

1

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

2

Song S. Qian,<sup>\*,†</sup> Maxwell D. Qian,<sup>‡</sup> and Sabrina Jaffe<sup>†</sup>

<sup>†</sup>*Department of Environmental Sciences, The University of Toledo, Toledo, OH 43606*

<sup>‡</sup>*Department of Economics and History, Vanderbilt University, Nashville, TN 27325*

E-mail: song.qian@utoledo.edu

3

## Abstract

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

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

## 21 **Introduction**

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

40 We present a statistical model to estimate the prevalence of the virus in each state  
41 using publicly available data. The quality of the model result depends on (1) our knowl-  
42 edge of the test (how well do we know the rates of false positives and false negatives)  
43 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 allow 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](https://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](https://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.

## 66 Statistical Methods

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

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

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

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

90 meaning of their test results.

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

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

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

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

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

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

100 As the three quantities of interest ( $\theta, f_p, f_n$ ) are probabilities, we use the beta distri-  
101 bution as their priors. We note that the likelihood function provides information on the  
102 products of  $\theta f_p$  and  $\theta f_n$ . Using independent forms of priors for  $\theta$ ,  $f_p$ , and  $f_n$  is unlikely  
103 to jointly estimate the three quantities. As a result, providing realistic informative  
104 priors for at least two of the three parameters is necessary.

## 105 Prior Specification

106 As the SARS-CoV2 is a new virus and only a relatively small number of tests are done  
107 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 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 ( $f_p \sim \text{beta}(1, 99)$  and  $f_n \sim \text{beta}(1, 99)$ , 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<sup>2-5</sup>, we constructed a prior distribution for false positive to be  $\text{beta}(3, 23)$  (95% credible interval of 0.025-0.26) and false negative  $\text{beta}(2, 22)$  (0.01-0.22)
- The worst-case scenario assumes that the test is as unreliable as the rapid influenza diagnostic test (RIDT)<sup>6</sup>. The RIDT test has a high false negative probability and we use  $\text{beta}(16, 24)$  (0.26-0.55) as the prior. The probability of a false positive is relatively low and we used  $\text{beta}(4, 45)$  (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  $\theta_j$  are exchangeable and impose a common prior. Advantages of using a hierarchical

133 formulation were explored elsewhere<sup>7,8</sup>. Expressing the model hierarchically, we have  
 134 the following model.

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

$$y_j \sim \text{Bin}(p_j, n_j)$$

137 where  $j$  represents the  $j$ th state,  $y_j$  and  $n_j$  are the observed number of positive  
 138 and total number of tests, and  $\text{Bin}$  represents the binomial distribution. The  
 139 probability of observing a positive result ( $p_j$ ) is a function of  $\theta_j$ ,  $f_p$ , and  $f_n$ :

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

- 140 2. To connect all states together, we used a common prior for state-level prevalence  
 141  $\theta_j$  (after a logit transformation).

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

142 Non-informative priors are used for  $\mu_0$  and  $\sigma_0^2$ .

- 143 3. Prior distributions of other parameters

$$f_n \sim \text{beta}(\alpha_n, \beta_n)$$

$$f_p \sim \text{beta}(\alpha_p, \beta_p)$$

144 The hyper-parameter  $\mu_0$  is the national average of prevalence (in logit scale) and  
 145  $\sigma_0^2$  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 state 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 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 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



173 results.

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

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

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

---

<sup>1</sup><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 positive and false negative is imperative for proper estimation of the disease prevalence in the population.

## Acknowledgement

## 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-Khedhairi, 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.

- 221 (5) thi Tham, N.; thi Ty Hang, V.; Khanh, T. H.; Viet, D. C.; Hien, T. T.; Farrar, J.;  
222 van Vinh Chau, N.; van Doorn, H. R. Comparison of the Roche RealTime ready  
223 Influenza A/H1N1 Detection Set with CDC A/H1N1pdm09 RT-PCR on samples  
224 from three hospitals in Ho Chi Minh City, Vietnam. *Diagnostic Microbiology and*  
225 *Infectious Disease* **2012**, *74*, 131–136.
- 226 (6) Center for Disease Control and Prevention, Rapid influenza diagnostic  
227 tests. 2016; [https://www.cdc.gov/flu/professionals/diagnosis/clinician\\_](https://www.cdc.gov/flu/professionals/diagnosis/clinician_guidance_ridt.htm)  
228 [guidance\\_ridt.htm](https://www.cdc.gov/flu/professionals/diagnosis/clinician_guidance_ridt.htm).
- 229 (7) Wu, R.; Qian, S.; Hao, F.; Cheng, H.; Zhu, D.; Zhang, J. Modeling Contaminant  
230 Concentration Distributions in China’s Centralized Source Waters. *Environmental*  
231 *Science and Technology* **2011**, *45*, 6041–6048.
- 232 (8) Qian, S.; Stow, C.; Cha, Y. Implications of Stein’s Paradox for Environmental  
233 Standard Compliance Assessment. *Environmental Science and Technology* **2015**,  
234 *49*, 5913–5920.

Table 1: Probability of infection given a positive result

Scenarios	$\theta_0$	$E(f_p)$	$E(f_n)$	$\Pr(v +)$
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

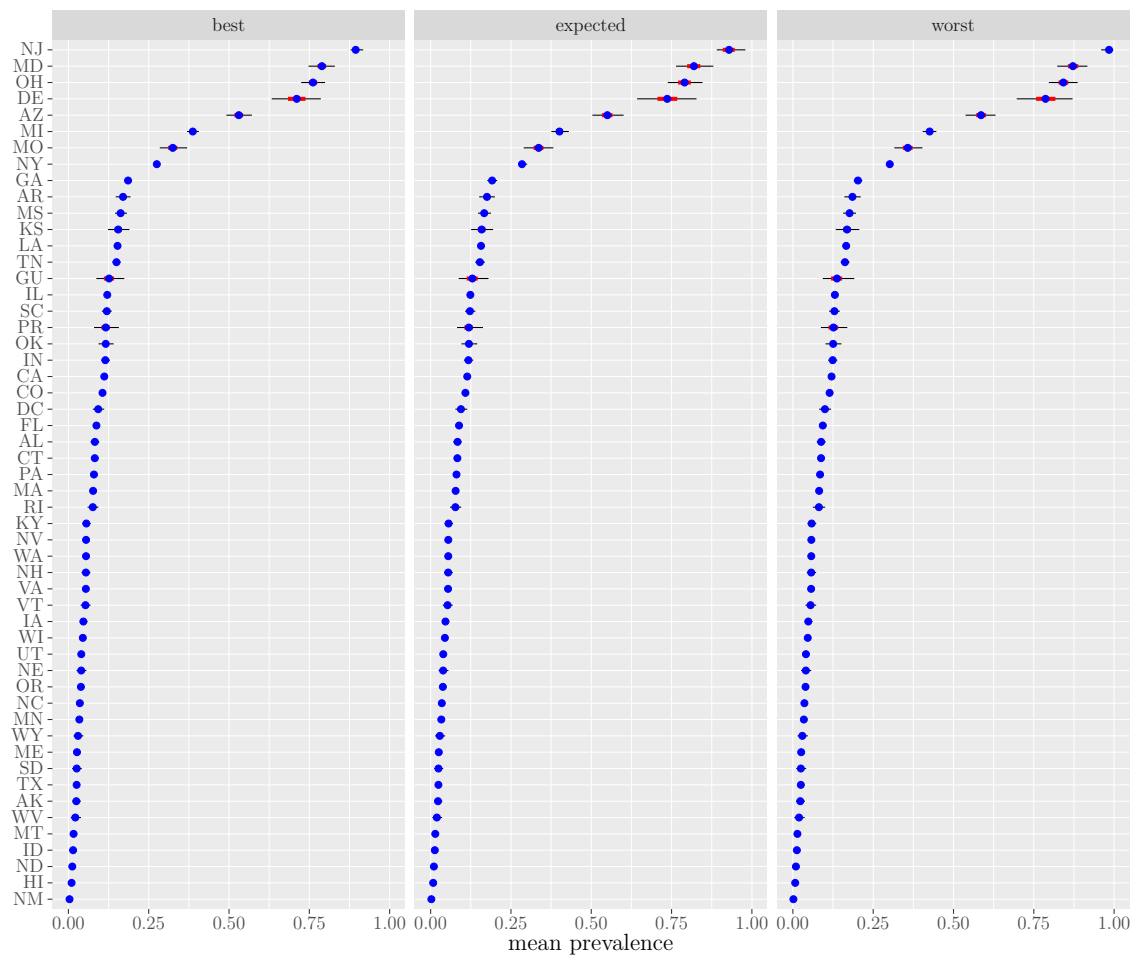


Figure 1: Hierarchical model estimated state-level SARS-CoV2 prevalence based on data reported by March 23, 2000. The blue dots are the estimated means, the red thick bars are the 50% credible intervals, and the thin black lines are the 95% credible intervals.

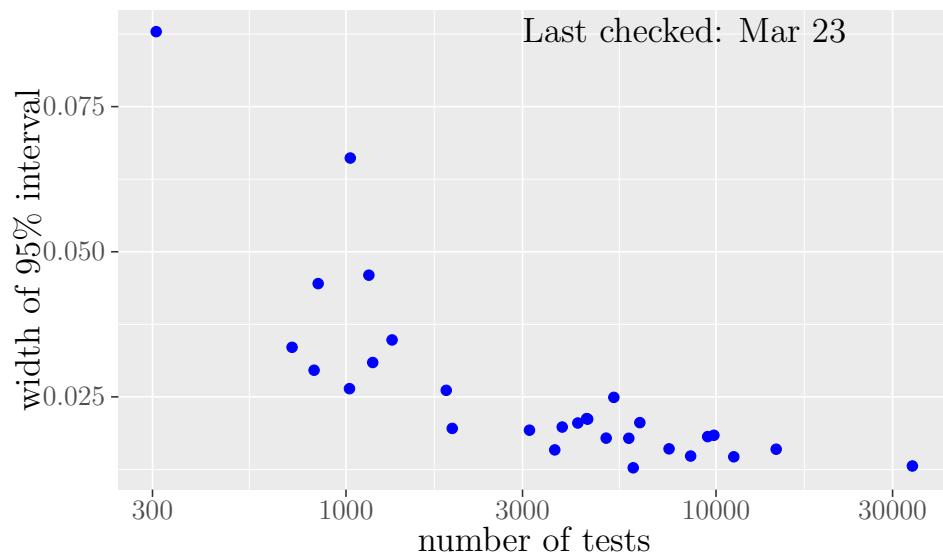


Figure 2: Uncertainty of the estimated state-level prevalence, measured as the width of the 95% credible interval (estimated using data from states with reliable data reporting system, see Figure 3), is a inversely related to the total number of tests performed.

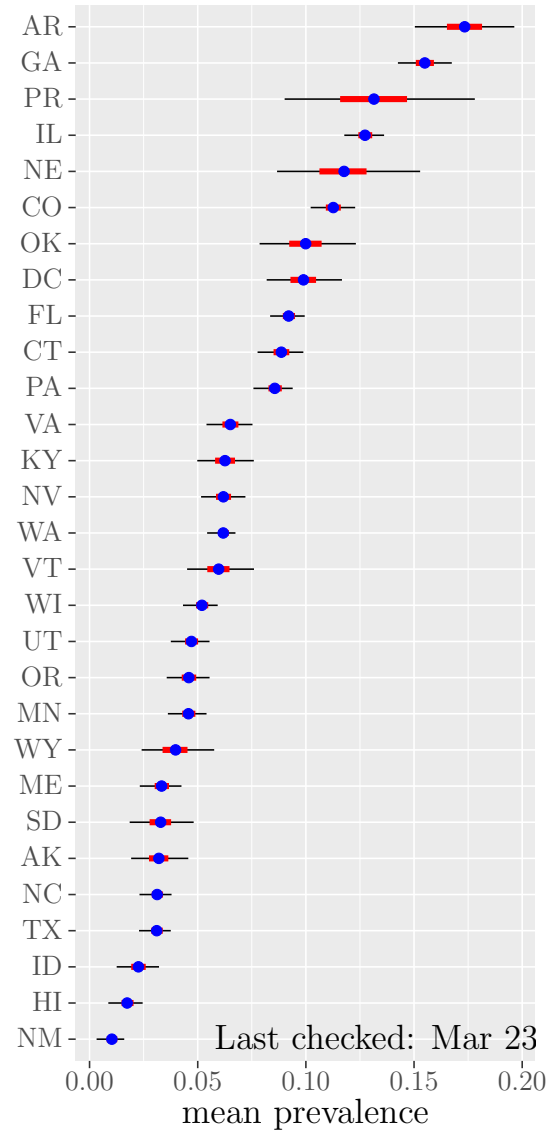


Figure 3: Estimated state-level prevalence using data from states with reliable data reporting system

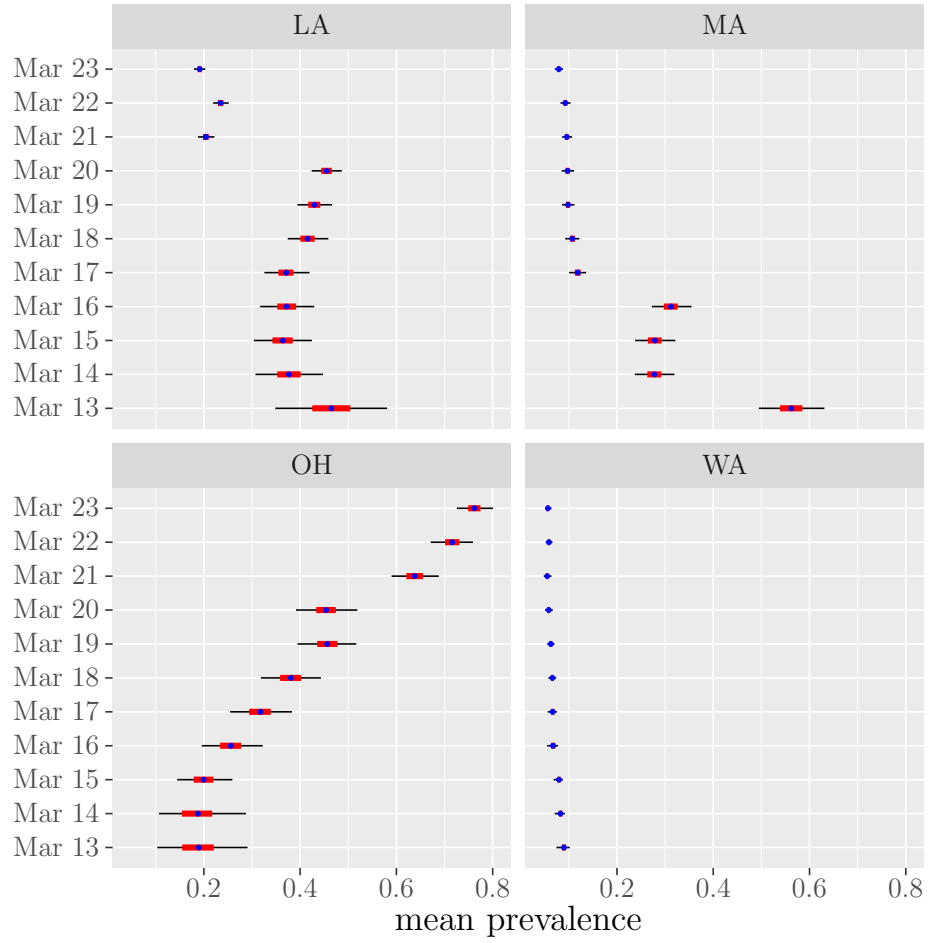


Figure 4: Estimated state-level prevalence for four states from March 13 to March 23, 2020



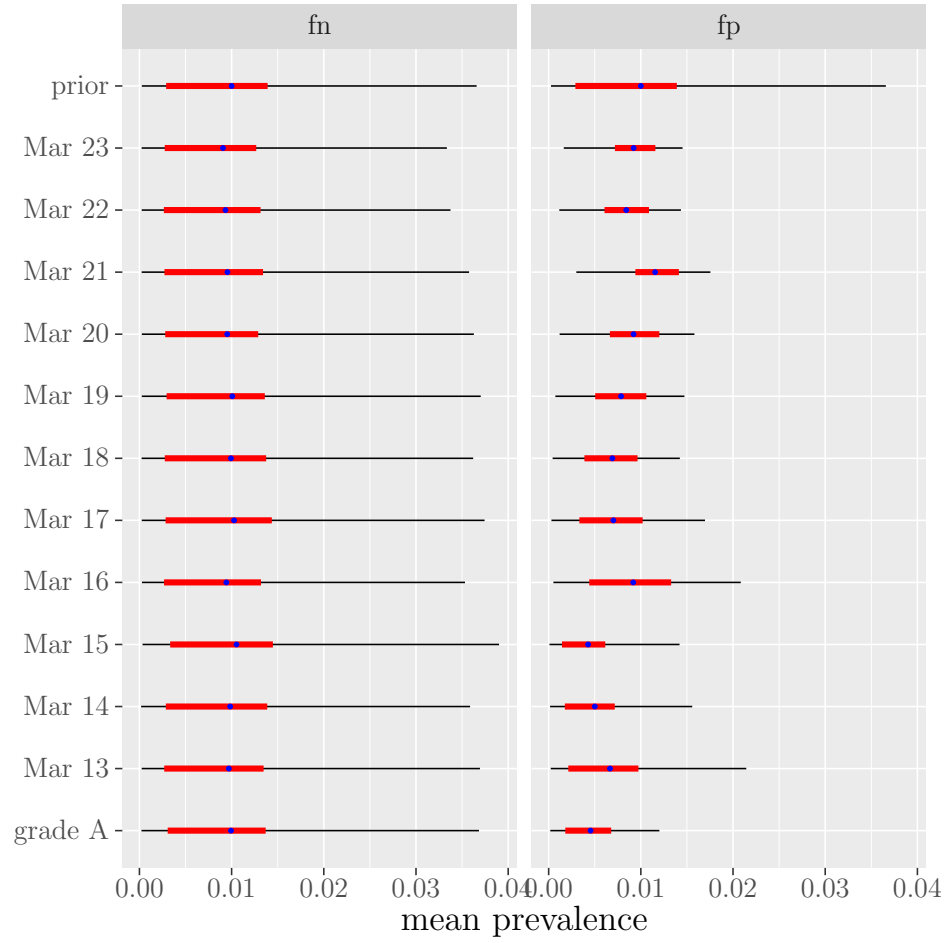


Figure 5: Comparison of priors and posteriors of  $f_p$  and  $f_n$  based on data from March 13 to 23, 2020