalizing  $R_t$  towards zero as the population ever-infected approaches 100%.

$$R_t = X_{R,t} \left( 1 - \frac{\sum_{i=0}^t A_i}{N} \right) \tag{2}$$

### Disease progression

We assumed that a fraction of individuals with asymptomatic disease ( $p_s$ ) progress to the *Symptomatic* state. The delay from infection to symptoms was assumed to follow a Gamma distribution, with  $\rho_{s,i}$  representing the fraction progressing between  $i$  and  $i+1$  days after infection, among those progressing to the symptomatic state. We tested the sensitivity of model outcomes to the choice of symptom onset delay (S5A Fig).

$$S_t = \sum_{i=0}^t A_{t-i} p_s \rho_{s,i} \tag{3}$$

Similarly, a fraction of individuals in the *Symptomatic* state ( $p_v$ ) were assumed to progress to the *Severe* state, with Gamma-distributed delay distribution  $\rho_{v,i}$ . A fraction of individuals with severe disease ( $p_{D,t}$ ) die, with Gamma-distributed delay distribution  $\rho_{D,i}$ . We tested the sensitivity of model outcomes to the choice of delay to severe disease (S5B Fig) and to death (S5C Fig).

$$V_t = \sum_{i=0}^t S_{t-i} p_v \rho_{v,i} \tag{4}$$

$$D_t = \sum_{i=0}^t V_{t-i} p_{D,t} \rho_{D,i} \tag{5}$$

With the exception of  $p_{D,t}$ , disease progression parameters were not allowed to vary over time. For  $p_{D,t}$  we assumed higher values applied in early 2020, reflecting higher case fatality among individuals with severe disease early in the epidemic due to later presentation and lower effectiveness of treatment at that time. We modeled the time trend in  $p_{D,t}$  as the product of  $p_{D,0}$  (the progression probability after early 2020) and  $OR_{p_{D,t}}$ , an odds ratio describing the elevated case fatality early in the epidemic (Eq 6).  $OR_{p_{D,t}}$  was operationalized using a declining sigmoid curve (1.0 minus the Normal cumulative distribution function  $\Phi$ ) with an inflection point on May 1 2020 (Eq 7). In this equation,  $\mu$  is equal to the number of days between the start of the model ( $t = 0$ ) and May 1<sup>st</sup> 2020,  $\sigma$  is equal to 21 days, and  $a_{p_D}$  represents an uncertain parameter for the *additional* mortality risk early in the epidemic. This formulation ensures that  $OR_{p_{D,t}}$  asymptotes toward 1.0 as  $t$  increases after May 1<sup>st</sup> 2020.

$$\frac{p_{D,t}}{1 - p_{D,t}} = \frac{p_{D,0}}{1 - p_{D,0}} OR_{p_{D,t}} \tag{6}$$

$$OR_{PD,t} = 1 + \left(1 - \Phi\left(\frac{t - \mu}{\sigma}\right)\right) a_{PD} \tag{7}$$

While vaccination would also affect disease progression probabilities, we assumed that vaccination coverage was insufficient to impact disease natural history during the study period.

### Infection fatality ratio

We assumed that the infection fatality ratio (IFR) differs across states and counties, reflecting differences in the age distribution of the epidemic and differences in the prevalence of medical risk factors for severe COVID-19 disease. First, we calculated the age distribution of infections for each state, based on the reported age distribution of COVID-19 deaths [27] and published age-specific IFRs. [28] Second, we used these age distributions to calculate an average IFR for each state, weighting the age-specific IFRs by the fraction of the population in each age group. This produced a national average IFR of 0.35, which we believe to be implausibly low; we rescaled state-level values to produce a national average IFR of 0.5%. [29] As the age-distribution of COVID-19 deaths was not available at the county-level, we estimated county-level IFR values by multiplying the state-average IFR by the prevalence of medical risk factors for severe COVID-19 disease in each county relative to the rest of the state. [30] To test the impact of this assumption, we performed a sensitivity analysis with a simpler approach that holds all county-level IFRs equal to the state level (S4 Fig). To understand the implications uncertainty in the IFR for modeled estimates of the infected population, we plotted the relationship between these two quantities in the fitted model outcomes.

### Diagnosis

We assumed that infected individuals could be diagnosed from the *Asymptomatic*, *Symptomatic*, or *Severe* states, and that diagnosis would not affect disease progression. To reduce model complexity, we assumed that diagnosis in the *Asymptomatic* state only occurs among individuals who will not progress to the *Symptomatic* state. The daily number of these diagnoses is denoted  $\hat{A}_t$  (with the ^ used to indicate quantities related to diagnosis). The fraction of these individuals diagnosed ( $q_{A,t}$ ) was assumed to vary over time, to allow for changes in case ascertainment over the course of the epidemic. The delay to diagnosis was defined by  $\hat{\rho}_{A,i}$ , which is described by a Gamma distribution.

$$\hat{A}_t = \sum_{i=0}^t A_{t-i} q_{A,t-i} \hat{\rho}_{A,i} (1 - p_S) \tag{8}$$

To estimate the number diagnosed from the *Symptomatic* state ( $\hat{S}_t$ ) we assumed a time-varying probability of diagnosis  $q_{S,t}$  and delay to diagnosis  $\hat{\rho}_{S,i}$ .

$$\hat{S}_t = \sum_{i=0}^t S_{t-i} q_{S,t-i} \hat{\rho}_{S,i} \tag{9}$$

The number diagnosed from the *Severe* state ( $\hat{V}_t$ ), was calculated based on a time-invariant probability of diagnosis ( $q_V$ ) and delay to diagnosis  $\hat{\rho}_{V,i}$ . These were applied after subtracting individuals developing severe disease who had been previously diagnosed at *Symptomatic* ( $\bar{V}_t$ ).

$$\bar{V}_t = \sum_{i=0}^t S_{t-i} q_{S,t-i} p_V \rho_{V,i} \tag{10}$$

$$\hat{V}_t = \sum_{i=0}^t (V_{t-i} - \bar{V}_{t-i}) q_V \hat{\rho}_{V,i} \tag{11}$$

Time-varying diagnosis probabilities ( $q_{A,t}$ ,  $q_{S,t}$ ) were calculated as a function of  $q_V$ :

$$q_{S,t} = q_V X_{q_{S,t}} \tag{12}$$

$$q_{A,t} = q_V X_{q_{S,t}} RR_{q_A} \tag{13}$$

In Eqs 12 and 13,  $X_{q_{S,t}}$  is operationalized as a cubic b-spline that has been logit-transformed to fall within the unit interval, with knots spaced 21 days apart, and with penalties on first and second differences of the spline parameters.  $RR_{q_A}$  is constrained to fall in the unit interval, so that that  $q_{A,t} \leq q_{S,t} \leq q_V$  for all  $t$ .

### Reporting

We assumed that all diagnosed COVID-19 cases were reported. The number of diagnoses reported on a given day ( $\dot{C}_t$ , with the ‘ $\dot{\cdot}$ ’ used to indicate quantities related to reporting) was calculated as the sum of diagnoses from *Asymptomatic*, *Symptomatic* and *Severe* states, with reporting delay  $\dot{\rho}_{C,i}$ .

$$\dot{C}_t = \sum_{i=0}^t (\hat{A}_{t-i} + \hat{S}_{t-i} + \hat{V}_{t-i}) \dot{\rho}_{C,i} \tag{14}$$

The reported number of COVID-19 deaths ( $\dot{D}_t$ ) were calculated from the number of diagnosed individuals who subsequently died ( $\hat{D}_t$ ).  $\hat{D}_t$  was calculated as the sum of deaths among individuals diagnosed from the *Symptomatic* and *Severe* states, represented by the first and second terms in Eq 15, respectively. We assumed that all deaths among diagnosed COVID-19 cases were reported, with reporting delay  $\dot{\rho}_{D,i}$ .

$$\hat{D}_t = \left( \sum_{i=0}^t \bar{V}_{t-i} p_{D,t} \rho_{D,i} \right) + \left( \sum_{i=0}^t (V_{t-i} - \bar{V}_{t-i}) q_V p_{D,t} \rho_{D,i} \right) \tag{15}$$

$$\dot{D}_t = \sum_{i=0}^t \hat{D}_{t-i} \dot{\rho}_{D,i} \tag{16}$$

We tested the sensitivity of model outcomes to the choice of reporting delays (S5D and S5E Figs).

### Data likelihood

We specified negative binomial likelihood functions to fit the model to observed cases ( $Y_{C,t}$ ) and death data ( $Y_{D,t}$ ).

$$Y_{C,t} \sim \text{NegBin}(\dot{C}_t, \phi_C)$$

$$Y_{D,t} \sim \text{NegBin}(\dot{D}_t, \phi_D)$$

To account for variation in daily reported cases and deaths, we fit the likelihood function using a seven-day moving average of input data. The negative binomial dispersion parameters ( $\phi_C$ ,  $\phi_D$ ) were estimated simultaneously, allowing for additional variance in the observed time series.

### Model parameters

Model parameters are shown in Table 1. The distributions of delays from infection to symptom onset, symptom onset to severe disease, and severe disease to death are used in the model

**Table 1. Model parameters.**

Model Parameter	Mean, std. Deviation	Distribution	Type	Source
Log of New Infections At T = 0 ( $A_0$ )	0,10	Normal(0,10)	prior	Assumed
$X_{R,t}$ Spline Parameters	0,3	Normal(0,3)	prior	Assumed
First Derivative of $X_{R,t}$ Spline Parameters	0,0.5	Normal(0,0.5)	prior	Assumed
Second Derivative of $X_{R,t}$ Spline Parameters	0,0.1	Normal(0,0.1)	prior	Assumed
Serial Interval	5.8, 0.5	Gamma(129.1, 22.25)	prior	[31]
Probability of Developing Symptoms If Infected	0.59, 0.16	Beta(5.14, 3.53)	prior	[3,32–34]
Probability of Becoming Severely Ill If Symptomatic	0.09, 0.06	Beta(1.89, 20.00)	prior	[35,36]
Probability of Death for All Infections (national average)	0.005, 0.001	Beta(15.9, 3167)	prior	[28,29]
Probability of Death for Severe Infections	0.15, 0.03	Beta(28.2, 162.3)	prior	[36]
Additional Risk of Mortality Prior to May 1 2020 ( $a_{pp}$ )	1.34, 0.39	Gamma(12.03, 8.99)	prior	Assumed
Rate Ratio, Diagnosis at Asymptomatic Vs. Symptomatic	0.1, 0.07	Beta(2,18)	prior	Assumed
Rate Ratio, Diagnosis at Symptomatic Vs. Severe	0.5, 0.22	Beta(2,2)	prior	Assumed
Probability of Diagnosis at Severe	0.72, 0.16	Beta(20,5)	prior	Assumed
Dispersion Parameter for Reported COVID-19 Cases ( $1/\sigma^2$ )	0.8, 0.6	Half-Normal(0,1)	prior	[37]
Dispersion Parameter for Reported COVID-19 Deaths ( $1/\sigma^2$ )	0.8, 0.6	Half-Normal(0,1)	prior	[37]
Scaling Factor: Time to Diagnosis Relative to Time in Symptomatic State	0.5, 0.22	Beta(2,2)	prior	Assumed
Scaling Factor: Time to Diagnosis Relative to Time in Severe State	0.5, 0.22	Beta(2,2)	prior	Assumed

<https://doi.org/10.1371/journal.pcbi.1010465.t001>

as fixed inputs and can be found in [Table 2](#). [S6 Fig](#) shows the comparison of prior and posterior distributions for key parameters listed in [Table 1](#).

## Model implementation

The model was implemented in R using the *rstan* package. [41] The model initializes 28 days before the first reported case or death. Given the delay from infection to death, we chose 28 days to allow the model to generate the necessary number of new infections to plausibly result in a death early in the observed time series. The model is fit to data from each county or state separately. For state-level results (including Washington, DC) we estimated outcomes using a Hamiltonian Monte Carlo algorithm. [42] The model ran for 3000 iterations (2000 burn-in) on 4 chains, and 3000 samples (across 3 chains) from the posterior were included in these results. Counties were fit using an optimization routine that reports the maximum a posteriori estimate, which represents an estimate of the mode of the posterior distribution of the model parameters.

## Covidestim package

The *covidestim* package is a package for the R programming language, suitable for public as well as research use. It can accommodate a number of data inputs. Users may enter a vector of

**Table 2. Delay Distributions.**

Delay	Mean, std. Deviation	Distribution	Source
Infected to Symptomatic (Days)	5.6, 3.1	Gamma(3.41, 0.61)	[38]
Symptomatic to Severe (Days)	7.5, 5.8	Gamma(1.72, 0.22)	[39]
Severe to Death (Days)	9.1, 6.3	Gamma(2.10, 0.23)	[40]
Case Reporting Delay	2.2, 1.5	Gamma(2.2, 1)	Assumed
Death Reporting Delay	2.2, 1.5	Gamma(2.2, 1)	Assumed

<https://doi.org/10.1371/journal.pcbi.1010465.t002>

daily case counts and/or daily death counts. These data sources can be used in combination, so long as they are the same length and cover the same time period; days with no observed events may be represented with zeroes.

The package contains default model priors for progression probabilities and delays, detection probabilities and delays, and reporting delays associated with each data type. Users have the ability to override these defaults, though we recommend that they only specify priors for reporting delays; we do not recommend that users change default priors on parameters related to the natural history of COVID-19.

### [Covidestim.org](https://covidestim.org) and code repositories

We produce daily estimates of COVID-19 infections and the effective reproduction number of SARS-CoV-2 at the state- and county-levels at <https://covidestim.org>. To allow for daily production of model estimates for all U.S. counties and states, we developed several tools. The *covidestim* Docker image is a container which allows for model execution in any HPC or cloud environment, and is the easiest way to begin using the *covidestim* R package. The *covidestim-sources* repository enables automated, version-controlled, reproducible data cleaning of four different case/death data sources by leveraging Git's submodules feature. Finally, the *dailyFlow* repository uses the Nextflow workflow engine [43] to clean the data, orchestrate 3200+ model runs within three supported execution environments (local, HPC, cloud), and export the results for research use and for web consumption. These repositories can be found at <https://github.com/covidestim>, and contain extensive documentation.

## Supporting information

### **S1 Text. Derivation of Eq 1.**

(DOCX)

### **S1 Fig. Model fit to data for four states—California, Florida, New York, South Dakota.**

(TIFF)

### **S2 Fig. Relationship between COVID-19 infection fatality rate (IFR) and modeled outcomes, using North Dakota as an example.**

(TIFF)

### **S3 Fig. Impact of spline knot width on $R_t$ estimates for four states—California, Florida, New York, South Dakota.**

(TIFF)

### **S4 Fig. County-level cumulative incidence estimates for December 31, 2020, comparing main analysis results to a sensitivity analysis in which county IFRs were fixed at the state average.**

(TIFF)

### **S5 Fig. Comparison of major model outcomes for selected states, under alternative assumptions for the average delay between infection and development of symptoms, for individuals developing symptomatic COVID-19.**

(TIFF)

### **S6 Fig. 95% Credible intervals for key parameters, across states.**

(TIFF)



## Acknowledgments

We thank Jeffrey Eaton for his thoughts on statistical analysis.

## Author Contributions

**Conceptualization:** Ted Cohen, Nicolas A. Menzies.

**Data curation:** Marcus Russi, Kenneth Gunasekera, Joshua Havumaki.

**Formal analysis:** Melanie H. Chitwood, Marcus Russi, Nicolas A. Menzies.

**Funding acquisition:** Joshua A. Salomon, Nicolas A. Menzies.

**Methodology:** Melanie H. Chitwood, Marcus Russi, Fayette Klaassen, Virginia E. Pitzer, Joshua A. Salomon, Nicole A. Swartwood, Joshua L. Warren, Daniel M. Weinberger, Ted Cohen, Nicolas A. Menzies.

**Software:** Melanie H. Chitwood, Marcus Russi, Fayette Klaassen.

**Supervision:** Ted Cohen.

**Validation:** Nicole A. Swartwood, Nicolas A. Menzies.

**Visualization:** Melanie H. Chitwood, Marcus Russi, Nicole A. Swartwood, Nicolas A. Menzies.

**Writing – original draft:** Melanie H. Chitwood, Nicolas A. Menzies.

**Writing – review & editing:** Melanie H. Chitwood, Marcus Russi, Kenneth Gunasekera, Joshua Havumaki, Fayette Klaassen, Virginia E. Pitzer, Joshua A. Salomon, Nicole A. Swartwood, Joshua L. Warren, Daniel M. Weinberger, Ted Cohen, Nicolas A. Menzies.

## References

1. Coronavirus in the U.S.: Latest Map and Case Count. Retrieved from <https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html>
2. Coronavirus. Retrieved from [https://www.washingtonpost.com/graphics/2020/national/coronavirus-us-cases-deaths/?itid=hp\\_pandemic%20test](https://www.washingtonpost.com/graphics/2020/national/coronavirus-us-cases-deaths/?itid=hp_pandemic%20test)
3. Oran DP and Topol EJ. Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. *Annals of internal medicine*. 2020; 173(5): 362–367. <https://doi.org/10.7326/M20-3012> PMID: 32491919
4. Hitchings MDT, Dean NE, Garcia-Carreras B, Hladish TJ, Huang AT, Yang B, et al. The Usefulness Of SARS-CoV-2 Test-Positive Proportion As A Surveillance Tool. *American Journal of Epidemiology*. 2021; 190(7):1396–1405. <https://doi.org/10.1093/aje/kwab023>
5. Pitzer VE, Chitwood MH, Havumaki J, Menzies NA, Perniciaro S, Warren JL, et al. The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. *American Journal of Epidemiology*. 2021; <https://doi.org/10.1093/aje/kwab089> PMID: 33831148
6. Weinberger D, Cohen T, Crawford F, Mostashari F, Olson D, Pitzer VE, et al., Estimating the early death toll of COVID-19 in the United States. [Preprint]. 2020 [Cited 2020 July 13] Available from: <https://doi.org/10.1101/2020.04.15.20066431>.
7. Gostic KM, McGough L, Baskerville E, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, Rt. *PLoS Comput. Biol.* 2021; 16(12): e1008409. <https://doi.org/10.1371/journal.pcbi.1008409>
8. Unwin HJT, Mishra S, Bradley VC, Gandy A, Mellan TA, Coupland H, et al. State-level tracking of COVID-19 in the United States. *Nat. Commun.* 2020; 11, 6189 <https://doi.org/10.1038/s41467-020-19652-6> PMID: 33273462
9. COVID-19 Portal, Center for the Ecology of Infection Diseases, University of Georgia [Cited 2021 July 10]. Available at: <https://www.covid19.uga.edu/nowcast.html>
10. Considerations for implementing and adjusting public health and social measures in the context of COVID-19. Geneva: World Health Organization; [https://apps.who.int/iris/bitstream/handle/10665/336374/WHO-2019-nCoV-Adjusting\\_PH\\_measures-2020.2-eng.pdf?sequence=1&isAllowed=y](https://apps.who.int/iris/bitstream/handle/10665/336374/WHO-2019-nCoV-Adjusting_PH_measures-2020.2-eng.pdf?sequence=1&isAllowed=y)

11. Ibarondo FJ, Fulcher JA, Goodman-Meza D, Elliot J, Hofmann C, Hausner MA, et al. Rapid Decay of Anti-SARS-CoV-2 Antibodies in Persons with Mild Covid-19. *N Engl J Med.* 2020; 383:1085–1087. <https://doi.org/10.1056/NEJMc2025179> PMID: 32706954
12. Seow J, Graham C, Merrick B, Acors S, Pickering S, Steel KJA. Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. *Nat Microbiol.* 2020; 5:1598–1607 <https://doi.org/10.1038/s41564-020-00813-8> PMID: 33106674
13. Bajema KL, Wiegand RE, Cuffe K, Patel SV, Iachan R, Lim T. Estimated SARS-CoV-2 Seroprevalence in the US as of September 2020. *JAMA Intern Med.* November 24, 2020; <https://doi.org/10.1001/jamainternmed.2020.7976> PMID: 33231628
14. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Inf. Dis.* May 1, 2020; 20(5):533–534. [https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1) PMID: 32087114
15. The COVID Tracking Project. [Cited 15 January 2021] Available at: <https://covidtracking.com/>
16. Nationwide Commercial Laboratory Seroprevalence Survey. [Cited 23 March, 2020] Available at: <https://data.cdc.gov/Laboratory-Surveillance/Nationwide-Commercial-Laboratory-Seroprevalence-Su/d2tw-32xv>
17. Woolf SH, Chapman DA, Sabo RT, Weinberger DM, Hill L. Excess Deaths From COVID-19 and Other Causes, March–April 2020. *JAMA.* 2020; 324(5):510–513. <https://doi.org/10.1001/jama.2020.11787> PMID: 32609307
18. Iuliano AD, Chang HH, Patel NN, Threlkel R, Kniss K, Reich J, et al. Estimating under-recognized COVID-19 deaths, United States, March 2020–May 2021 using an excess mortality modelling approach. *Lancet Reg Health Am.* 2021 Sep; 1:100019. <https://doi.org/10.1016/j.lana.2021.100019> PMID: 34386789
19. Wajnberg A, Amanat F, Firpo A, Altman DR, Bailey MJ, Mansour M. Robust neutralizing antibodies to SARS-CoV-2 infection persist for months. *Science.* 2020; 370(6521):1227–1230. <https://doi.org/10.1126/science.abd7728> PMID: 33115920
20. Qureshi AI, Baskett WI, Huang W, Lobanova I, Naqvi SH, Shyu C. Reinfection with Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in Patients Undergoing Serial Laboratory Testing. *Clinical Infectious Diseases.* 2021; <https://doi.org/10.1093/cid/ciab345>
21. Li T, White LF. Bayesian back-calculation and nowcasting for line list data during the COVID-19 pandemic. *PLOS Computational Biology.* 2021; 17(7): e1009210. <https://doi.org/10.1371/journal.pcbi.1009210> PMID: 34252078
22. Flaxman S, Mishra S, Gandy A, Unwin JT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature.* 2020; 584: 257–261 <https://doi.org/10.1038/s41586-020-2405-7> PMID: 32512579
23. Leung K, Wu JT & Leung GM. Real-time tracking and prediction of COVID-19 infection using digital proxies of population mobility and mixing. *Nat Commun.* 2021; 12: 1501 <https://doi.org/10.1038/s41467-021-21776-2> PMID: 33686075
24. Kishore N, Taylor AR, Jacob PE, Vembar N, Cohen T, Buckee CO, et al. The relationship between human mobility measures and SAR-CoV-2 transmission varies by epidemic phase and urbanicity: results from the United States. [Preprint] 2021. [Cited 15 July 2021]. Available from: <https://www.medrxiv.org/content/10.1101/2021.04.15.21255562v1>
25. Nouvellet P, Bhatia S, Cori A, Ainslie KEC, Baguelin M, Bhatt S, et al. Reduction in mobility and COVID-19 transmission. *Nat Commun.* 2021; 12: 1090 <https://doi.org/10.1038/s41467-021-21358-2> PMID: 33597546
26. COVID Data Tracker. “Variant Proportions” [Cited 8 June 2021] Available at: <https://covid.cdc.gov/covid-data-tracker/#variant-proportions>
27. National Center for Health Statistics. “Provisional COVID-19 Death Counts by Sex, Age, and State” [Cited 15 January 2021] Available at: <https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku>
28. O’Driscoll M, Ribeiro Dos Santos G, Wang L, Cummings DAT, Azman AS, Paireau J, et al. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature.* 2021; 590: 140–145 <https://doi.org/10.1038/s41586-020-2918-0> PMID: 33137809
29. Levin AT, Hanage WP, Owusu-Boaitey N, Cochran KB, Walsh SP, Meyerowitz-Katz G. Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. *Eur J Epidemiol.* 2020; 35: 1123–1138. <https://doi.org/10.1007/s10654-020-00698-1> PMID: 33289900
30. Razzaghi H, Wang Y, Lu H, Marshall KE, Dowling NF, Paz-Bailey G, et al. Estimated County-Level Prevalence of Selected Underlying Medical Conditions Associated with Increased Risk for Severe

- COVID-19 Illness—United States, 2018. *MMWR Morb Mortal Wkly Rep* 2020; 69:945–950. <https://doi.org/10.15585/mmwr.mm6929a1> PMID: 32701937
31. He X, Lau EH, Wu P, Marshall KE, Dowling NF, Paz-Bailey G, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature medicine*. 2020; 26(5): pp.672–675. <https://doi.org/10.1038/s41591-020-0869-5> PMID: 32296168
  32. Poletti P, Tirani M, Cereda D, Trentini F, Guzzetta G, Sabatino G, et al. Probability of symptoms and critical disease after SARS-CoV-2 infection. 2020. arXiv preprint arXiv:2006.08471.
  33. Buitrago-Garcia D, Egli-Gany D, Counotte MJ, Hossmann S, Imeri H, Ipekci AM, et al. Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLOS Medicine*. 2020; 17(9): e1003346. <https://doi.org/10.1371/journal.pmed.1003346> PMID: 32960881
  34. Byambasuren O, Cardona M, Bell K, Clark J, McLaws M, Glasziou P (2020). “Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis.” *JAMMI*. 5(4): 223–234. <https://doi.org/10.3138/jammi-2020-0030>
  35. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*. 2020; 20(6): 669–677. [https://doi.org/10.1016/S1473-3099\(20\)30243-7](https://doi.org/10.1016/S1473-3099(20)30243-7) PMID: 32240634
  36. CDC COVID-19 Response Team. “Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19)—United States, February 12–March 16, 2020.” *MMWR. Morbidity and Mortality Weekly Report* (2020) 69(12), 343–346. ISSN 0149-2195, 1545-861X, <https://doi.org/10.15585/mmwr.mm6912e2> PMID: 32214079
  37. Gelman A. Prior Choice Recommendations. Github. 2020 April 17 [Cited 2021 July 10]. Available from: <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>
  38. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng, HR Q, Meredith et al. “The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application.” *Annals of Internal Medicine* (2020). 172(9), 577–582. ISSN 0003-4819, 1539–3704, <https://doi.org/10.7326/M20-0504> PMID: 32150748
  39. 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), 1054–1062. ISSN 01406736, [https://doi.org/10.1016/S0140-6736\(20\)30566-3](https://doi.org/10.1016/S0140-6736(20)30566-3) PMID: 32171076
  40. Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S, et al. “Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data.” *Journal of Clinical Medicine*. (2020). 9(2), 538. ISSN 2077-0383, <https://doi.org/10.3390/jcm9020538> PMID: 32079150
  41. Stan Development Team. RStan: the R interface to Stan. 2018 R package version 2.17.3. <http://mc-stan.org>
  42. Hoffman MD and Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res*. 2012; 15(1): 1593–1623.
  43. Di Tommaso P, Chatzou M, Floden E, Barja PP, Palumbo E, Noterdame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 2017; 35, 316–319. <https://doi.org/10.1038/nbt.3820> PMID: 28398311