Using test positivity and reported case rates to estimate state-level COVID-19 prevalence in the United States

Weihsueh A. Chiu, Ph.D[1], Martial L. Ndeffo-Mbah, Ph.D[1,2]

[1]Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77845

[2]Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M University, College Station, TX 77845

Corresponding authors contact information:
Martial Ndeffo-Mbah, PhD
Address: Veterinary & Biomedical Education Complex, 660 Raymond Stotzer Pkwy, College Station, TX 77843
Phone: +1 979 845 5646
Email: mndeffo@cvm.tamu.edu,

Weihsueh Chiu, PhD
Address: Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences Texas A&M University VMA Building, Rm. 107 4458 TAMU College Station, TX 77843-4458, USA
Tel: +1 979-845-4106
Fax: +1 979-847-8981
Email: wchiu@tamu.edu

Keywords: COVID-19; Reported cases; test positivity rate; semi-empirical model; Mathematical modeling

Words count: 3130 words

Abstract

Accurate estimates of the prevalence of infection are essential for evaluating and informing public health responses to the ongoing COVID-19 pandemic in the United States (US), but reliable, timely prevalence data based on representative population sampling are unavailable, and reported case and test positivity rates may provide strongly biased estimates. A single parameter semi-empirical model was developed, calibrated, and validated with prevalence predictions from two independent data-driven mathematical epidemiological models, each of which was separately validated using available cumulative infection estimates from recent state-wide serological testing in 6 states. The analysis shows that individually, reported case rates and test positivity rates may provide substantially biased estimates of COVID-19 prevalence and transmission trends in the U.S. However, the COVID-19 prevalence for U.S. states from March-July, 2020 is well approximated, with a 7-day lag, by the geometric mean of reported case and test positivity rates averaged over the previous 14 days. Predictions of this semi-empirical model are at least 10-fold more accurate than either test positivity or reported case rates alone, with accuracy that remains relatively constant across different US states and varying testing rates. The use of this simple and readily-available surrogate metric for COVID-19 prevalence, that is more accurate than test positivity and reported case rates and does not rely on mathematical modeling, may provide more reliable information for decision makers to make effective state-level decisions as to public health responses to the ongoing COVID-19 pandemic in the US.

Introduction

Accurate or reliable estimates of the prevalence of infection are essential for evaluating and informing public health responses to the ongoing COVID-19 pandemic. The gold standard method to empirically measure disease prevalence is to conduct periodic large-scale surveillance testing via random sampling [1]. However, this approach may be time- and resource-intensive, and only a handful of such surveillance studies has been conducted so far in the United States (US) [2–4]. Metrics such as the test positivity rate has been commonly used by public health officials to infer the level of COVID-19 transmission in a population and/or the adequacy of testing [5,6]. The use of the test positivity rate for these purposes is often justified by referring to a WHO recommendation which was not necessarily intended to be used as such [7]. In the absence of high-quality data and rigorous criteria for measuring COVID-19 prevalence and adequacy of testing, public health officials have relied on reported cases, test positivity rate, fatality rates, hospitalization rates, and epidemiological models' predictions to inform COVID-19 responses. Reported cases and test positivity rate, though readily available and well-understood by public health officials, are very likely to substantially underestimate and overestimate, respectively, disease transmission/prevalence [1]. Hospitalization and death rates are also similarly readily available, but tend to lag infections by several weeks and only reflect the most severe outcomes [1]. Finally, epidemiological models are generally complex mathematical, computational, or statistical models that require extensive data and information for model

training, and are perceived as a 'black box' by most public health practitioners and decision makers [8–10].

Here, we develop a simple semi-empirical model to estimate the prevalence of COVID-19 at the US state level based only on reported cases, test positivity rate, and testing rate. Specifically, we hypothesized that the preferential nature of diagnostic testing for individuals at higher risk of infection in the US is a convex function of the overall testing rate due to the "diminishing return" from expanding general population testing. We modeled this convexity using a negative power function, and calibrated the power parameter using prevalence predictions from two independent, data-driven epidemiological models [8,11,12] that were separately validated using available cumulative infection estimates from recent state-wide serological testing. Using this single fitted power parameter, we found that the state-level prevalence of COVID-19 in the US can be well-approximated by the *geometric mean* of the reported cases and test positivity rates. We evaluated how disease prevalence varies with changes in reported cases and test positivity, and the implications of applying this simple model on informing public health decision-making to the COVID-19 pandemic in the US.

Methods

Conceptual basis of a semi-empirical model for the prevalence of COVID-19 infection

Test positivity rate $P_{+,\tau}(t) = N_{+,\tau}(t)/N_{test,\tau}(t)$ is defined as the percentage of positive diagnostic tests administered over a given period τ between $t-\tau$ and t. We first hypothesize that, because testing is currently "preferential" in that only those considered more likely to be infected due to symptoms, contacts, etc., are tested, $P_{+,\tau}(t)$ is correlated to the lagged prevalence $I(t-t_{lag})/N$ of COVID-19-infected persons in the population with a time-dependent bias parameter $b(t)$:

$$P_{+,\tau}(t) = b(t) \times \frac{I(t-t_{lag})}{N} \qquad (1)$$

Conceptually, this relationship is shown in Figure 1A. We next hypothesize that the bias parameter $b(t)$ is inversely related to the testing rate $\Lambda_\tau(t) = N_{test,\tau}(t)/N$ over the same period τ. Clearly, at a testing rate of 1, where everyone is tested, there is no bias, so $b = 1$. On the other hand, for very low testing rates, the bias is likely to be high, as only the most severely ill will be tested. For instance, if all COVID-19 deaths are tested daily, but no one else, then the bias will be equal to the inverse of the infected fatality rate (IFR), $b = 1/IFR$; and of course, in the limit of zero testing, the bias will be infinite. Moreover, at low testing rates, increasing the testing rate will likely preferentially increase the infected population testing rate relative to the general population testing rate, so $b(t)$ will decline more rapidly than at higher testing rates, as there is "diminishing return" from increased testing. Thus, it is reasonable to assume that $b(t)$ is a convex function of $\Lambda_\tau(t)$. We therefore model the bias as a negative power function of $\Lambda_\tau(t)$:

$$b(t) = \left[\frac{N_{test,\tau}(t)}{N}\right]^{-n} \equiv \Lambda_\tau(t)^{-n} \qquad (2)$$

Combining equations (1)-(2), and re-arranging leads to the following relationship between test positivity and the infectious population:

$$\frac{I(t - t_{lag})}{N} = P_{+,\tau}(t) \times \Lambda_\tau(t)^n \qquad (3)$$

Additionally, because test positivity and the testing rate share a term $N_{test,\tau}(t)$, equation (3) can be further rearranged as

$$\frac{I(t - t_{lag})}{N} = P_{+,\tau}(t)^{1-n}\left[\frac{N_{+,\tau}(t)}{N}\right]^n \equiv P_{+,\tau}(t)^{1-n} \times C_{+,\tau}(t)^n \qquad (4)$$

where the last term is the reported cases per capita $C_{+,\tau}(t) = N_{+,\tau}(t)/N$. Thus, our hypothesis predicts that the infectious population is proportional to a weighted geometric mean of the positivity rate and the reported case rate, with $n = 1/2$ corresponding to equal weighting.

We can also rearrange equation (1) and view the bias parameter as the relative efficacy of testing infected individuals compared to the general population:

$$b(t) = \frac{N_{+,\tau}(t)/I(t - t_{lag})}{N_{test,\tau}(t)/N} \equiv \frac{\Lambda_{I,\tau,t_{lag}}(t)}{\Lambda_\tau(t)}. \qquad (5)$$

Here, $\Lambda_{I,\tau,t_{lag}}(t)$ is the daily rate of testing of infectious individuals (with averaging time τ and lag t_{lag}), whereas $\Lambda_\tau(t)$ is the daily rate of testing of the general population, as previously defined. Thus, the bias reflects the extent to which infectious individuals are "preferentially" tested.

Fitting the power parameter of the semi-empirical model

For a given time point t_i, we define the "effective" power parameter $n_{eff}(t_i)$ by solving equation (3) for n:

$$n_{eff}(t_i) = \frac{\log\left[I(t_i - t_{lag})/N\right] - \log P_{+,\tau}(t_i)}{\log \Lambda_\tau(t_i)} \qquad (6)$$

The positivity $P_{+,\tau}(t_i)$ and the testing rate $\Lambda_\tau(t_i)$ are calculated from data obtained from the COVID Tracking Project [13], averaged of the past τ days. Because I is generally unobserved, we use estimates derived from data-driven epidemiological models. Two models of SARS-CoV-2 transmission in U.S. states are considered here: a Bayesian extended-SEIR model [12] (run through July 22, 2020) and a Bayesian semi-mechanistic model by Imperial College [14] (run through July 20, 2020), which applies the model published by Flaxman et al., 2020 [8] to the US. These models are calibrated at the US state-level to reported deaths and/or cases, but do not utilize test positivity or testing rate as an input. Both of these are Bayesian

models, so incorporate multi-parameter uncertainty, which is accounted for in our analysis by utilizing random samples of the posterior distributions for I.

The result of this procedure is a time-series in $n_{eff}(t_i)$, which we split approximately in half into a training set ($t_i \leq$ May 31, 2020) and a validation set ($t_i \geq$ June 1, 2020). Using intercept-only fixed and random effects linear models, we calculated the average power parameter with the training set, using an averaging time $\tau = 14$ days and a lag $t_{lag} = 7$ days. For fixed effects, a single power parameter was derived, whereas for random effects, state-specific power parameters were derived. We calculated the power parameter separately based on the extended-SEIR model only, the Imperial model only, and based on combining the two models (i.e., treating the two models as two independent measurements of I).

For each set of power parameter estimates, we evaluated predictive accuracy and precision for both the training set and the validation set by comparing the prevalence predicted using equation (4) to the predictions of the original epidemiological models. Specifically, the prediction errors (residuals) were calculated by comparison, after $\log_{10}$ transformation, to a random posterior sample from the extended-SEIR or Imperial model. Predictive accuracy was characterized by both the mean and median residual; predictive precision was characterized using both the usual root mean squared error (RMSE) [15], which assumes normally distributed errors, as well as the median absolute deviation (MAD), which is robust to outliers [16]. Furthermore, the fraction of predictions within the posterior 95% Credible Interval (CrI) of the extended-SEIR or Imperial model (F_{95}) as a metric of whether the semi-empirical-based model estimates of prevalence are consistent with the epidemiologic model-based estimates. These statistics were calculated for both overall as well as on a state-specific basis. For comparison, the same metrics for accuracy and precision were calculated based on reported cases or test positivity individually, as well as for the geometric mean of reported cases and positivity (corresponding to a power parameter n=1/2). Additionally, it should be noted that overall precision is limited by the posterior uncertainty of the original extended-SEIR and Imperial models, which had, respectively, RMSE of 0.19 and 0.09 and MAD of 0.18 and 0.10.

Epidemiological model validation

We independently validated the extended-SEIR and Imperial models by comparing their predictions to empirical data on the cumulative number of COVID-19 infections, which is available for 6 states, Connecticut (CT), Indiana (IN), Louisiana (LA), Missouri (MO), New York (NY), and Utah (UT), on a state-wise basis [2–4]. For 4 states (CT, LA, MO, UT), data were downloaded from the Interactive Serology Dashboard for Commercial Laboratory Surveys [17]. We only included data from states where the entire state was sampled or where adjustments for convenience sampling bias were integrated into the analysis.

Sensitivity Analyses

To assess the sensitivity of our results to the averaging time τ of 14 days, the time lag t_{lag} of 7 days, and the split between training and validation data at May 31, 2020, we also

considered an averaging time period τ of 7 days, time lag t_{lag} of 0-14 days, and training data cutoffs of April 30 and June 30, 2020.

Software

All analyses were performed using the R statistical software (R version 3.6.1) in RStudio (Version 1.2.1335). Random effects models were fit using the lme4 package (version 1.1-21). We developed a publicly available online dashboard, using R and Shiny, to analyze and visualize state-level predictions of COVID-19 prevalence from our semi-empirical model. The dashboard self-updates once daily (https://wchiu.shinyapps.io/COVID-19-Prevalence-and-Trends-Alpha/).

Results

Validation of epidemiological model predictions

The comparison between epidemiological model predictions and empirical cumulative infections data are shown in Figure 2. For all states, the empirical data were consistent with the 95% CrI of the extended-SEIR model predictions. The Imperial model predictions were less consistent, with overprediction of CT and underprediction of NY and to some extent UT.

Semi-empirical model calibration and predictions

The power parameter of the semi-empirical model was estimated by fitting equation (5) to the models' predictions, resulting in a power parameter $n \approx 0.5$, regardless of whether the estimate of the bias parameter was derived from the extended-SEIR, Imperial, or both models combined (Figure S1A). Similar parameter estimates were obtained when equation (5) was fitted to both training and validation data combined (Figure S1B). n_{eff} remained relatively constant over time, and the random effects estimate of n_{eff} remained close to the fixed effects estimate of ≈ 0.5 for nearly all states (Figure S2).

Given the power parameter, equation (4) was then used to predict the prevalence of infection based on the test positivity rate and reported case rates. We showed that the semi-empirical model predictions were consistent with those of the epidemiological models (Figures 3 & S3). Quantitative measures of accuracy and precision are summarized in Table 1 and Figure S4. The accuracy of the semi-empirical model, as measured by the mean and median residuals, ranged from -0.10 to 0.45 log-10 units, with the random effects estimate from the combined datasets having values of 0.07 and 0.04, respectively, corresponding to biases of < 20%. By comparison, mean or median residuals for reported cases and test positivity alone were < -1.17 and > 1.34, respectively, corresponding to biases of over 10-fold. The precision of the semi-empirical model over training and validation periods showed an overall MAD and RMSE in log-10 units ranging from 0.20-0.31 for the extended-SEIR model-based analysis, 0.16-0.28 for the Imperial model-based analysis, and 0.23-0.30 for the combined analysis (Figure S4). Precision for reported cases and test positivity were in the same range, with MAD and RMSE values from 0.2-0.3. The overall coverage F_{95} of the Imperial model was lower (0.3-0.8) as compared to the SEIR

model (> 0.9), and uniformly zero for reported cases and test positivity (Table 1, Figure S4). The accuracy and precision were mostly similar for the fixed and random effects models (Figure S4). The simplified approach of using a geometric mean of test positivity and reported case rates to estimate prevalence was as accurate and precise as random effects models. Sensitivity analysis was conducted with different averaging times (7 or 14 days), lag times (1 to 14 days), and training data date range ($t \leq$ April 30, May 31, or June 30). Our results were shown to be robust for both epidemiological models (Figure S5).

The pitfalls of relying on reported cases or test positivity rate alone to estimate the course of the epidemic are illustrated for four states, MN, MT, VA, and WI, where reported cases and positivity trends were in opposite directions in May and/or August (Figure 4). Specifically, in May, reported cases were rising substantially in MN, VA, and WI at the same time that the test positivity rate was declining, testing rate was increasing, and the model predicted prevalence was flat or decreasing (Figure 4). By contrast, in August, the states of MN, MT, and WI all showed declining reported case rates while positivity was increasing. Our model predicts that COVID-19 prevalence was either flat or increasing during this time. In both scenarios, the increase (decrease) in reported cases was due to expanded (declining) testing rates, respectively.

With our semi-empirical model, an individual state or municipality can easily utilize reported cases and positivity rates to better estimate their COVID-19 prevalence and its recent trend (Figure 5). We observed a positive correlation between state-level test positivity rate and reported cases (Figure 5A); however, states with the same reported case rate may differ greatly in their test positivity rates, resulting in substantial differences in prevalence. For example, thus was the case with DC and SD, which both had a reported case rate of ~10 cases/100,000 on August 15, 2020, but a test positivity rate differing by almost 10-fold, resulting in a 3-fold difference in predicted prevalence. We showed that prevalence decreases (increases) when test positivity rate and reported case rate both decrease (increase), respectively (Figure 5B). However, when the trends in test positivity and reported cases are in opposite directions, the prevalence trend was shown to be highly dependent on the relative change in the value of the test positivity and reported case rates (Figure 5B).

Discussion

Reported case rates and test positivity rates have been widely used to inform or justify public health decisions as to increasing or relaxing intervention measures for the control of the COVID-19 pandemic in the US [18,19]. A recent report of the National Academies of Sciences and Engineering Medicine (NASEM) has urged caution about the reliability/validity of directly using data such as reported case rates and test positivity rates to inform decision making for COVID-19 [1]. Though these data are, typically, readily available, the NASEM report concludes that they are likely to substantially underestimate or overestimate the real state of disease spread [1]. It is therefore urgent and paramount to develop simple and more reliable data-driven metrics/approaches to inform public health decision-making.

We have developed a semi-empirical approach to estimate the prevalence of COVID-19 infections in a population using only reported cases and testing rates. Fitting only a single parameter, we find that the COVID-19 prevalence, with a 1-week lag, is well-approximated by the *geometric mean* of the positivity rate and the reported case rate averaged over the last 2 weeks. The prevalence predictions of our semi-empirical model were shown to compare favorably to those from two data-driven epidemiological models whose predictions were validated against U.S. state-level empirical data of COVID-19 prevalence. These predictions were at least 10-fold more accurate than estimates from using positivity or reported case rates alone. Although the prediction of our semi-empirical model has an uncertainty of about 3-fold in either direction, this degree of uncertainty is reduced by more than 10-fold compared to using positivity or reported case rates alone, and does not require developing and maintaining a complex data-driven mathematical model.

Our analysis suggests that public health policy on relaxing or reinstating of social distancing and other non-pharmaceutical interventions may be informed by available data in three main ways:

- First, prevalence should be inferred as decreasing only if both reported cases and positivity rates are continuously declining without decreasing testing rates. This criterion can be taken as a surrogate for $R_{eff}(t) < 1$, since declining prevalence is a sufficient condition for an effective reproduction number below one.

- Two, the gap between reported case and test positivity rates, measured on the same scale (e.g., per 100,000), should be decreasing to ensure improvement in the sufficiency of testing. For most of the U.S., this gap has hovered around 1000-fold. For example, in MN, MT, VA, and WI (Figure 5), as testing ramped up in May, the gap narrowed; however, the gap widened in August, indicating that testing rates need to be brought back up. Indeed, as noted by the NASEM report, absolute testing rates, irrespective of positivity rates, may be a better indicator of the sufficiency of testing.

- Finally, reported cases, test positivity, and testing rates should be publicly reported at the county or municipal level in order to provide local governments, health agencies, medical personnel, and the public with the necessary information to evaluate the local state of the pandemic. Currently, only reported cases are routinely provided at the local level, with positivity and testing rates comprehensively aggregated only at the state level.

As with any model, ours has a number of limitations. The most significant limitation is the reliance on epidemiologic model-based estimates of prevalence. However, we believe this concern is mitigated by our use of two independent estimates with completely different model structures, one of which is a more traditional extended-SEIR model, and other of which is a "semi-mechanistic" model partially statistical in nature. Moreover, both models were compared to the limited empirical prevalence data available. The other important limitation is the relatively limited range of test positivity rate and/or testing rate observations for most U.S. states. For this reason, we cannot necessarily guarantee that our results can be easily extrapolated to substantially higher testing rates (e.g., if wide-spread immunoassay-based testing were implemented). However, with higher testing rates, the

difference between test positivity rates and reported case rates would decrease and reduce the effect of greater uncertainty in the degree of bias between the test positivity rate and the lagged prevalence. Finally, we did not address the sensitivity and specificity of diagnostic testing; however, the impact of imperfect test accuracy is likely to have a minimal impact on our results.

In conclusion, we found that the COVID-19 prevalence is well-approximated by the *geometric mean* of the positivity rate and the reported case rate. This predictor is 10-fold more accurate than either rate alone, and consistent with prevalence estimates from much more complex mathematical epidemiologic models. The use of this simple, reliable, and easy-to-communicate measure of COVID-19 prevalence will improve the ability to make public health decisions that effectively respond to the ongoing COVID-19 pandemic in the U.S.

Funding: National Science Foundation (NSF DEB RAPID 2028632) and National Institutes of Health, National Institute of Environmental Health Sciences (P30 ES029067). The funders have no role in the design of the study, collection, analysis, and interpretation of data.

Conflict of Interest Disclosure: The authors declare no conflicts of interest

Ethical approval: Ethical approval was not required for this work.

Data sharing: The codes and data used to generate our results are available on GitHub at https://github.com/wachiuphd/COVID-19-US-Semi-Empirical

References

1. National Academies of Sciences and Medicine E. *Evaluating Data Types: A Guide for Decision Makers Using Data to Understand the Extent and Spread of COVID-19*. Washington, DC: The National Academies Press; 2020. doi:10.17226/25826

2. Havers FP, Reed C, Lim T, et al. Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. *JAMA Intern Med*. 2020. doi:10.1001/jamainternmed.2020.4130

3. Menachemi N, Yiannoutsos CT, Dixon BE, et al. Population Point Prevalence of SARS-CoV-2 Infection Based on a Statewide Random Sample — Indiana, April 25–29, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(29):960-964. doi:10.15585/mmwr.mm6929e1

4. Rosenberg ES, Tesoriero JM, Rosenthal EM, et al. Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York. *Ann Epidemiol*. 2020;48:23-29.e4. doi:10.1016/j.annepidem.2020.06.004

5. Coronavirus (COVID-19) Testing - Statistics and Research - Our World in Data.

https://ourworldindata.org/coronavirus-testing#the-positive-rate-a-crucial-metric-for-understanding-the-pandemic. Accessed August 28, 2020.

6. How Bad Is the Coronavirus Outbreak? Here's a Key Number. - The Atlantic. https://www.theatlantic.com/technology/archive/2020/04/us-coronavirus-outbreak-out-control-test-positivity-rate/610132/. Accessed August 28, 2020.

7. Public health criteria to adjust public health and social measures in the context of COVID-19. https://www.who.int/publications/i/item/public-health-criteria-to-adjust-public-health-and-social-measures-in-the-context-of-COVID-19. Accessed August 28, 2020.

8. Flaxman S, Mishra S, Gandy A, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*. 2020;584(7820):257-261. doi:10.1038/s41586-020-2405-7

9. COVID-19. https://covid19.healthdata.org/united-states-of-america?view=total-deaths&tab=trend. Accessed August 28, 2020.

10. COVID-19 Mathematical Modeling | COVID-19 | CDC. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html. Accessed August 28, 2020.

11. State-level tracking of COVID-19 in the United States - Imperial College London. https://mrc-ide.github.io/covid19usa/#/. Accessed August 28, 2020.

12. Chiu W, Fischer R, Ndeffo-Mbah M. State-level needs for social distancing and contact tracing to contain COVID-19 in the United States. doi:10.21203/rs.3.rs-40364/v2 (Accepted for Publication in Nature Human Behavior)

13. The COVID Tracking Project | The COVID Tracking Project. https://covidtracking.com/. Accessed May 18, 2020.

14. Unwin HJT, Mishra S, Bradley VC, et al. State-level tracking of COVID-19 in the United States. *medRxiv*. 2020;2(May):2020.07.13.20152355. doi:10.1101/2020.07.13.20152355

15. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?-Arguments against avoiding RMSE in the literature. *Geosci Model Dev*. 2014;7:1247-1250. doi:10.5194/gmd-7-1247-2014

16. Högel J, Schmid W, Gaus W. Robustness of the Standard Deviation and Other Measures of Dispersion. *Biometrical J*. 1994;36(4):411-427. doi:10.1002/bimj.4710360403

17. Commercial Laboratory Seroprevalence Surveys | Coronavirus | COVID-19 | CDC. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html#interpreting-serology-results. Accessed August 28, 2020.

18. US travel restrictions state by state | CNN Travel.

https://www.cnn.com/travel/article/us-state-travel-restrictions-covid-19/index.html. Accessed August 28, 2020.

19. of Disease Contol C. *CDC Activities and Initiatives Supporting the COVID-19 Response and the President's Plan for Opening America Up Again*.; 2020.

Table

Table 1. Accuracy and precision of different estimates of prevalence

Prevalence Estimator	Epidemiologic model calibration	Validation Accuracy (log_{10} units)		Validation Precision (log_{10} units)		
		Mean residual	Median residual	RMSE	MAD	F95
Reported cases*	Extended SEIR	-1.47	-1.46	0.21	0.20	0.00
	Imperial	-1.22	-1.17	0.24	0.21	0.00
	Combined	-1.35	-1.34	0.26	0.28	0.00
Test positivity*	Extended SEIR	1.34	1.35	0.24	0.24	0.00
	Imperial	1.60	1.64	0.29	0.24	0.00
	Combined	1.47	1.48	0.29	0.30	0.00
Fixed effects	Extended SEIR	0.02	0.04	0.20	0.20	1.00
	Imperial	0.43	0.45	0.15	0.16	0.19
	Combined	0.22	0.20	0.32	0.39	0.52
Random effects	Extended SEIR	-0.10	-0.09	0.20	0.19	0.97
	Imperial	0.15	0.18	0.15	0.16	0.40
	Combined	0.07	0.04	0.32	0.39	0.53
GM of reported cases* and positivity*	Extended SEIR	-0.06	-0.06	0.20	0.20	0.97
	Imperial	0.19	0.23	0.25	0.21	0.29
	Combined	0.06	0.07	0.26	0.28	0.63

*14-day moving average

RMSE = Root mean squared error; MAD = median absolute deviation; F95 = coverage of 95% CrI from epidemiological model; one log-10 unit corresponds to an order of magnitude, and 0.1 log-10 units corresponds to 20 - 25%.

Figures

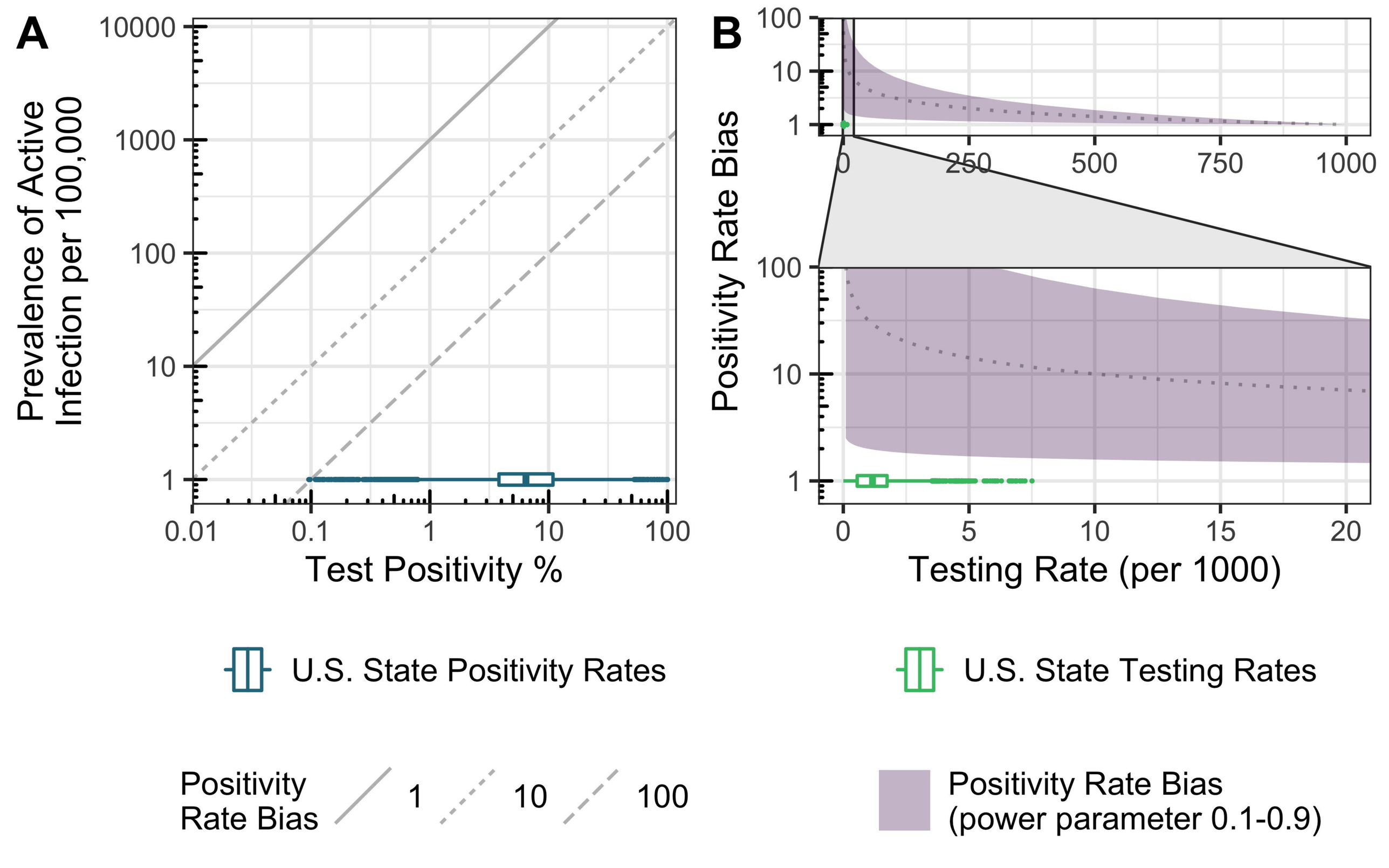
*Figure 1. Conceptual model for relationship between test positivity, prevalence of infection, and testing rate. We assumed the test positivity rate is correlated to delayed disease prevalence with a bias parameter $b(t)$. In **A**) the diagonal lines represent different values of the bias parameter and the U.S. state positivity rates from March-July are shown for reference. B) Illustrate the relationship between testing rate and bias parameter represented by equation (2). Here the shaded region represents different powers n ranging from 0.1 to 0.9, the dotted line represents $n = 0.5$, and the U.S. state testing rates from March-July are shown for reference.*

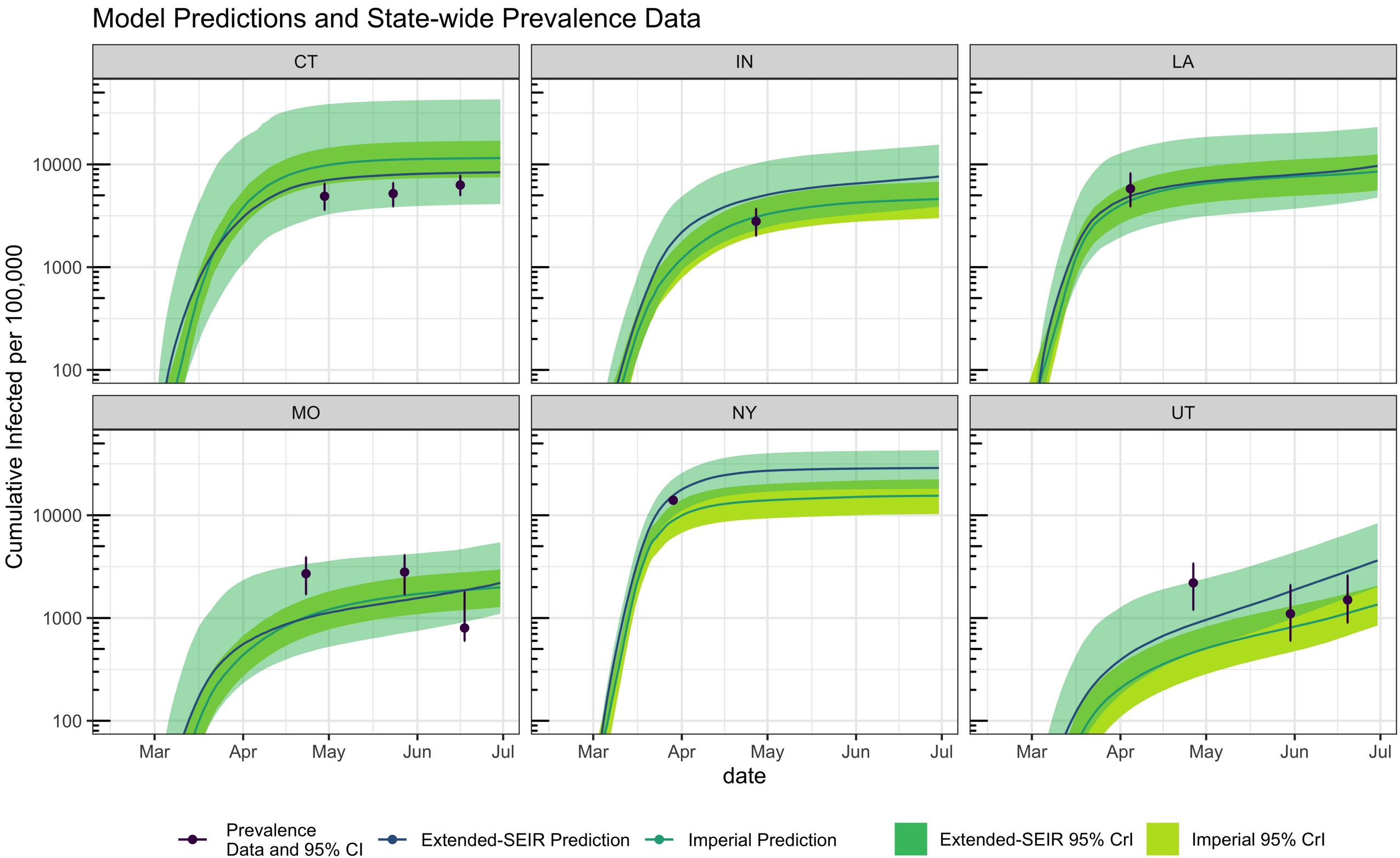
Figure 2. Validation of data-driven SEIR and Imperial epidemiological models using cumulative infections data.

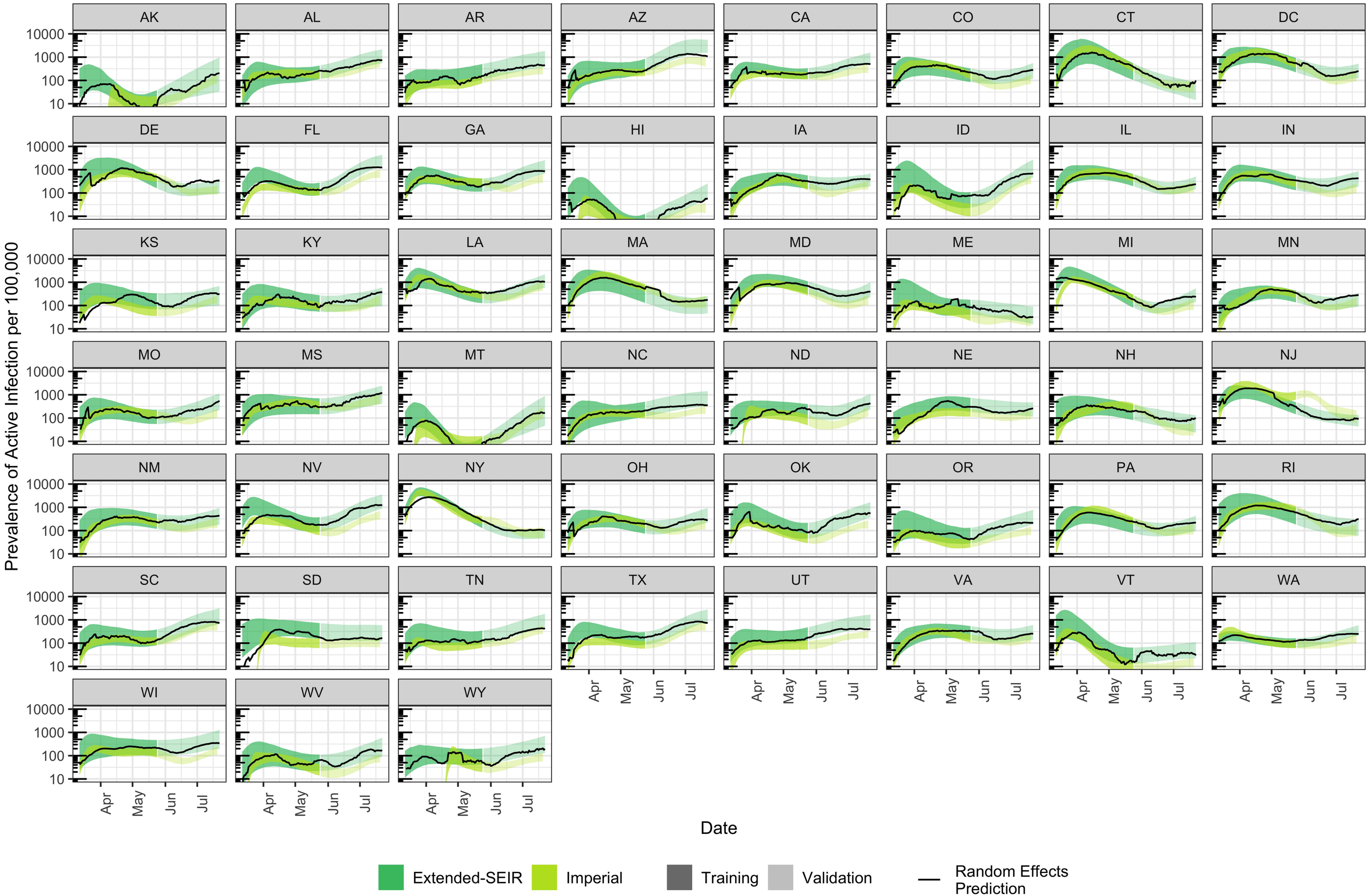
Figure 3. Semi-empirical estimate of the infected population for each state, using random effects estimates for the power parameter estimated from both the extended-SEIR and Imperial-based datasets combined. Bands show the 95% CrI for the extended-SEIR and Imperial predictions.

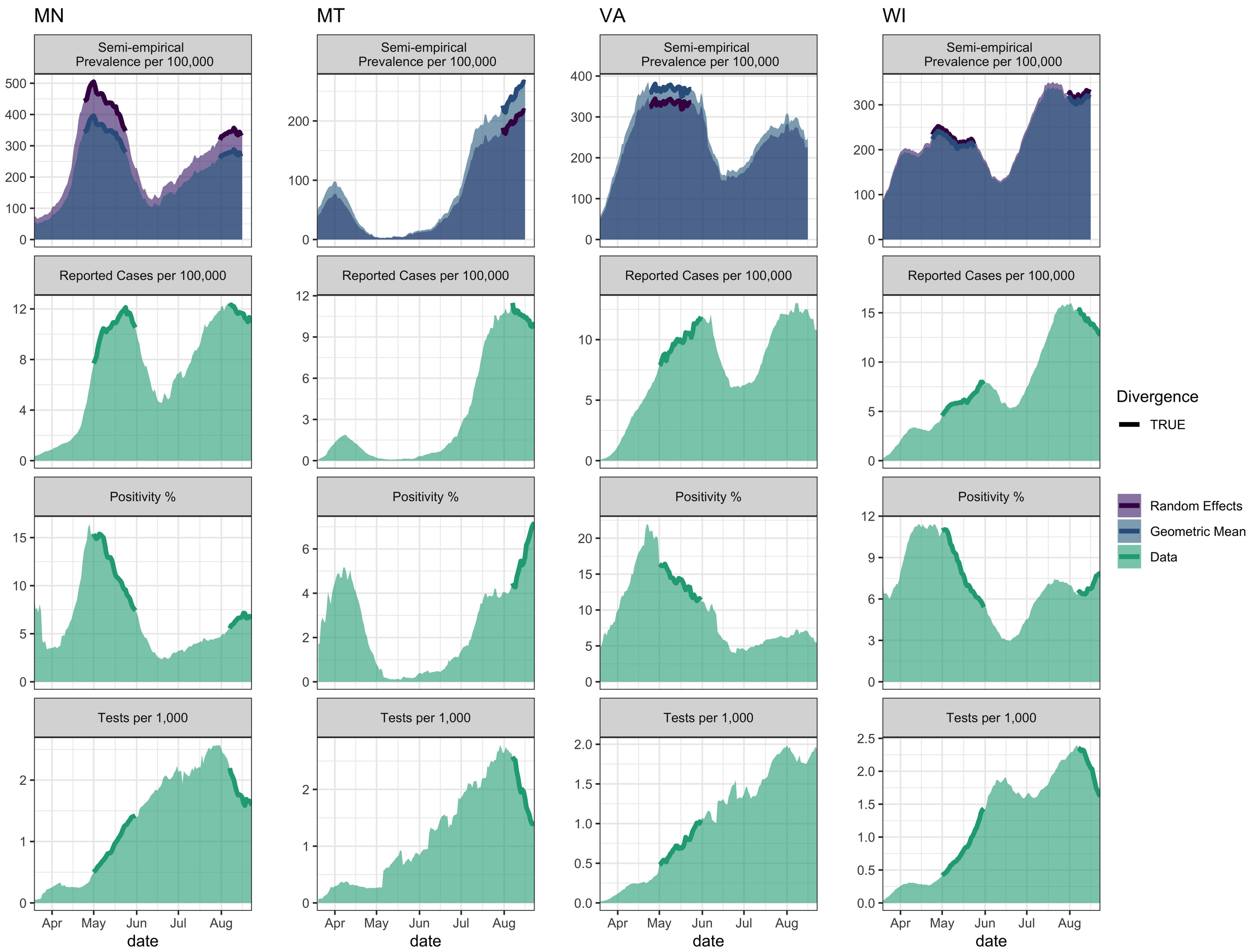
Figure 4. Examples of divergences between reported case rates and positivity rates for four states. For each state, the top panel is the prevalence as predicted by the semi-empirical model including a 7-day lag, whereas the bottom three panels show the reported case, positivity, and testing rates, each averaged over the last 14 days.

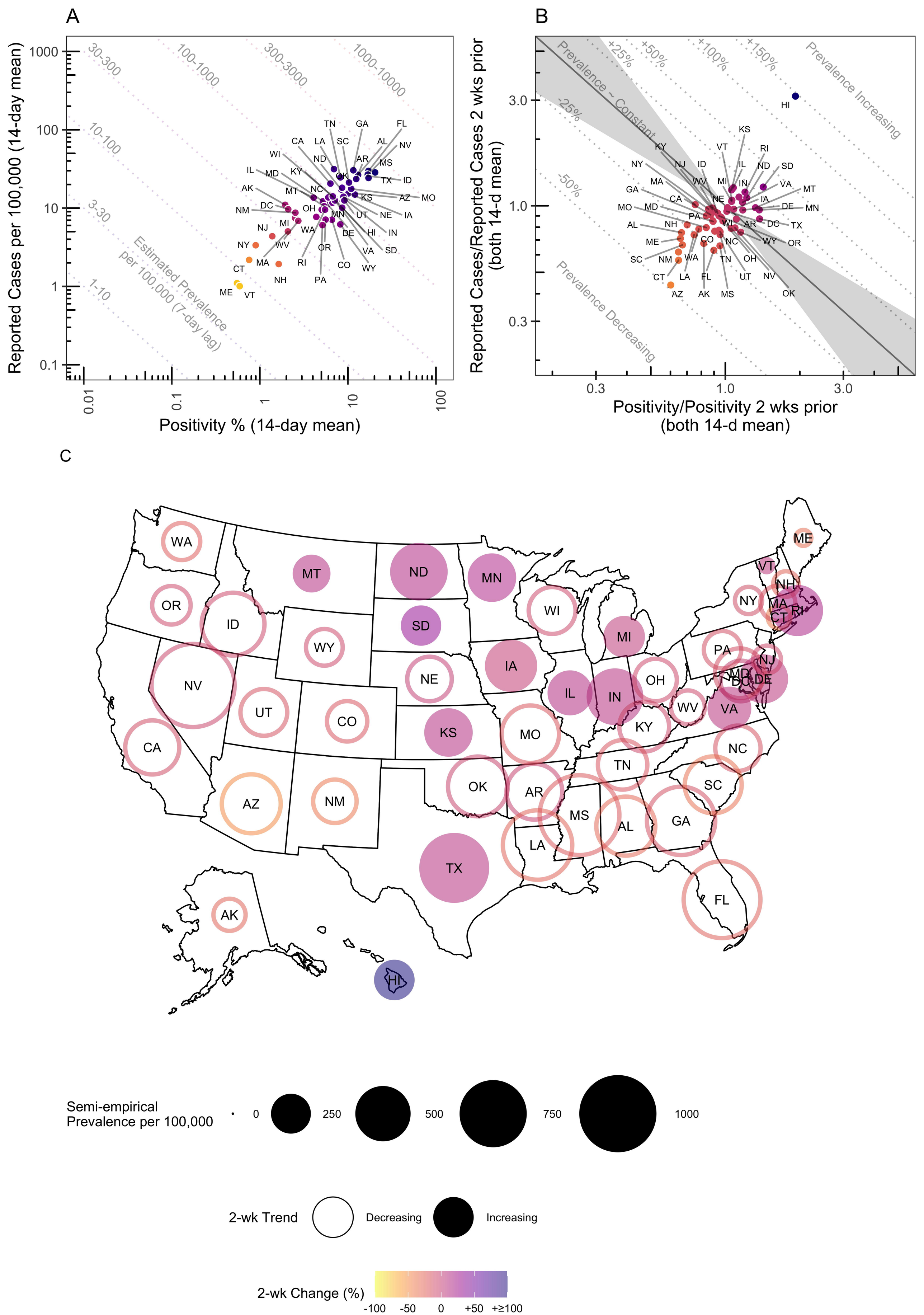
Figure 5. A: Prevalence estimated for each state based on scatter plot of reported case rates and positivity rates. Prevalence increases to the upper right, and is expressed as a range to reflect the residual error estimated from the model. B: Two-week prevalence trend for each state, based on scatter plot of two-week changes in reported case rates and positivity rates. The shaded region expresses the uncertainty in the power parameter, and demarks where prevalence is predicted to be decreasing from where it is predicted to be increasing. C: Map of estimated prevalence and transmission trend. All rates are two-week averages evaluated on August 15, 2020, and all prevalence predictions are for a one-week lag.


A
Prevalence of Active Infection per 100,000
10000
1000
100
10
1
0.01
0.1
1
10
100
Test Positivity %
B
Positivity Rate Bias
100
10
1
0
250
500
750
1000
100
10
1
0
5
10
15
20
Testing Rate (per 1000)
U.S. State Positivity Rates
U.S. State Testing Rates
Positivity Rate Bias
1
10
100
Positivity Rate Bias (power parameter 0.1-0.9)


Model Predictions and State-wide Prevalence Data
CT
IN
LA
MO
NY
UT
Cumulative Infected per 100,000
10000
1000
100
Mar
Apr
May
Jun
Jul
date
Prevalence Data and 95% CI
Extended-SEIR Prediction
Imperial Prediction
Extended-SEIR 95% CrI
Imperial 95% CrI


AK
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
10000
1000
100
10
Prevalence of Active Infection per 100,000
Apr
May
Jun
Jul
Date
Extended-SEIR
Imperial
Training
Validation
Random Effects Prediction


MN
MT
VA
WI
Semi-empirical Prevalence per 100,000
Reported Cases per 100,000
Positivity %
Tests per 1,000
Apr
May
Jun
Jul
Aug
date
Divergence
TRUE
Random Effects
Geometric Mean
Data


A
Reported Cases per 100,000 (14-day mean)
Positivity % (14-day mean)
Estimated Prevalence per 100,000 (7-day lag)
30-300
100-1000
300-3000
1000-10000
10-100
3-30
1-10
B
Reported Cases/Reported Cases 2 wks prior (both 14-d mean)
Positivity/Positivity 2 wks prior (both 14-d mean)
Prevalence ~ Constant
Prevalence Increasing
Prevalence Decreasing
+25%
+50%
+100%
+150%
-25%
-50%
C
Semi-empirical Prevalence per 100,000
0
250
500
750
1000
2-wk Trend
Decreasing
Increasing
2-wk Change (%)
-100
-50
0
+50
+≥100