Analysis: The prevalence of SARS-CoV 2 in the US estimated from imperfect testing

Song S. Qian[1,3], Maxwell D. Qian[2] & Sabrina Jaffe[1]

[1] *Department of Environmental Sciences, The University of Toledo, Toledo, OH, USA*

[2] *Department of Economics and History, Vanderbilt University, Nashville, TN, USA*

[3] *Corresponding author, email: song.qian@utoledo.edu*

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 lack of sufficient testing capacity in the United States hinders the effort of stopping the spread of the virus. Furthermore, the US lacks a coherent reporting system where concerned citizens can find accurate information. The lack of information not only increases the anxiety of the American people, but also prevents the proper interpretation of a test result. Without an accurate estimation 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 delayed-release in the United States of the test for the novel coronavirus that emerged in Wuhan, China (SARS-CoV2) at the end of 2019 brought attention to the need for more testing. The need for more testing is especially urgent because the test for detecting SARS-CoV2 is imperfect, and it will occasionally produce false-positive or false-negative results. As a result, 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 allow us 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.

We used a Bayesian hierarchical modeling approach, which requires proper evaluation of existing information on the three quantities we want to estimate (prior distributions). Using the hierarchical model structure, we assumed that state-level prevalence are exchangeable (different but otherwise uncertain) and imposed a non-informative prior on the hyper-parameters (national-level prevalence). We don't have reliable information on the rates of false positives and false negatives for the current test. Based on literature information, we proposed three scenarios and parameterized the priors using the commonly used beta distribution [2]. The best-case scenario assumes that the test is highly accurate and reliable. The average rates of false positive and false negative are both 1% and stable. The worst-case scenario is to assume that the test is as reliable as the Rapid Influenza Diagnostic Test (RIDT)[3], which has a false-negative rate between 26-55% and a false-positive rate of 2.3-17%. The middle (or the expected) scenario assumes that the current test is similar to the RT-PCR test used during SARS and MERS outbreaks, with a false-positive rate ranging from 2.5 to 26% and a false-negative rate ranging from 1 to 22% [4–7].

Under the best-case scenario, the national mean prevalence is 0.096 (with a 95% credible interval of between 0.069 and 0.128) based on reported data as of March 15, 2020. The lowest state-level prevalence is 0.014 (0.0038-0.030) and the highest is 0.443 (0.377-0.508). Under the worst-case scenario, the national average prevalence is 0.153 (0.090-0.241), and the state-level prevalence ranging from 0.0144 (0.0017-0.038) to 0.756 (0.569-0.971) (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). A comparison of the estimated prevalence between the states of Washington and Massachusetts is highly revealing. As of March 15, 2020, a total of 7,764 and 490 people were tested (with results) in Washington and Massachusetts, respectively. Although the high number of fatalities in Washington is widely reported, the large number of tests performed in the state allowed a more accurate estimate of the prevalence of the disease in Washington (0.080 with a 95% credible interval between 0.069 and 0.088). The estimated prevalence is not particularly high and the uncertainty of the estimate is among the lowest. In contrast, the estimated prevalence in Massachusetts is among the highest in the nation (0.279) and is highly uncertain (0.238-0.321). The high level of estimation uncertainty is largely a result of the small number of tests and the unusually high prevalence could be a result of the highly selected screening process, as reported by the New York Times [8]. As of March 16, 2020, Georgia, Hawaii, Texas, and Wyoming did not report negative counts. New Jersey changed the way it reports negatives on March 16th (number of negatives dropped to 20 from the previous day's 120).

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. To evaluate the effectiveness of the current prevention practices, we need to estimate the prevalence in the general population to monitor its trend. As a result, more testing is needed to include people in the general population.

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 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`).

Statistical Methods The basic statistical approach is documented by [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 probability 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:

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

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 the 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 [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 [1]:

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

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.

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 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.

- Estimates assuming a highly accurate test Assuming 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).

- Estimates based on similar 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 a study of the MERS and SARS virus tests [4–7], 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)

- Estimates assuming the test is as unreliable as the rapid influenza diagnostic test (RIDT) [3] This test has a high false negative probability and we use $beta(16, 24)$ (0.26-0.55). 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 50 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. 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. The probability of observing a positive result (p_j) is a function of θ_j, f_p, and f_n:

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

2. To connect all states together, we used a common prior for state-level prevalence θ_j.

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

   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.

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* **6**, e03571 (2020).

2. Gelman, A. *et al. Bayesian Data Analysis* (CRC Press, Boca Raton, Florida, 2014), 3rd edn.

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

4. 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* **56**, 153–160 (2006).

5. Binsaeed, A. A. *et al.* 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* **31**, 351–355 (2011).

6. Rainer, T. H. *et al.* The spectrum of severe acute respiratory syndrome–associated coronavirus infection. *Annals of Internal Medicine* **140**, 614–619 (2004).

7. thi Tham, N. *et al.* 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* **74**, 131–136 (2012).

8. Stockman, F. Sick people across the u.s. say they are being denied the coronavirus test. *The New York Times* (March 12, 2020).

Competing Interests The authors declare that they have no competing financial interests.

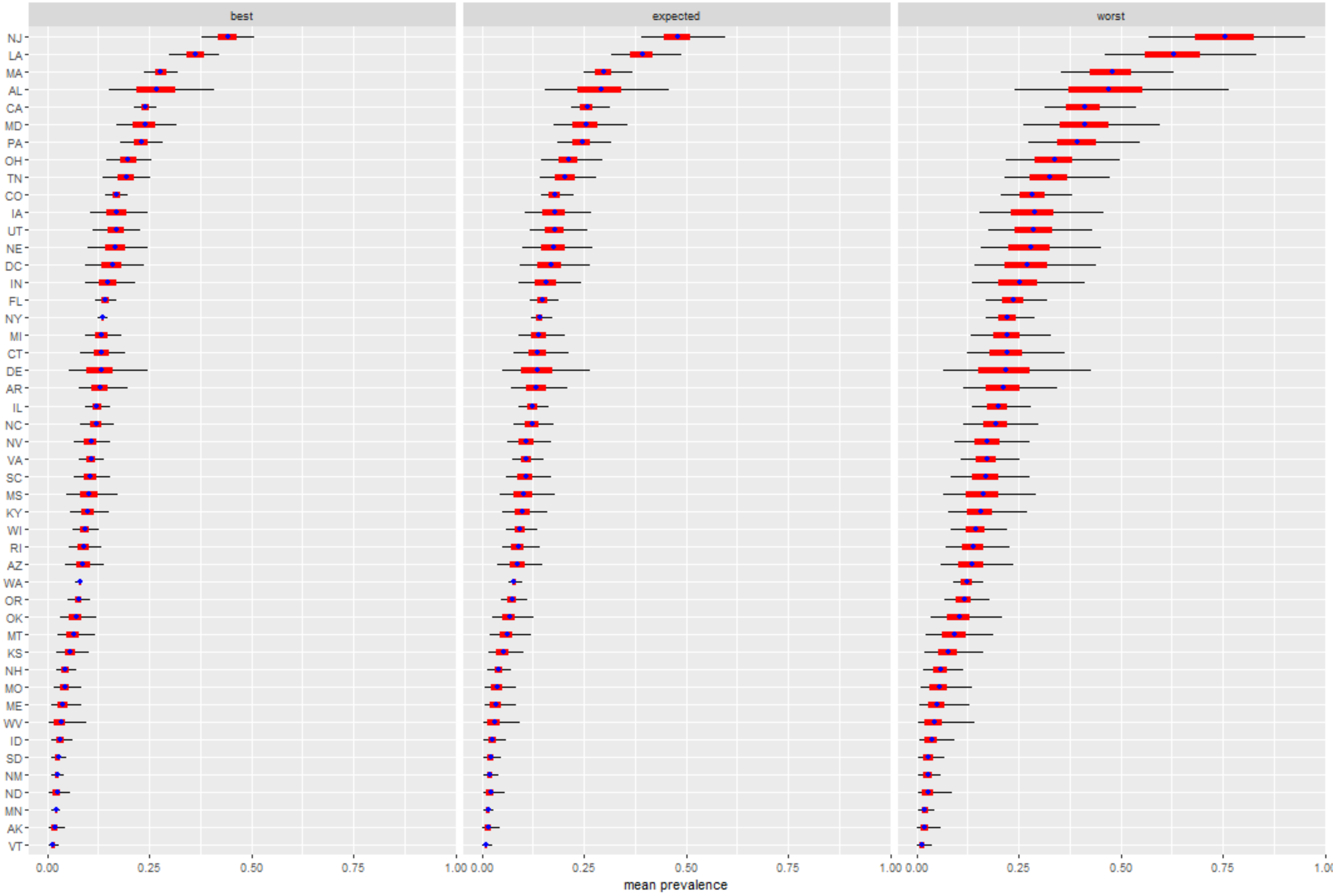
Correspondence Correspondence and requests for materials should be addressed to S.S.Q.

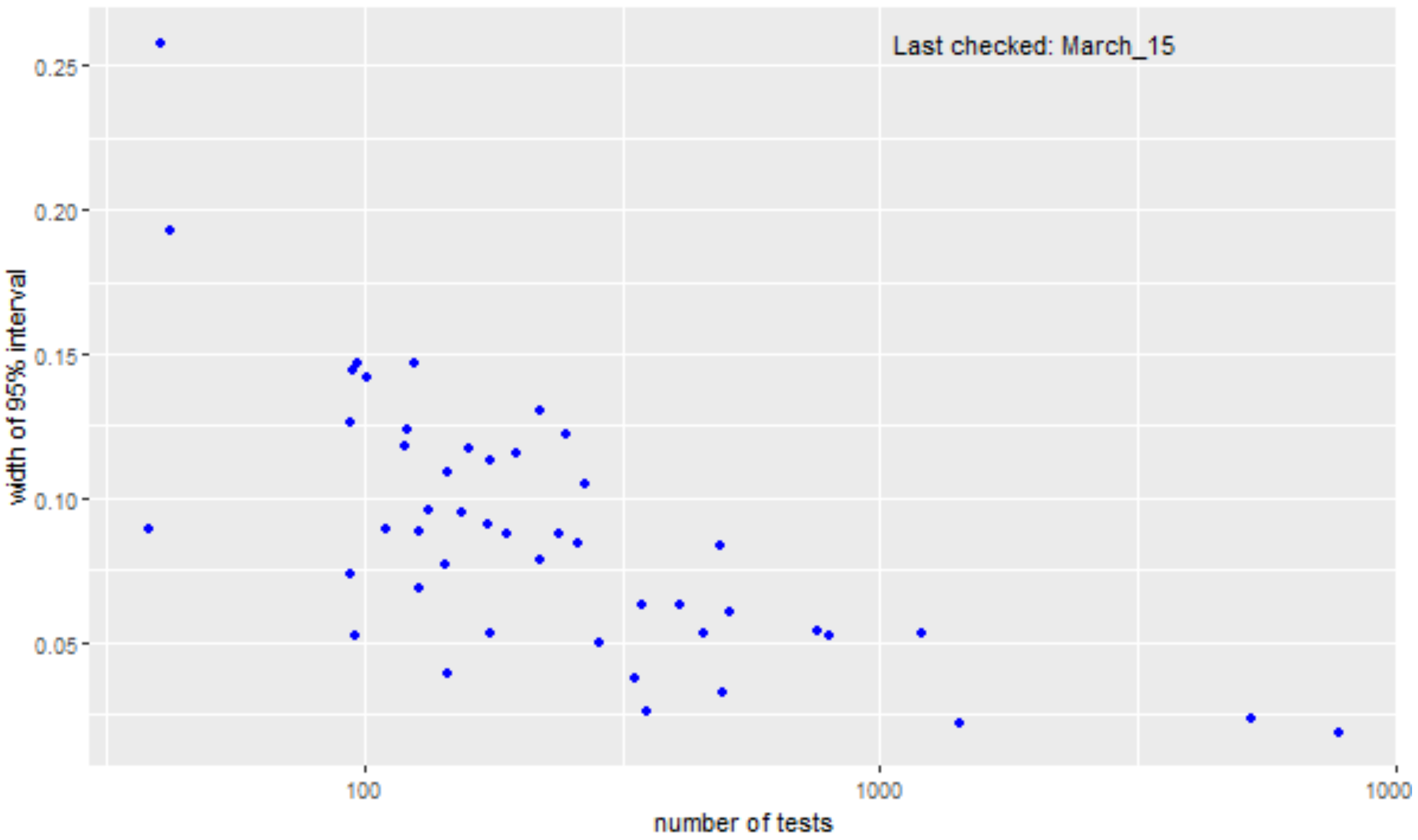
(email: song.qian@utoledo.edu).

Additional details and computer code are available at `https://github.com/songsqian/covid19`

Figure 1 Estimated state-level prevalence under three scenarios of test quality. Thin black lines represent the 95% credible intervals, red boxes represent the 50% intervals, and blue dots are the estimated means.

Figure 2 Estimation uncertainty measured as the width of 95% credible interval. Prevalence estimation uncertainty is inversely related to the number of tests performed.


best
expected
worst
NJ
LA
MA
AL
CA
MD
PA
OH
TN
CO
IA
UT
NE
DC
IN
FL
NY
MI
CT
DE
AR
IL
NC
NV
VA
SC
MS
KY
WI
RI
AZ
WA
OR
OK
MT
KS
NH
MO
ME
WV
ID
SD
NM
ND
MN
AK
VT
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.0
mean prevalence



Last checked: March_15
width of 95% interval
0.25
0.20
0.15
0.10
0.05
100
1000
1000
number of tests