The prevalence of SARS-CoV2 in the US estimated from imperfect testing

Song S. Qian,[*,†] Maxwell D. Qian,[‡] and Sabrina Jaffe[†]

†*Department of Environmental Sciences, The University of Toledo, Toledo, OH 43606*

‡*Department of Economics and History, Vanderbilt University, Nashville, TN 27325*

E-mail: song.qian@utoledo.edu

Abstract

An accurate estimate of the prevalence and distribution of the novel coronavirus (SARS-CoV2) in the United States is essential for an effective response to the COVID-19 pandemic. The initial lack of sufficient testing capacity in the United States hinders the effort to stop the spread of the virus. Furthermore, the US lacks a coherent reporting system where concerned citizens can find accurate information. Without an accurate estimate of the prevalence of the infection in the population, we cannot properly determine how likely a positive test result is to be a false positive and how likely a negative result is to be a false negative. Here we present a statistical model for estimating the prevalence of SARS-CoV2 infection in the nation and in each state by pooling data from all reported state-level testing results. Our results show that accurate reporting (of both positive and negative results) is necessary to properly understand the spread of the virus. The estimated national average prevalence is about 10% when using all available data. When states without a consistent record of reporting negative results are removed from the analysis, the average is about 6%.

Key Words and Terms: Bayesian statistics, binary tests, COVID-19, false negative, false positive, hierarchical mdoel, prevalence

Introduction

Tests with binary outcomes to indicate a binary state of nature (e.g., presence or absence of a disease agent) are common. Nearly all tests are imperfect: they produce occational false positive and false negative results. Here we use testing data for the noval coronavirus (SARS-CoV2) in the United States to demonstrate the use of the Bayesian hierarchical modeling framework to better understand the prevalence of the disease. Our analysis highlights the need for more testing because the test for detecting SARS-CoV2 is imperfect. When a patient is tested we are uncertain whether a positive result is a reliable indicator of the presence of the virus. Likewise, we cannot rule out a negative result being a false negative. The problem of interpreting results from an imperfect test is not new. Recently, Qian et al.[1] provided a summary of the underlying statistical issues of interpreting imperfect test results. The interpretation and use of imperfect test results depend on the purpose of the test. For testing of SARS-CoV2, the purposes of the test are (1) diagnosing individual patients and (2) estimating the prevalence of the virus in a population. For the diagnostic purpose, whether a positive test result is indicative of an infection depends on (1) the quality of the test measured by the rates of false positives and false negatives and (2) the prevalence of the virus in the population. At this point, we do not have a good understanding of these three quantities. They must be estimated from testing data.

We present a statistical model to estimate the prevalence of the virus in each state using publicly available data. The quality of the model result depends on (1) our knowledge of the test (how well do we know the rates of false positives and false negatives) and (2) the number of people tested (sample size). The better we can characterize the test and the more people who are tested, the more accurate the estimated population prevalence. A large number of tests allows us to better estimate not only the prevalence but also the test's false-positive and false-negative rates. A better understanding of the test and the prevalence makes the test a better diagnostic tool. Using data reported in the public domain (testing results and estimates of rates of false positives and false negatives for similar tests), we developed a computer program to automatically retrieve data from the internet and estimate state-level prevalence. Model estimated prevalence can be readily updated when more data are made available.

Materials and Methods

Source of Data

Unfortunately, the US Center for Disease Control and Prevention (CDC) is not publishing complete testing results. We retrieved data from the COVID Tracking Project (`covidtracking.com`), a joint effort led by Jeff Hammerbacher of Related Sciences (`https://www.related.vc`) and Robinson Meyer and Alexis Madrigal of *The Atlantic* (`theatlantic.com`). A full list of the team is on the project's webpage (`https://covidtracking.com/about-team`).

Data quality for each state was evaluated and graded (A-D) by the COVID Tracking Project team based on whether the state reports (1) positive results reliably, (2) negative results reliably, and (3) commercial testing results. All states (except NV) and territories reports positive result reliably. However, many states do not report negative results consistently. For example, the State of Ohio stopped reporting negative results after March 15, 2020. Some states do not report commercial testing results.

Statistical Methods

The basic statistical approach is documented by Qian et al.[1], along with the computational details (`github.com/songsqian/imperfect`). The basic concept is that the test result is uncertain because of the inevitable false positive and false negative outcomes. To properly interpret the test result, either by a patient or by the state health authority, we must translate the result into the relevant quantity. Because of the inevitable false positive, a positive result cannot be equated to the presence of the virus; likewise, the possibility of false negatives makes a negative result less reassuring. As a result, a positive (or negative) result should be interpreted in terms of the probability of infection (or non-infection), specifically, the conditional probability of infection given a positive result. Let $+$ (or $-$) represent a positive (or negative) result, v represent the presence of the virus, and a represent the absence of the virus. When observing $+$ for an individual patient, we want to know $\Pr(v|+)$, which reads "the conditional probability of v given $+$." This conditional probability is calculated by the Bayes Theorem:

$$\Pr(v|+) = \frac{\Pr(v)\Pr(+|v)}{\Pr(v)\Pr(+|v) + \Pr(a)\Pr(+|a)} \tag{1}$$

In other words, to learn about the meaning of a positive result, we need to know three more quantities: $\Pr(+|v)$ (probability of a positive result when the virus is present, or 1 minus the probability of a false negative), $\Pr(+|a)$ (the probability of a false positive), and $\Pr(v)$ the prevalence of the virus infection in the population. Interpretation of $\Pr(v)$ depends on the definition of the population[1].

Probabilities of false positive and false negative are features of the test and the prevalence of the infection is what we, as a society, want to learn from repeated testing. At this point, we have no definite knowledge of these three quantities. Therefore, from the perspective of government health authorities, we want to use test results to learn about these quantities so that individual patients can better understand the meaning of their test results.

Following the notation of Qian et al.[1], let θ be the prevalence ($\theta = \Pr(v)$), f_p the false positive rate, and f_n false negative rate, the statistical model for updating the probability distribution of θ is the continuous variable version of the Bayes theorem.

$$\pi(\theta|y,n) = \frac{\pi(\theta)L(\theta|y,n)}{\int \pi(\theta)L(\theta|y,n)d\theta}$$

where y and n are numbers of positive results and total tests and $L(\theta|y,n)$ is the likelihood function (representing the probability of observing y positives out of n tests). The likelihood is derived based on the binomial distribution assumption of y, and it is a function of θ, f_p, and f_n

$$\begin{array}{rcl} p_+ & = & \theta(1-f_n) + f_p(1-\theta) \\ L(\theta, f_p, f_n|y,n) & \propto & p_+^y(1-p_+)^{n-y} \end{array}$$

where p_+ is the probability of observing a positive result. Because we don't know f_p and f_n, we use the Bayes theorem to update them as well:

$$\pi(\theta, f_p, f_n|y,n) = \frac{\pi(\theta)\pi(f_p)\pi(f_n)L(\theta, f_p, f_n|y,n)}{\int_\theta \int_{f_p} \int_{f_n} \pi(\theta)\pi(f_p)\pi(f_n)L(\theta, f_p, f_n|y,n)d\theta df_p df_n}$$

As the three quantities of interest (θ, f_p, f_n) are probabilities, we use the beta distribution as their priors. We note that the likelihood function provides information on the products of θf_p and θf_n. Using independent forms of priors for θ, f_p, and f_n is unlikely to jointly estimate the three quantities. As a result, providing realistic informative priors for at least two of the three parameters is necessary.

Prior Specification

As the SARS-CoV2 is a new virus and only a relatively small number of tests are done in the US (only for people with specific symptoms), we don't have a basis for specifying a more informative prior for the prevalence. As a result, we used the hierarchical modeling approach and imposed non-informative prior distributions on the hyper-parameters. For the qPCR test used for detecting SARS-CoV2, we haven't seen studies to quantify f_p and f_n. However, the basic principle of the test is well known. We can use reported f_p and f_n for similar types of tests to develop the priors. We estimate state-level prevalence under three scenarios.

- The best-case scenario assumes that the test has a false positive probability of 1% and a false negative probability of 1% and both are stable ($f_p \sim beta(1, 99)$ and $f_n \sim beta(1, 99)$, with a 95% credible interval of 0.00026 0.036).

- The expected scenario assumes that the current test is similar to tests for other corona-viruses (SARS, MERS, H1N1). Existing studies on tests of similar viruses have a range of false positive and false negative rates. Using studies of the MERS and SARS virus tests[2–5], we constructed a prior distribution for false positive to be $beta(3, 23)$ (95% credible interval of 0.025-0.26) and false negative $beta(2, 22)$ (0.01-0.22)

- The worst-case scenario assumes that the test is as unreliable as the rapid influenza diagnostic test (RIDT)[6]. The RIDT test has a high false negative probability and we use $beta(16, 24)$ (0.26-0.55) as the prior. The probability of a false positive is relatively low and we used $beta(4, 45)$ (0.023, 0.17).

A Hierarchical Formulation

We have data from nearly all states and territories. We assumed that f_p and f_n are the same for all states because they use the same test. However, the prevalence can vary by region. But we have no information to separate one state from another other than the testing data. As a result, we assumed that the prevalence for each state θ_j are exchangeable and impose a common prior. Advantages of using a hierarchical formulation were explored elsewhere[7,8]. Expressing the model hierarchically, we have the following model.

1. At the observational level, data from each state (numbers of positive and negative) are modeled by the binomial distribution

$$y_j \sim Bin(p_j, n_j)$$

   where j represents the jth state, y_j and n_j are the observed number of positive and total number of tests, and Bin represents the binomial distribution. The probability of observing a positive result (p_j) is a function of θ_j, f_p, and f_n:

$$p_j = \theta_j(1 - f_n) + (1 - \theta_j)f_p$$

2. To connect all states together, we used a common prior for state-level prevalence θ_j (after a logit transformation).

$$\text{logit}(\theta_j) \sim N(\mu_0, \sigma_0^2)$$

   Non-informative priors are used for μ_0 and σ_0^2.

3. Prior distributions of other parameters

$$\begin{aligned} f_n &\sim beta(\alpha_n, \beta_n) \\ f_p &\sim beta(\alpha_p, \beta_p) \end{aligned}$$

   The hyper-parameter μ_0 is the national average of prevalence (in logit scale) and σ_0^2 is among state variance of the logit transformed state-specific prevalence.

Results and Discussions

The estimated state-level prevalence varies by a wide range. Under the best-case scenario, the national mean prevalence is 0.092 (with a 95% credible interval of between 0.058 and 0.132) based on data reported on March 23, 2020. The lowest state-level prevalence is 0.0040 (0.0004-0.011) and the highest is 0.894 (0.879-0.919). For the expected scenario, the national average prevalence is 0.091 (0.056-0.135), and the state-level prevalence ranges from 0.0023 (0.0002-0.0066) to 0.929 (0.889-0.981). Finally, under the worst-case scenario, the national average prevalence is 0.100 (0.0595-0.151), and the state-level prevalence ranging from 0.0019 (0.00014-0.0059) to 0.984 (0.960-0.997) (Figure 1). In all three scenarios, the uncertainty of the estimated prevalence (measured as the width of the 95% credible interval) decreases as the number of tests increases (Figure 2). Because many states did not properly report negative results (e.g., Ohio's number of negative result has stopped at 140 since March 15, 2020 and New Jersey may have changed how it reports negatives on March 16th, 2020), we rerun the model using data from states with a data quality grade (from the COVID Tracking Project) of A. The estimated national average is 0.060 (0.043, 0.079) and the state-level prevalence ranges from 0.0093 (0.003, 0.014) to 0.167 (0.149, 0.185) (Figure 3).

The importance of testing a large number of people and consistently reporting the results is illustrated by comparing the estimation results for four states: Louisiana, Massachusetts, Ohio, and Washington (Figure 4). The State of Washington tested the largest number of people early on and reported all test results. The total number of tested people exceeded 1000 on March 9, 2020. Louisiana and Massachusetts had a slow start and were initially inconsistent in reporting negative results. Once their numbers of tests exceed 1000, the estimate prevalence for these two states reached more stable levels. Ohio tested far fewer people compared to the other three states. Furthermore, Ohio stopped reporting negative test results on March 15, 2020. As a result, the steady increase in the estimated prevalence for Ohio is an artifact of the missing negative results.

As the nation contemplates resuming normal economic activities based on perceived risk of infection, our model estimated prevalence is a natural measure of the risk. The model can be used by each state to estimate county or regional prevalence to provide daily assessment. However, the current restriction on testing is limiting such application because results from a statistical model reflect the population represented by the data. In this case, when testing is highly selective, the estimated prevalence is relevant to the sub-population who meet the screening criteria. The population represented by the data are people who met CDC guidelines. The guidelines posted on CDC web page [1] include hospitalized patients with COVID-19 symptoms, symptomatic people in high risk groups (older adults and individuals with chronic medical conditions and/or an immunocompromised state), and healthcare providers with contact with a suspected or confirmed COVID-19 patient. It is unclear whether the SARS-CoV2 prevalence in this sub-population is higher than the prevalence in the general population. To evaluate the effectiveness of the current prevention practices, we need to estimate the prevalence in the general population to monitor the trend.

For individual patients, a positive result translates to a probability of true infection of 0.9, 0.52, or 0.146 using equation (1) based on the best-case, expected, or the worst-case scenarios, respectively (Table 1). The difference between the three scenarios is largely represented in the rate of false negatives. Under all scenarios, the estimated national average prevalence is relatively stable. The interpretation of the test result in this case lies largely with the quality of the test. If the current test is on par with the qPCR tests used in previous outbreaks of similar viruses, a positive test result to a patient means that the likelihood of infection is only marginally higher than 0.5; a re-test follwing a positive result is likely necessary.

The posterior distributions of f_p and f_n are similar to their respective prior distri-

[1] https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-criteria.html

butions (Figure 5). This result suggests that properly characterizing the test's rates of false positives and false negatives is imperative for proper interpretation of a positive result for individual patients.

Supporting Information

R code and link to source data are posted at `https://github.com/songsqian/COVID19`.

References

(1) Qian, S. S.; Refsnider, J. M.; Moore, J. A.; Kramer, G. R.; Streby, H. M. All tests are imperfect: Accounting for false positives and false negatives using Bayesian statistics. *Heliyon* **2020**, *6*, e03571.

(2) Alvarez-Martínez, M. J. et al. Sensitivity and specificity of nested and real-time PCR for the detection of *Pneumocystis jiroveci* in clinical specimens. *Diagnostic Microbiology and Infectious Disease* **2006**, *56*, 153–160.

(3) Binsaeed, A. A.; Al-Khedhairy, A. M., A. A.and Mandil; Shaikh, R., S. A.and Qureshi; Al-Khattaf, A. S.; Habib, H. A.; Alam, A. A.; Al-Ansary, L. A.; Al-Omran, M. A validation study comparing the sensitivity and specificity of the new Dr. KSU H1N1 RT-PCR kit with real-time RT-PCR for diagnosing influenza A (H1N1). *Annals of Saudi medicine* **2011**, *31*, 351–355.

(4) Rainer, T. H.; Chan, P. K.; Ip, M.; Lee, N.; Hui, D. S.; Smit, D.; Wu, A.; Ahuja, A. T.; Tam, J. S.; Sung, J. J.; Cameron, P. The Spectrum of Severe Acute Respiratory Syndrome–Associated Coronavirus Infection. *Annals of Internal Medicine* **2004**, *140*, 614–619.

(5) thi Tham, N.; thi Ty Hang, V.; Khanh, T. H.; Viet, D. C.; Hien, T. T.; Farrar, J.;

van Vinh Chau, N.; van Doorn, H. R. Comparison of the Roche RealTime ready Influenza A/H1N1 Detection Set with CDC A/H1N1pdm09 RT-PCR on samples from three hospitals in Ho Chi Minh City, Vietnam. *Diagnostic Microbiology and Infectious Disease* **2012**, *74*, 131–136.

(6) Center for Disease Control and Prevention, Rapid influenza diagnostic tests. 2016; `https://www.cdc.gov/flu/professionals/diagnosis/clinician_guidance_ridt.htm`.

(7) Wu, R.; Qian, S.; Hao, F.; Cheng, H.; Zhu, D.; Zhang, J. Modeling Contaminant Concentration Distributions in China's Centralized Source Waters. *Environmental Science and Technology* **2011**, *45*, 6041–6048.

(8) Qian, S.; Stow, C.; Cha, Y. Implications of Stein's Paradox for Environmental Standard Compliance Assessment. *Environmental Science and Technology* **2015**, *49*, 5913–5920.

Tables

Table 1: Probability of infection given a positive result

Scenarios	θ_0	$E(f_p)$	$E(f_n)$	$\Pr(v\|+)$
Best-case	0.092	0.01	0.01	0.90
Expected	0.091	0.12	0.08	0.52
Worst-case	0.100	0.08	0.40	0.15

Figure

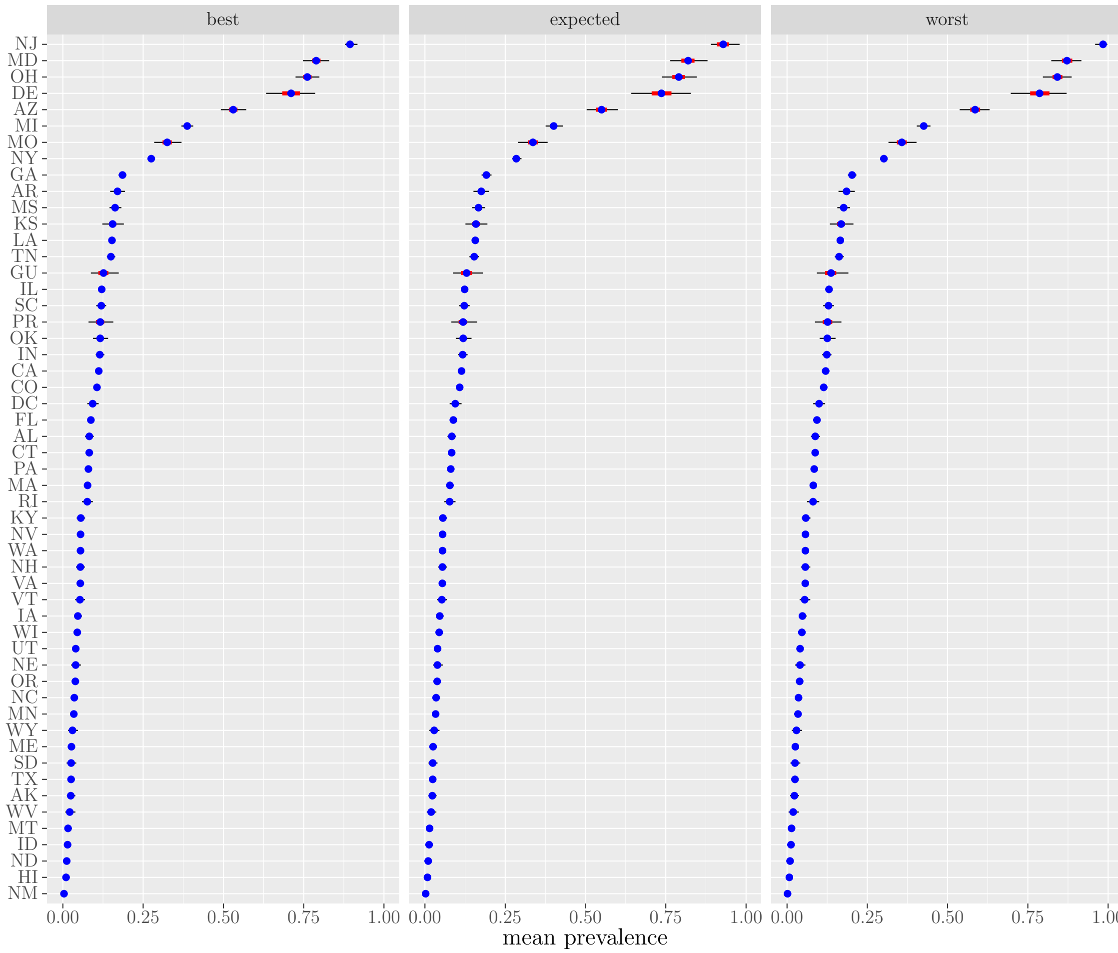

Figure 1: Hierarchical model estimated state-level SARS-CoV2 prevalence based on data reported by March 23, 2000. The blue dots are the estimated means, the red thick bars are the 50% credible intervals, and the thin black lines are the 95% credible intervals.

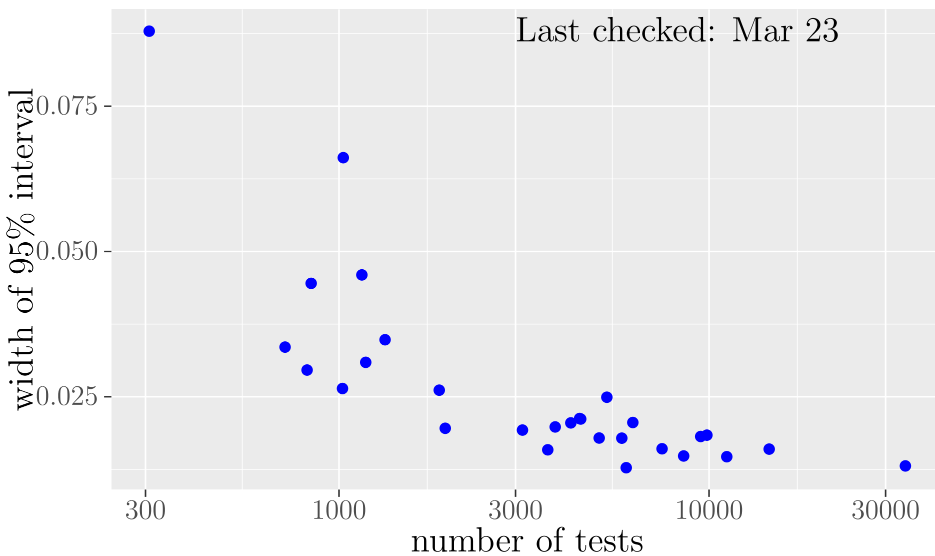

Figure 2: Uncertainty of the estimated state-level prevalence, measured as the width of the 95% credible interval (estimated using data from states with reliable data reporting system, see Figure 3), is a inversely related to the total number of tests performed.

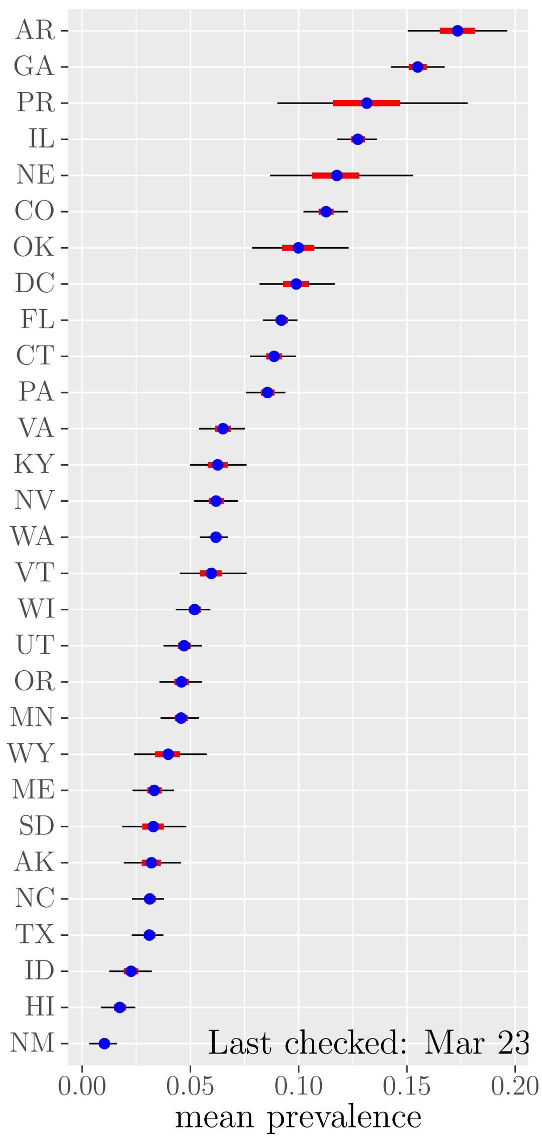

Figure 3: Estimated state-level prevalence using data from states with reliable data reporting system

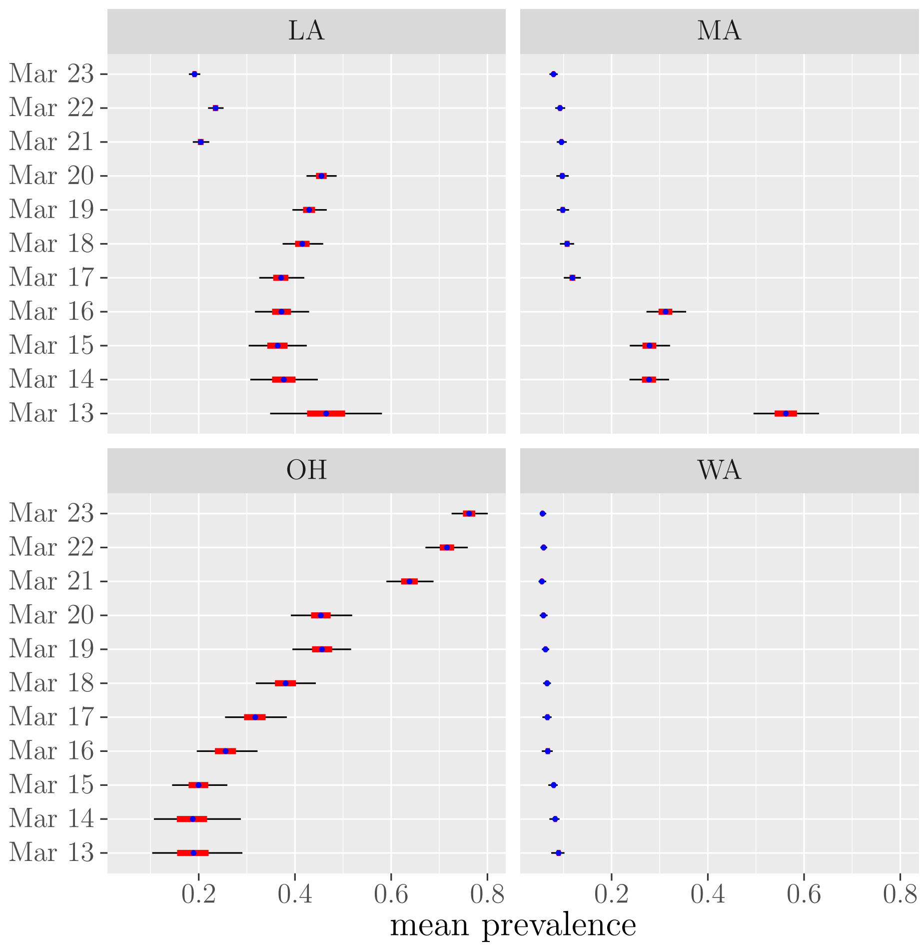

Figure 4: Estimated state-level prevalence for four states from March 13 to March 23, 2020

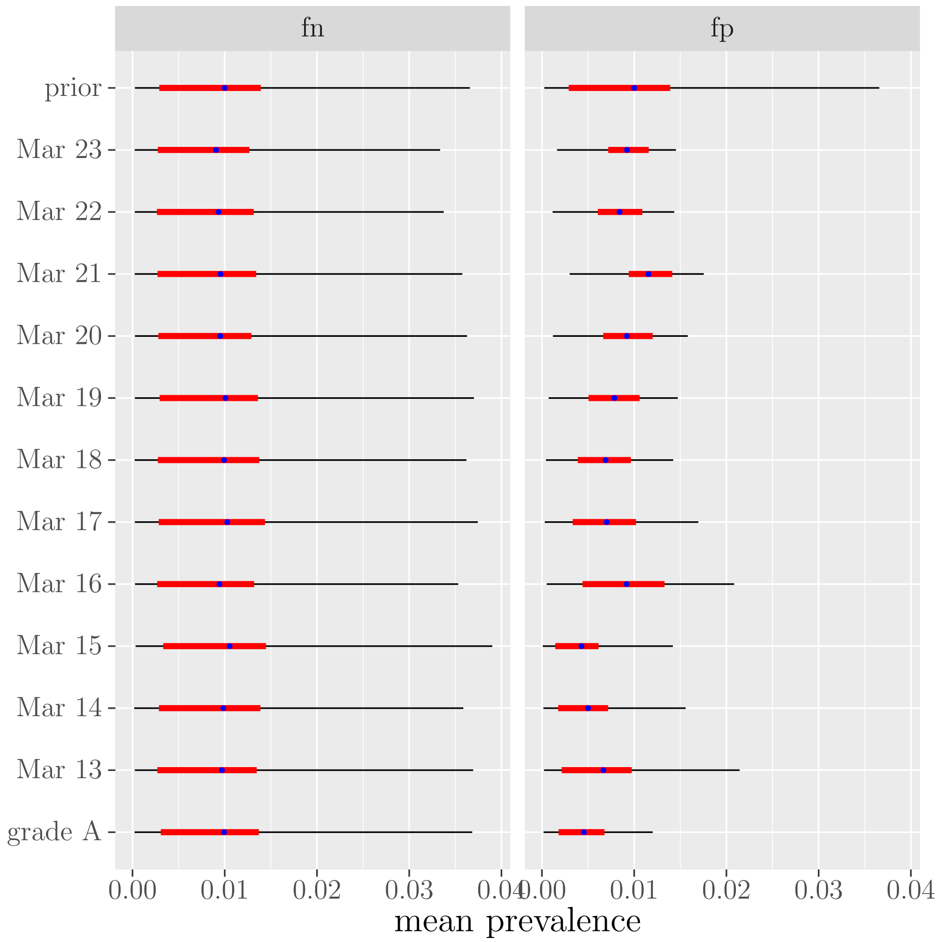

Figure 5: Comparison of priors and posteriors of f_p and f_n based on data from March 13 to 23, 2020