medRxiv preprint doi: https://doi.org/10.1101/2020.06.17.20133983; this version posted June 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Title: Bayesian nowcasting with adjustment for delayed and incomplete reporting to estimate COVID-19 infections in the United States

Authors: Melanie H. Chitwood[1†*], Marcus Russi[1†], Kenneth Gunasekera[1], Joshua Havumaki[1], Virginia E. Pitzer[1], Joshua L. Warren[2], 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, CT USA
[2]Department of Biostatistics and Public Health Modeling Unit, Yale School of Public Health, Yale University, New Haven, CT USA
Department of Global Health and Population, Harvard T.H. Chan School of Public Health, [3]Harvard University, Boston, MA USA
[†]These authors contributed equally to this work
[‡]These authors share senior authorship
[*]Corresponding authors

Abstract: Real-time estimates of the true size and trajectory of local COVID-19 epidemics are key metrics to guide policy responses. We developed a Bayesian nowcasting approach that explicitly accounts for reporting delays and secular changes in case ascertainment to generate real-time estimates of COVID-19 epidemiology on the basis of reported cases and deaths. Using this approach, we estimate time trends in infections, symptomatic cases, and deaths for all 50 US states and the District of Columbia from early-March through June 11, 2020. At the beginning of June, our best estimates of the effective reproduction number (R_t) are close to 1 in most states, indicating a stabilization of incidence, but there is considerable variability in the level of incidence and the estimated proportion of the population that has already been infected.

One Sentence Summary: A new method to track epidemiologic measures of COVID-19, available in the *covidestim* package for R.

Main Text:

The number of newly-diagnosed cases and confirmed COVID-19 deaths are the most easily observed measures of the health burden associated with COVID-19. However, the early response to COVID-19 in the United States (US) was hindered by limited access to reliable diagnostics (*1*). Because of these limitations, the number of reported cases and deaths over this period will have underestimated the true size of the epidemic, with some individuals developing, and potentially dying from, COVID-19 without receiving a lab-confirmed diagnosis (*2*). Furthermore, while the number of cases and deaths describes the magnitude of health effects, they are a lagged indicator of the transmission dynamics of the pathogen, affected by delays associated with the disease incubation period, care-seeking behavior of symptomatic individuals, diagnostic processing times, and reporting practices.

Near real-time estimates of the size and trajectory of the local COVID-19 epidemic are key metrics to guide state-level policy about stay-at-home orders and physical distancing recommendations. Ideally, information describing trends in transmission, rather than trends in reported cases or deaths, would be available to track the impact of disease control efforts. A key metric for describing transmission changes is the effective reproduction number (R_t), which represents the average number of secondary cases caused by a case at a given point in time. An

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.17.20133983; this version posted June 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

R_t less than 1 indicates a declining epidemic, while an R_t greater than 1 indicates a growing epidemic (*3*). The delayed and incomplete nature of reported cases and deaths means that these observable outcomes are not ideal for producing reliable estimates of R_t. Moreover, rapid changes in testing practices can bias R_t estimates generated from raw surveillance data, making it difficult to understand the current trajectory of local outbreaks and the effects of interventions (*4*).

Here, we present a method to estimate the daily incidence of new SARS-CoV-2 infections from observed case notification and death reports. We apply our model to publicly available COVID-19 case, death, and testing data for all 50 U.S. states and the District of Columbia. The model is available on GitHub and through the *covidestim* package for the R programming language.

We estimate COVID-19 epidemiological outcomes from reported time-series of cases and deaths, using a mechanistic model to account for changes in case ascertainment that result from fluctuating availability of diagnostics, as well as delays associated with disease progression, diagnosis, and reporting systems. We estimate infections, cases (symptomatic infections), deaths, and detected and reported cases (Fig. 1) and deaths (Fig. S1). Ascertainment of COVID-19 cases was estimated to vary considerably across states. We estimate that the percentage of detected cases (estimated cumulative cases divided by cumulative detected cases) ranged from 13.2% (95% uncertainty interval: 6.0%, 28.2%) in Arizona to 41.7% (22.4%, 62.5%) in Wyoming as of June 11, 2020. The estimated fraction of COVID-19 deaths detected ranged from 77.4% (59.4%, 85.7%) in Wyoming to 82.0% (42.3%, 90.0%) in New York as of June 11, 2020.

The state with the greatest number of estimated new infections in a single day was New York, with 112,838 (58,816, 281,042) on March 17, equivalent to 580 (302, 1,445) new infections per 100,000 population (Fig. 2). The majority of states saw a peak in new infections sometime between mid-March and early-April and have experienced flat or declining incidence between mid-April and early-June. Notable exceptions to this trend include Arizona, where we estimate new infections have been increasing steadily since early April, and Georgia, where we estimate new infections plateaued at the end of March (Fig. 2).

We use the modeled incident infection series to estimate R_t and seroprevalence while propagating the uncertainty associated with each stage of the analysis. We estimated R_t using a 5-day moving average (Fig. 3). In early March, median R_t estimates across states ranged from 0.99 (0.68, 1.38) in Washington to 4.4 (3.45, 5.34) in New York. For the period June 7 – June 11, the last period for which R_t could be estimated, we estimate that the median value of R_t was less than 1 in 31 states, though all credible intervals crossed 1. We found that R_t has been steadily increasing in a number of states over the past few weeks, suggesting COVID-19 case and deaths may increase in the coming weeks. Estimates of R_t reflect the greater uncertainty in modeled incidence (i.e., the "true" underlying source of transmission) compared to observed cases (Fig. S2).

We compare our estimates of COVID-19 incident deaths (detected and undetected) to estimates of excess all-cause mortality (*2*) (Fig. 4). We present comparisons to states where all-cause mortality was higher at each weekly timepoint from March 21 – May 9 than the expected number of deaths (based on all-cause mortality from 2015 – 2019). We expect COVID-19 deaths to follow similar trajectories to, but be less than, estimated all-cause mortality, which will include additional non-COVID-19 deaths linked to the social disruption caused by the pandemic (*2*). On average, the median modeled estimate of COVID-19 deaths is less than estimates of excess all-cause mortality, though uncertainty bounds around COVID-19 mortality estimates typically include estimates of excess all-cause mortality.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.17.20133983; this version posted June 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Additionally, we estimate cumulative incidence of COVID-19 infections as of June 11, 2020 (Fig. 4), as well as at each time point since the beginning of the epidemic. Ideally, these estimates would be compared with population-representative seroprevalence studies, although few state-wide serosurveys have yet been completed. We estimate seroprevalence in New York State to be 10.8% (5.55%, 24.3%) on April 19 and 11.3% (5.78%, 25.3%) on April 28. A recent seroprevalence survey (*5*) suggested New York State's seroprevalence in that period was 14.0% (95% confidence interval: 13.3%, 14.7%). Serosurvey estimates are dependent on test characteristics such as sensitivity and specificity, which are not precisely known, as well as an unbiased representative population sample; the authors report possible alternative seroprevalence estimates of 9.8% (9.1% - 10.5%) and 15.0% (14.3%, 15.7%).

As a sensitivity analysis, we ran the model in Delaware, Iowa, and New Jersey without adjustments for secular changes in case and death ascertainment over the course of the epidemic. We chose these states because they had highly variable reported fraction positive, which we used to estimate the completeness of ascertainment (Fig. S3). Removing this adjustment had a small effect on estimates of new infections and R_t, particularly early in the time series when the fraction of positive tests was most unstable (Fig. S4). Additionally, for the main analysis we estimated outcomes using time-series of cases and deaths indexed by date of report. However, some state departments of health have made detailed COVID-19 data publicly available which provide better information on the timing of events. For example, the Massachusetts Department of Public Health regularly publishes daily case counts by date of symptom onset or test and daily death counts by date of death (*6*). With minor adjustments to the model (described in Materials and Methods), we produced model estimates based on these data. Data type did not noticeably change fit to data and had a small impact on estimates of incident infections and R_t (Fig. S5).

Our approach makes a number of simplifying assumptions related to delays in disease progression and detection. We assume that sojourn time in each health state follows a fixed distribution, and that diagnostic and reporting delays are constant over the course of the epidemic, although we allow for secular changes in the probability that a symptomatic case will be diagnosed. The model is parameterized to estimate only one geographic unit at a time, and we do not model spatial spread of the disease between states. We also make the simplifying assumption that case detection is limited to individuals who are currently symptomatic. This assumption will not hold in situations in which PCR and antibody test results are reported in aggregate or if many asymptomatic individuals are tested for infection, which may lead to biased estimates of COVID-19 case detection. As of June 1, 2020, Arizona and Florida are reporting PCR and antibody tests in aggregate; these test results have previously been reported in aggregate in Iowa, Maine, Michigan, Mississippi, Missouri, New Hampshire, Texas, Virginia, and West Virginia. To date, antibody tests have constituted a small fraction of all reported tests, and so biases in our state-level estimates will be minimal.

We used publicly available data from the COVID Tracking Project (*7*) to produce the estimates reported here. Because states report cases and deaths in different ways, there are a number of potential inconsistencies in these data. Confirmed and probable cases are reported in aggregate in Idaho, Kansas, and Wyoming as of June 1, 2020, leading to potential upward bias in our estimates of new infections and seroprevalence in these states. Tests are reported as the total number of specimens tested (rather than the total number of individuals tested) in Arizona, California, Connecticut, Georgia, Idaho, Illinois, Maine, Massachusetts, Michigan, Mississippi, New Jersey, Oklahoma, South Carolina, Texas, Virginia, and Wyoming as of June 1, 2020. Our estimates of the probability of diagnosis conditional on symptoms may be upwardly biased in these states, though the size of this bias is likely to be small. Finally, data are subject to

medRxiv preprint doi: https://doi.org/10.1101/2020.06.17.20133983; this version posted June 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

occasional audits and revisions. Revisions to cumulative counts of cases and deaths have frequently been implemented as a single-day change in the cumulative count, 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.

As compared to other approaches for estimating R_t and COVID-19 epidemiological outcomes for US states (*8*,*9*) and other countries (*10*), our approach allows for time-trends in diagnostic coverage, which in turn accounts for under-ascertainment of cases and deaths over the course of the epidemic. Our approach utilizes information on the time-course of the natural history of infection and known lags in reporting systems, in combination with a metric that adjusts for changes in diagnostic test availability, to produce R_t estimates. This provides a coherent framework for simultaneously estimating the trend in disease incidence and the fraction of the population that has already been infected, providing key information on the current status and past severity of state-wide epidemics. These methods can also be applied to more local (e.g., county-level) data to produce a more refined understanding of local epidemics. To that end, we have open-sourced and documented our *covidestim* R package on GitHub, so that others can easily apply these methods to their own data.

References and Notes:

1. O Dyer, Covid-19: US testing ramps up as early response draws harsh criticism. *BMJ*; 368:m1167 (2020).
2. 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. *Medrxiv* preprint DOI: https://doi.org/10.1101/2020.04.15.20066431 (2020).
3. RM Anderson, RM May, *Infectious Diseases of Humans* (Oxford University Press, New York, 1991).
4. VE Pitzer, MH Chitwood, J Havumaki, NA Menzies, S Perniciaro, JL Warren, et al., The impact of changes indiagnostictesting practices on estimates of COVID-19 transmission in the United States. *Medrxiv* preprint doi: https://doi.org/10.1101/2020.04.20.20073338 (2020).
5. ES Rosenberg, JM Tesoriero, EM Rosenthal, R Chung, MA Barranco, LM Stey, et al., Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York. *Medrxiv* preprint doi: https://doi.org/10.1101/2020.05.25.20113050 (2020).
6. COVID-19 Response Reporting. Available at: https://www.mass.gov/info-details/covid-19-response-reporting [Accessed June 11, 2020].
7. The COVID Tracking Project. Available at: https://covidtracking.com/ [Accessed June 11, 2020]
8. 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), doi: https://doi.org/10.25561/79231 (2020).
9. COVID-19 Portal, Center for the Ecology of Infection Diseases, University of Georgia. Available at: https://www.covid19.uga.edu/nowcast.html
10. 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. doi:https://doi.org/10.25561/77731 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2020.06.17.20133983; this version posted June 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

11. SF McGough, MA Johansson, M Lipsitch, NA Menzies, Nowcasting by Bayesian smoothing: a flexible, generalizable model for real-time epidemic tracking. *PLOS Comp. Bio.* 16(4): e1007735 (2020).
12. 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).
13. H Nishiura, T Kobayashi, T Miyama, A Suzuki, S Jung, K Hayashi, Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *Int J Inf Dis* 94: 154 – 155 (2020)
14. O Byambasuren, M Cardona, K Bell, J Clark, M McLaws, P Glasziou, Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. *Medrxiv* preprint doi: https://doi.org/10.1101/2020.05.10.20097543 (2020).
15. 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).
16. 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).
17. A Basu (2020) Estimating the Infection Fatality Rate Among Symptomatic COVID-19 Cases in The United States. *Health Affairs* 39(7) (2020).
18. SA Lauer, KH Grantz, Q Bi, FK Jones, Q Zheng, HR Meredith, et al., The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med* 172(9):577-582 (2020).
19. F Zhou, R Du, G Fan, Y Liu, Z Liu, J Xiang, et al., Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* 395: 1054-62 (2020).
20. NM Linton, T Kobyayshi, Y Yang, K Hayashi, AR Akhmetzhanov, 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* 9(2): 538 (2020).
21. Santa Clara County COVID-19 Testing Dashboard. Available at: https://www.sccgov.org/sites/covid19/Pages/dashboard.aspx#testing [Accessed April 8, 2020]
22. Stan Development Team. RStan: the R interface to Stan. 2018 R package version 2.17.3. http://mc-stan.org
23. Stan Development Team. Stan Modeling Language Users Guide and Reference Manual. 2018. Version 2.18.0. http://mc-stan.org
24. A Cori, S Cauchemez, NM Ferguson, C Fraser, E Dahlqwist, PE Demarsh, et al. (2004) EpiEstim: Estimate Time Varying Reproduction Numbers from Epidemic Curves. R package version 2.2-3.
25. H Nishiura, NM Linton, AR Akhmetzhanov, Serial interval of novel coronavirus (COVID-19) infections. *Int J Inf Dis* 93: 284 – 86 (2020).

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

Funding:
National Institutes of HealthT32 GM007205

medRxiv preprint doi: https://doi.org/10.1101/2020.06.17.20133983; this version posted June 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

National Institute of Allergy and Infectious Diseases R01 AI112970
National Institute of Allergy and Infectious Diseases R01 AI137093
National Institute of Allergy and Infectious Diseases R01 AI146555
National Institute of Allergy and Infectious Diseases R01 AI112438
Fogarty International Center D43 TW010540

Author contributions: TC and NAM conceived the project; MHC and NAM developed the methodology; MR designed the software; KG and JH curated data; MHC and MR visualized results; MHC, TC, NAM, VEP, JLW, DMW contributed to the analysis; MHC prepared the original manuscript draft; all authors contributed to reviewing and editing.

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.

Data and materials availability: All data used in the main analysis are available from The Covid Tracking Project at https://covidtracking.com/. Data from Massachusetts used in the sensitivity analysis are available at https://www.mass.gov/info-details/covid-19-response-reporting. The *covidestim* package is available for download at https://github.com/covidestim/covidestim.

Supplementary Materials:

Materials and Methods

Figures S1-S6

Table S1

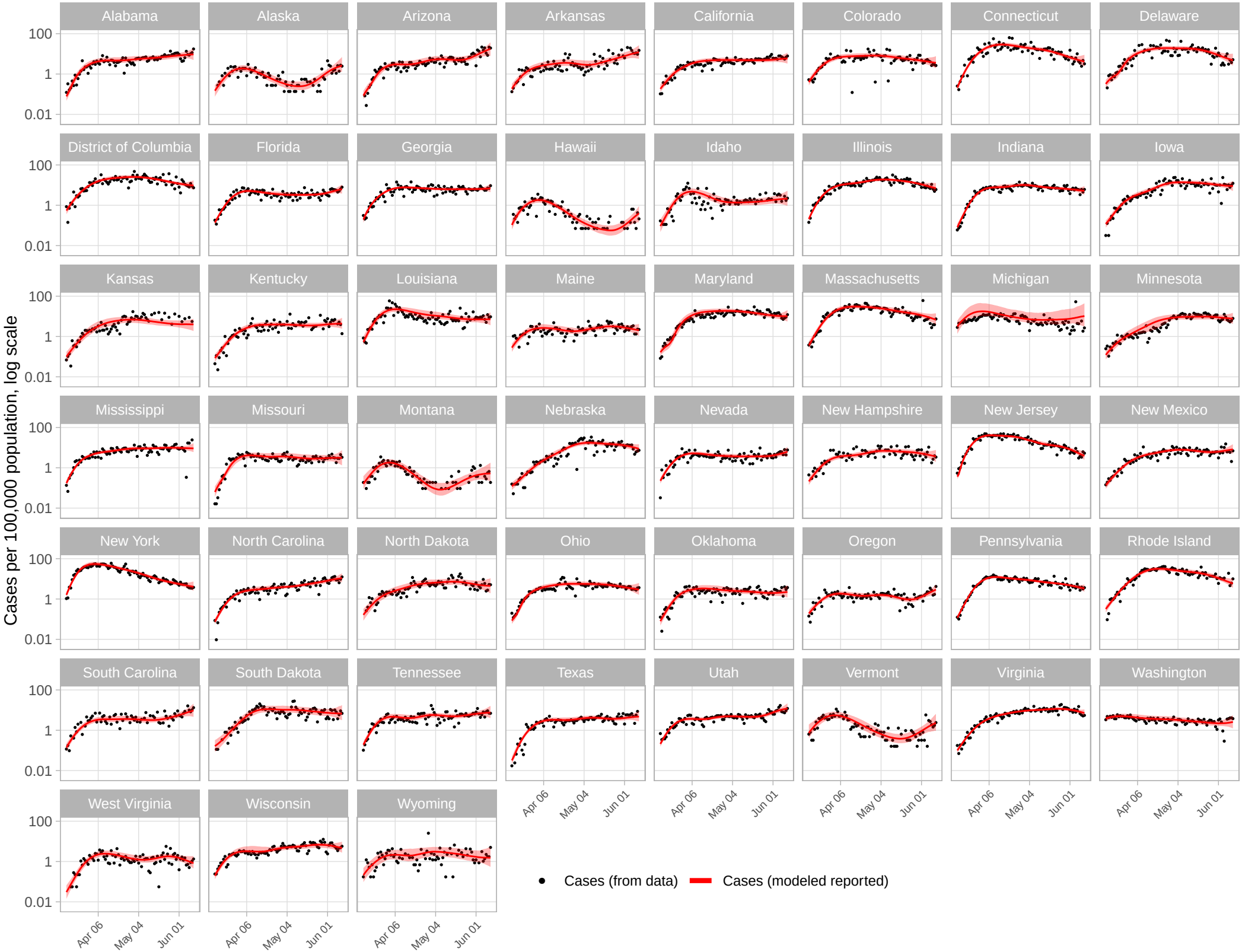
References (*11-25*)

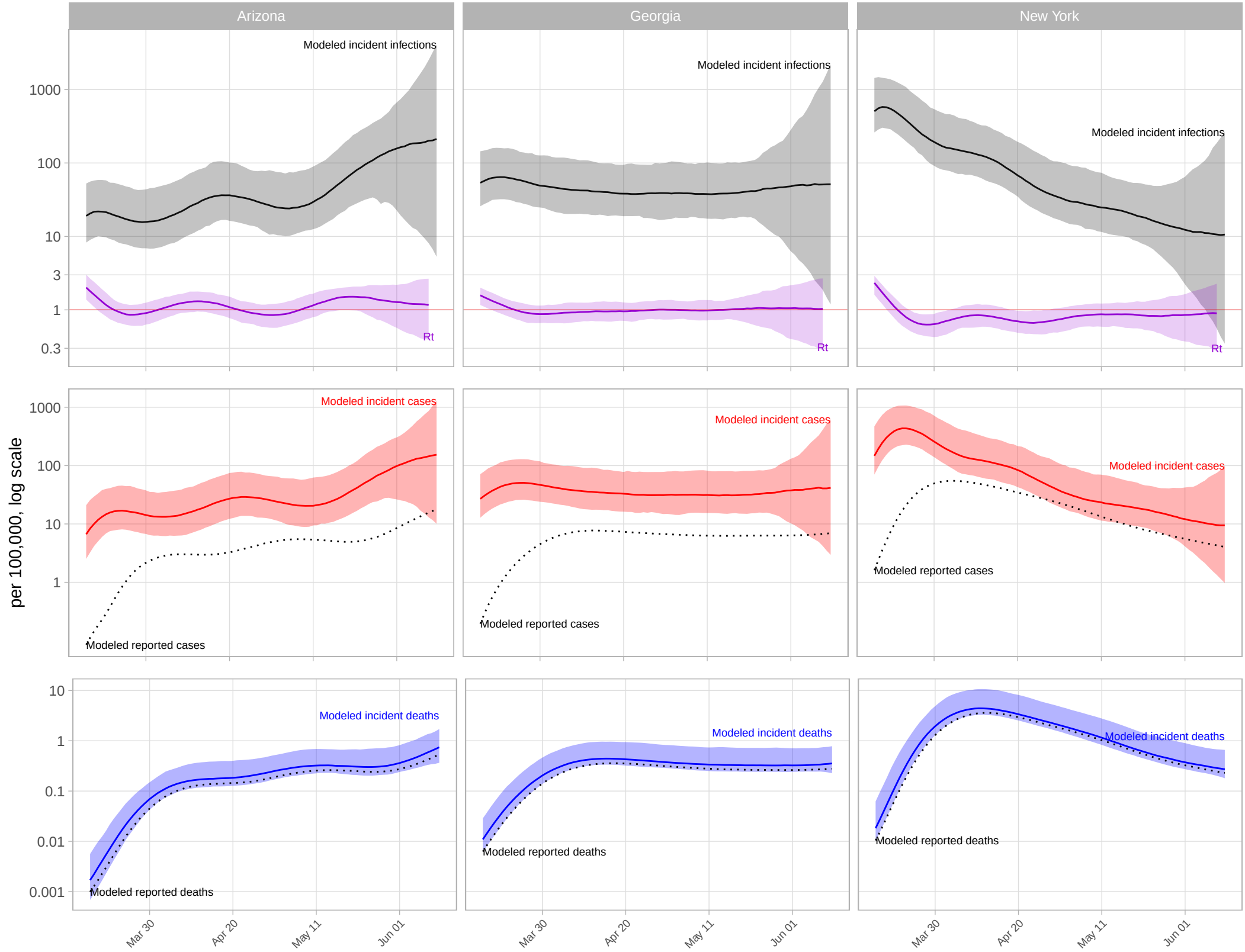
Fig. 1. Model estimates and empirical data for reported COVID-19 cases by state, log scale. Values are per 100,000 population, from March 15 to June 11, 2020. Empirical values not shown for days with zero reported cases.

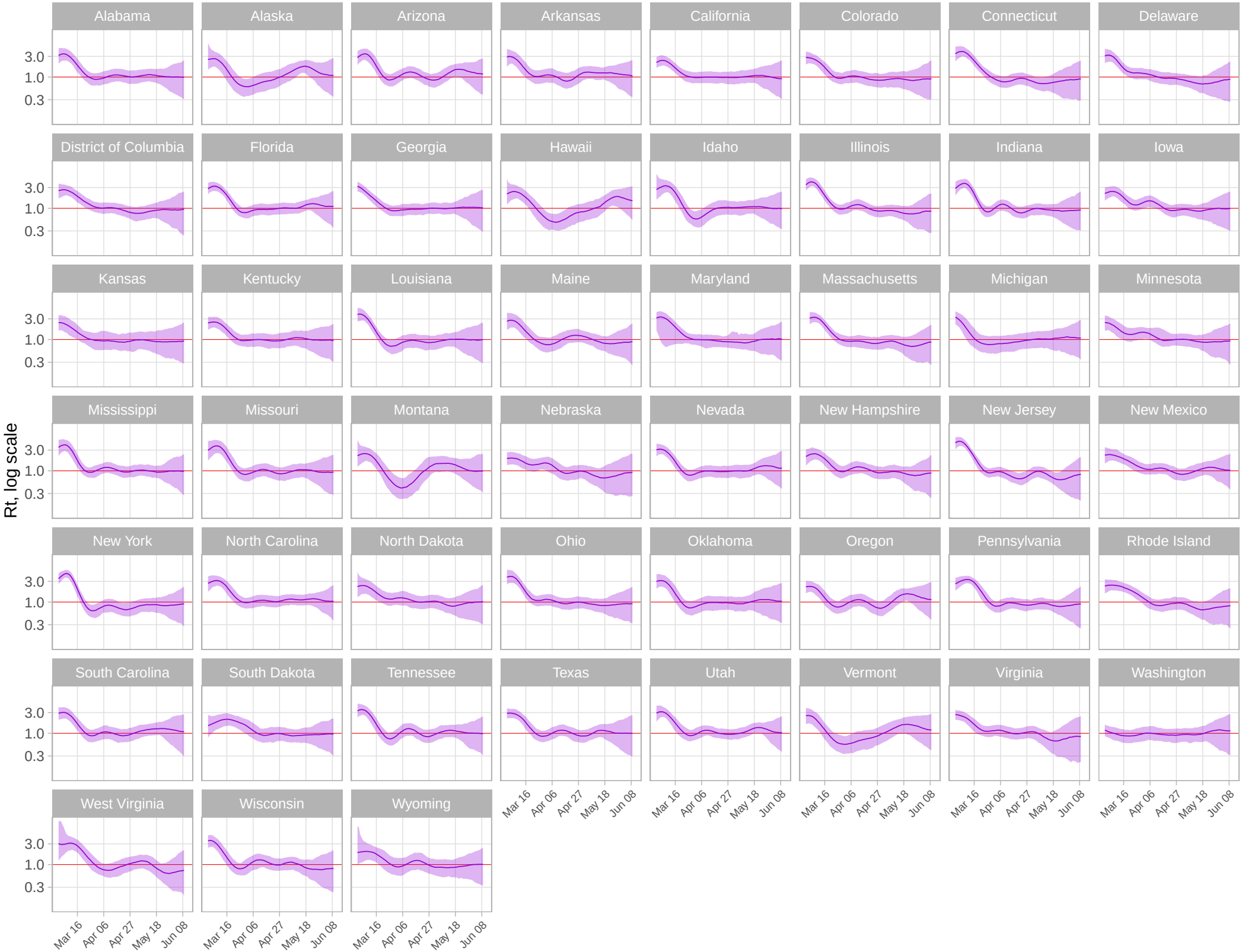
Fig. 2. Modeled COVID-19 epidemiology from selected states. Top: modeled estimates for incident infections per 100,000 population and R_t, log scale. Middle: model estimates for incident cases and reported cases, per 100,000 population, log scale. Bottom: model estimates for incident deaths and reported deaths, per 100,000 population, log scale. Estimates from March 15 to June 11, 2020 are shown.

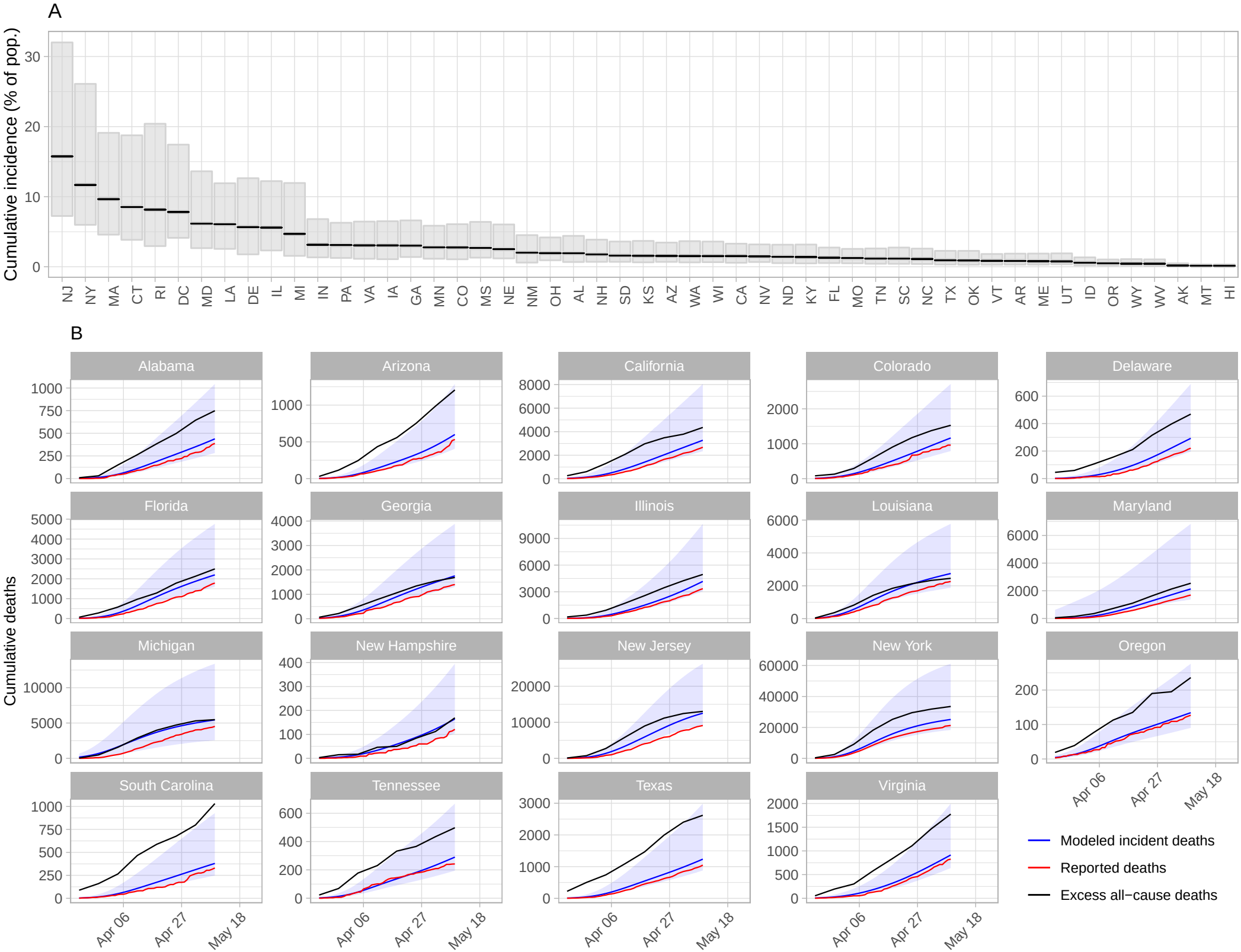
Fig. 3. Effective reproduction number (R_t) estimates by state, log scale. Estimates from March 1 to June 11, 2020 are shown.

Fig. 4. Seroprevalence and cumulative death estimates. Top: seroprevalence estimates (cumulative incident infections divided by state population) for June 11, 2020. Bottom: Comparison of modeled incident deaths, modeled reported deaths, estimated excess all-cause mortality.


Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
Cases per 100,000 population, log scale
100
1
0.01
Apr 06
May 04
Jun 01
Cases (from data)
Cases (modeled reported)


Arizona
Georgia
New York
Modeled incident infections
Rt
Modeled incident cases
Modeled reported cases
Modeled incident deaths
Modeled reported deaths
per 100,000, log scale
1000
100
10
3
1
0.3
0.1
0.01
0.001
Mar 30
Apr 20
May 11
Jun 01


Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
Rt, log scale
3.0
1.0
0.3
Mar 16
Apr 06
Apr 27
May 18
Jun 08


A
Cumulative incidence (% of pop.)
0
10
20
30
NJ NY MA CT RI DC MD LA DE IL MI IN PA VA IA GA MN CO MS NE NM OH AL NH SD KS AZ WA WI CA NV ND KY FL MO TN SC NC TX OK VT AR ME UT ID OR WY WV AK MT HI
B
Cumulative deaths
Alabama
Arizona
California
Colorado
Delaware
Florida
Georgia
Illinois
Louisiana
Maryland
Michigan
New Hampshire
New Jersey
New York
Oregon
South Carolina
Tennessee
Texas
Virginia
Apr 06
Apr 27
May 18
Modeled incident deaths
Reported deaths
Excess all−cause deaths