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

Authors: Melanie H. Chitwood[1¶*], Marcus Russi[1¶], Kenneth Gunasekera[1], Joshua Havumaki[1], Virginia E. Pitzer[1], Joshua A. Salomon[2], Nicole Swartwood[3], Joshua L. Warren[4], Daniel M. Weinberger[1], Ted Cohen[1&] and Nicolas A. Menzies[3&]

Affiliations:
[1] Department of Epidemiology of Microbial Diseases and Public Health Modeling Unit, Yale School of Public Health, Yale University, New Haven, Connecticut USA
[2] Department of Medicine, Stanford University, Palo Alto, California USA
[3] Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts US
[4] Department of Biostatistics and Public Health Modeling Unit, Yale School of Public Health, Yale University, New Haven, Connecticut USA

[¶]These authors contributed equally to this work
[&]These authors share senior authorship
[*]Corresponding author
Email: melanie.chitwood@yale.edu

Word Count: 3572

Abstract

Estimating the true magnitude of the United States (US) SARS-CoV-2 epidemic is crucial for understanding disease dynamics and, ultimately, for determining the effectiveness of interventions intended to interrupt transmission. We developed a Bayesian evidence synthesis model that explicitly accounts for reporting delays and secular variation in case ascertainment to generate estimates of incident COVID-19 infections on the basis of reported cases and deaths. We estimate time trends in COVID-19 epidemiology for every US state and county, from the first reported case (January 13, 2020) through January 1, 2021. Across counties, we estimate considerable variability in the level and pattern of incidence, producing major differences in the estimated proportion of the population infected by the end of 2020. Our estimates of COVID-19 deaths are consistent with independent estimates of excess mortality, and our estimates of cumulative incidence of infection are consistent with seroprevalence estimates from available antibody testing studies.

38 # Introduction

39 The number of newly diagnosed cases and confirmed COVID-19 deaths are the most easily
40 observed measures of the health burden associated with COVID-19, and have been widely used
41 to track the trajectory of the epidemic.[1,2] However, there are at least three important limitations
42 of using reported case and death counts to track the epidemic. First, testing is primarily
43 organized to identify symptomatic individuals, and most individuals with asymptomatic infection
44 will fail to be diagnosed. As asymptomatic disease represents a large fraction of all infections,[3]
45 reported diagnoses will substantially underestimate the true incidence of infection. Second, the
46 degree to which case counts undercount the true burden of disease is sensitive to the availability
47 of diagnostic testing, which has also varied over time and geography.[4,5,6] For this reason, it is
48 difficult to distinguish true disease trends from changes in testing practices. Third, case and death
49 counts are lagging indicators of the transmission dynamics of the pathogen, as they are affected
50 by delays associated with the disease incubation period, care-seeking behavior of symptomatic
51 individuals, diagnostic processing times, and reporting practices.

52 Complete estimates of COVID-19 cases and deaths provide insight into the size and scope of the
53 United States (US) epidemic. From total infections, one can estimate the case ascertainment and
54 how it has changed over the course of the epidemic, as well as the percentage of the population
55 that has ever been infected with SARS-CoV-2. In addition, information describing trends in
56 transmission can inform current and future COVID-19 control policies. A key metric for
57 describing transmission changes is the effective reproduction number (R_t), which represents the
58 average number of secondary infections caused by an infected individual at a given point in
59 time.[7] R_t can signal short-term changes in transmission in response to policy and behavioral

60 changes. However, estimates of R_t will be biased when the gaps and lags in reporting are
61 inconsistent over time,[5] weakening its usefulness as a measure of transmission.

62 Here, we present detailed estimates of COVID-19 trends for all US states and counties, based on
63 a Bayesian evidence synthesis model that estimates SARS-CoV-2 infection patterns from
64 observed case notifications and death reports. We apply our model to publicly available COVID-
65 19 case and death data and report on the trajectory of the epidemic from the first reported case
66 (January 13, 2020) until January 1, 2021. The model is available on GitHub (https://github.com/
67 covidestim/covidestim/) as a package for the R programming language (*covidestim*)."

68 # Results

69 **Main Findings**

70 From the start of the epidemic until January 1, 2021 we estimated COVID-19 epidemiological
71 outcomes from the time-series of cases and deaths in each US state and county, using a
72 mechanistic model to account for changes in case ascertainment as well as delays associated with
73 disease progression, diagnosis, and reporting systems. Counties and states are fit using an
74 optimization routine that reports the *maximum a posteriori* estimate, which represents a best
75 estimate of each outcome of interest, based on the available data.

76 *Incidence and R_t*

77 Nationally, the COVID-19 epidemic has been characterized by three 'waves'; however, different
78 trends are observed at the state- and county-level (Figure 1). While the third wave was larger (in
79 terms of total reported cases) nationally, the first wave was characterized by large localized
80 epidemics. The largest number of new incident infections per capita in a single day nationally

occurred on January 1 2021, when there were 255 new infections per 100,000. However, at the state level, incident infections peaked at 842 infections per 100,000 on March 24, 2020, in New York (Figure 2). On that day, 43.7% of all incident infections in the US occurred in New York State (163,732 total infections) and 29.1% of all US infections occurred in the five boroughs of New York City (108,739 infections). At the state level, the worst days of the pandemic (in terms of new infections per 100,000 residents) occurred between March 20 and April 1 in New York and New Jersey.

In contrast, the largest number of incidence infections per capita in a single day during the second wave (June 6, 2020 – August 20, 2020) occurred in Arizona on June 29, 2020, when there were 284 new infections per 100,000. On this day, 9.6% of all US infections occurred in Arizona. The largest number of incidence infections per capita in a single day during the third wave (beginning October 1, 2020) occurred in South Dakota on November 8, 20201, when there were 649 new infections per 100,000. On this day, just 1.3% of all US infections occurred in South Dakota.

Figure 1: 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) and reported diagnoses (blue) per 100,000 people per day from March 1, 2020 to January 1, 2021. We assume that there were no reported cases or deaths prior to the first date indicated in the data.

Figure 2: Incident infections per 100,000 residents per day (blue, left y-axis) and the percentage of the population ever-infected (red) for each US state from March 1, 2020 to January 1, 2021. Numbers in red indicate the percent ever-infected on January 1, 2021.

At the county-level, the largest number of new infections per 100,00 residents in a single day occurred on November 29, 2020 in Comanche County, KS (population 17,000), when we estimate there were 2342 new infections per 100,000 (40 new infections). Restricting this analysis to the 143 most populous counties where 50% of the US population lives,[8] the date with the highest incidence occurred on March 28, 2020 in Queens County, NY, when there were 1765 new infections per 100,000 population (39,791 new infections). The largest estimated number of new infections in a single day was 68,033 in Los Angeles County, California on December 28. From December 9, 2020 to January 1, 2021, Los Angeles County had the highest daily estimate of new SARS-CoV-2 infections of any county over the course of the epidemic.

The median state-level estimate of R_t on the first day a case was reported in each state was 3.63 (range: 1.93 in Washington to 8.26 in New York). By April, R_t estimates had dropped substantially; from April 1, 2020 to January 1, 2021, state-level estimates of R_t were between 0.56 and 1.68 (Figure 3). At the county level, R_t estimates were more variable after April 1, 2020, ranging from 0.35 (Wibaux County, MT on January 1, 2021) to 5.65 (Dakota County, NE on April 1, 2020).

Figure 3: 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 $R_t = 1$.

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 (Figure 4). Nationally, we estimate that 26.7% of the US population had been infected with SARS-CoV-2 by January 1,

125 2021. Across states, the percent ever-infected ranged from 6.4% (Hawaii) to 45.9% (Arizona)
126 (Figure 2). Across counties, the estimated percentage of the population ever-infected ranged
127 from 0.8% (Kauai, HI) to 76.2% (Lamb County, TX). By January 1 2021, we found that the
128 percent of the population ever-infected exceeded 50% in 241 (7.7%) counties and exceeded two-
129 thirds of the population in just 22 (0.7%) counties. Conversely, the percent ever-infected was less
130 than 10% in 145 (4.6%) counties and less than 5% in 34 (1.1%) counties.

131 On January 1, 2021, the US had reported 349,247 cumulative COVID-19 deaths.[9] For the same
132 date, our analysis estimated 375,123 cumulative COVID-19 deaths, 7.4% greater than
133 cumulative reported deaths and approximately 0.11% of the US population in 2019.

134 **Figure 4:** Percentage of the population ever infected with SARS-CoV-2 as of January 1, 2021.

135 *Probability of Diagnosis*

136 In March and April 2020, the median probability of diagnosis for infected individuals across
137 states was 11.8% (range: 3.4%, 49.0%) and 11.1% (range: 5.6%, 41.7%), respectively (Figure 5).
138 Case ascertainment was lower than the national median in New York and New Jersey, where
139 6.9% and 4.9%, respectively, of infections were diagnosed and reported in March. Over the
140 period March 1 to April 31, 2020, the median probability of diagnosis across the five boroughs
141 of New York City was 5.3%; for every 1 reported case in New York City during this period,
142 there were 20 SARS-CoV-2 infections.

143 **Figure 5:** The probability that a person infected with SARS-CoV-2 on a given day will be
144 diagnosed (black) and the 7-day moving average of the fraction of COVID-19 antigen tests that
145 are positive (red), for each US state from March 1, 2020 to January 1, 2021.

146 From the introduction of SARS-CoV-2 in the US until July 1, 2020, there were approximately 7
147 infections for every reported case of SARS-CoV-2 nationally (14.0% of cases were diagnosed).
148 Over the month of July, the median state probability of diagnosis was 32.4% (range: 12.6%,
149 49.34%). In areas with high incidence of SARS-Cov-2 infections during the second wave,
150 ascertainment was lower than the national median; for example, the median probability of
151 diagnosis in July was 16.6% (range: 11.8%, 18.2%) in Arizona and 13.6% (range: 10.6%,
152 24.9%) in Texas.

153 From July 1, 2020 through January 1, 2021 the median state probability of diagnosis was 32.5%
154 (range: 17.7%, 50.2%). From the beginning of the epidemic until January 1, 2021, we estimate
155 that the median probability that a SARS-CoV-2 infection is diagnosed is 26.3%. We estimate
156 that there have been 88.4 million SARS-CoV-2 infections as of January 1, 2021, far greater than
157 the 20.0 million reported cases of COVID-19.

158 Figure 5 compares trends in the probability of diagnosis (estimated in this analysis) and the
159 reported fraction of all COVID-19 antigen tests that were positive. These results indicate that the
160 relationship between test positivity and probability of diagnosis is not a simple linear
161 relationship, making test positivity an imperfect proxy for the level of undiagnosed disease in the
162 community.

163 **Comparisons to External Estimates**

164 We compared our estimates of the percent ever infected with SARS-CoV-2 to U.S. Centers for
165 Disease Control (CDC) seroprevalence estimates based on commercial laboratory data.[10]
166 Derived from a convenience sample of blood specimens collected for reasons unrelated to
167 COVID-19, these estimates provide state-level evidence on SARS-CoV-2 antibody test positivity

168 at multiple time points (Figure 6). However, these estimates are incomplete in some states (e.g.
169 South Dakota) and the series deviate implausibly from the expectation of increasing
170 seroprevalence over time in others (e.g. New York). Model estimates of the percent ever infected
171 are higher than the CDC seroprevalence estimates for most states. Comparing these estimates to
172 other reported indicators of cumulative disease burden on January 1 2021, the modelled
173 estimates of the percent ever infected were more strongly correlated with cumulative
174 hospitalizations (Spearman rank correlation (ρ) = 0.63) and cumulative reported deaths (ρ =
175 0.83) than the CDC seroprevalence estimates (ρ = 0.41 and 0.37 for hospitalizations and deaths
176 respectively).

177 **Figure 6:** Comparison of the estimated percent ever infected with SARS-CoV-2 (blue) to CDC
178 seroprevalence estimates from commercial laboratory data (red) for each US state from August
179 1, 2020 to January 1, 2021.

180 In addition, we compared model estimates of cumulative COVID-19 deaths (detected and
181 undetected) to state-level estimates of excess all-cause mortality[6] (Figure 7) at each weekly
182 timepoint from March 7 to December 19, 2020. We expect COVID-19 deaths to follow
183 trajectories similar to, but less than, estimated excess all-cause mortality, which also includes
184 non-COVID-19 deaths linked to social disruption caused by the pandemic.[6] On average, the
185 modeled estimate of cumulative COVID-19 deaths is less than or approximately equal to
186 estimates of excess all-cause mortality. Notably, three states (Alaska, Hawaii, Maine) have
187 extended periods where the all-cause mortality was not estimated to exceed all-cause mortality
188 from previous years (i.e. excess mortality was negative); in periods where all-cause mortality is
189 higher than expected, our estimates of COVID-19 deaths approximate the excess mortality
190 estimates. Additionally, model estimates of cumulative COVID-19 deaths exceed estimates of

excess all-cause mortality in two states (Massachusetts and Rhode Island). Estimates of excess all-cause mortality were not available for Connecticut, North Carolina, or West Virginia.

Figure 7: 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.

Discussion

We present detailed estimates of the time-course of SARS-CoV-2 infections in US states and counties over 2020, accounting for under-ascertainment and delays in reported outcomes. Case ascertainment varied over space and time; the probability that an infected individual is diagnosed was three times higher over July 1, 2020 to January 1, 2021 that it was over March 1, 2020 to July 1, 2020. Even as case ascertainment improved over the course of the epidemic, the probability of diagnosis typically decreased in jurisdictions experiencing a surge of new infections; a possible explanation is that testing availability still lags behind demand. Overall, we estimated that official case counts represent a fraction of total infections. As of January 1, 2021, we estimate that 26.7% of the US population has been infected with SARS-CoV-2 and 0.11% of the US population has died from their infection. The modeled estimates of the percent ever infected are higher than CDC seroprevalence estimates and more strongly correlated with cumulative hospitalizations and deaths across states, potentially reflecting biases in the empirical seroprevalence estimates. Specific limitations in the observed seroprevalence series include the use of non-representative samples,[11] and reduced sensitivity associated with waning of antibody titers, as has been reported for some tests.[12,13] The comparison between modeled estimates and seroprevalence data suggest that the model may add valuable information on the fraction of the

population that has been infected over time, via synthesis of multiple data sources with adjustment for biases in various observation mechanisms.

Our approach makes a number of simplifying assumptions related to delays in disease progression and detection. The model is parameterized to estimate only one geographic unit at a time, and we do not model spread of disease between counties or states. We deliberately use the same model assumptions and data for all geographies; we do not make use of additional data that might improve estimation (such as hospitalizations or wastewater-based surveillance) because these data were not available for all jurisdictions. Because we anchor the analysis on death data (under the assumption that deaths were more consistently reported over the course of the epidemic), model estimates are sensitive to infection fatality risk (IFR) estimates, which are themselves uncertain. Furthermore, we assume the completeness of death reporting is the same across jurisdictions. Finally, while our analysis describes epidemiological changes, it does not provide insight into the causes of these changes, such as changes to "safer-at-home" policies, indoor dining regulations, and mask mandates.

Because states report cases and deaths in different ways, there are there are a number of potential inconsistencies in these data, including exclusion of antigen test results, reporting the number of positive tests (as opposed to number of individuals who have tested positive), and inconsistent lags between detection and reporting. Rarely do available metadata indicate the presence, absence, or nature of these inconsistencies, which prevented specialized model configurations. Additionally, data are subject to occasional revisions, which have frequently been implemented as a single-day change in the cumulative count of cases or deaths, without matching revisions to the historical time series. This leads to additional variance in the reported data, and a reduction in the precision of reported estimates.

235 The Bayesian estimation approach used for this analysis provides a coherent framework for
236 simultaneously estimating the trend in SARS-CoV-2 incidence and the fraction of the population
237 that has previously been infected, providing key information on the current status and past extent
238 of state or county epidemics. In contrast to other approaches for estimating R_t and COVID-19
239 epidemiological outcomes for US states[14,15] and other countries,[16] our approach allows for time-
240 trends in diagnostic coverage, which in turn allows for the extent of under-ascertainment to vary
241 over the course of the epidemic. We have open-sourced and documented our *covidestim* R
242 package on GitHub, so that others can easily apply these methods to their own data.

243 Mathematical modelling has filled critical gaps in the evidence base for COVID-19 decision-
244 making, and has been used to weigh the benefits of competing mitigation strategies, plan for the
245 deployment of healthcare resources, and infer key features of COVID-19 natural history.[14] For
246 the current study, we used modelling as a quantitative framework to synthesis multiple data
247 sources and describe COVID-19 burden and trends for a large number of locations in the United
248 States. A key finding of this work is that the has been no single US COVID-19 epidemic, but
249 instead a large number of related epidemics, differing in their timing and magnitude even within
250 individual states. While our study does not provide an explanation for the patterns of this
251 epidemic, it demonstrates the need for explanations that are relevant at a fine geographic and
252 temporal scale. The deployment of effective vaccines, and the spread of more transmissible
253 COVID-19 variants, represent reasons for both hope and concern about future COVID-19 trends
254 in the United States. Ongoing, locally-relevant evidence on these trends will continue to be
255 important for both governmental and individual decision-making.

256 **Methods**

Data

For every state and county in the US, we extracted data on reported COVID-19 cases and deaths. We used state-level data on cumulative total cases and deaths compiled by the Covid Tracking Project[18] and county-level on cumulative total cases and deaths compiled by Johns Hopkins University.[9] We calculated new cases and deaths as the difference between the cumulative counts reported each day. For days in which the difference is below zero (for example, a data audit resulting in a downward revision of the cumulative count), we adjusted the difference to zero until the cumulative count rose above the previous maximum cumulative count, such that the cumulative count increases monotonically. For counties, we maintained this monotonic property, but allowed incident cases or deaths to rise immediately after a downward revision in the cumulative count.

We used state-specific age distributions of COVID-19 related deaths[19] and estimates of age-specific infection fatality risks[20] to estimate state-specific IFRs. Additionally, we used county-level estimates of the prevalence of one or more risk factors for severe COVID-19[21] to create county-specific IFRs, by multiplying the state IFR by the ratio of risk factor prevalence in each county relative to their state.

Analytic Overview

We developed a mathematical model to estimate epidemiological measures COVID-19 based on the empirical data, accounting for under-ascertainment of cases and delays associated with care-seeking, disease progression, and reporting. We parameterized this model using available evidence, and used it to estimate SARS-CoV-2 infections and COVID-19 outcomes for all states

and counties in the US. We constructed visualizations to help reveal patterns in these outcomes and a web interface to make the estimates broadly available.

Newly infected and asymptomatic individuals may recover or progress to symptomatic disease with or without diagnosis. Individuals who progress to symptomatic, non-severe disease may recover or progress to severe disease with or without diagnosis. Finally, individuals with severe disease will either die or recover with or without diagnosis; only individuals who transit through severe disease can die. The model assumes fixed delays associated with disease progression and diagnosis, a time-varying probability of diagnosis, and complete reporting of diagnosed cases (with a fixed delay). Figure S1 shows modeled health states and possible transitions.

Infection and Natural History

The model estimates the time-series of SARS-CoV-2 effective reproduction numbers (R_t) using a cubic b-spline, allowing for flexibility in the evolution of the epidemic curve over time. We estimated the change in new infections based on the value of R_t on a given day and the serial interval. Specifically, the change in log(new infections) between two days was calculated as $\log(R_t)$ divided by the serial interval. We also adjusted this value for the fraction of the population not yet infected, penalizing high values of R_t in settings where a large fraction of the population is already infected. For an individual newly entering a given health state, their sojourn time in the health state is assumed to follow a fixed gamma distribution, and additional parameters describe the probability of further progression (versus recovery), upon exiting the health state.

Modeling Case Detection and Reporting

300 We assumed that for a case to be reported, an individual must first be tested and the positive test
301 result subsequently entered into the surveillance system. We assumed that for a death to be a
302 confirmed COVID-19 death, an individual must be tested prior to their death and their death
303 must be subsequently entered into the surveillance system. We modeled the delays associated
304 with each of these steps separately – from symptom onset to test specimen collection, from
305 symptom onset to death, from specimen collection to reporting, and from death to reporting.
306 Because data on reporting delays were not widely available, we assumed a fixed delay for the
307 reporting of COVID-19 cases and deaths.

308 The probability of diagnosis conditional on entering the *Severe* state is constant over the course
309 of the epidemic. We assumed the probability of diagnosis in the *Symptomatic* state can change
310 over time and is always less than the probability of diagnosis in the *Severe* state. We modeled a
311 rate ratio of diagnosis at *Symptomatic* compared to *Severe* using a cubic b-spline; the time-
312 varying probability of diagnosis at *Symptomatic* is the product of this rate ratio and the
313 probability of diagnosis at *Severe*. Similarly, the probability of diagnosis in the *Asymptomatic*
314 state is the product of the probability of diagnosis in the *Symptomatic* state and the time-varying
315 rate ratio of diagnosis at *Asymptomatic* compare to *Symptomatic*, also modeled with a cubic b-
316 spline. An individual in a given state on day i can be tested on day $i+j$, determined by a
317 probability of receiving a test and an associated delay, which is a modeled proportion of sojourn
318 time in that state. We assume that testing does not occur postmortem.

319 *Model priors*

320 Model priors and inputs are presented in Table S1. We used informative priors and fixed
321 distributions for parameters relating to the natural history of the disease, such as probability of

disease progression and sojourn time in each state, respectively. The reporting delay is poorly identified by currently available case report data in most locations; we adopted log-normal prior distributions for the shape and rate parameters of these gamma distributions

Model implementation

The model was implemented in R using the rstan package.[22] 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 separately. Counties are 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. The optimization routine does not produce samples, and, therefore, we do not report credible intervals around model estimates.

Acknowledgments: We thank Jeffrey Eaton for his thoughts on statistical analysis.

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
3. Oran, D.P. and Topol, E.J., 2020. Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. *Annals of internal medicine*, *173*(5), pp.362-367.
4. MDT Hitchings, NE Dean, B Garcia-Carreras, TJ Hladish, AT Huang, B Yang, DAT Cummings. The Usefulness Of SARS-CoV-2 Test-Positive Proportion As A Surveillance Tool. *American Journal of Epidemiology*, 2021, https://doi.org/10.1093/aje/kwab023
5. VE Pitzer, MH Chitwood, J Havumaki, NA Menzies, S Perniciaro, JL Warren, et al., The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. [Preprint]. 2020 [Cited 2020 July 13]. Available from: https://doi.org/10.1101/2020.04.20.20073338.
6. D Weinberger, T Cohen, F Crawford, F Mostashari, D Olson, VE Pitzer, 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. KM Gostic, L McGough, E Baskerville, S Abbott, K Joshi, C Tedijanto, et al. Practical considerations for measuring the effective reproductive number, Rt. PLoS Comput. Biol. 16(12): e1008409. https://doi.org/10.1371/journal.pcbi.1008409
8. H El Nasser. More Than Half of U.S. Population in 4.6 Percent of Counties. October 4, 2017. Retrieved from: https://www.census.gov/library/stories/2017/10/big-and-small-counties.html
9. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. [Cited 15 January 2021] Available at: https://coronavirus.jhu.edu/map.html
10. Nationwide Commercial Laboratory Seroprevalence Survey. [Cited 23 March, 2020] Available at: https://data.cdc.gov/Laboratory-Surveillance/Nationwide-Commercial-Laboratory-Seroprevalence-Su/d2tw-32xv
11. Bajema KL, Wiegand RE, Cuffe K, et al. Estimated SARS-CoV-2 Seroprevalence in the US as of September 2020. JAMA Intern Med. November 24, 2020. doi:10.1001/jamainternmed.2020.7976
12. FJ Ibarrondo, JA Fulcher, D goodman-Meza, et al. Rapid Decay of Anti-SARS-CoV-2 Antibodies in Persons with Mild Covid-19. N Engl J Med 2020; 383:1085-1087. DOI: 10.1056/NEJMc2025179
13. J Seow, C Graham, B Merrick, *et al.* Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. *Nat Microbiol* **5,** 1598–1607 (2020). https://doi.org/10.1038/s41564-020-00813-8
14. HJT Unwin, S Mishra, VC Bradley, A Gandy, M Vollmer, T Mellan, et al., Report 23: state-level tracking of COVID-19 in the United States - version 2 (28-05-2020). 2020. doi: https://doi.org/10.25561/79231
15. COVID-19 Portal, Center for the Ecology of Infection Diseases, University of Georgia [Cited 2020 July 13]. Available at: https://www.covid19.uga.edu/nowcast.html
16. S Flaxman, S Mishra, A Gandy, HJT Uwin, H Coupland, TA Mellan, et al, Report 13: estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. 2020. doi: https://doi.org/10.25561/77731.
17. LP James, JA Salomon, CO Buckee, NA Menzies. The Use and Misuse of Mathematical Modeling for Infectious Disease Policymaking: Lessons for the OCIVD-19 Pandemic. Med. Decis. Making. 2021 Feb 3;272989X21990391. doi: 10.1177/0272989X21990391
18. The COVID Tracking Project. [Cited 15 January 2021] Available at: https://covidtracking.com/
19. 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
20. O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L. *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590,** 140–145 (2021). https://doi.org/10.1038/s41586-020-2918-0
21. Razzaghi H, Wang Y, Lu H, 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. DOI: http://dx.doi.org/10.15585/mmwr.mm6929a1
22. Stan Development Team. RStan: the R interface to Stan. 2018 R package version 2.17.3. http://mc-stan.org

Funding:
KG reports grant from National Institutes of Health T32 GM007205 and Fogarty International Center D43 TW010540
VEP reports grants from National Institute of Allergy and Infectious Diseases R01 AI137093
DMW reports grants from National Institute of Allergy and Infectious Diseases R01 AI137093
JLW reports grants from National Institute of Allergy and Infectious Diseases R01 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 though the Council of State and Territorial Epidemiologists (NU38OT000297-02) and the National Institute on Drug Abuse (3R37DA01561217S1).

Author contributions:

MHC developed the methodology, visualized the results, contributed to analysis, prepared the original manuscript, and edited the manuscript. MR designed the software, visualized results, prepared the original manuscript, and edited the manuscript. KG curated data and edited the manuscript. JH curated data and edited the manuscript. VEP contributed to the analysis and edited the manuscript. JAS contributed to the analysis and edited the manuscript. NS contributed to the analysis, prepared the original manuscript, and edited the manuscript. JLW contributed to the analysis and edited the manuscript. DMW contributed to the analysis and edited the manuscript. TC conceived the project, contributed to the analysis, prepared the original manuscript, and edited the manuscript. NAM conceived the project, developed the methodology, contributed to the analysis, prepared the original manuscript, and edited the manuscript.

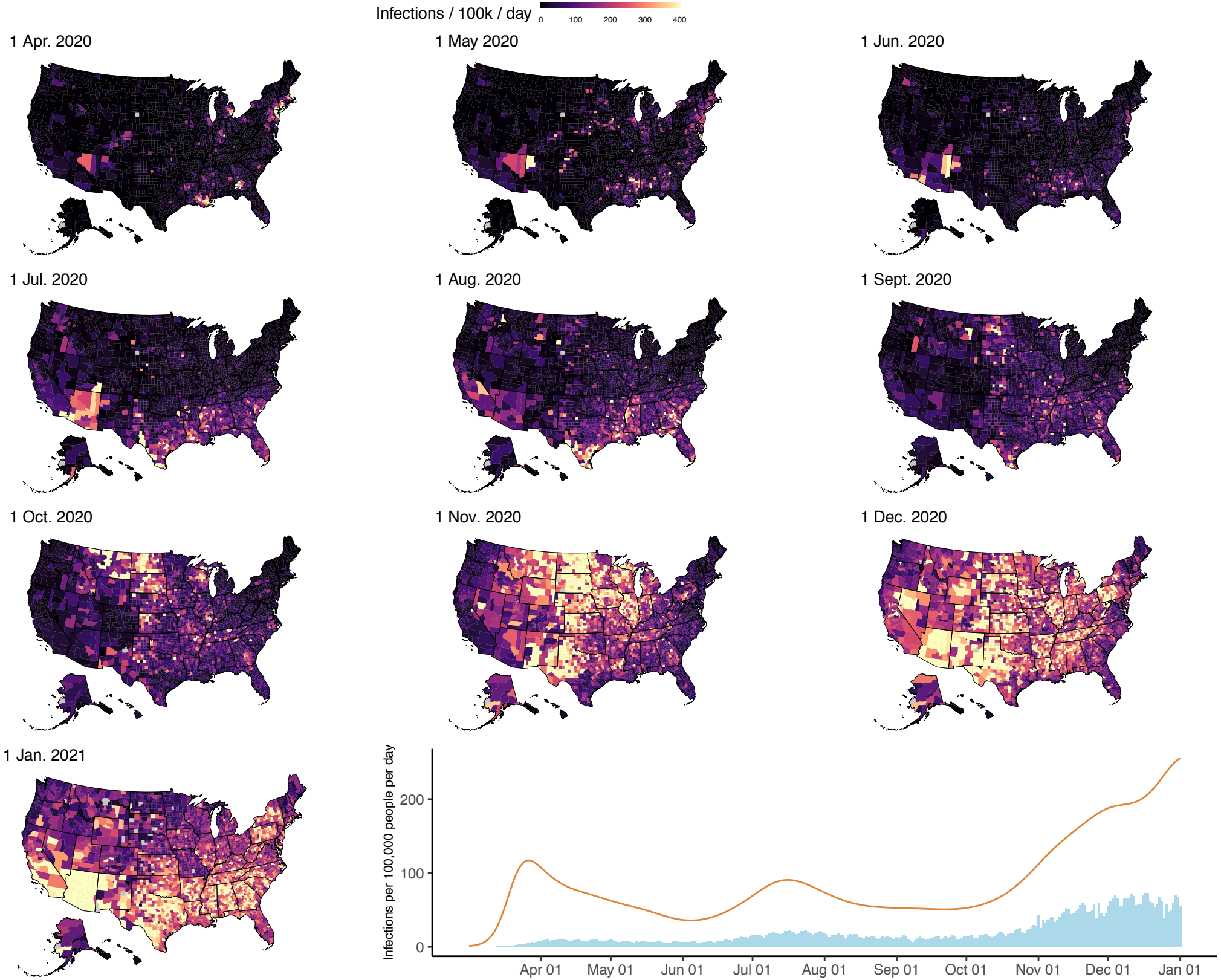
Competing interests: DMW has received consulting fees from Pfizer, Merck, GSK, and Affinivax for topics unrelated to this manuscript and is Principal Investigator on a research grant from Pfizer on an unrelated topic. VEP has received reimbursement from Merck and Pfizer for travel expenses to Scientific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

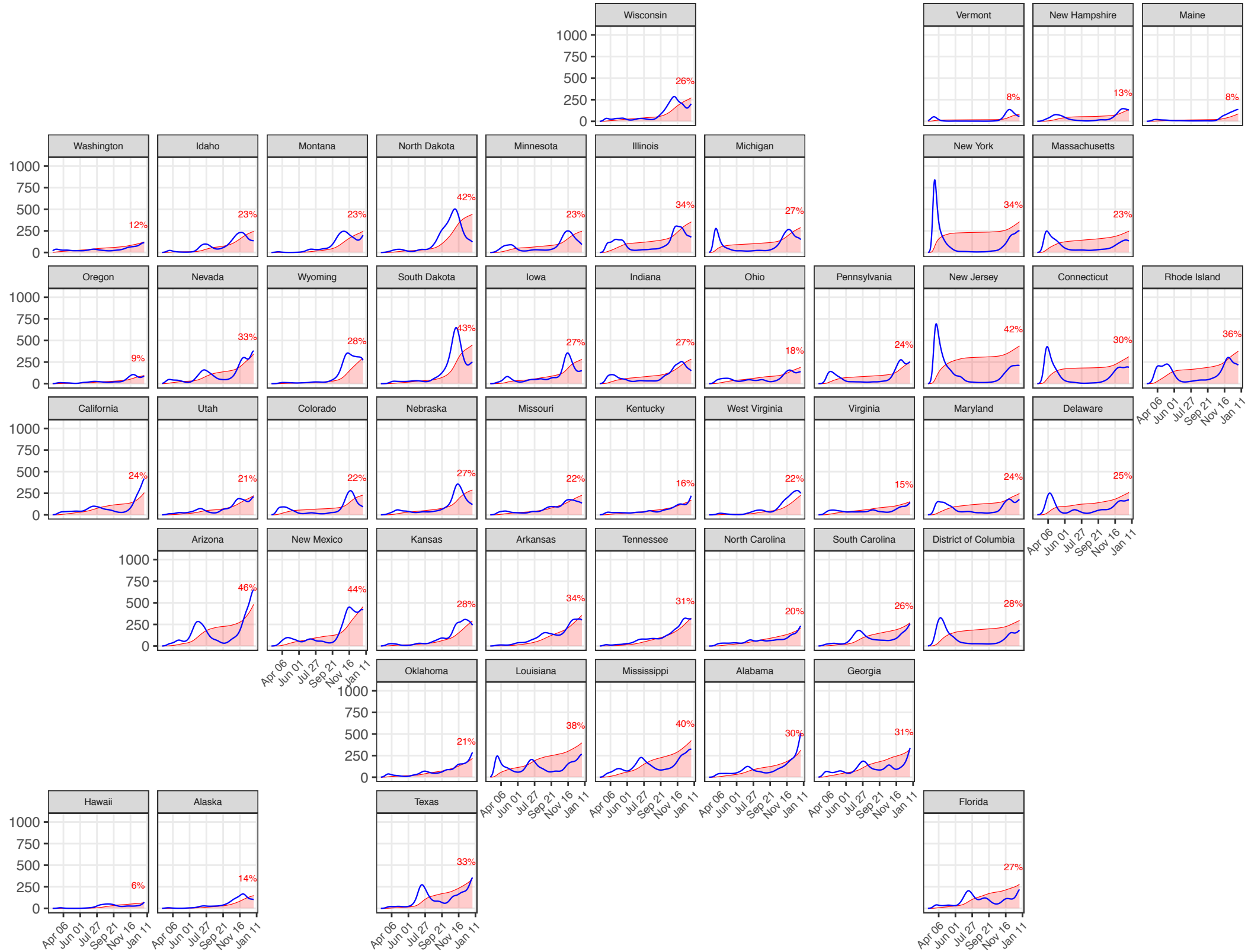
Data and materials availability: All data used in the main analysis are available for use at https://github.com/covidestim/covidestim-sources. The *covidestim* package is available for download at https://github.com/covidestim/covidestim.

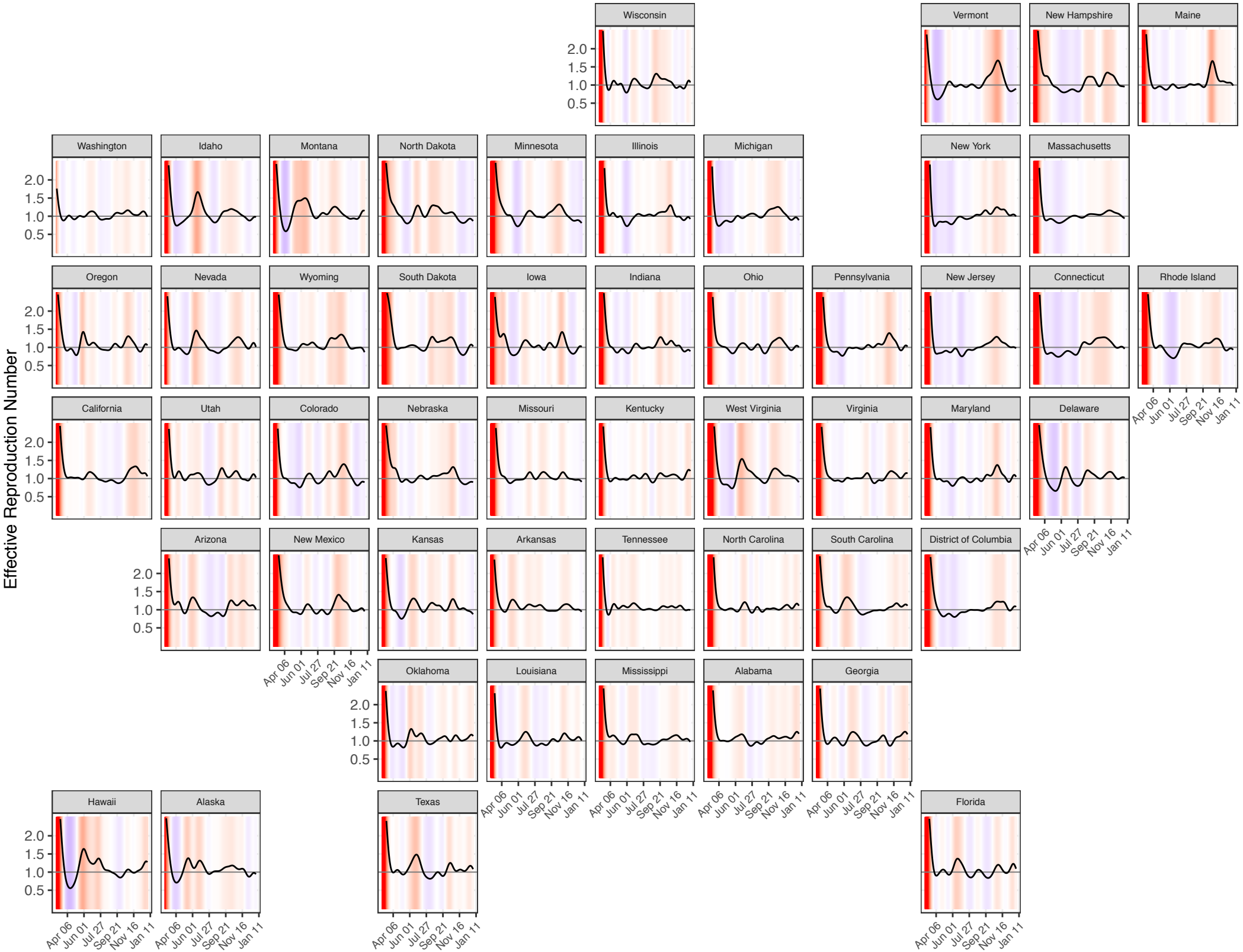
Supplementary Materials:

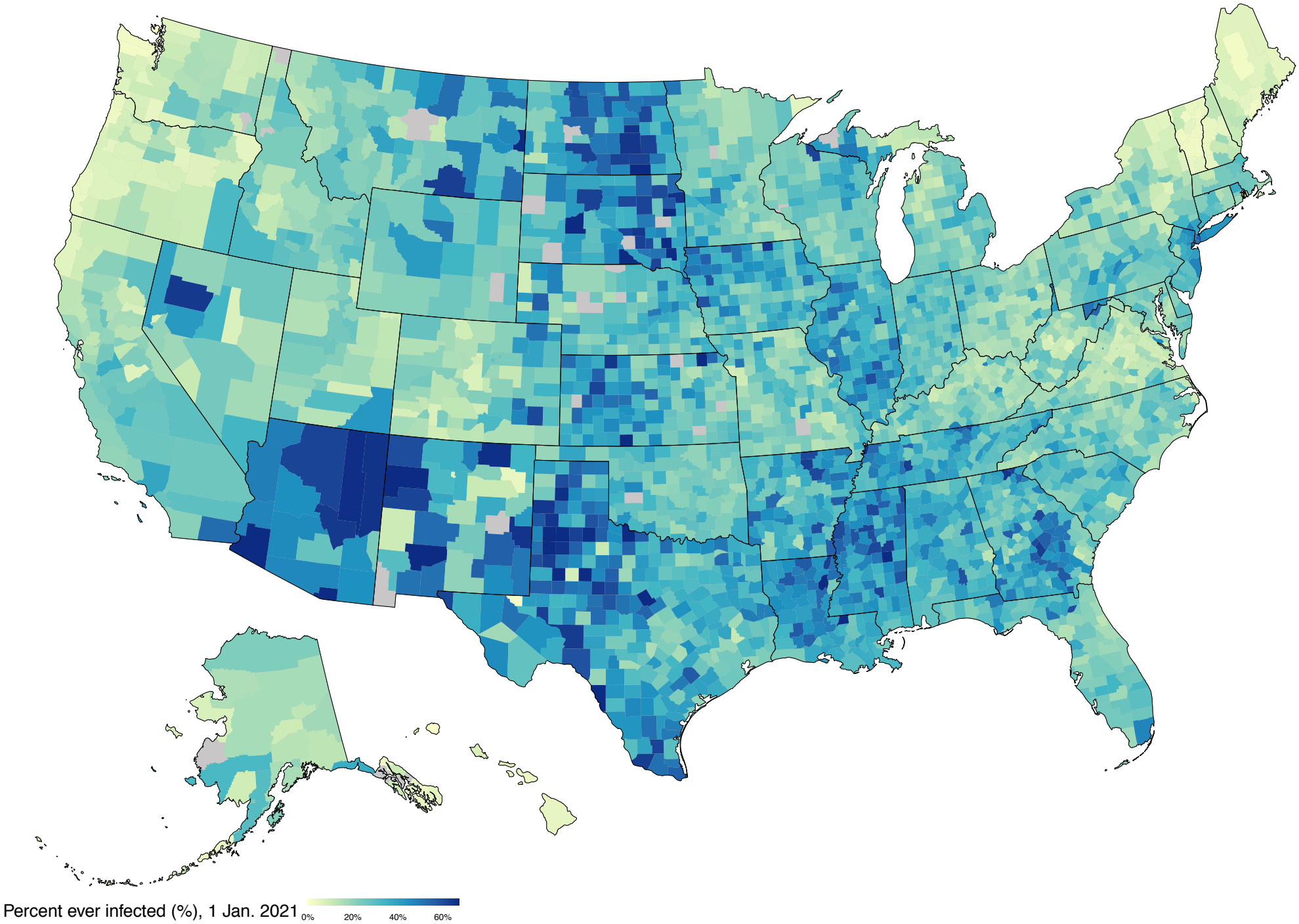
Supplementary Methods

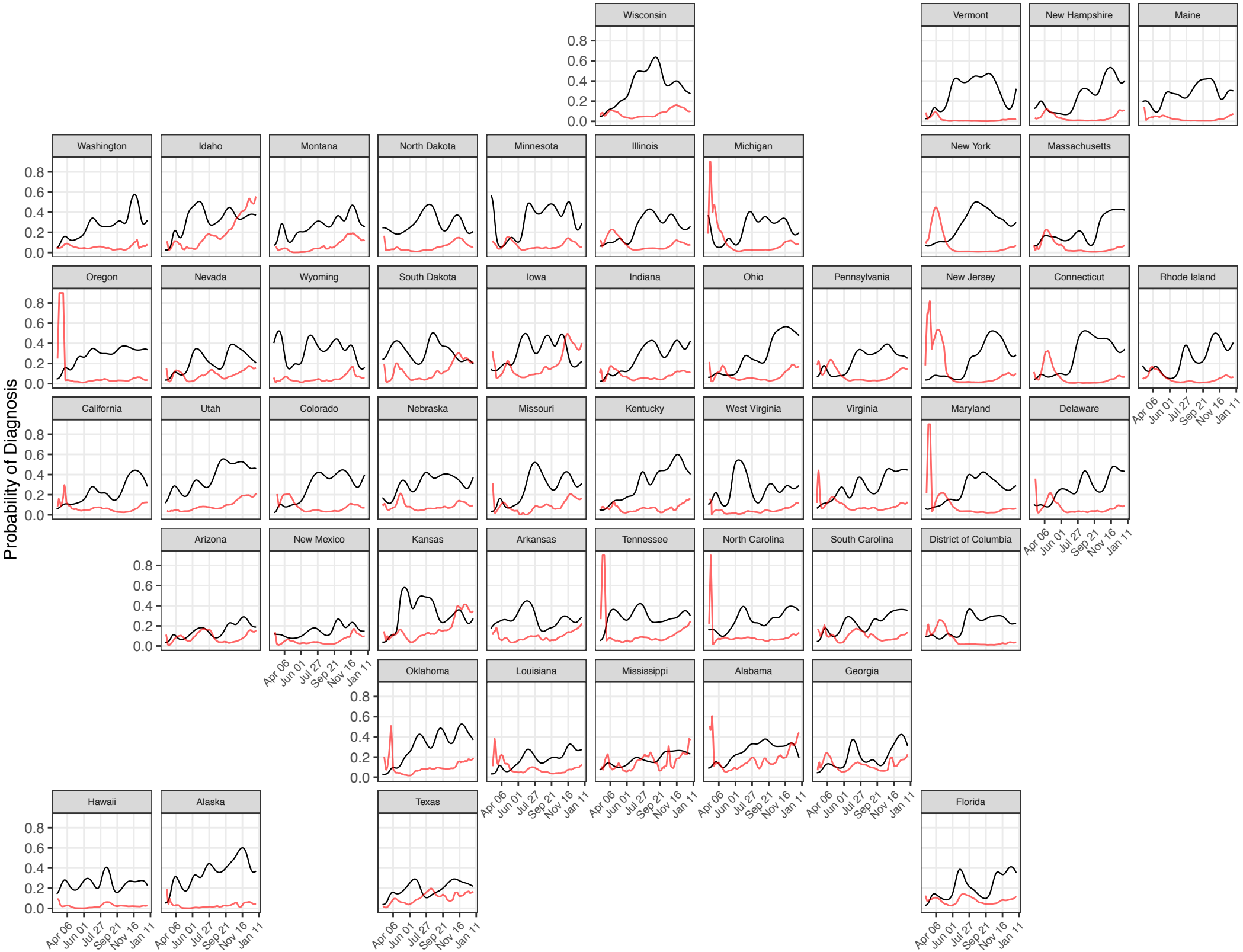
Table S1. Model Priors and Inputs

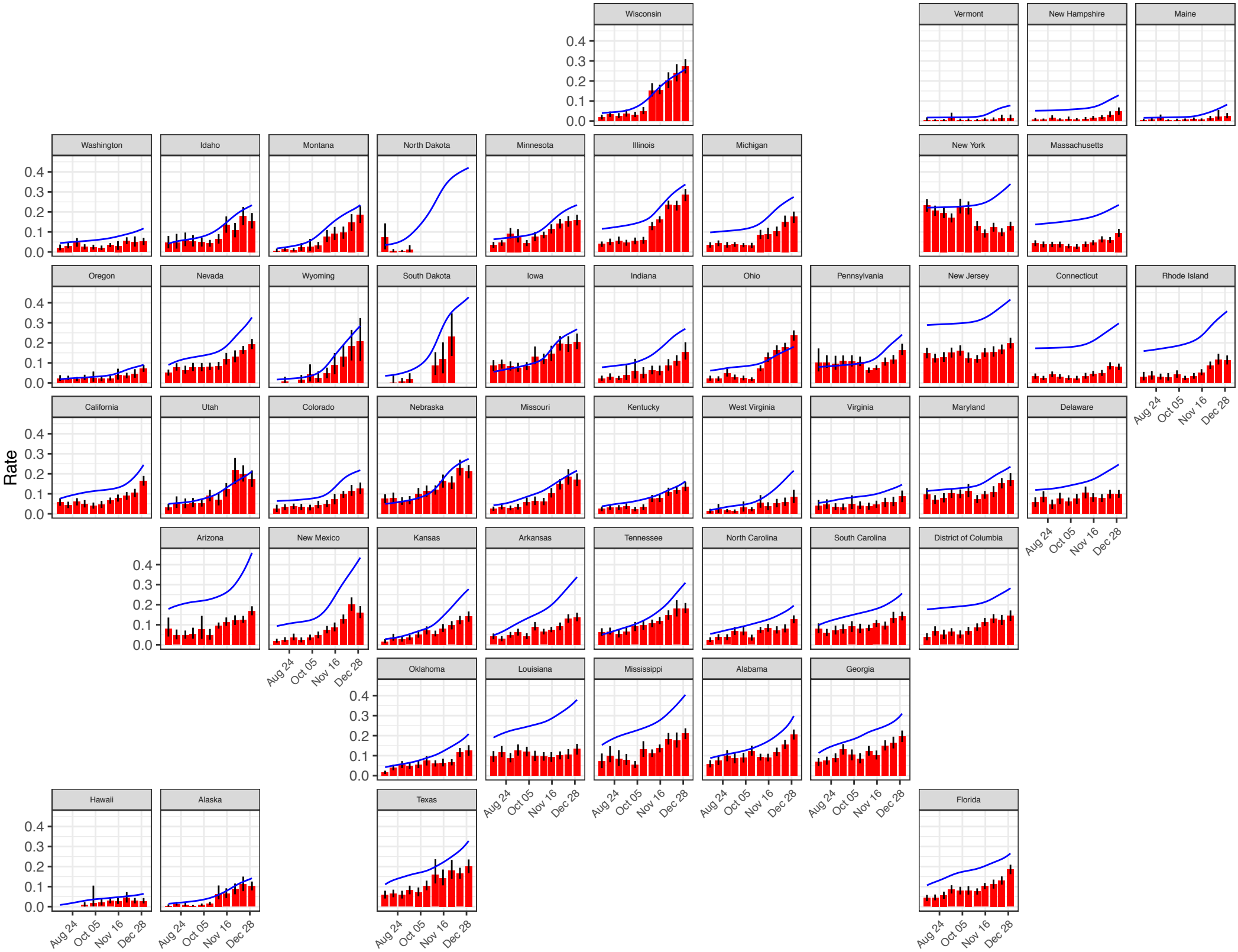

Infections / 100k / day
0
100
200
300
400
1 Apr. 2020
1 May 2020
1 Jun. 2020
1 Jul. 2020
1 Aug. 2020
1 Sept. 2020
1 Oct. 2020
1 Nov. 2020
1 Dec. 2020
1 Jan. 2021
Infections per 100,000 people per day
0
100
200
Apr 01
May 01
Jun 01
Jul 01
Aug 01
Sep 01
Oct 01
Nov 01
Dec 01
Jan 01

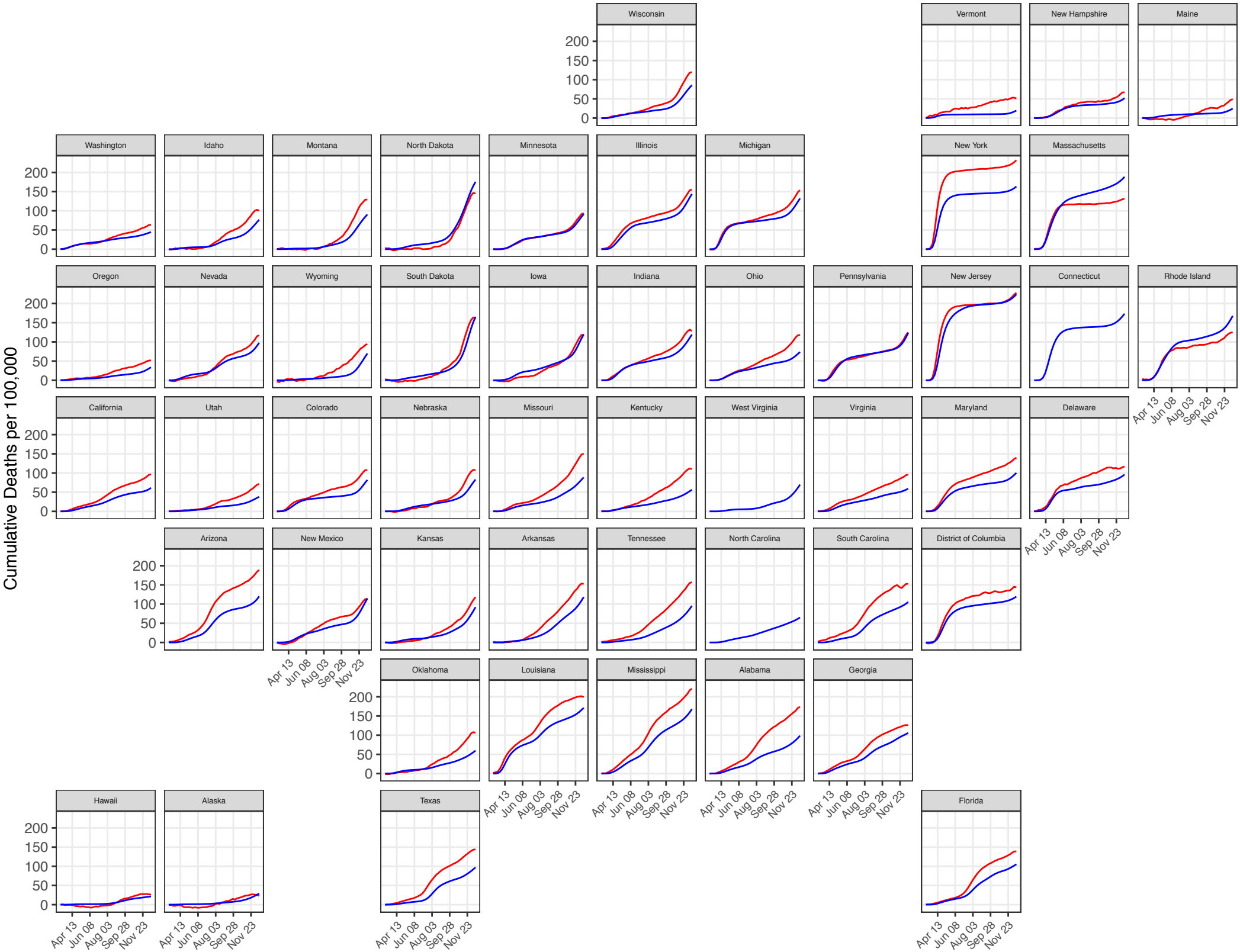

Wisconsin
26%
Vermont
8%
New Hampshire
13%
Maine
8%
Washington
12%
Idaho
23%
Montana
23%
North Dakota
42%
Minnesota
23%
Illinois
34%
Michigan
27%
New York
34%
Massachusetts
23%
Oregon
9%
Nevada
33%
Wyoming
28%
South Dakota
43%
Iowa
27%
Indiana
27%
Ohio
18%
Pennsylvania
24%
New Jersey
42%
Connecticut
30%
Rhode Island
36%
California
24%
Utah
21%
Colorado
22%
Nebraska
27%
Missouri
22%
Kentucky
16%
West Virginia
22%
Virginia
15%
Maryland
24%
Delaware
25%
Arizona
46%
New Mexico
44%
Kansas
28%
Arkansas
34%
Tennessee
31%
North Carolina
20%
South Carolina
26%
District of Columbia
28%
Oklahoma
21%
Louisiana
38%
Mississippi
40%
Alabama
30%
Georgia
31%
Hawaii
6%
Alaska
14%
Texas
33%
Florida
27%
1000
750
500
250
0
Apr 06
Jun 01
Jul 27
Sep 21
Nov 16
Jan 11


Effective Reproduction Number
Wisconsin
Vermont
New Hampshire
Maine
Washington
Idaho
Montana
North Dakota
Minnesota
Illinois
Michigan
New York
Massachusetts
Oregon
Nevada
Wyoming
South Dakota
Iowa
Indiana
Ohio
Pennsylvania
New Jersey
Connecticut
Rhode Island
California
Utah
Colorado
Nebraska
Missouri
Kentucky
West Virginia
Virginia
Maryland
Delaware
Arizona
New Mexico
Kansas
Arkansas
Tennessee
North Carolina
South Carolina
District of Columbia
Oklahoma
Louisiana
Mississippi
Alabama
Georgia
Hawaii
Alaska
Texas
Florida
2.0
1.5
1.0
0.5
Apr 06
Jun 01
Jul 27
Sep 21
Nov 16
Jan 11


Percent ever infected (%), 1 Jan. 2021
0%
20%
40%
60%


Probability of Diagnosis
Wisconsin
Vermont
New Hampshire
Maine
Washington
Idaho
Montana
North Dakota
Minnesota
Illinois
Michigan
New York
Massachusetts
Oregon
Nevada
Wyoming
South Dakota
Iowa
Indiana
Ohio
Pennsylvania
New Jersey
Connecticut
Rhode Island
California
Utah
Colorado
Nebraska
Missouri
Kentucky
West Virginia
Virginia
Maryland
Delaware
Arizona
New Mexico
Kansas
Arkansas
Tennessee
North Carolina
South Carolina
District of Columbia
Oklahoma
Louisiana
Mississippi
Alabama
Georgia
Hawaii
Alaska
Texas
Florida
0.0
0.2
0.4
0.6
0.8
Apr 06
Jun 01
Jul 27
Sep 21
Nov 16
Jan 11


Rate
Wisconsin
Vermont
New Hampshire
Maine
Washington
Idaho
Montana
North Dakota
Minnesota
Illinois
Michigan
New York
Massachusetts
Oregon
Nevada
Wyoming
South Dakota
Iowa
Indiana
Ohio
Pennsylvania
New Jersey
Connecticut
Rhode Island
California
Utah
Colorado
Nebraska
Missouri
Kentucky
West Virginia
Virginia
Maryland
Delaware
Arizona
New Mexico
Kansas
Arkansas
Tennessee
North Carolina
South Carolina
District of Columbia
Oklahoma
Louisiana
Mississippi
Alabama
Georgia
Hawaii
Alaska
Texas
Florida
0.0
0.1
0.2
0.3
0.4
Aug 24
Oct 05
Nov 16
Dec 28


Cumulative Deaths per 100,000
Wisconsin
Vermont
New Hampshire
Maine
Washington
Idaho
Montana
North Dakota
Minnesota
Illinois
Michigan
New York
Massachusetts
Oregon
Nevada
Wyoming
South Dakota
Iowa
Indiana
Ohio
Pennsylvania
New Jersey
Connecticut
Rhode Island
California
Utah
Colorado
Nebraska
Missouri
Kentucky
West Virginia
Virginia
Maryland
Delaware
Arizona
New Mexico
Kansas
Arkansas
Tennessee
North Carolina
South Carolina
District of Columbia
Oklahoma
Louisiana
Mississippi
Alabama
Georgia
Hawaii
Alaska
Texas
Florida
0
50
100
150
200
Apr 13
Jun 08
Aug 03
Sep 28
Nov 23

Supplementary Materials

Supplementary Methods

State and county-level IFR estimates

We estimated the IFR for each state based on the distribution of recorded COVID-19 deaths in the state[1] and age-specific IFR estimates.[2] For each state we divided the number of deaths in each age group by the IFR estimate for that age group, to produce the implied age distribution of infections. The state-average IFR was then calculated by averaging the age-specific IFR values, weighted by the implied fraction of infections in each age group. We rescaled these state-level values to produce a national average IFR value of 0.5%. 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 comorbidities that predispose to COVID-19 mortality in the county[3] relative to the rest of the state. In addition, for all modelled locations we assumed higher IFR values earlier in the epidemic, reflecting deficiencies in diagnosis and care at this time. We parameterized this as a rate ratio applied to the probability of death among individuals with severe disease. This rate ratio declined from 2.34 [1.69, 3.19] at the beginning of 2020 to a value of 1.00 by the middle of the year, based on the ratio of reported COVID-19 deaths to hospitalizations prior to May 1 2020 compared to the subsequent 6 months. The time trend in this function was calculated as 1.0 minus the Normal Cumulative Distribution Function, centered on May 1 2020 with a standard deviation of 3 weeks.

Model details

We modeled the flow of infected individuals through four health states: asymptomatic and pre-symptomatic infection (*Asymptomatic*), symptomatic disease that is mild to moderate and would

not require hospitalization (*Symptomatic*), severe disease that would likely require hospitalization (*Severe*), and finally death (*Death*). Upon infection, all individuals enter the *Asymptomatic* state. Gamma density functions describing the probability that an individual exits a health state *j* days after entering it are given by $\theta_{AsySym,j}$, $\theta_{SymSev,j}$, $\theta_{SevDie,j}$ for *Asymptomatic* (*Asy*), *Symptomatic* (*Sym*), and *Severe* (*Sev*) states, respectively (where θ_j represents the probability mass obtained by integrating the gamma density between *j* and *j+1*). The probability that an individual will progress from *Asymptomatic* to *Symptomatic*, from *Symptomatic* to *Severe*, and from *Severe* to *Death* are given by *P(Sym|Asy)*, *P(Sev|Sym)*, and *P(Die|Sev)*, respectively:

$$Sym_{i+j} = \sum_{i,j} Asy_i * P(Sym|Asy) * \theta_{AsySym,j}$$

$$Sev_{i+j} = \sum_{i,j} Sym_i * P(Sev|Sym) * \theta_{SymSev,j}$$

$$Death_{i+j} = \sum_{i,j} Sev_i * P(Die|Sev) * \theta_{SevDie,j}$$

The number of individuals entering a health state on day *i+j* is equal to the product of the number of individuals on day *i* who could enter that state, the probability of entering the state, and the probability that they entered the health state on day *j*, summed over all *i,j* combinations. We assume a maximum possible sojourn time of 30 days for each disease state. The model is initialized 28 days prior to the first day for which data are available, given that observed data (diagnoses and deaths) are lagged relative to incidence.

Individuals can be diagnosed from any health state except *Death*. We assumed a time-invariant probability of diagnosis for individuals entering the *Severe* state without a diagnosis, and time-varying probabilities of diagnosis for undiagnosed individuals in the *Asymptomatic* and

Symptomatic states. The time-varying probabilities were modeled with a rate-ratio and were restricted such that:

$$P(Dx|Sev) > P(Dx|Sym)_i > P(Dx|Asy)_i$$

Where $P(Dx|Asy)_i$, $P(Dx|Sym)_i$, and $P(Dx|Sev)$ are the probability of diagnosis for undiagnosed individuals in each state (i.e. total probability of diagnosis before they exit the health state). The number of newly-diagnosed individuals for a given health state on day $i+j$ is equal to the product of the number of undiagnosed individuals entering a health state on day i, the probability of diagnosis from that state, and the probability that an individual will have been diagnosed on day j, summed over all i,j combinations:

$$DxAsy_{i+j} = \sum_{i,j} \mathrm{Asy_i} * P(Dx|Asy)_{i+j} * \rho_{Asy,j}$$

$$DxSym_{i+j} = \sum_{i,j} \left(\mathrm{Sym_i} - \left(\sum_{k,l} Asy_k * P(Dx|Asy)_{k+l} * P(Sym|Asy) * \theta_{AsySym,l} \right) \right) * P(Dx|Sym)_{i+j} * \rho_{Sym,j}$$

$$DxSev_{i+j} = \sum_{i,j} \left(\mathrm{Sev_i} - \left(\sum_{k,l} Sym_k * P(Dx|Sym)_{k+l} * P(Sev|Sym) * \theta_{SymSev,l} \right) \right) * P(Dx|Sev) * \rho_{Sev,j}$$

where *DxAsy DxSym,* and *DxSev* are new diagnoses at *Asymptomatic*, *Symptomatic* and *Severe*, respectively, and $\rho_{Asy,j}$, $\rho_{Sym,j}$ and $\rho_{Sev,j}$ are the diagnostic delays. We model the cascade of outcomes after testing by the health state at which the individual was tested. Therefore, we estimated detected cases on day i as the sum of cases diagnosed in each health state on a given

day i and detected deaths as the sum of cases diagnosed at any health state on any previous day that progress to *Death* on day i.

Data likelihood

We used a negative binomial likelihood function for observed cases and deaths, with the mean values of these functions being the modeled estimate for each quantity. We estimated the expected number of cases reported on a given day as the convolution of the time series of detected cases and the reporting delay distribution for cases. A similar approach was used to estimate the expected number of deaths reported on a given day:

$$ObservedCases_i \sim NegBin\left(\sum_j DetectedCases_{i-j} * \psi_{Case,j}, \phi_{Cases}\right)$$

$$ObservedDeaths_i \sim NegBin\left(\sum_j DetectedDeaths_{i-j} * \psi_{Death,j}, \phi_{deaths}\right)$$

where ψ_j is the probability of reporting j days after the event (testing or death) and ϕ is the negative binomial dispersion parameter. To account for variation in daily reported cases and deaths, we used a seven-day moving average of input data and modeled values in the likelihood function, and divided the value of the log-likelihood by the number of days in the moving average to avoid misrepresenting the strength of evidence.

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 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 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 virtually any HPC or cloud environment, and is the easiest way to begin using the *covidestim* R package, especially at scale. The *covidestim-sources* repository enables automated, version-controlled, and 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[4] 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, have proved stable over roughly 400,000 cpu-hours of production use, and contain extensive documentation.

Citations

1. NCHS. Provisional COVID-19 Death Counts by Sex, Age, and State. Retrieved from: https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku

2. O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L. *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590,** 140–145 (2021). https://doi.org/10.1038/s41586-020-2918-0
3. Razzaghi H, Wang Y, Lu H, 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. DOI: http://dx.doi.org/10.15585/mmwr.mm6929a1
4. Di Tommaso, P., Chatzou, M., Floden, E. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35,** 316–319 (2017). https://doi.org/10.1038/nbt.3820

Table S1. Model Priors and Inputs

DESCRIPTION	MEAN, STD. DEVIATION	DISTRIBUTION	TYPE	SOURCE
New infections intercept (log scale)	0,10	Normal	prior	Assumed
R_t intercept (log scale)	0,3	Normal	prior	Assumed
First derivative of R_t spline (log scale)	0,0.5	Normal	prior	Assumed
Second derivative of R_t spline (log scale)	0,0.1	Normal	prior	Assumed
Serial interval	5.8, 0.5	Gamma	prior	*1*
Probability of developing symptoms if infected	0.59 0.16	Beta	prior	*2-4*
Probability of becoming severely ill if symptomatic	0.09, 0.06	Beta	prior	*5*
Probability of death for all infections	Variable	Beta	prior	*5,6*
Probability of death for severe infections	0.15, 0.03	Beta	prior	*7*
Time from infected to symptomatic (days)	5.6, 3.1	Gamma	fixed	*8*
Time from symptomatic to severe (days)	7.5, 5.8	Gamma	fixed	*9*
Time from severe to death (days)	9.1, 6.3	Gamma	fixed	*9*
Scaling factor: time to diagnosis relative to time in symptomatic state	0.5, 0.22	Beta	prior	Assumed
Scaling factor: time to diagnosis relative to time in severe state	0.5, 0.22	Beta	prior	Assumed
Rate Ratio, diagnosis at asymptomatic vs. symptomatic	0.1, 0.07	Beta	prior	Assumed
Rate Ratio, diagnosis at symptomatic vs. severe	0.5, 0.22	Beta	prior	Assumed
Probability of diagnosis at severe	0.72, 0.16	Beta	prior	Assumed
Reporting delay*	2.2, 1.5	Lognormal	prior	*10*
Dispersion parameter $(1/\phi)^2$ *	0,1	Normal	prior	Assumed

* *separate model parameters for each reported event type*

Citations

1. He, X., Lau, E.H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y.C., Wong, J.Y., Guan, Y., Tan, X. and Mo, X. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. Nature medicine, 26(5), pp.672-675.
2. K Mizumoto, K Kagaya, A Zarebski, G Chowell, Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eur Surveill. 5(10): 2000180 (2020).
3. Poletti, P., Tirani, M., Cereda, D., Trentini, F., Guzzetta, G., Sabatino, G., Marziano, V., Castrofino, A., Grosso, F., Del Castillo, G. and Piccarreta, R., 2020. Probability of symptoms and critical disease after SARS-CoV-2 infection. arXiv preprint arXiv:2006.08471.
4. Oran, D.P. and Topol, E.J., 2020. Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. Annals of internal medicine, 173(5), pp.362-367.
5. R Verity, LC Okell, I Dorigatti, P Winskill, C Whittaker, N Imai, et al., Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases 20(6): 669 – 677 (2020).
6. O'Driscoll, M., Dos Santos, G.R., Wang, L., Cummings, D.A., Azman, A.S., Paireau, J., Fontanet, A., Cauchemez, S. and Salje, H., 2020. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature, pp.1-6.'
7. COVID-19 Pandemic Planning Scenarios. Retrieved from: https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html
8. McAloon, C., Collins, Á., Hunt, K., Barber, A., Byrne, A.W., Butler, F., Casey, M., Griffin, J., Lane, E., McEvoy, D. and Wall, P., 2020. Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. BMJ open, 10(8), p.e039652.
9. CDC COVID-19 Response Team, Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020. MMWR Morb Mortal Wkly Rep; 69:343-346 (2020).
10. Santa Clara County COVID-19 Testing Dashboard. Available at: https://www.sccgov.org/sites/covid19/Pages/dashboard.aspx#testing [Accessed April 8, 2020]