Approximating the Number of COVID-19 Infections in the United States Across 2021

Quinn White

2023-07-23

Abstract

Introduction

Results

State-level Estimates

At both the state and county levels, we considered multiple implementations of the probabilistic bias analysis to compare estimates produced using different specification of the priors using survey data from the COVID-19 Trends and Impact Survey[1].

However, for simplicity, at the state-level we focus on the implementation where the priors do not vary by state or date. This also allows us to consider the entirety of 2021, since survey data is only available for dates after March 20, 2021. A full comparison of implementations is included in Supplementary Figure covidestim-concordance-state.

In Figure 1, we consider three distinct two-week intervals during waves of the pandemic in 2021.

Although incidence of COVID-19 was highest during the time interval during the Omicron wave, as is clear in Figure 2, we see that the ratio of estimated infections to observed infections is higher during the time intervals in the alpha and delta waves. This distinction is explained by the differences in testing rates during these period: on average, the testing rate during this two-week interval during the omicron wave was 2.4 times that of the alpha wave and 4.9 times that of the delta wave. This particular time interval during the delta wave; June 16, 2021 through July 1st, 2021; had the maximum ratio of estimated to observed infections, and corresponded to the minimum testing rate (Supplementary Figure testrate-low-summer).

Several states consistently have among the highest or lowest ratios of estimated to observed infections. In particular, there are 6 states with among the lowest 10 ratios of estimated infections to observed infections, and as such the highest case ascertainment rates, for more than 80% of time intervals considered. These states were Rhode Island, Massachusetts, District of Columbia, Alaska, New York, and Vermont. Meanwhile, states that had the highest ratios, and equivalently the lowest case ascertainment rates, include Mississippi, South Dakota, Oklahoma, Nebraska, and Tennessee.

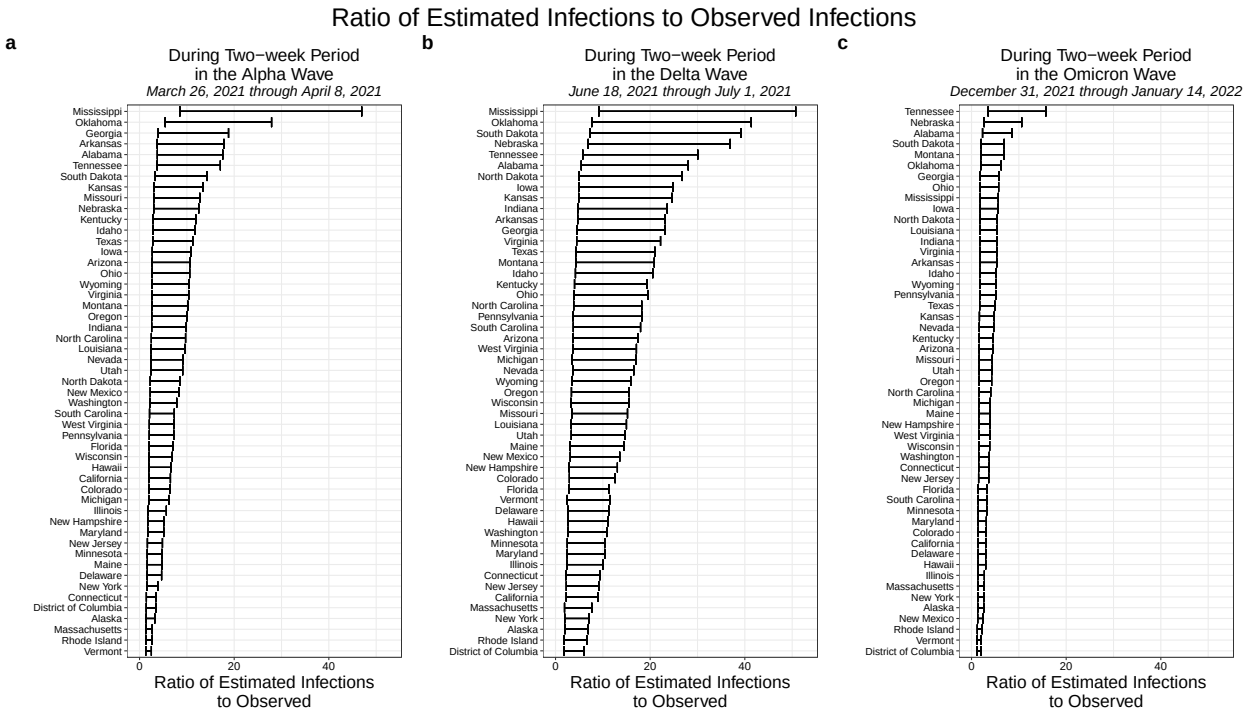


Figure 1: The ratio of estimated infections to observed infections for three time intervals of interest: one during the alpha wave, one during the delta wave, and one during the omicron wave. Although the prevalence of COVID-19 was highest during the omicron wave, the ratios of estimated to observed infections are higher for the time intervals during the alpha and delta waves, a difference that was driven by lower testing rates during these times. The trend we see in these three time intervals where Mississippi, South Dakota, Oklahoma, Nebraska, and Tennessee have among the highest ratios of estimated infections to observed infections, and as such the lowest case ascertainment rates, is consistent across the full set of time intervals considered from January of 2021 to March of 2022.

Comparison to Covidestim

Because there is no established ground truth for the true number of infections for any time-interval, a useful source of comparison is other models that also estimate the true number of incident infections. Covidestim is a notable Bayesian nowcasting model that has been maintained throughout the course of the pandemic, sharing publicly available estimates at the state and county levels[2]. Consequently, we compare our estimates to Covidestim, with recognition that the Covidestim model is not a ground truth, and this comparison serves primarily to gain insight into the times and locations where estimates are concordant or discordant rather than to validate the approach. Given the true number of infections is unobserved, we are to some extent operating in the unknown, and must rely on assumptions based on available data.

In Figure 2, we see the simulation intervals for all two-week intervals and all states as well as the 95% Covidestim credible intervals summed to be on the same time scale. We see agreement is much higher before the time period spanning December 2021 through January 2022, where Covidestim intervals tend to be higher than the probabilistic bias intervals. This period corresponds to the Omicron wave of the pandemic (Supplementary Figure michigan-variant). This was a major shift in the pandemic: the Omicron variant is highly transmissible compared to previous variants, and it has immune invasion capacity, which means vaccines provides less protection against infection[3]. Omicron also is associated with a lower infection fatality ratio[4], as well as milder infection overall, which may influence testing behavior. Omicron's rise was rapid too, presumably as a result of this enhanced transmissibility and immune escape; it became the dominant variant over the course of a single month.

As a result of these dramatic changes, Chitwood *et al.*[2] made substantial changes to Covidestim, as noted in the model change log. Because the variant causes much milder infections, the infection fatality for Omicron infections is lower than previous variants, and death counts were much lower relative to the number of infections. To handle this change, rather than fitting model with deaths, they switched to using hospitalizations. They also allowed for the possibility of reinfections, since although reinfections were more rare with previous variants, Omicron is associated with higher reinfection rates[5]. The changes in the model may contribute to the differences we see between the probabilistic bias intervals and Covidestim intervals during the Omicron wave.

Another difference we see is in the months of the summer of 2021, where the increase in Covidestim estimates appears lagged in comparison to the probabilistic bias simulation intervals. The difference we see between the Covidestim estimates and probabilistic bias intervals is likely a result of the way these approaches treat incomplete testing. The focus of the probabilistic bias analysis is to correct for incomplete testing, and as such the method is sensitive to changes in the total number tested and the positivity rate. By contrast, while Covidestim models probabilities of diagnosis by symptom state to allow for variation in case ascertainment, the total number of tests is not an input into the model, so model estimates are not affected by changes in testing rate. This is particularly relevant during the summer of 2021, when an increase in test positivity preceded the increase in observed tests, leading to an increase in the probabilistic bias simulation intervals that precedes the increase in Covidestim estimates.

Figure 2: Simulation intervals for each 2-week interval considered, for all states. For any given state, each vertical bar shows the 2.5% percentile and 97.5% percentile for the total number of infections in that two-week interval. Covidestim intervals summed over the same two-week time-scale are shown in red. The scale on the y-axis is distinct across states to highlight differences across time within each state.

County Level Estimates in Massachusetts

At the county level, this approach to approximating the true number of infections is only possible in a subset of the counties in the United States due to the need for both positive tests and total tests. While positive tests are frequently reported at the county level, total tests are not.

We focus on Massachussetts because this state is the only state that reports both positive and total tests and has the most counties where wastewater data was reported throughout the entire time period of interest. Wastewater data here is useful in the sense that it is inherently not subject to the same biases as voluntary testing data, since it captures anyone in the wastewater catchment area, and as such is not influenced by more symptomatic people being more likely to be tested or differences in access to testing.

In Figure 3, we present simulation intervals for two of the four implementations of probabilistic bias analysis. For most time intervals, the other two implementations fall between those shown in the figure.

In the three out of the four implementations, the prior distributions specified for the bias parameters vary by state and date. In particular, we use survey data from the COVID-19 Trends and Impact Survey[1] to inform the prior distributions for one to two of the bias parameters: the probability of having symptoms among the untested population ($\Pr(S_1|\text{untested})$), and the parameter β, which represents the ratio of the test positivity among members of the untested population who have no symptoms to the overall test positivity. More detail on these implementations is provided in the Methods section. In another, more simple, implementation, which is the only implementation presented at the state-level (Figure 1 and 2), the prior distributions are informed by the distributions from the survey data across all dates and states, but the same prior distributions are used for all states and time intervals considered.

While there is substantial overlap among all of the versions, the implementation allowing β and $\Pr(S_1|\text{untested})$ to vary by state and date tends to be the highest, followed by that only allowing β to vary by state and date, followed by that only allowing $\Pr(S_1|\text{untested})$ to vary by state and date, and lastly the implementation that does not vary the priors by state or date tends to produce the lowest estimates (Supplementary Figure sim-intervals-faceted-county).

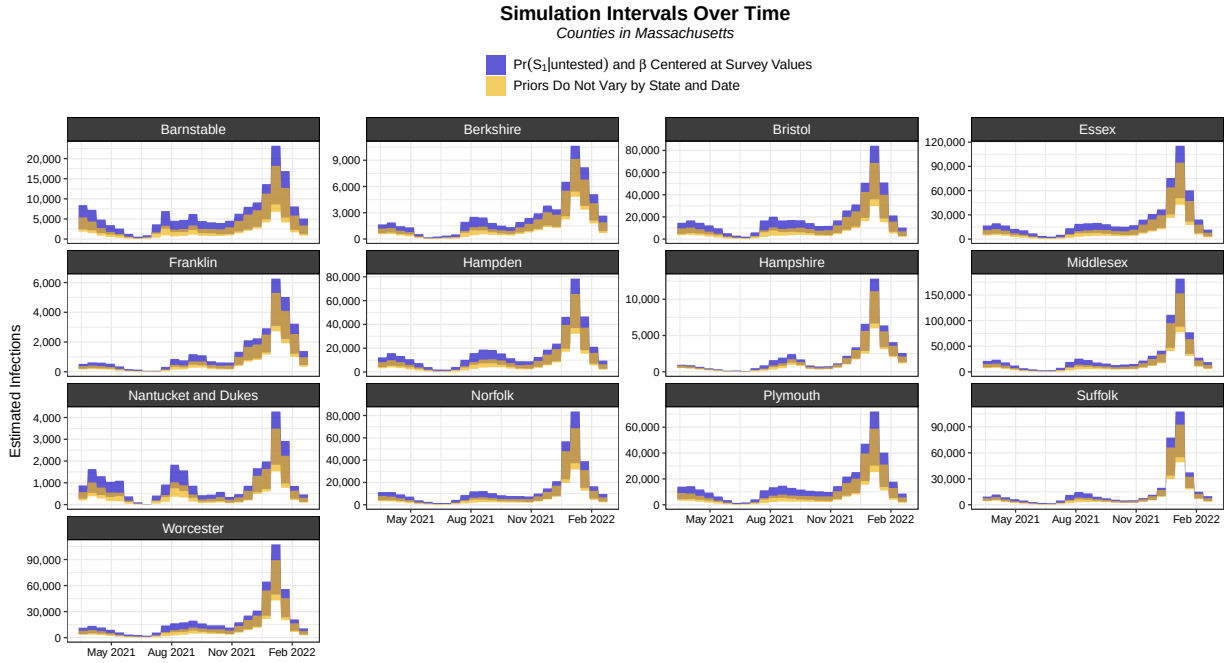


Figure 3: Simulation intervals for counties in Massachusetts, colored by the implementation of probabilistic bias analysis. Only the two implementations that were consistently the highest and lowest among the implementations are included for clarity. In the In the three out of the four implementations, the priors vary by state and date, while for the fourth, they are the same for all states and time intervals considered. For the first implementation, we center the distribution of β at the ratio of the screening test positivity to the overall test positivity from the survey, and we center the distribution of $\Pr(S_1|\text{untested})$ at the percentage of the population experiencing COVID-19-like illness from the survey for each two-week interval. The second implementation, we center only β at the aformentioned value, and the for the next we only center $\Pr(S_1|\text{untested})$ at the aforementioned value. The last implementation corresponds to the implementation where we specify priors that are the same for all dates. The implementation that centers both $\Pr(S_1|\text{untested})$ and β at the survey values is consistently the highest among the implementations, followed by the implementation that centers only β at the survey value, followed then by the implementation only centering $\Pr(S_1|\text{untested})$ at the survey value, and then the lowest among the implementations is that where the priors do not vary by date.

Also of interest beyond the total numbers of infections is the ratio of the infections that are estimated to those that were observed, as this gives us a sense for the case ascertainment at different time periods in the pandemic. We see in Figure 4 that the two-week interval estimated to have the highest ratio of estimated to observed infections was July 2, 2021 through July 30, 2021, during the delta wave. This corresponds to a similar time frame that we saw at the state-level in Figure 1. Hampshire County and Suffolk County were consistently among the lowest with regard to the ratio of estimated to observed infections, which may be explained by the screening testing occurring at universities in these counties.

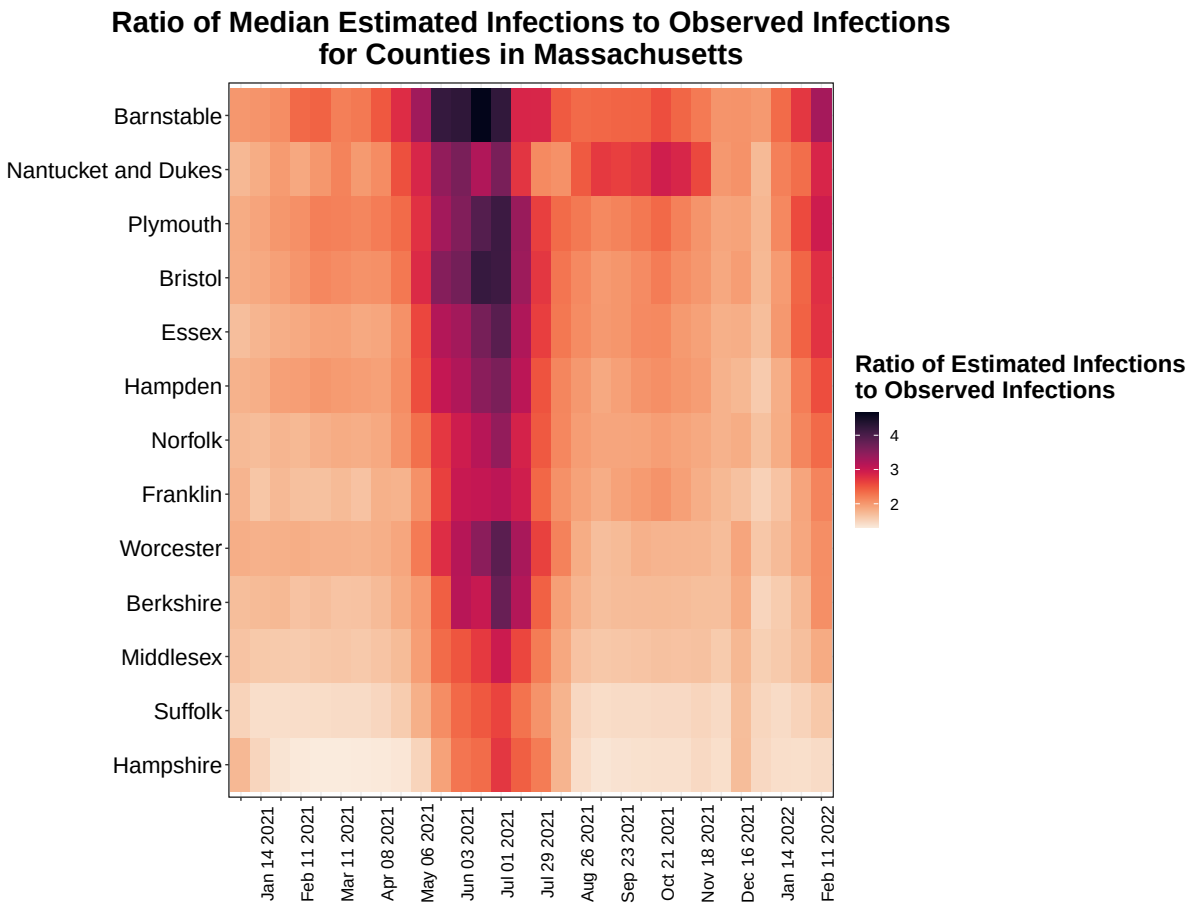


Figure 4: The ratio of estimated to observed infections across time for counties in Massachusetts. Counties are ordered by the median ratio across time intervals, from the highest ratio (Barnstable) to the lowest (Hampshire). Similar to what we see at the state level, the highest ratios were during the summer of 2021 during the Delta wave – a period of decreased testing. The span of time with the highest ratio of estimated to observed infections was July 2, 2021 through July 30, 2021.

As we see in Figure 5, the nature of the relationship between the testing rate and the ratio of estimated infections to observed infections depends on whether we allow β and $\Pr(S_1|\text{untested})$ to vary by location and date. In particular, when we sample from the same priors for every correction (the first panel of Figure 5, we see there is little variability in the relationship between the testing rate and median estimated infections, because the form of the correction is identical for each two-week interval and state considered. Allowing β and/or $\Pr(S_1|\text{untested})$ to vary by time and location introduces additional variability in the relationship between the ratio of estimated infections to observed and testing rate.

The nonlinearity of the relationship between the testing rate and the ratio of estimated infections to observed infections is more clear when we consider how we calculate the positives among the untested population. To do this, we split up the population N into the number tested, $N_{untested}$, and untested, N_{tested}, for that two-week interval.

Denoting N^* to be the number who would test positive for COVID-19 if they were tested, on the y-axis, the ratio of estimated infections to observed infections is approximately[1] $\frac{N^*_{\text{tested}}+N^*_{\text{untested}}}{N^*_{\text{tested}}}$, where we calculate

[1]This isn't exactly the estimated infections, because for simplicity of notation we are not writing out the correction for test inaccuracy.

N^*_{untested} using the specified priors, and N^*_{untested} is the observed positive tests.

On the x-axis, we have the number tested over the population size, $\frac{N_{\text{tested}}}{N}$. Thus, we see the trend in each panel where for small changes in testing rate when the testing rate is very low, the ratio of unobserved to unobserved is very high since N^*_{untested} will be large relative N^*_{tested}, that is, a large proportion of infections are going undetected. However, with higher testing rates, N^*_{untested} will be smaller relative N^*_{tested}, and the ratio of estimated to observed infections nears one.

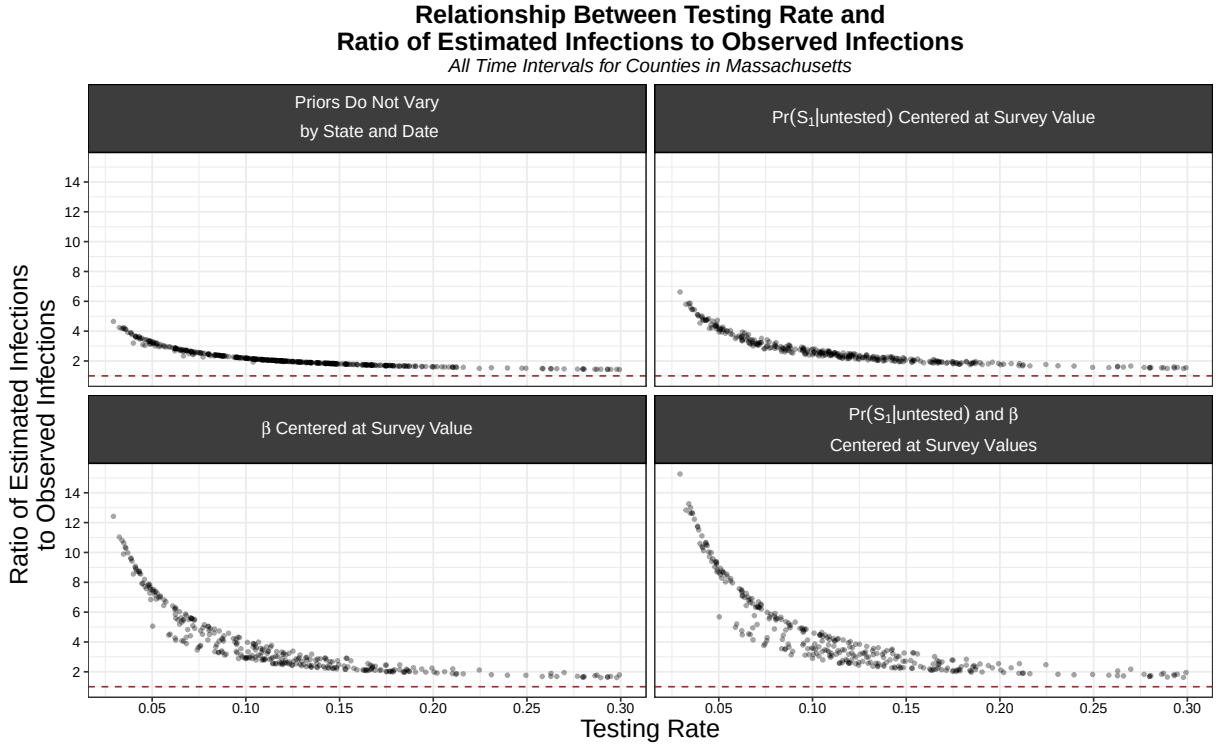


Figure 5: The ratio of the median estimated infections to observed infections plotted against the testing rate, where the testing rate is calculated as the total number tested in a two-week interval over the population size. When the priors are the same for all time intervals, there is minimal variability relationship between the testing rate and the ratio of estimated to observed infections, since the correction for incomplete testing and diagnostic test inaccuracy is identical for each time-interval. However, when we allow β or $\Pr(S_1|\text{untested})$ to vary over time, there is more variability in the relationship. A horizontal line in red at 1 is included to reference; a ratio of exactly 1 would indicate no infections went unobserved.

Comparison to Wastewater Data and Covidestim Estimates

As with the state level results, we also compare estimates at the county-level to Covidestim estimates. However, at the county-level, we proceeded with Massachusetts specifically with the aim of providing another source of comparison: wastewater concentrations.

We see in Figure 6a) that for all implementations, the probabilistic bias estimates are highly correlated with wastewater concentrations, and in Figure 6b) we see they also highly correlated with the Covidestim estimates. The correlations in both cases are very similar between implementations, differing by less than 0.01.

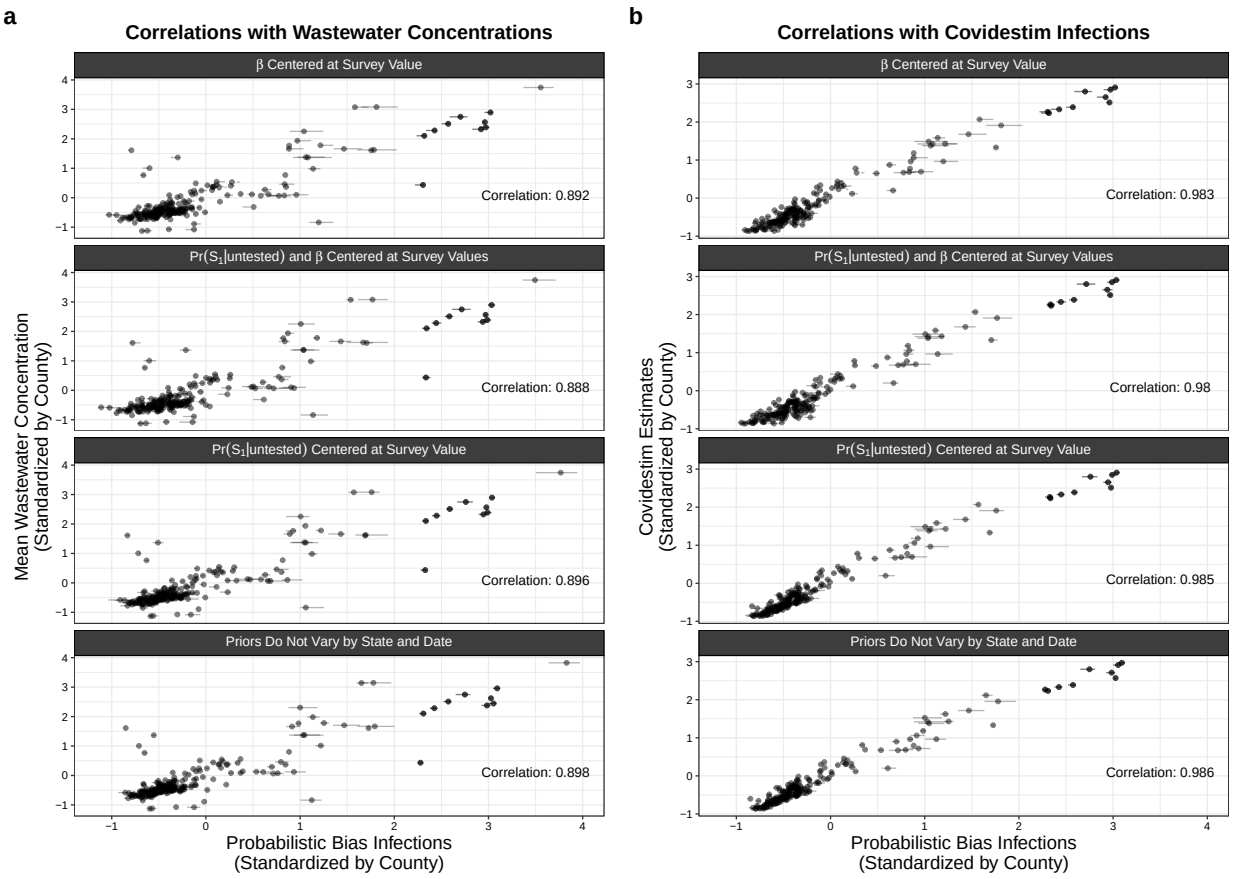


Figure 6: Considering the correlations between the probabilistic bias estimates and wastewater concentrations (a) and between probablistic bias estimates and Covidestim estimates (b). We see that all implementations considered are highly correlated with both wastewater concentrations and Covidestim estimates. In both panels, each point is a county-biweek, where the value on the x-axis is the median of the simulation interval for that county in that two-week interval. For (a), the value on the y-axis is the mean wastewater concentration for that county and two-week interval. For (b), the value on the y-axis is the median of the Covidestim credible interval for that county, summed to be on the same two-week-interval time scale. The correlations are highly similar between implementations, differing by less than 0.01.

Table 1 is

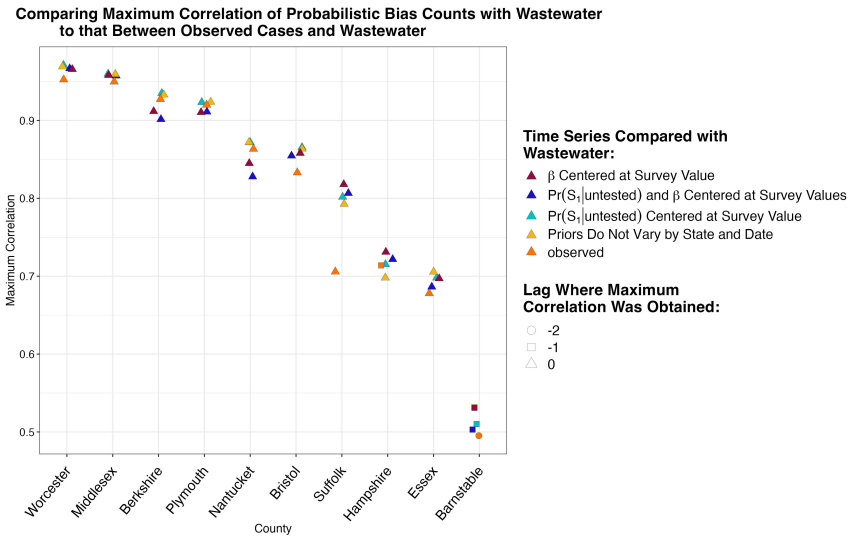
Table 1: Coverage of Covidestim Medians at the County and State Levels

Implementation	Percent Below Interval	Percent Contained in Interval	Percent Above Interval
County			
$Pr(S_1$|untested) Centered at Survey Value	1.282	87.500	11.218
Priors Do Not Vary by State or Date	1.075	74.194	24.731
β Centered at Survey Value	29.487	63.782	6.731
$Pr(S_1$|untested) and β Centered at Survey Values	34.615	62.179	3.205
State			
$Pr(S_1$|untested) Centered at Survey Value	0.889	79.111	20.000
Priors Do Not Vary by State or Date	5.882	71.569	22.549
β Centered at Survey Value	16.296	70.963	12.741
$Pr(S_1$|untested) and β Centered at Survey Values	22.222	66.815	10.963

The percent of simulation intervals where the Covidestim median falls below, within, or above the interval,when considering all simulation intervals for that implementation and geographic scale.Implementations are ordered from those with the highest proportion of Covidestim medians contained in the interval, to the lowest.

Also of interest is whether there is a lag between the wastewater concentrations and bias corrected estimates. Although wastewater concentrations can be a leading indicator since people may shed viral material before developing symptoms and subsequently getting tested, the lead time is a result of multiple epidemiological factors, including viral shedding dynamics that may differ with evolution of the virus, access to testing, and testing behavior[6]. Indeed, the lead or lag time between wastewater concentrations and hospitalizations or cases has varies substantially between waves, where the lead time is higher in earlier waves and near zero at later waves, a difference that may be attributable to changes in diagnostic test availability over time[7,8].

In Figure **??**, we see that among all counties except one in Massachsuetts, the maximum correlation between mean wastewater concentration and biweekly infection estimates was obtained at a lag of zero, that is, there typically there was no lag among the biweekly infection estimates and mean wastewater concentrations. This is to be expected given the lead time for wastewater concentrations, when observed, is less than 2 weeks[6]. In two counties, Hamsphire and Barnstable, the lag between the probabilistic bias counts and wastewater concentration was less than that between the observed positive tests and wastewater concentrations.



Methods

Data

Massachusetts County Level

State Level

Survey Data

The COVID-19 Trends and Impact Survey was run in collaboration by . . .

Wastewater Data

Biobot analytics . . .

Statistical Methods

Sample from priors on $\Pr(S_1|\text{untested}), \alpha, \beta$

$\downarrow$

Constrain priors with Bayesian melding to obtain constrained distributions

for $\Pr(S_1|\text{untested}), \alpha, \beta,$ and $\Pr(S_0|\text{test}_+, \text{untested})$

$\downarrow$

For each geographic unit, use sampled α and β to calculate:

$$\Pr(\text{test}_+|S_1, \text{untested}) = \alpha \ \Pr(\text{test}_+|\text{tested})$$
$$\Pr(\text{test}_+|S_0, \text{untested}) = \beta \ \Pr(\text{test}_+|\text{tested})$$

$\downarrow$

$$N^*_{\text{untested},S_0} = N_{\text{untested}}(1 - \Pr(S_1|\text{untested}))\Pr(\text{test}_+|S_0, \text{untested})$$
$$N^*_{\text{untested},S_1} = N_{\text{untested}}(\Pr(S_1|\text{untested}))\Pr(\text{test}_+|S_1, \text{untested})$$

$\downarrow$

Estimate total unobserved positive tests as

$$N^*_{\text{untested}} = N^*_{\text{untested},S_0} + N^*_{\text{untested},S_1}$$

$\downarrow$

Take the sum to acquire total positive tests

$$N^* = N^*_{\text{untested}} + N^*_{\text{tested}}$$

$\downarrow$

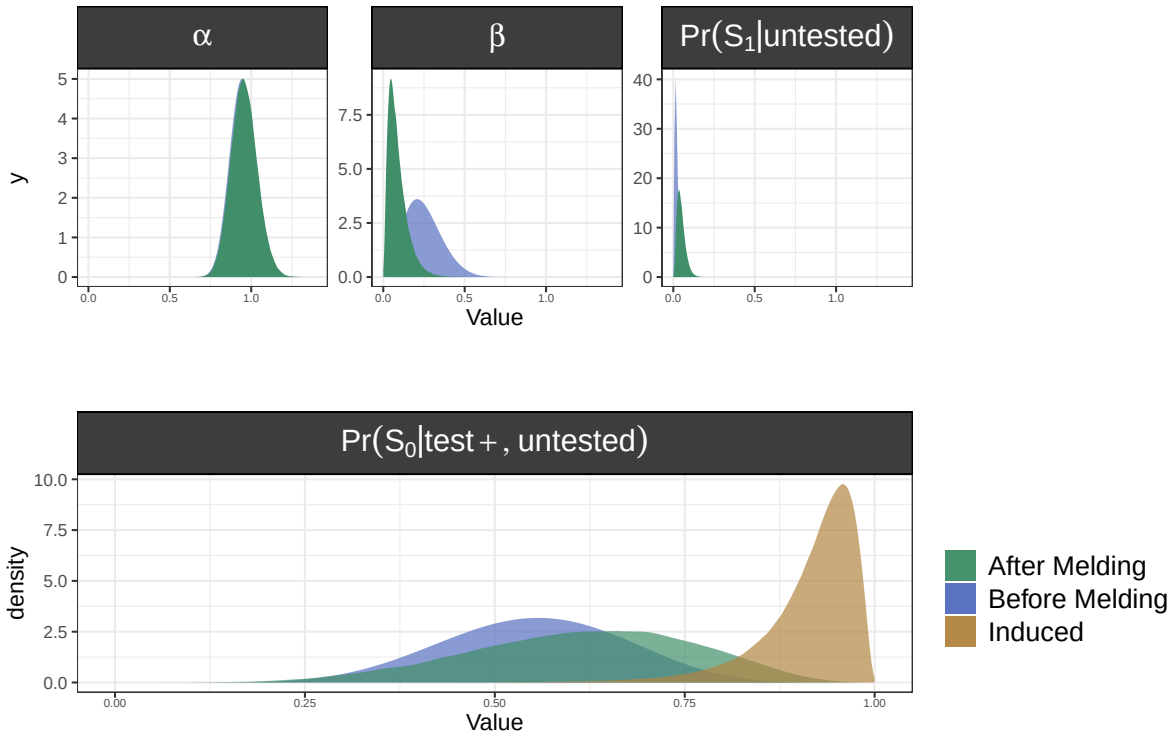
Correct for diagnostic test inaccuracy

$$N^+ = \frac{(N^* - (1 - S_p)N)}{(S_e + S_p - 1)}$$

Probabilistic bias analysis

Bayesian Melding

Specification of Priors



α

β

$\Pr(S_1|\textbf{untested})$

$\Pr(S_0|\textbf{test}_+, \textbf{untested})$

There is substantial heterogeneity in estimates of the percent of infections that are asymptomatic. This is in part due to distinct study populations and selection criteria. In particular, estimates from screening studies may be better estimates of the asymptomatic rate among the untested population: estimates from studies where the population was not screened, and as such was comprised of individuals that sought out a PCR test, may include a higher proportion of symptomatic individuals, biasing estimates of the asymptomatic rate downwards.

One meta-analysis included studies across the globe as of February 4, 2021, and estimated the pooled percent of asymptomatic infections among confirmed infections to be 40.50% (95% CI 33.50%-47.50%)[9]. This analysis did not restrict to screening studies. Another meta-analysis, when restricting to screening studies, found

the pooled asymptomatic percentage to be 47.3% (95% CI, 34.0 - 61.0%)[10]. Both meta-analyses noted the substantial amount of heterogeneity in the percent of asymptomatic infections.

In a large screening study where the sample was individuals arriving from overseas, the asymptomatic rate was 76.8%[11]. A screening study among children admitted to a pediatric emergency department between May 2020 and January 2021 found the asymptomatic rate to be 51.7%[12].

Several studies were conducted among university students. Among students at the University of Arizona in the fall semester of 2020, including students who sought testing and who were required to test, the asymptomatic rate of infection was 79.2%. A study at the University of Notre Dame distinguished between presymptomatic infection and asymptomatic infection, and found found 32% to be asymptomatic throughout the entire course of infection, 27.0% to be presymptomatic, and 40.5% to be symptomatic. The asymptomatic rate among nonresidential students participating in the surveillance testing system at Clemson University was 69%.The generalizability of these studies is limited given that students are typically healthy and as such may be more likely to experience asymptomatic infection.

Vaccine coverage also may influence the asymptomatic rate. In a study in Israel on the effectiveness of the Pfizer–BioNTech mRNA COVID-19 vaccine BNT162b2, 55.7% (49,138 out of 88203) of infections were asymptomatic in the unvaccinated group, and 68.2% of infections (3632 out of 5324) of infections were asymptomatic in the vaccinated group.

Numerous additional factors contribute to the heterogeneity we see among estimates of the percent of infections that are asymptomatic, including community prevalence, the study population, and the time period when the study population was tested. The use of different definitions also may contribute. This includes the definition of a symptomatic infection, since our understanding of the clinical presentation of COVID-19 has evolved over time[10], as well as the definition of asymptomatic, since some define asymptomatic to include presymptomatic cases, where people that had no symptoms upon testing positive but may have went on to develop symptoms at a later date, and truly asymptomatic cases, where an infected individual never goes on to develop symptoms.

Because of the heterogeneity in estimates of the percent of infections that are asymptomatic, for this prior we specified a beta distribution with the majority of the density between 0.3 and 0.8, with a mean of 0.55 and standard deviation of 0.12.

Limitations

- Validation:
  - Wastewater encatchments do not fall exactly on county lines, so aggregation here contains limited information
  - Covidestim – sum of daily medians != median for entire two-week time interval
- Repeat testing may bias estimates in places where this form of testing comprises a substantial proportion of total testing, but data on people-viral-total and people-viral-positive is not widely available

References

1. Salomon, J. A. *et al.* The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2111454118 (2021).
2. Chitwood, M. H. *et al.* Reconstructing the course of the COVID-19 epidemic over 2020 for US states and counties: Results of a Bayesian evidence synthesis model. *PLoS Comput Biol* **18**, e1010465 (2022).
3. Andrews, N. *et al.* Covid-19 Vaccine Effectiveness against the Omicron (B.1.1.529) Variant. *N Engl J Med* **386**, 1532–1546 (2022).

4. Liu, Y., Yu, Y., Zhao, Y. & He, D. Reduction in the infection fatality rate of Omicron variant compared with previous variants in South Africa. *International Journal of Infectious Diseases* **120**, 146–149 (2022).

5. Pulliam, J. R. C. *et al.* Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science* **376**, eabn4947 (2022).

6. Olesen, S. W., Imakaev, M. & Duvallet, C. Making waves: Defining the lead time of wastewater-based epidemiology for COVID-19. *Water Research* **202**, 117433 (2021).

7. Xiao, A. *et al.* Metrics to relate COVID-19 wastewater data to clinical testing dynamics. *Water Research* **212**, 118070 (2022).

8. Hopkins, L. *et al.* Citywide wastewater SARS-CoV-2 levels strongly correlated with multiple disease surveillance indicators and outcomes over three COVID-19 waves. *Science of The Total Environment* **855**, 158967 (2023).

9. Ma, Q. *et al.* Global Percentage of Asymptomatic SARS-CoV-2 Infections Among the Tested Population and Individuals With Confirmed COVID-19 Diagnosis: A Systematic Review and Meta-analysis. *JAMA Netw Open* **4**, e2137257 (2021).

10. Sah, P. *et al.* Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2109229118 (2021).

11. Fang, L.-L. *et al.* PCR combined with serologic testing improves the yield and efficiency of SARS-CoV-2 infection hunting: A study in 40,689 consecutive overseas arrivals. *Front. Public Health* **11**, 1077075 (2023).

12. Ford, J. S. *et al.* Use of an Asymptomatic COVID-19 Testing Protocol in a Pediatric Emergency Department. *The Journal of Emergency Medicine* **63**, 332–338 (2022).