

Using Sensitivity Analyses to Approximate Total COVID-19 Infections: State and  
County level in the United States, March 2021 - March 2022

Quinn White

Submitted to the Department of Statistical and Data Sciences  
of Smith College  
in partial fulfillment  
of the requirements for the degree of  
Bachelor of Arts

Ben Baumer, Primary Faculty Advisor  
Nicholas Reich, Secondary Faculty Advisor

May 2023



# **Acknowledgements**

Will add



# Table of Contents

<b>Chapter 1: Introduction</b> . . . . .	<b>1</b>
<b>Chapter 2: Background</b> . . . . .	<b>3</b>
2.1 Probabalistic Bias Analysis . . . . .	5
2.2 Background for the Approach . . . . .	5
2.2.1 Simple Discrete Example . . . . .	5
2.2.2 General Solution for the Discrete Case . . . . .	5
2.2.3 General Solution for the Continuous Case . . . . .	5
2.2.4 Implementation through the Sampling-Importance-Resampling Algorithm . . . . .	5
2.3 Bayesian Melding Applied to COVID-19 Misclassification . . . . .	5
2.3.1 Distribution of $\theta = \{\alpha, \beta, P(S_1 \text{untested})\}$ . . . . .	5
2.3.2 Direct Prior and Induced Prior Distributions for $P(S_0 \text{test}_+, \text{untested})$ . . . . .	5
2.3.3 Pooling . . . . .	5
2.3.4 Motivating Example . . . . .	5
2.3.5 Derivation of $M$ . . . . .	5
2.4 Sampling-Importance-Resampling Algorithm . . . . .	5
2.4.1 Overview . . . . .	5
2.4.2 Proof that Algorithm Obtains Approximate Sample from Target Distribution . . . . .	5
2.4.3 Obtaining Logarithmic Pooled Distribution with the Sampling-Importance-Resampling Algorithm . . . . .	5
2.4.4 Implications of the Sample Size and Resample Size . . . . .	5
2.5 LOESS Smoothing . . . . .	5
2.5.1 Introduction . . . . .	5
2.5.2 Fitting the LOESS Curve . . . . .	5
2.6 Kernel Density Estimation . . . . .	5
2.6.1 Overview . . . . .	5
2.6.2 Bounded Density Estimation . . . . .	5
<b>Chapter 3: Definition of Prior Distributions for Bias Parameters</b> . . . . .	<b>7</b>
3.1 Background on the Beta Distribution . . . . .	8
3.2 Background on the Gamma Distribution . . . . .	8
3.3 Definition of Prior Distributions for Incomplete Testing Correction . . . . .	8
3.3.1 Defining $P(S_1 \text{untested})$ . . . . .	8

3.3.2	Defining $\alpha$	8
3.3.3	Defining $\beta$	8
3.3.4	Defining $P(S_0 \text{test}_+, \text{untested})$	8
3.4	Definition of Priors for Test Inaccuracy Correction	8
3.4.1	Defining Test Sensitivity ( $S_e$ )	8
3.5	Defining Test Specificity ( $S_p$ )	8
3.6	Exploration of the Implications of Changes in the Bias Parameters	8
3.7	Correction for Incomplete Testing	8
3.8	Correction for Diagnostic Test Inaccuracy	8
3.8.1	Derivation of Formula for Correction for Diagnostic Test Inaccuracy	8
<b>Chapter 4: Details of Implementation</b>		<b>9</b>
4.1	Version 1: Priors do Not Vary by State or Date	9
4.2	Version 2-4: Allowing Some Prior Parameters to Vary	9
<b>Chapter 5: Results</b>		<b>11</b>
5.1	Comparison to the Covidestim Model	11
5.1.1	Overview	11
5.1.2	The Covidestim Model	11
5.1.3	Assumptions	12
5.1.4	Comparison to Serological Data	13
5.1.5	Limitations of this Comparison	13
5.2	State-level Results	13
5.2.1	Example Where July Peak is Larger in Probabilistic Bias Analysis Intervals	18
5.2.2	Example Where Covidestim Intervals Lag Probabilistic Bias Intervals	20
5.3	Ratio Between Estimated to Observed Infections	23
5.4	Relationship Between the Ratio of Estimated to Observed Infections Compared to Testing Rate	23
5.5	County-level Results	27
5.5.1	Massachusetts	27
5.5.2	Michigan	33
5.6	Cross Correlation Comparison	37
5.6.1	Background	37
5.6.2	Cross Correlation Results Comparing Bias Corrected Counts, Covidestim Estimates, and Wastewater Concentrations	40
<b>Chapter 6: Conclusion</b>		<b>45</b>
<b>Chapter 7: Appendix</b>		<b>47</b>
7.1	Smoothing Span	47
7.1.1	Changing SPAN for LOESS Smoothing of $\beta$	47
7.2	Changing Mean and Variance for Prior Distribution Specifications	47

7.3	Relationship Between $(X + Y)_\alpha$ and $X_\alpha + Y_\alpha$ for Dependent Variables $X, Y$ . . . . .	47
7.3.1	Simulation: Bivariate Normal . . . . .	47
7.3.2	Derivation of the Distribution of X+Y for Bivariate Normal . .	47
7.4	Total Tests and Positive Tests Across All States . . . . .	47
<b>References</b>	. . . . .	<b>49</b>



# Abstract

As we have navigated the COVID-19 pandemic, case counts have been a central source of information for understanding transmission dynamics and the effect of public health interventions. However, because the number of cases we observe is limited by the testing effort in a given location, the case counts presented on local or national dashboards are only a fraction of the true infections. Variations in testing rate by time and location impacts the number of cases that go unobserved, which can cloud our understanding of the true COVID-19 incidence at a given time point and can create biases in downstream analyses. Additionally, the number of cases we observe is impacted by the sensitivity and specificity of the diagnostic test. To quantify the number of true infections given incomplete testing and diagnostic test inaccuracy, this work implements probabilistic bias analysis at a biweekly time scale from January 1, 2021 through February 2022. In doing so, we can estimate a range of possible true infections for a given time interval and location. This approach can be applied at the state level across the United States, as well as in some counties where the needed data are available.



# **Dedication**

You can have a dedication here if you wish.



# **Chapter 1**

## **Introduction**

Placeholder



# **Chapter 2**

## **Background**

Placeholder



## 2.1 Probabalistic Bias Analysis

## 2.2 Background for the Approach

### 2.2.1 Simple Discrete Example

### 2.2.2 General Solution for the Discrete Case

### 2.2.3 General Solution for the Continuous Case

### 2.2.4 Implementation through the Sampling-Importance-Resampling Algorithm

## 2.3 Bayesian Melding Applied to COVID-19 Misclassification

### 2.3.1 Distribution of $\theta = \{\alpha, \beta, P(S_1|\text{untested})\}$

### 2.3.2 Direct Prior and Induced Prior Distributions for $P(S_0|\text{test}_+, \text{untested})$

### 2.3.3 Pooling

### 2.3.4 Motivating Example

### 2.3.5 Derivation of $M$

## 2.4 Sampling-Importance-Resampling Algorithm

### 2.4.1 Overview

### 2.4.2 Proof that Algorithm Obtains Approximate Sample from Target Distribution

Example 1:

Example 2:

### 2.4.3 Obtaining Logarithmic Pooled Distribution with the Sampling-Importance-Resampling Algorithm

### 2.4.4 Implications of the Sample Size and Resample Size

## 2.5 LOESS Smoothing

### 2.5.1 Introduction

### 2.5.2 Fitting the LOESS Curve

## 2.6 Kernel Density Estimation



# **Chapter 3**

## **Definition of Prior Distributions for Bias Parameters**

Placeholder

**3.1 Background on the Beta Distribution****3.2 Background on the Gamma Distribution****3.3 Definition of Prior Distributions for Incomplete Testing Correction****3.3.1 Defining  $P(S_1|\text{untested})$** **3.3.2 Defining  $\alpha$** **3.3.3 Defining  $\beta$** **3.3.4 Defining  $P(S_0|\text{test}_+, \text{untested})$** **3.4 Definition of Priors for Test Inaccuracy Correction****3.4.1 Defining Test Sensitivity ( $S_e$ )****3.5 Defining Test Specificity ( $S_p$ )****3.6 Exploration of the Implications of Changes in the Bias Parameters****3.7 Correction for Incomplete Testing****3.8 Correction for Diagnostic Test Inaccuracy****3.8.1 Derivation of Formula for Correction for Diagnostic Test Inaccuracy**

# **Chapter 4**

## **Details of Implementation**

Placeholder

### **4.1 Version 1: Priors do Not Vary by State or Date**

### **4.2 Version 2-4: Allowing Some Prior Parameters to Vary**



# Chapter 5

## Results

### 5.1 Comparison to the Covidestim Model

#### 5.1.1 Overview

One challenge in correcting for biases in general is that although we may have some information about the influence of possible biases, we do not have a ground truth for comparison. However, one approach to handle the fact that the true cases are unobserved is comparing our estimates to those from other approaches seeking to estimate a similar quantity. In particular, if other approaches make different assumptions and come to a similar result, this can give us more confidence in our estimates.

One notable project seeking to estimate the true infection burden at the county-level over time is the COVIDestim project. In this work, Chitwood et al. proposed a mechanistic model that includes states for asymptomatic/pre-symptomatic infection, symptomatic but mild infection, severe COVID-19 presentations, and death. This approach also enables the estimation of  $R_t$ , the number of secondary infections a single infected individual causes at time  $t$ . This is a useful quantity to estimate, but is sensitive to reporting delays and changes in testing practices (Pitzer et al., 2021).

#### 5.1.2 The Covidestim Model

Chitwood *et al.* propose a Bayesian evidence synthesis model to correct for reporting delays and time varying case ascertainment testing rate in the estimation of incident infections and  $R_t$ .

To estimate the expected cases and deaths at a particular point in time, the model uses a convolution of the time series of observed cases and deaths and reporting delay distributions that are specific to the health state categories. This enables the model to account for the fact that reporting delay is different. For any health state, for example, asymptomatic, the individual can either transition to the next health state (symptomatic) or recover. Thus, with each transition between a defined health state, for example, asymptomatic, there is a probability of transitioning to the next

health state (in this case, asymptomatic to symptomatic); the complement of this probability is the probability of recovery.

Each of these transitions is defined by a delay distribution. For example, the distribution for moving from asymptomatic to symptomatic represents the probability an individual moves to the symptomatic state at a point in time. The probabilities asymptomatic to symptomatic and symptomatic to severe are modeled as not varying with time. Meanwhile, the probability of transitioning from severe to death was defined to be higher in 2020 due to higher case fatalities early in the pandemic. The infection fatality rates, adjusted to be specific to a given state or county based on age distributions and the prevalence of risk factors for COVID-19, are used to inform the probability of moving from the severe category to the death category.

The change in daily infections from the previous day (i.e., the new infections) is calculated as a function of the estimated effective reproductive number  $R_t$  and the mean serial interval, where serial interval is the time from the onset of infection of a primary case to the time of onset of infection in the secondary case.  $R_t$  is estimated using a log-transformed cubic spline, under the assumption individuals can only be infected once.

They also defined a distribution for the delay to diagnosis, which was distinct by health state category to reflect differences in diagnosis delays that occur depending on the disease severity. The probability of diagnosis among different health states was allowed to vary by time to reflect changing testing rates throughout the pandemic.

A separate distribution models the reporting delay to correct the total number of diagnoses on a given day for the fact that these diagnoses correspond to past infections.

The observed cases and death data for each state to the model were fitted using negative binomial likelihood functions.

### 5.1.3 Assumptions

This approach relies on infection fatality ratios and death counts to estimate the true case counts. Thus, it is sensitive to estimates of infection fatality rate, with higher infection fatality ratio estimates resulting in lower estimated infections. The infection fatality ratio is defined as the proportion of COVID-19 infections that lead to death, which means there is uncertainty in estimating both the numerator and the denominator of the ratio. The true cumulative incidence depends on the same uncertainties in estimating the true case burden at any point in time. Estimating the infection fatality ratio itself is a challenging task.

The COVIDestim model uses age-specific estimates of IFR produced by O'Driscoll et al. (2021). This group used national-level age-stratified, and when possible sex-stratified, COVID-19 death counts and cumulative infection estimates from seroprevalence studies. Of note, the estimates of infection fatality ratio are assumed to be constant over time, which may not be the case due to improving treatments (e.g., Paxlovid), different variants leading to less severe presentations, or changes in the demographics of individuals being infected. However, reliable

estimates of infection fatality ratio that vary with time difficult to acquire; COVIDestim assumed a higher case fatality in 2020 given the novelty of the virus and consequent lack of available treatments.

### 5.1.4 Comparison to Serological Data

There are known issues with seroprevalence estimates. For one, these samples are drawn from a convenience (i.e. nonrandom) sample of individuals with blood specimens taken for purposes other than COVID-19 antibody detection (“Commercial Laboratory Seroprevalence Surveys | Coronavirus | COVID-19 | CDC,” n.d.). Secondly, while a positive serological test is evidence for infection, a negative serological test is less clear to interpret. The person may have been infected but not yet have developed antibodies, or their immune system may not have produced antibodies at a detectable level (CDC, 2020).

Indeed, Chitwood et al. found limited concordance between their estimates and seroprevalence data. However, there was a stronger correlation between estimates of cumulative infection and cumulative hospitalizations and cumulative deaths<sup>1</sup>.

### 5.1.5 Limitations of this Comparison

At this point in the pandemic, there is no true gold standard to compare to. Covides-tim is one model, among many, that makes key assumptions about aspects of the virus. Another note is that estimates from the Covides-tim model are reported on the daily timescale for counties, while the probabilistic approach we implemented here is at the biweekly time scale.

To ensure the comparisons are on the same time scale, we sum the reported 95% credible intervals for the days in each 2-week interval. These intervals do not represent a 95% credible interval for the 2-week interval, and while such an interval would be ideal for the comparison, computation of a 95% credible interval for the two-week interval is not feasible because of the model structure. Due to the correlation between observations for each day for a given location, summing the intervals yields an estimate that is likely to be more conservative than a true 95% credible interval for the two-week interval would be. More detail on this assumption is in [the appendix](#).

## 5.2 State-level Results

There is much more comprehensive testing data at the state-level than at the county level, as John Hopkins has tracked state-level testing data throughout the pandemic.

---

<sup>1</sup>The correlation employed here is the Spearman rank correlation, which measures the strength of the monotonic relationship rather than the strength of the linear relationship, in which case the Pearson correlation coefficient is the usual choice. The Spearman rank correlation is equivalent to the Pearson correlation of the rank values rather than the values themselves.

As a result, we can apply probabilistic bias analysis across the United States at the state-level.

However, since versions 2-4 of the analysis utilize empirical estimates of  $P(S_1|\text{untested})$  or  $\beta$  from the COVID-19 Trends and Impact Survey, these versions are only possible when there is sufficient data for the state. As we see in Figure 5.1, some states have very little data for the empirical estimate of beta. However in Figure 5.2 we see that there is more consistent reporting of the prevalence of COVID-19-like symptoms, which we can use as an estimate of  $P(S_1|\text{untested})$ .

For versions 2-4, we did not attempt the probabilistic bias analysis for states where more than 60% of observations were missing at the daily time step. Missing values for states with sufficient data were imputed before summarizing to the biweek level using a linear weighted moving average.

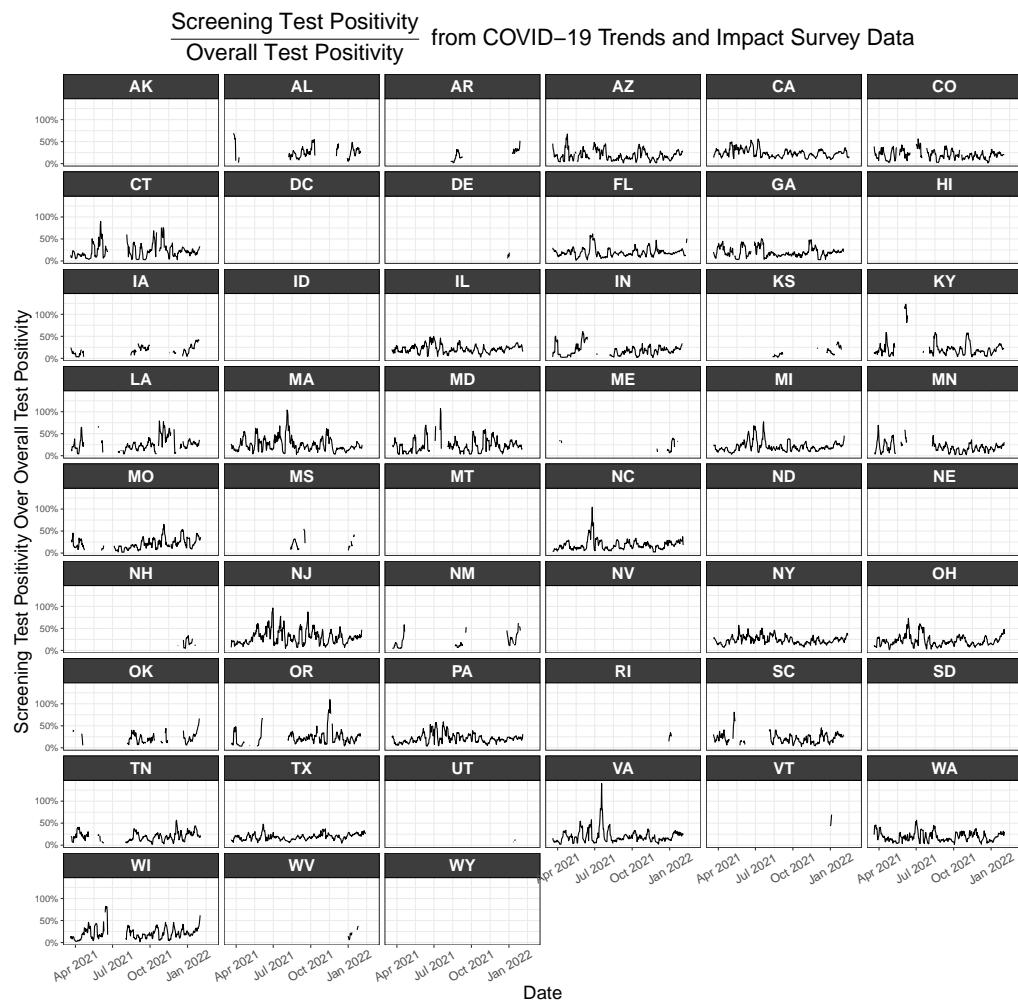


Figure 5.1

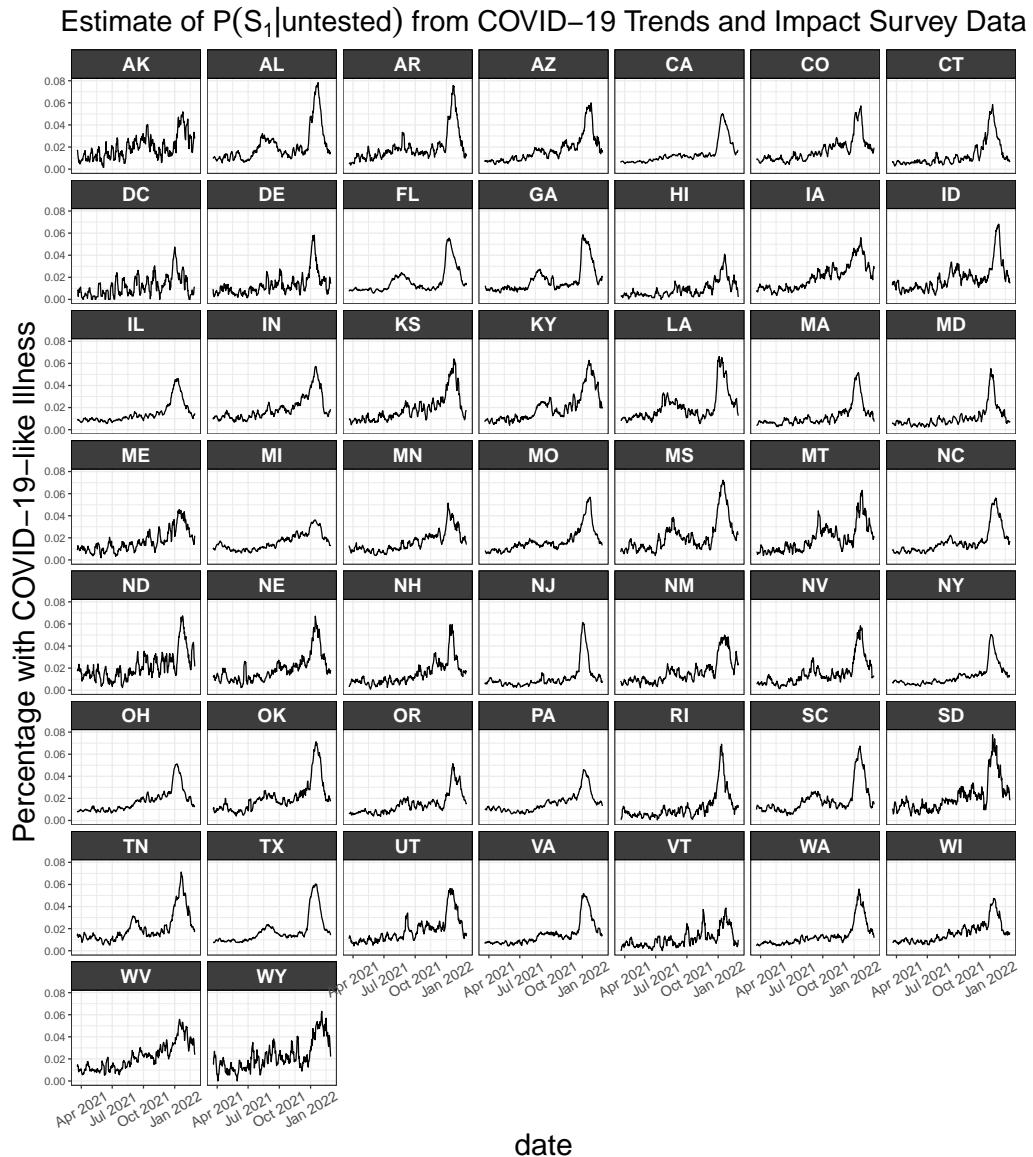


Figure 5.2

In Figures 5.3 and 5.4, we compare the 95% credible intervals of Covidestim summed to be on the biweek time scale to the probabilistic bias analysis intervals<sup>2</sup>. We see correspondence 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. From Figure 5.5, we see that this period corresponds to the Omicron wave of the pandemic.

<sup>2</sup>As mentioned in the [limitations](#) section, the two week intervals cannot be interpreted as 95% credible intervals. However, the comparison is still useful to contextualize the results of the probabilistic bias analysis.

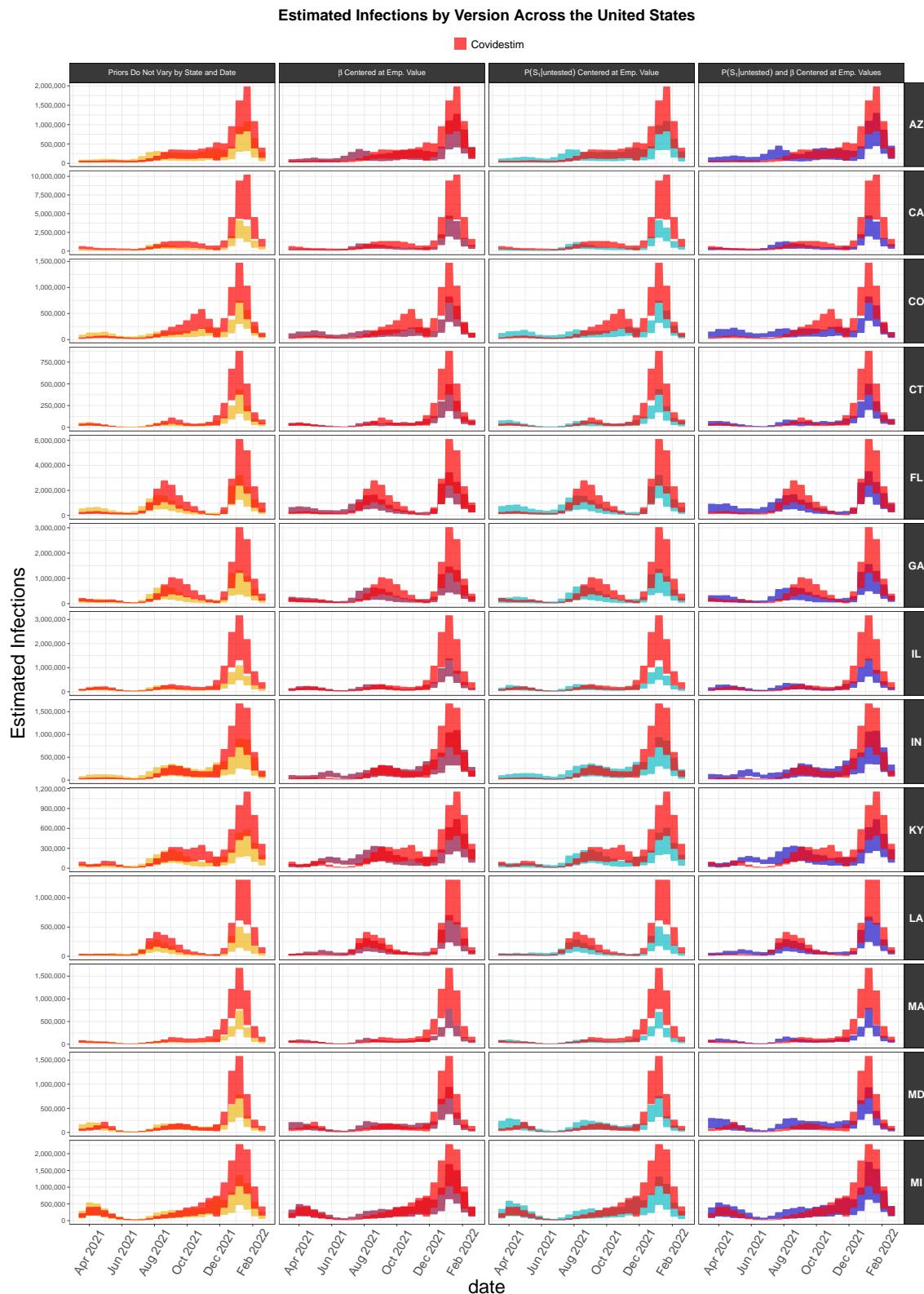


Figure 5.3

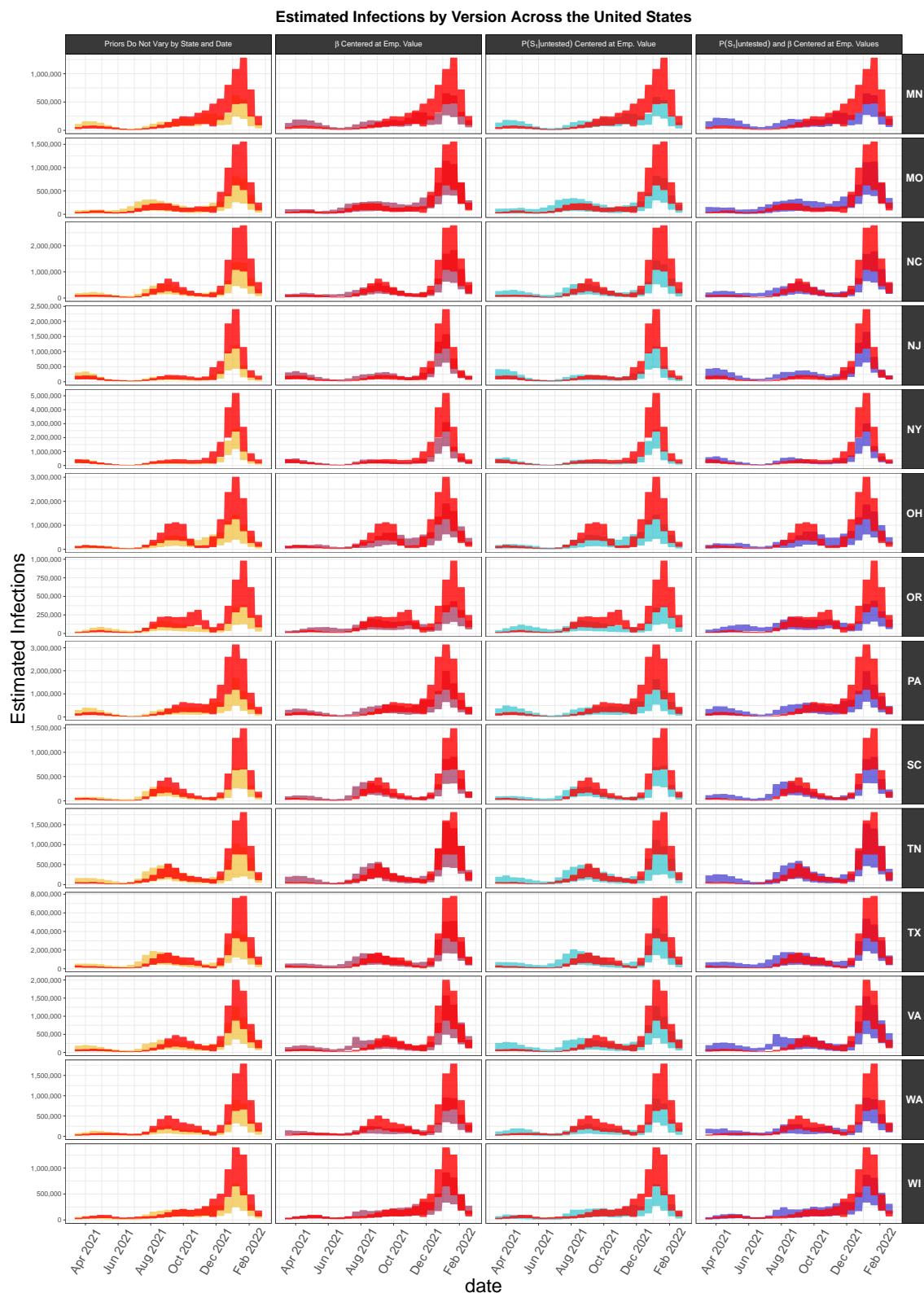


Figure 5.4

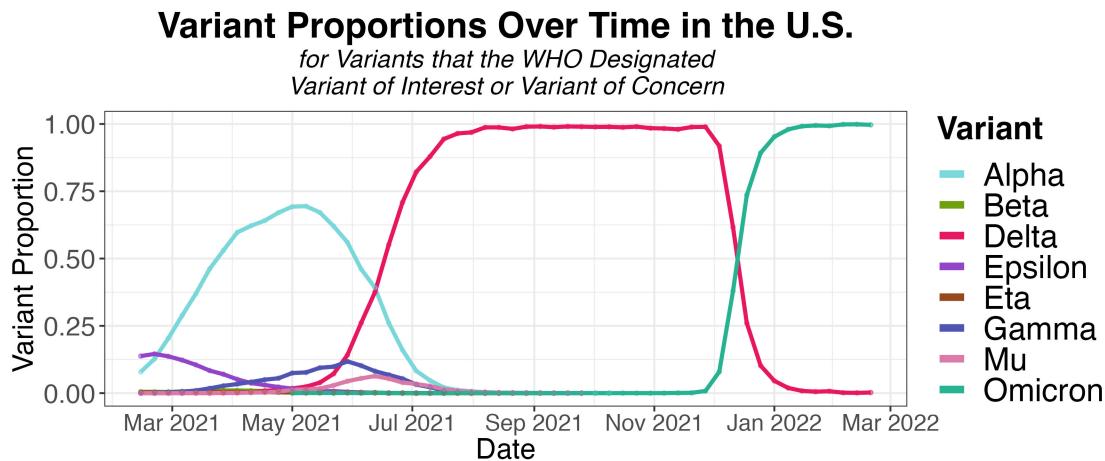


Figure 5.5: Variant proportions for variants designated as Variant of Interest or Variant of Concern by the World Health Organization. We see Delta begins its rise in May 2021 and reaches an estimated proportion of 0.98 of sequenced infections by mid-August 2021. Omicron's rise occurs more quickly, beginning its increase throughout December and becoming almost entirely dominant by mid-January. Data source: SARS-CoV-2 Variant Proportions from Data.CDC.gov.

This period of disagreement between the approaches in December 2021 through January 2022 corresponds to the Omicron wave of the pandemic, which we can see in Figure 5.5. This was a major shift in the pandemic: the Omicron variant is highly transmissible compared to previous variants, and it has immune invasion capacity that means vaccine efficacy provides less protection (<https://doi-org.libproxy.smith.edu/10.1038/s41591-022-01753-y?>). It also is associated with a lower infection fatality ratio ([liu2022?](#)).

Beyond the clear differences during the Omicron wave, we also see in several states that there is an additional rise in July with the probabilistic bias analysis implementations that include data from the COVID-19 Trends and Impact Survey (columns 2 through 4), while this peak is smaller in the Covidestim intervals. States exhibiting this trend include Arizona, Maryland, Michigan, Minnesota, Missouri, New Jersey, New York, and Wisconsin. Meanwhile, there are some states where there is a peak in the Covidestim intervals that lags the peak in the probabilistic bias analysis, for example, Colorado, Kentucky, Texas, and Virginia.

The peak in July corresponds to the Delta wave, where the more transmissible B.1.617.2 (delta) variant of concern displaced the previously dominant B.1.351 (alpha) variant ([delrio2021?](#)). The peak in December corresponds to the Omicron wave, which began the dominant variant over the course of a single month.

### 5.2.1 Example Where July Peak is Larger in Probabilistic Bias Analysis Intervals

We can take a closer look at Michigan to see the trends in the probabilistic intervals in relation to changes in circulating variants (Figure 5.6). In particular, we see the moments where the trend differs from Covidestim are in a peak around late July

and in December. Across states, these periods tended to be the most discordant (Figure 5.3 and Figure 5.4) .

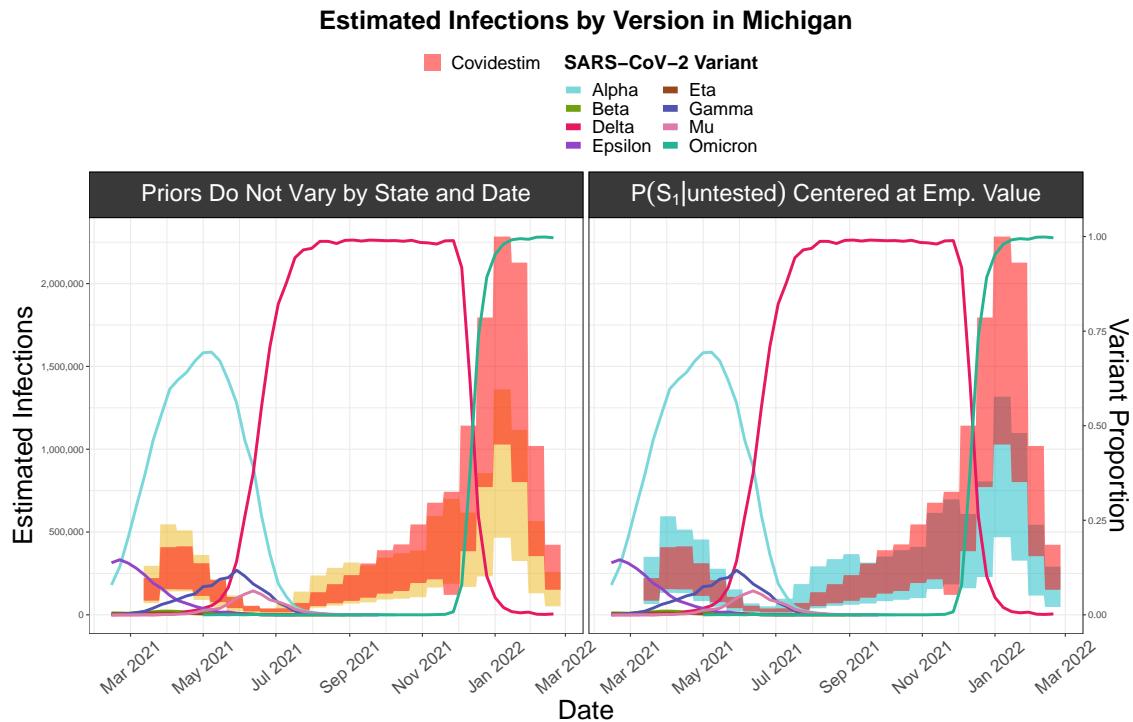


Figure 5.6: Presenting the probabilistic intervals along with the variant proportions for variants designated as Variant of Interest or Variant of Concern by the World Health Organization. We see the rise in Omicron clearly in the probabilistic bias intervals in December of 2021.

As discussed in the Covidestim model change log, Chitwood *et al.* made substantial changes to handle the rise of the Omicron variant. Because the variant causes much milder infections, the infection fatality for Omicron infections is lower than previous variants. To handle the changes in the infection fatality ratio and very low death counts, 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 ([pulliam2022?](#)). The changes in the model may contribute to the differences we see between the probabilistic bias intervals and Covidestim intervals during the Omicron wave.

In July, meanwhile, 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 in the time interval where

we observe a larger peak in the probabilistic bias intervals. While the trend in observed tests follows a more steady upward trend like that we see in the Covidestim estimates, the probabilistic bias intervals are influenced by the relationship between positive tests and total tests in this time period.

In Figure 5.7, consider the number of positive and total tests reported in Michigan. In the period from July through September, the positive tests increase at a higher rate than the total tests. Because the number of infections among the untested population is calculated based on the observed test positivity rate, the higher test positivity rate (resulting from positive tests increasing more than total tests) results in the peak we see in this time interval in the probabilistic bias intervals. The same trend is true across states (Supplementary Figure ??).

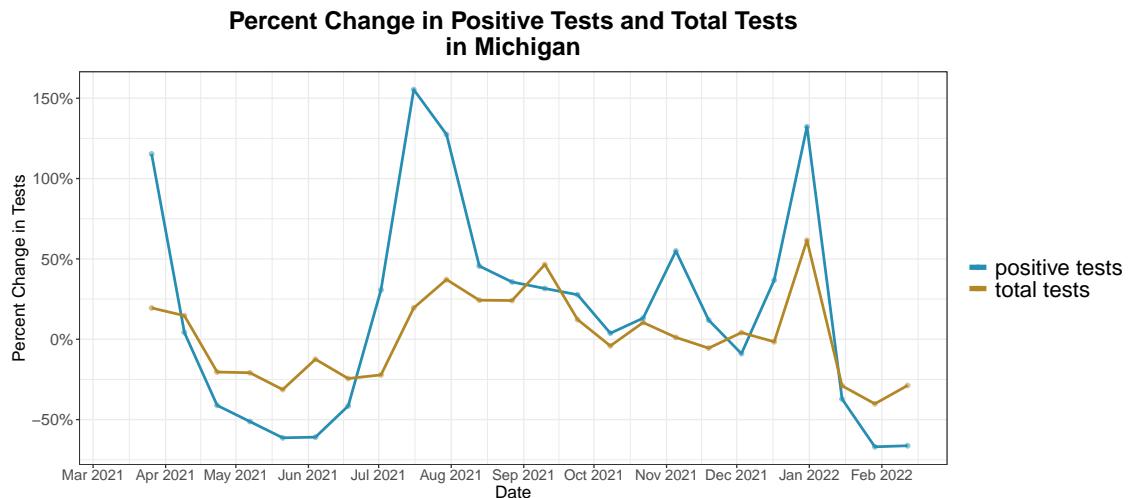
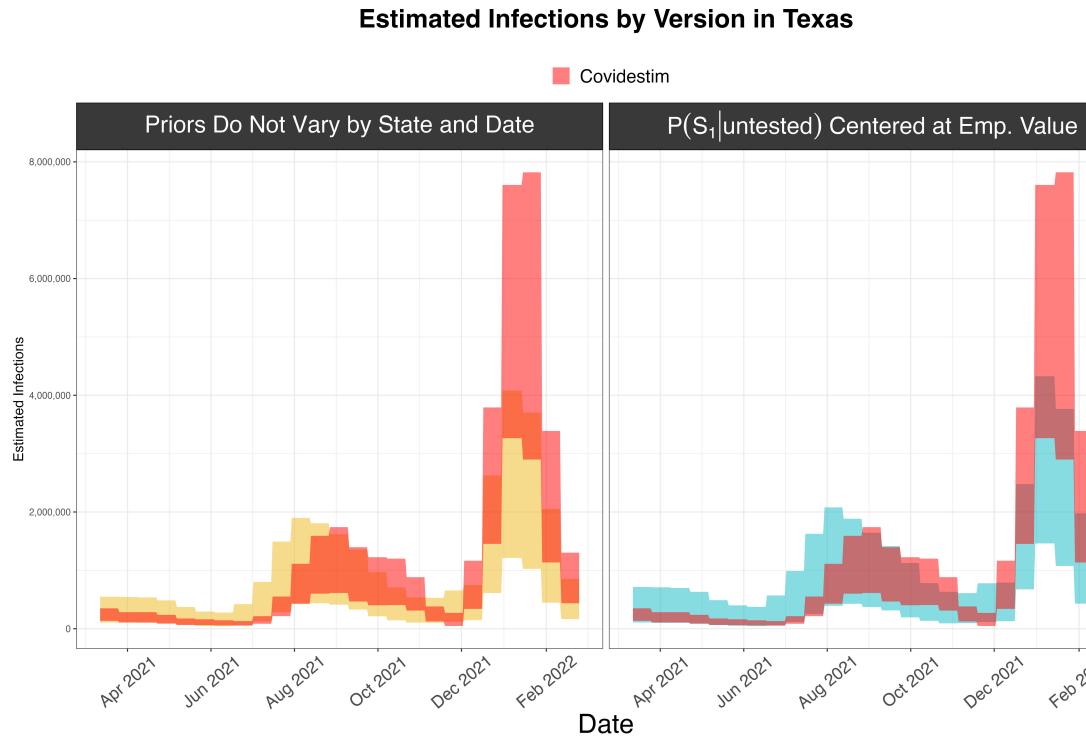


Figure 5.7: Percent change in positive tests and the percent change in total tests from March 2021 through the end of February 2022. We see in the period from July through September that the positive tests increase at a higher rate than the total tests, which relates to the increase we see beginning in July in the probabilistic bias intervals.

## 5.2.2 Example Where Covidestim Intervals Lag Probabilistic Bias Intervals

Texas is one state where there the peak corresponding to the Delta wave is lagged in the Covidestim intervals compared to the probabilistic bias intervals (Figure 5.2.2).



\begin{figure}  
\caption{ Probabilistic bias analysis intervals for Texas across time. Each probabilistic bias interval is a 95% simulation interval for the estimated infections for that entire 2-week period. In red are Covidestim estimates summed to be on the same biweekly time scale. Texas is one state where the Covidestim estimates appear to lag the probabilistic bias analysis for the Delta wave beginning in July.} \end{figure}

To consider the reasons we see this lag, we can first compare the data sources, since for the probabilistic bias analysis we use the CDC's reports of positive and total tests by state, while the Covidestim model uses the Johns Hopkins Center for Systems Science and Engineering (CSSE) case counts. In Figure

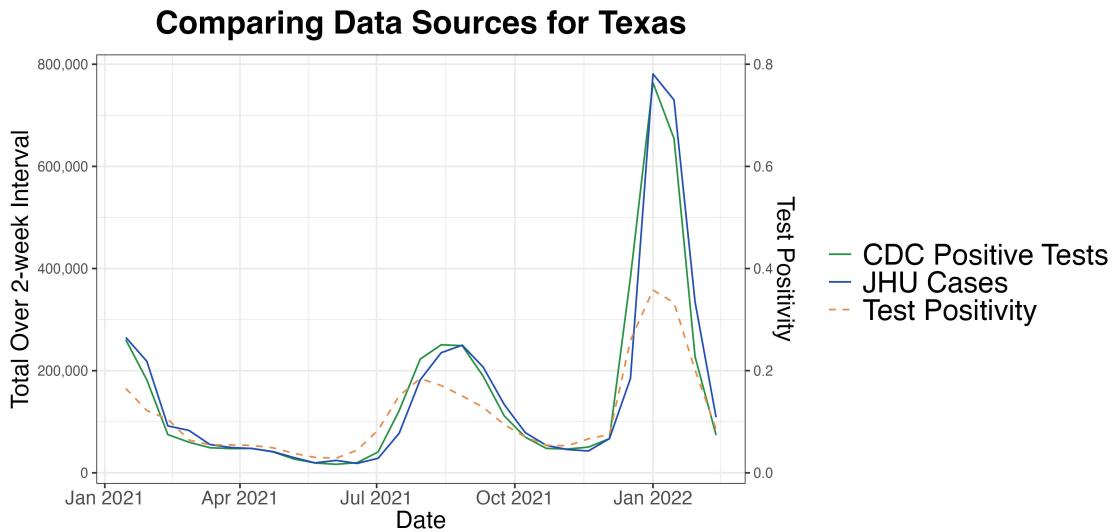


Figure 5.8: Comparing the data sources for the probabilistic bias analysis and the Covidestim model to look into the source of the lag observed in Texas, as shown in Figure `effig:texas`. The two solid lines are the data sources: the data source for the probabilistic analysis is PCR tests from the CDC's Diagnostic Laboratory Testing Time series, while the data source for the Covidestim model is Johns Hopkins CSSE case counts. Each source is summed on the biweekly time scale. The test positivity is the positive tests from the CDC source over total tests for the two-week interval. The Johns Hopkins CSSE and CDC positive tests are largely concordant, so differences in data source do not appear to explain the lag for the delta wave in Texas. The test positivity, however, does clarify the lag. The test positivity begins increasing in early June, before we see positive tests or cases increasing, due to a drop in total tests. This increase in test positivity is reflected in the probabilistic bias intervals in `effig:texas-source`, where increases corresponding to the delta wave start beginning earlier, in early to mid June, compared to July, where the increase is at about a two-week lag.

## 5.3 Ratio Between Estimated to Observed Infections

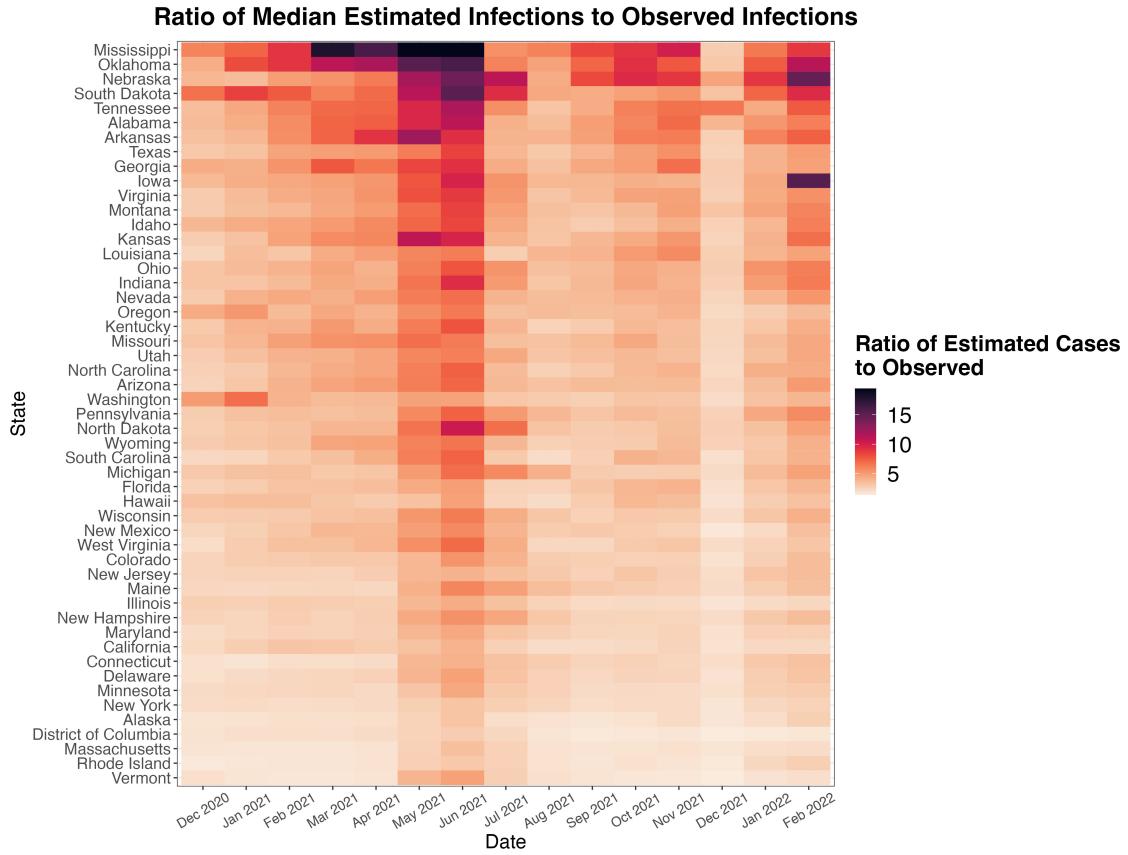


Figure 5.9: The ratio of estimated infections to observed infections for each 2-week interval and each state, only considering the version of the priors that does not vary by state or time. States are ordered from highest to lowest by median ratio of estimated infections to observed infections over all 2-week intervals considered. The two-week intervals with highest ratios are in June 18, 2021 through July 16, 2021.

## 5.4 Relationship Between the Ratio of Estimated to Observed Infections Compared to Testing Rate

One question to consider is the relationship between the testing rate and the ratio of estimated cases to observed cases, as we would expect higher testing rates would lead to less underestimation of the true number of infections.

As we see in Figure 5.10, the nature of the relationship between the testing rate and the ratio of estimated cases to observed cases depends on whether we allow  $\beta$  and  $P(S_1|\text{untested})$  to vary by location and date. In particular, when we sample from the same priors for every correction (the first panel of Figure 5.10), we see there is little variability in the relationship between the testing rate and median estimated cases, because the form of the correction is identical for each two-week interval and state considered. Allowing  $\beta$  and/or  $P(S_1|\text{untested})$  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 cases to observed cases is more clear when we think back to the correction formulas.

Denoting  $N^*$  again to be the number who would test positive, on the  $x$ -axis, we have approximately<sup>3</sup>

$$\frac{N_{\text{tested}}^* + N_{\text{untested}}^*}{N_{\text{tested}}^*},$$

where we calculate  $N_{\text{untested}}^*$  using the other priors:

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

$$\Pr(\text{test}_+ | S_0, \text{untested})(1 - (\Pr(S_1 | \text{untested}))(N_{\text{untested}})).$$

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  $N_{\text{untested}}^*$  will be large relative  $N_{\text{tested}}^*$ . However, with higher testing rates,  $N_{\text{untested}}^*$  will be large relative  $N_{\text{tested}}^*$ , and the ratio of estimated to observed infections nears one.

---

<sup>3</sup>This isn't exactly the estimated infections, because for simplicity of notation we are not writing out the correction for test inaccuracy.

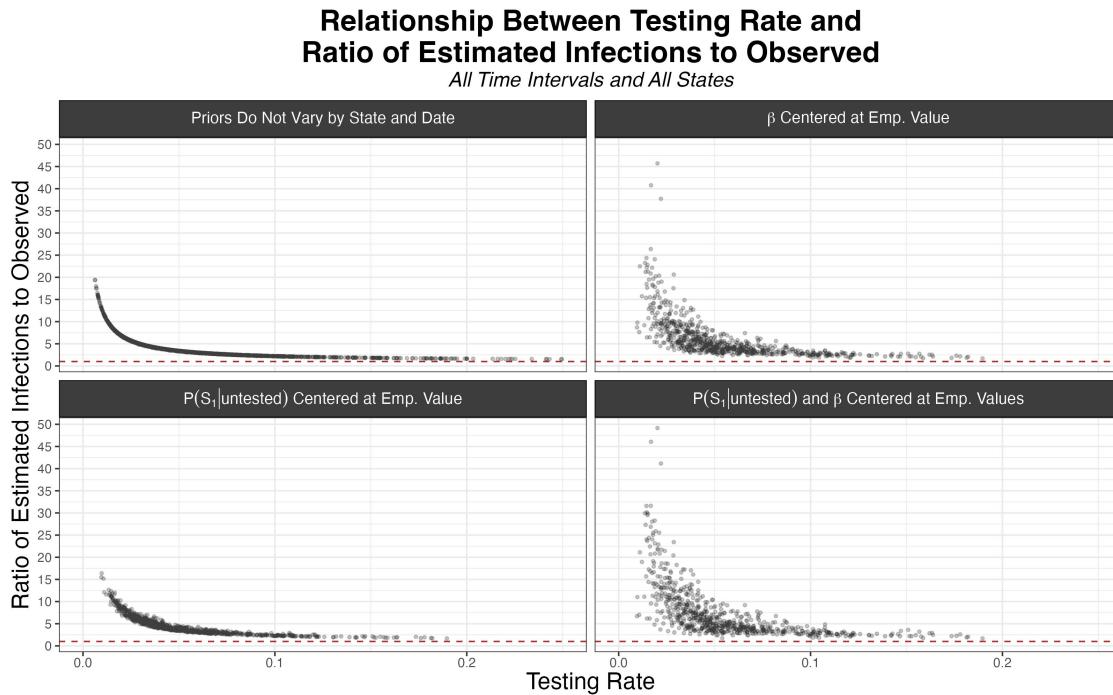


Figure 5.10: The ratio of the median estimated infections plotted against the testing rate, where the testing rate is calculated as the total number tested in a two-week interval over the population size. When the priors are the same for all time intervals and states, there is minimal variability relationship 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 and location. However, when we allow  $\beta$  or  $P(S_1|\text{untested})$  to vary by state and time interval, there is much variability in the relationship. A horizontal line in red at 1 is included to reference; a ratio of exactly one would indicate no infections went unobserved.

It is worth noting that in some cases, the ratio of total number of tests for a two-week interval over the population size is incredibly high. For example, there are two-week intervals in Vermont, Rhode Island, and DC where the total number of tests over the census population size exceeds 0.25. There are two major factors to consider when we interpret these really high testing rates.

For one, there is repeat testing. If people are part of a regular PCR screening program where they are tested multiple times in a 2-week time interval, this will go toward higher numbers of tests relative to the census population size. While some states also report the number of people who tested positive and people who tested total<sup>4</sup> (rather than total positive tests and total tests), this level of granularity was not universal.

Additionally, the census population is not exactly the tested population. This is in part because universities are a major source of PCR testing, and students who are out of state are not reflected in the census counts for the location where they are tested.

When we look at the testing rates (taken as the total number of tests over the

<sup>4</sup>A data dictionary on the John Hopkins COVID-19 testing repository provides more detail on these definitions.

census population size) in Figure 5.11, we see that the states with the highest biweekly testing rates are Vermont, Massachusetts, Rhode Island, and the District of Columbia. These are states that have substantial student populations, who also may be tested more than once in a single two-week interval.

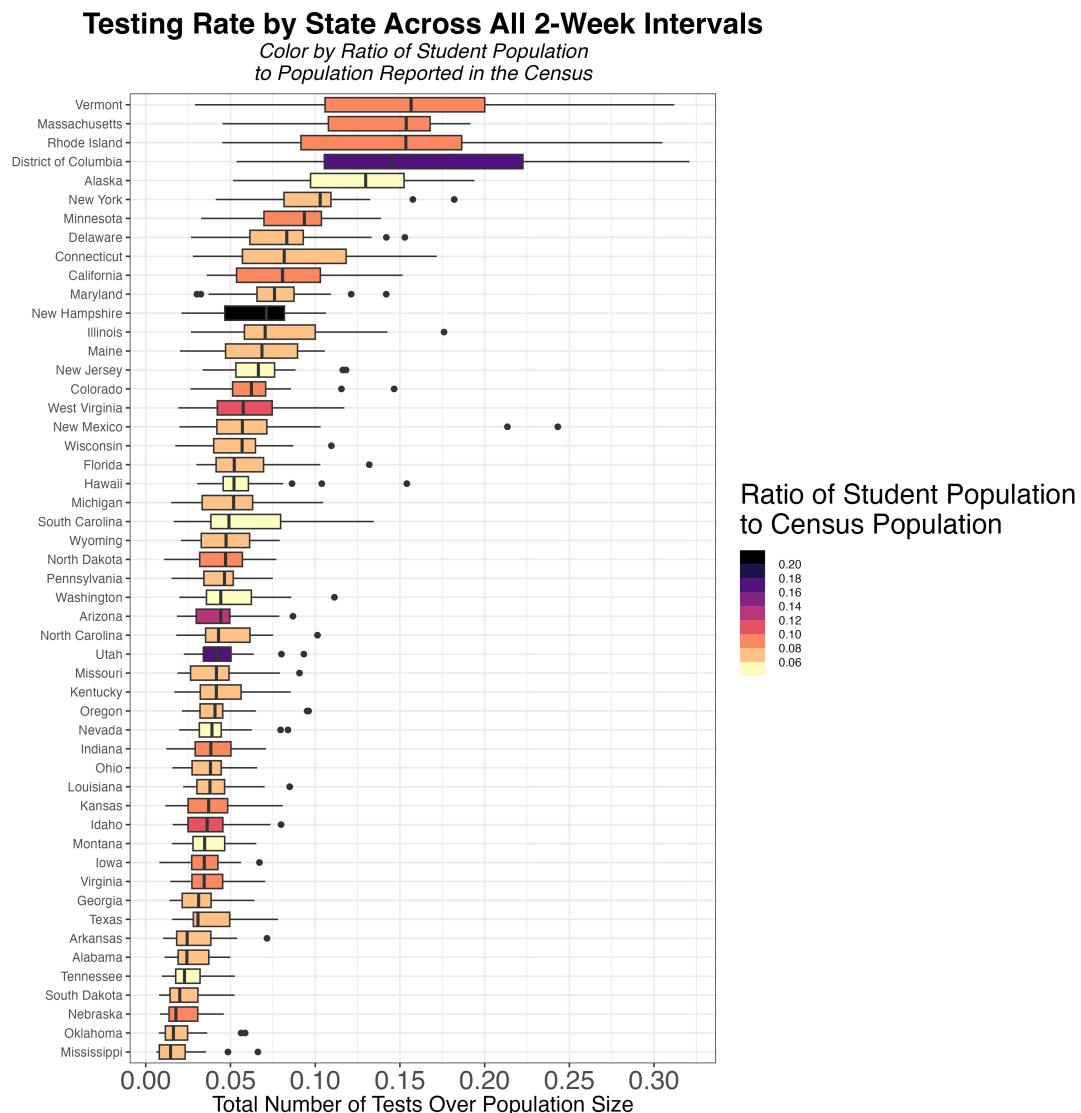


Figure 5.11: Comparing the biweekly testing rate by state across all 2-week intervals considered, where color is by the ratio of the student population to the census population. States are ordered by median testing rate. Data source for student population by state: the U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS), 12-month Enrollment component provisional data for 2020 to 2021. Student population counts include Title IV postsecondary institutions.

## 5.5 County-level Results

We performed county-level probabilistic bias analysis for Michigan and Massachusetts. This work can be expanded to consider other states as well where the needed data is available. In particular, we need both county-level positive PCR tests and county-level total PCR tests. Because the assumptions of the bias correction are related to test positivity, it does not make sense to apply the method to a positive cases count that includes positive PCR tests lumped together with probable cases. In some states, this is the only value reported.

While positive tests are more regularly reported, total tests are reported less frequently at the county level.

### 5.5.1 Massachusetts

Figure 5.12 shows the bias corrected estimates for each implementation, as well as the observed cases. We note that the lower bounds of the bias corrected estimates are always above the observed cases because adding (unobserved) infections among the untested population to the observed positives among the population never results in a decrease in the estimated infections. In theory such a decrease could be possible since we do correct for differences due to imperfect test accuracy, and if the false positive rate was high enough, we might estimate the lower bound of cases as lower than the observed cases. However, the false positive rate of the COVID-19 PCR test is so low that in practice we do not see lower bounds lower than the number of infections.<sup>5</sup>

We can see although the trends are broadly similar between versions for each county, centering the distribution at the empirical value of  $\beta$  leads to peaks not present in the version where priors do not vary by county and date. However, only centering  $P(S_1|\text{untested})$  at the empirical value leads to a distribution that is highly similar to the version where priors do not vary by county and date.

These results make sense when we consider that this analysis is much more sensitive to the choice of  $\beta$  than  $P(S_1|\text{untested})$ . This follows from the fact we calculate the number of positive infections among those who are untested and *asymptomatic* as

$$\begin{aligned} N_{\text{untested}, S_0}^+ &= P(\text{test}_+|S_0, \text{untested})(N_{S_0, \text{untested}}) \\ &= (\beta P(\text{test}_+|\text{tested})) N_{\text{untested}}(1 - P(S_1|\text{untested})) \end{aligned}$$

and the number of positive infections among those who are untested and *symptomatic* as

---

<sup>5</sup>The false positive rate differs by platform and laboratory, but multiple analyses estimated that it is less than 0.10% (Chandler, Bourassa, Mathias, & Greninger, 2021).

$$\begin{aligned}
N_{\text{untested}, S_1}^+ &= P(\text{test}_+ | S_1, \text{untested})(N_{S_1, \text{untested}}) \\
&= (\alpha P(\text{test}_+ | \text{tested})) N_{\text{untested}}(P(S_1 | \text{untested})).
\end{aligned}$$

Since  $N_{S_1, \text{untested}}$  is so much larger than  $N_{S_1, \text{untested}}$  for any of the specified values of  $P(S_1 | \text{untested})$  (since the bulk of this distribution is less than 5%),  $\beta$  has a larger impact on the number of estimated infections.

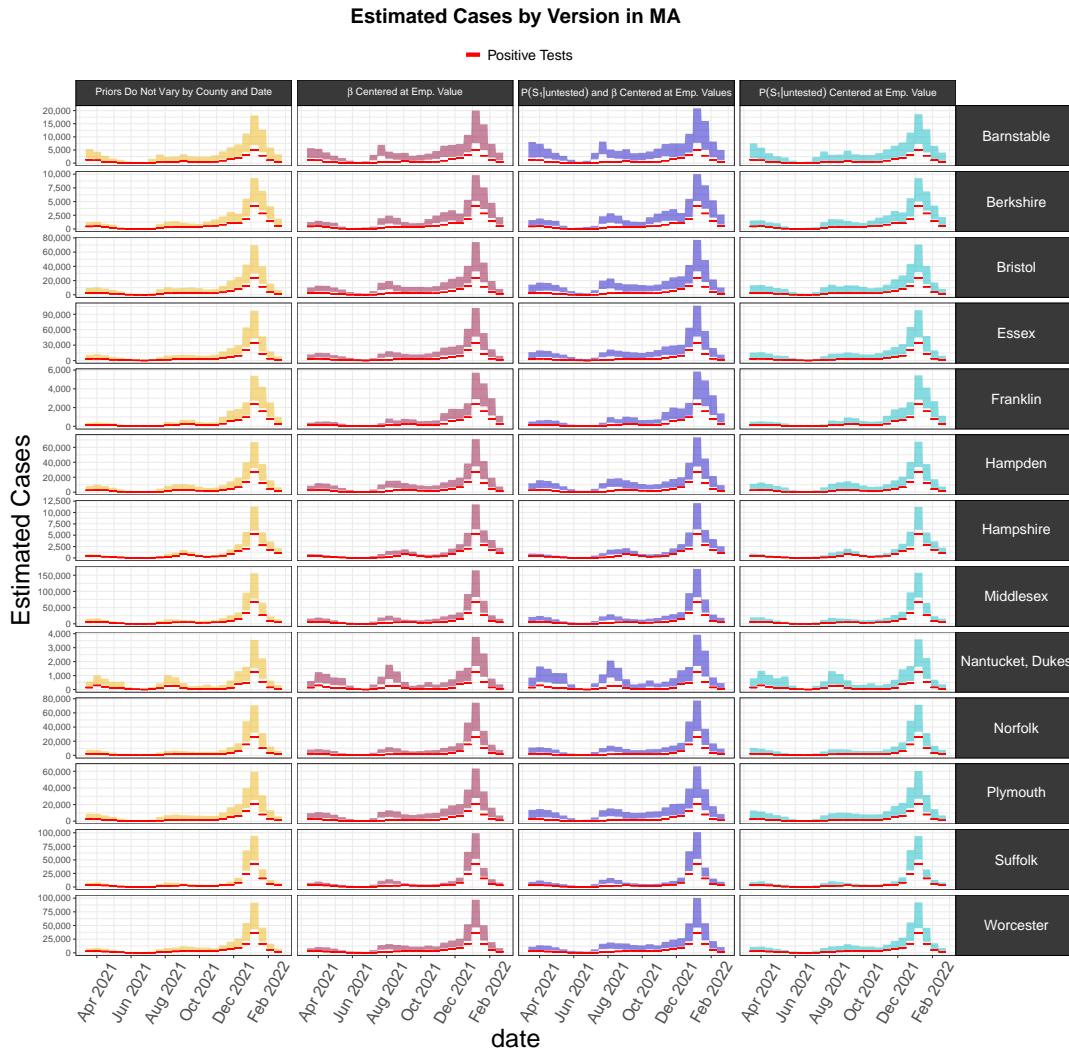


Figure 5.12

To better see the overlap between versions, in Figure 5.13 we can look at the versions together. This allows us to see more clearly how the version with both  $P(S_1 | \text{untested})$  and  $\beta$  centered at their empirical values is consistently the highest. Meanwhile, the version with only  $P(S_1 | \text{untested})$  centered at its empirical value corresponds so closely to the version that does not vary by date or location that

there is no part of the intervals for the version not varying by date or location that do not overlap with the  $P(S_1|\text{untested})$  version.

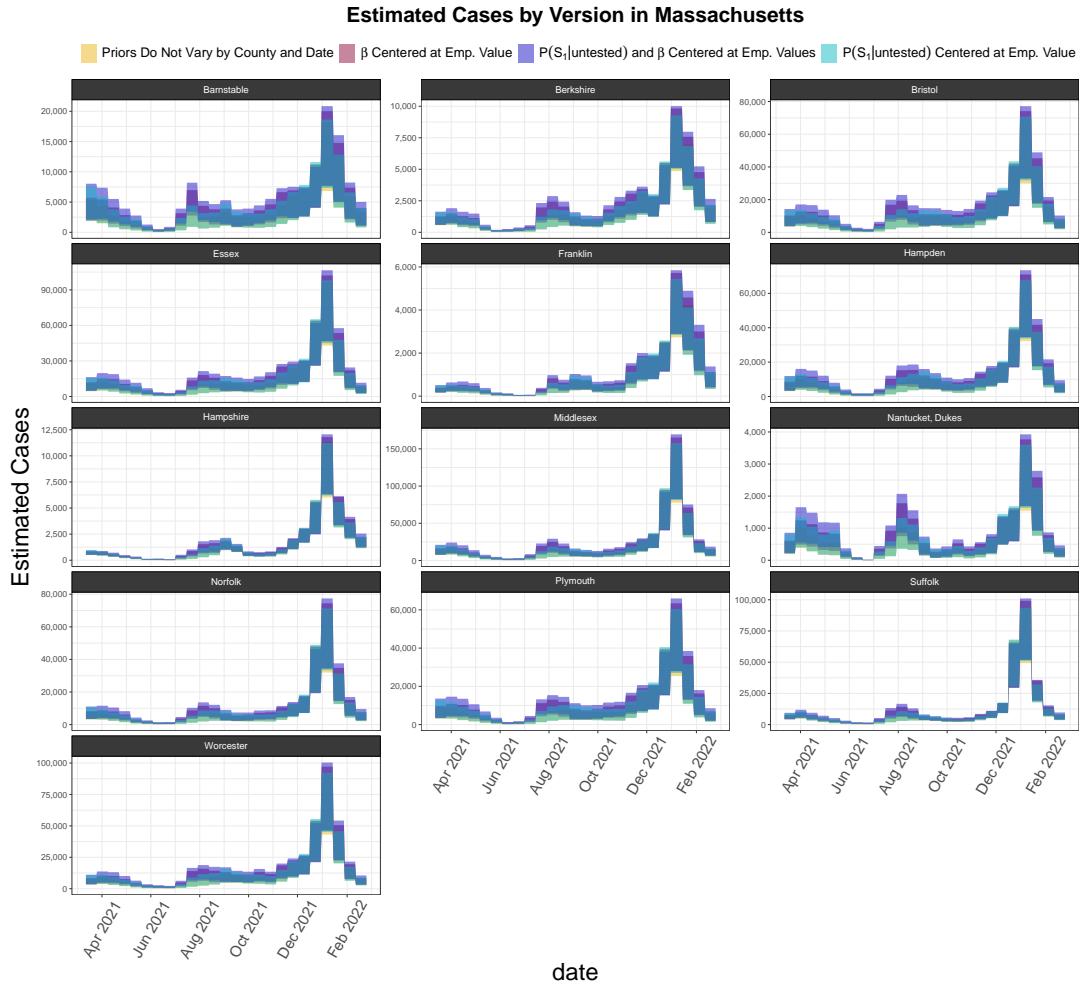


Figure 5.13

Now, we can also compare the extent to which these intervals agree with the Covidestim estimates. In 5.14, we see that when Covidestim estimates fall outside the probabilistic bias intervals, they fall below the intervals, estimating that there are less infections. The exception here is Hampshire County, which actually is very likely to be related to the high amount of asymptomatic testing done among the five colleges. Smith College alone, with a student body of about 2,500 and mandatory PCR testing twice a week, would contribute a substantial amount of tests to the test count in Hampshire County. Given the low infection burden, the vast majority of these tests were negative throughout most of 2021, reducing the positivity rate (Figure 5.15). The extensive amount of asymptomatic testing done here would mean that means that our correction factor  $\beta$  in probabilistic analysis<sup>6</sup>,

<sup>6</sup>Recall  $P(\text{test}_+|S_0, \text{untested}) = \beta P(\text{test}_+|\text{tested})$ .

which we use to estimate the positivity rate among the untested asymptomatic population, would be too low. This is because the observed test positivity rate when so much asymptomatic testing is conducted would be closer to what expect the positivity rate among the asymptomatic test positivity rate to be among the untested population. Because the assumptions of the probabilistic bias analysis here did not assume such extensive asymptomatic testing was contributing to the test positivity rate, we are likely underestimating the true number of unobserved infections if there is a substantial amount of screening testing. It would be simple, however, to adjust  $\beta$  to account for cases where there is increased asymptomatic testing.

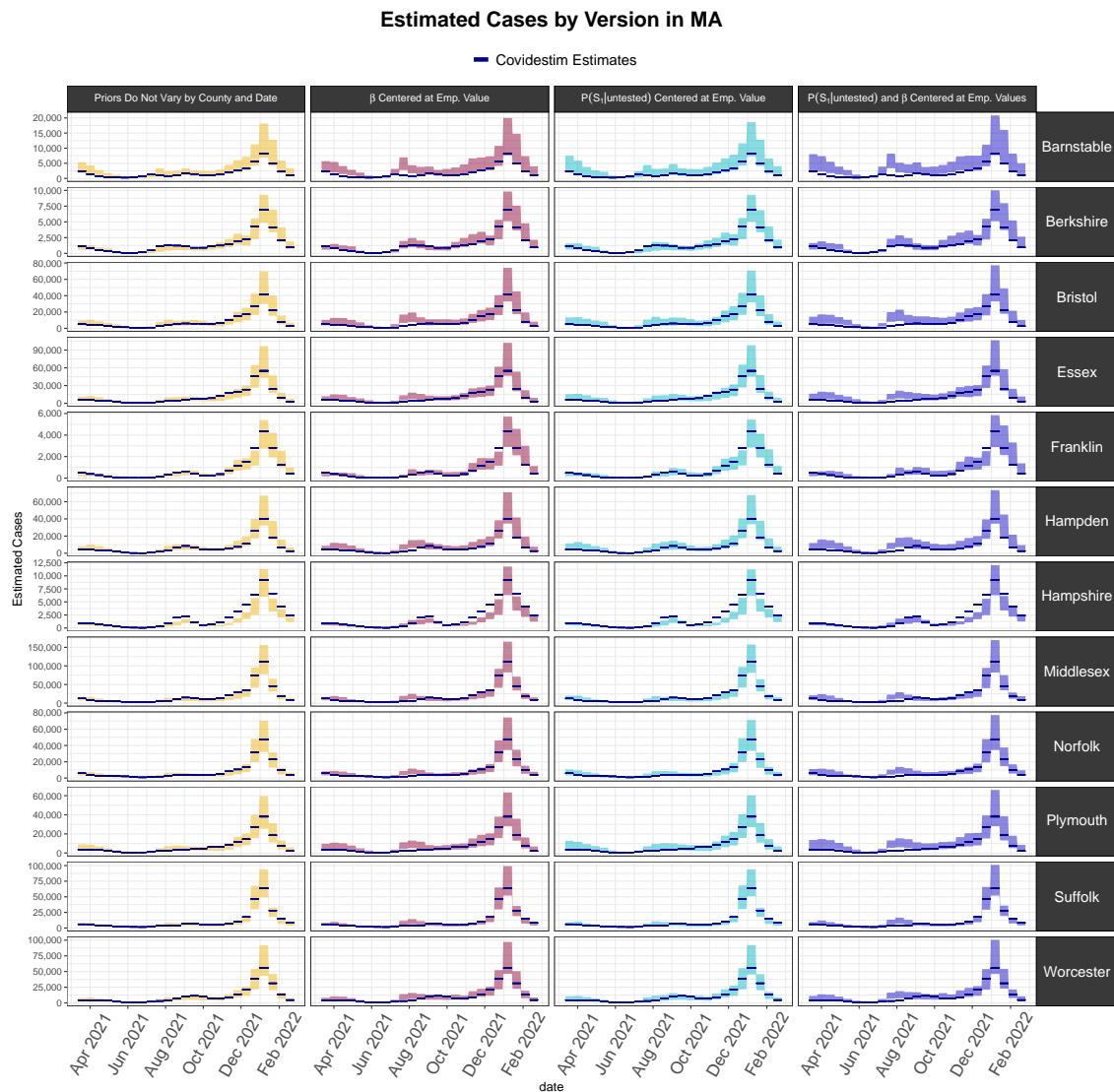


Figure 5.14

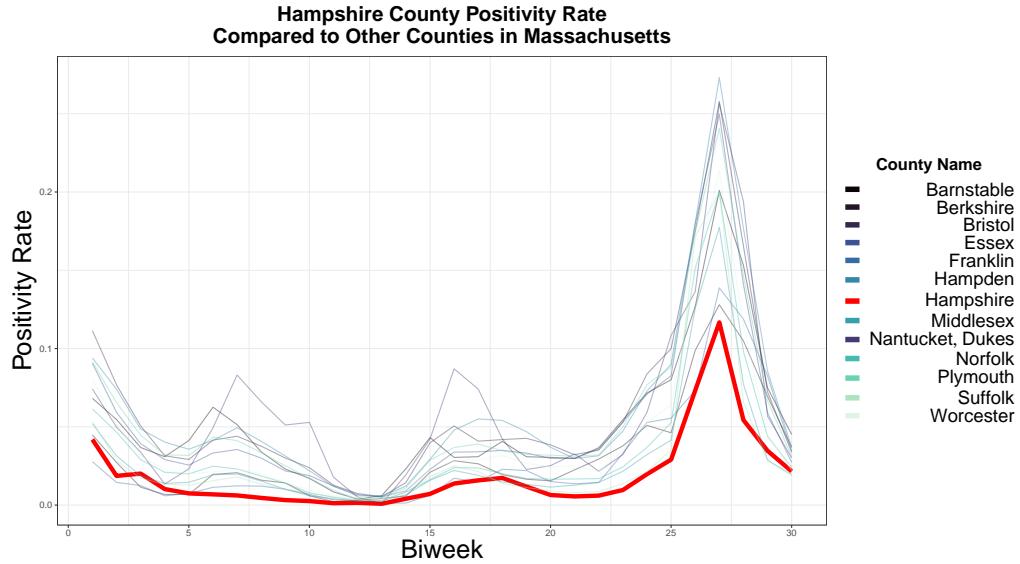


Figure 5.15: We see that Hampshire County has among the lowest test positivity rates among counties in Massachusetts in the time period considered.

To summarize which versions are most concordant with the Covidestim estimates, we can consider the proportion of the biweeks where the interval contains the Covidestim estimate. As we see in Figure 5.16, the implementation with the prior for  $P(S_1|\text{untested})$  centered at the empirical value tends to agree most with the Covidestim estimates.

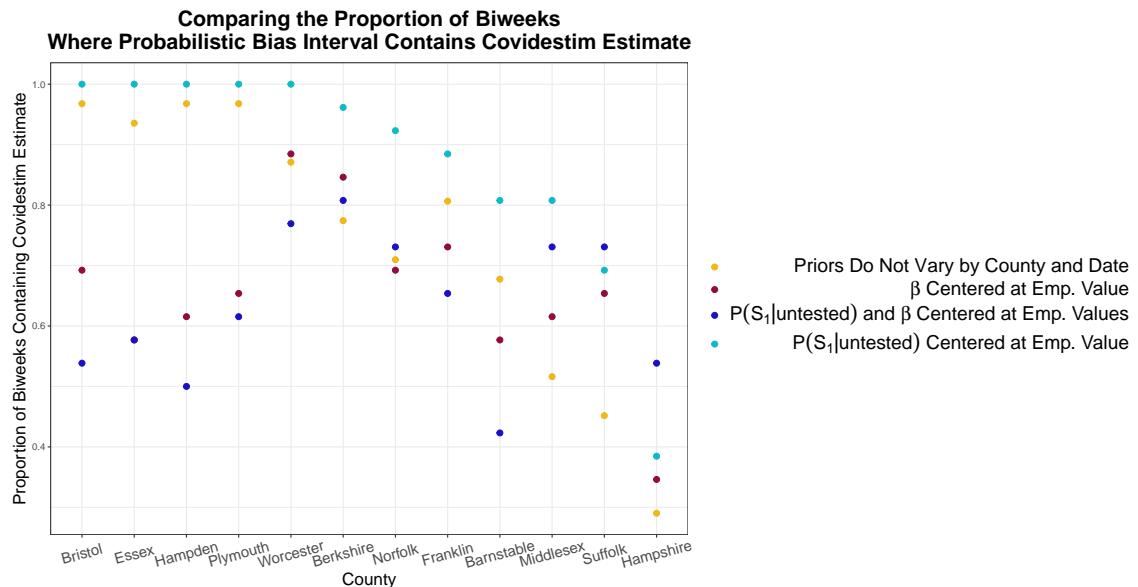


Figure 5.16



### 5.5.2 Michigan

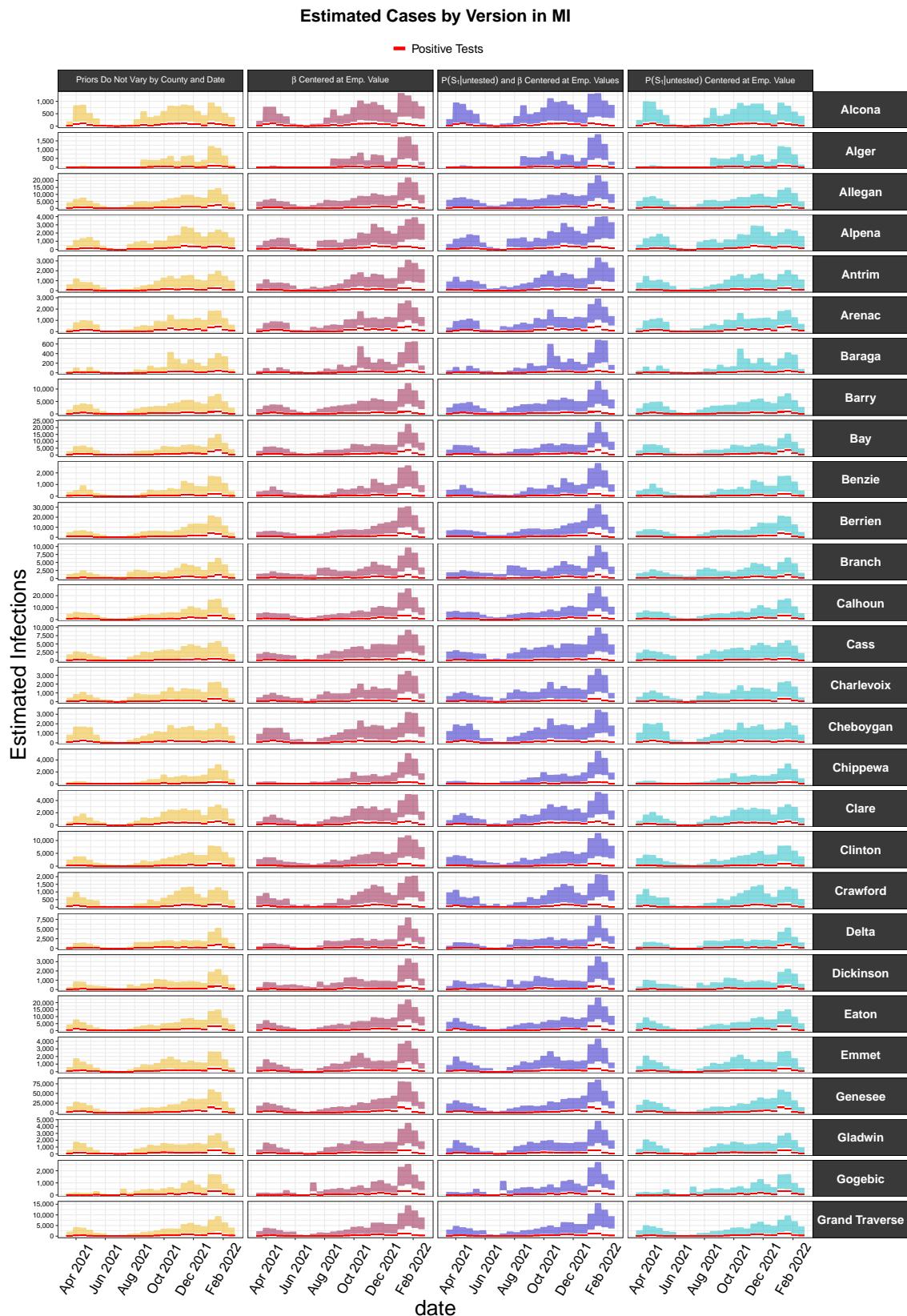


Figure 5.17

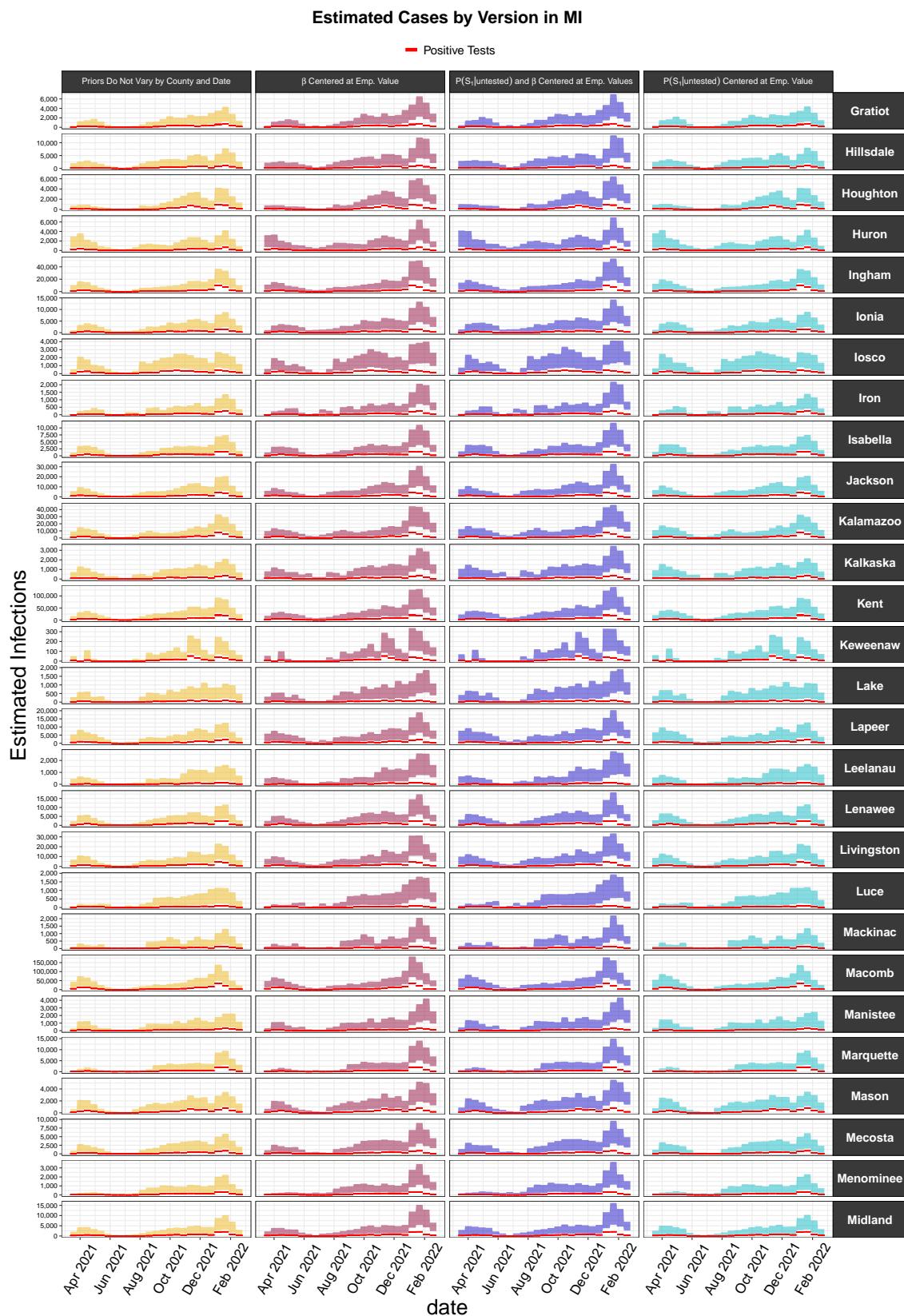


Figure 5.18

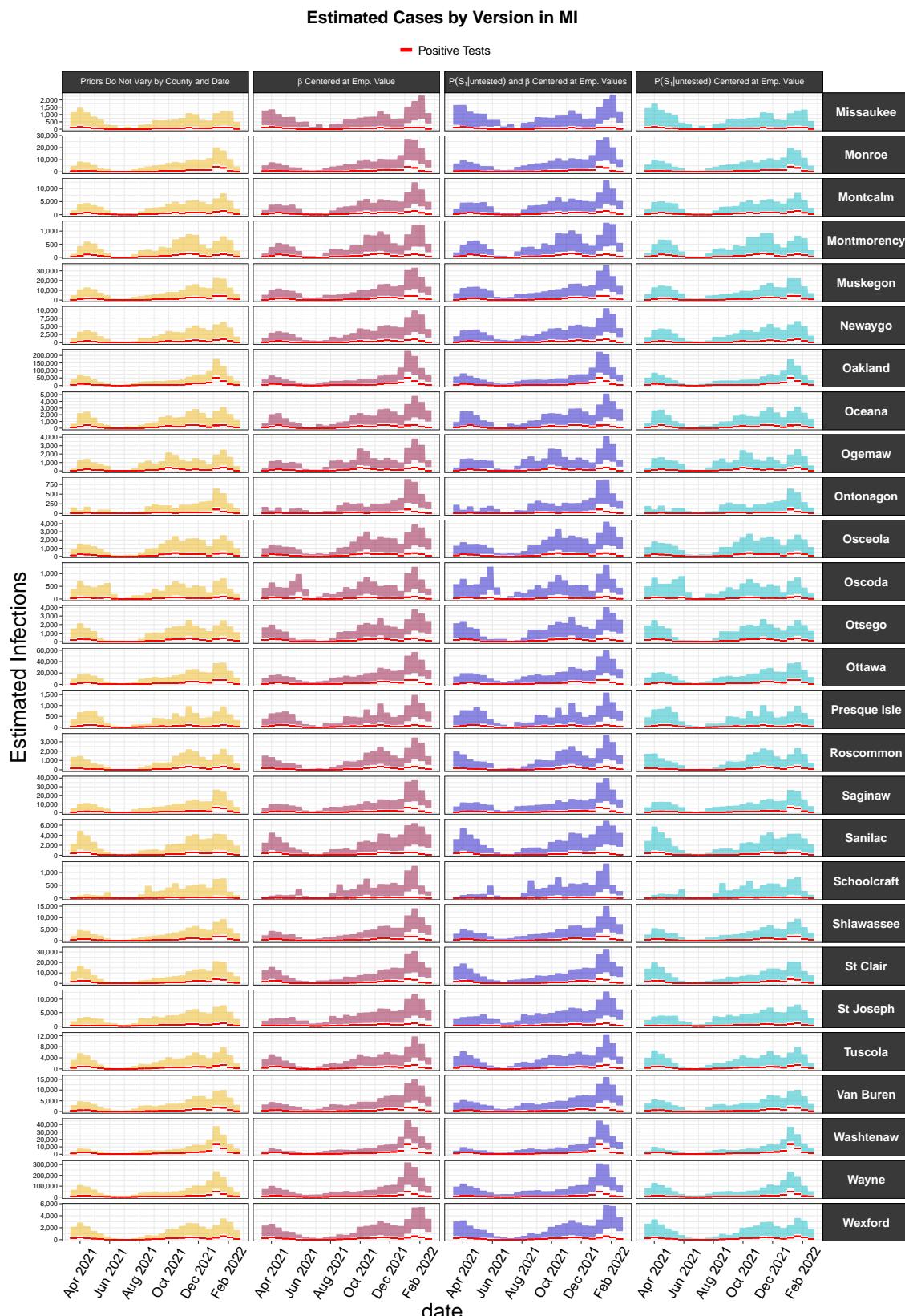
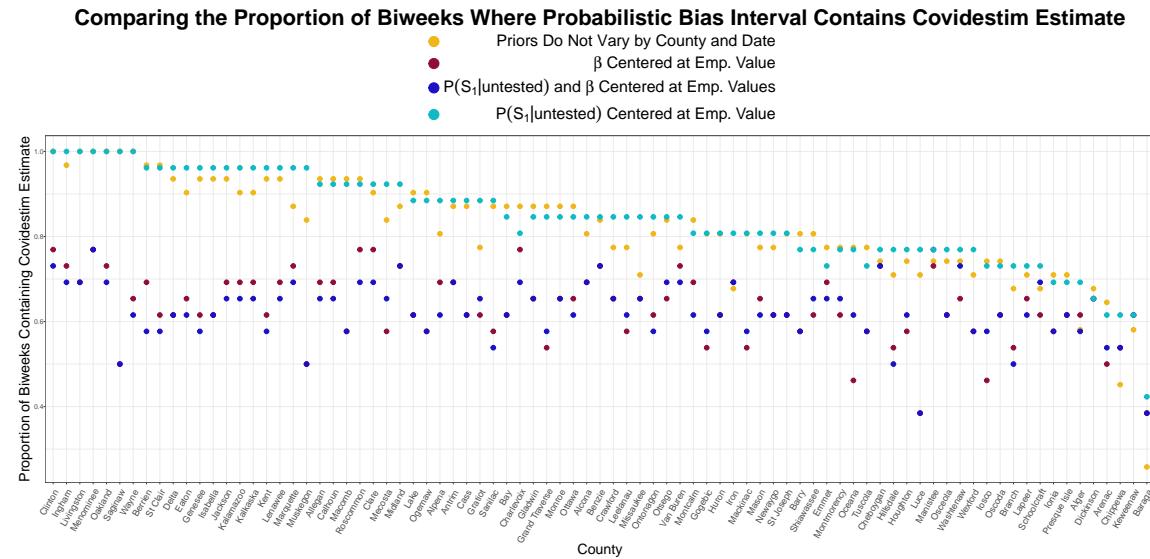


Figure 5.19



## 5.6 Cross Correlation Comparison

### 5.6.1 Background

An ongoing challenge for assessing the quality of the probabilistic bias intervals is that there is no ground truth to compare to. To broaden our comparison beyond Covidestim, we can look at wastewater data.

Wastewater data is a source of data that has been of rising interest throughout the pandemic, in part due to its cost effectiveness in assessing community-level burden, but also due to the fact it represents a much more unbiased sample than COVID-19 testing does.

That said, there are challenges in relating wastewater concentrations to the true number of infections, in part because of the same issue we face here of the lack of ground truth for the true number of infections in any location. The choice of normalization of the viral RNA concentrations of SARS-CoV-2 is important for understanding how these concentrations scale to the number of infections, since the concentration of virus (in genome copies per liter) in a sample will be influenced by various factor unrelated to the true prevalence COVID-19, such as processing differences between treatment plants or trends in water usage. One common choice is to normalize against the concentration of a virus that has a relatively stable population in wastewater, such as Pepper Mild Mottle Virus (PMMoV) (Zhan et al., 2022).

Wastewater testing has become increasingly widespread throughout the pandemic as the technology and analysis approaches have evolved, as well as the demand for a source of data on the presence of COVID-19 that is less reliant on access to tests (or symptoms strong enough to warrant a test, which differ by the variants circulating). A comprehensive source of wastewater data across the United States is provided by Biobot Analytics, which is the institution partnering with the CDC for the National Wastewater Surveillance System (NWSS) (Duvallet et al., 2022). Biobot Analytics provide wastewater concentrations aggregated at the county scale by using a weighted average of the concentrations at sampling locations within the county, weighted by the size of the corresponding sewershed populations. This data is publicly available on a public github repository.

Most notable for this work, several counties in Massachusetts have reported wastewater data for a substantial period throughout 2021 to 2022. This allows us to compare the bias-corrected estimates – as well as the Covidestim estimates – to the wastewater concentrations.

Wastewater concentrations are typically a leading indicator of observed cases, though there may be some variability in the lead time during different waves of the pandemic (Hopkins et al., 2023). In particular, the lead time was strongest in the earliest waves of the pandemic, and has since declined (Xiao et al., 2022). Various factors can create the changes we see in lead time over the course of the pandemic; for example, the lead time can be impacted by differences in viral shedding, diagnostic testing turnaround times, and testing capacity and behavior (Olesen, Imakaev, & Duvallet, 2021).

Since the correlation between the time series as well as the lag at which the maximum correlation occurs are both of interest, we assessed the cross correlation between the series.

First, we define autocorrelation since the definition of cross correlation is very similar. The definition here uses the notation of Shumway & Stoffer (2011).

### Definition: Autocorrelation

Denote the set of time points of a time series  $T$ . For any time series  $(x_t)_{t \in T}$ , we define the auto-correlation function (ACF) as

$$\rho_{XX}(\tau) = \frac{E[(X_{t+\tau} - \mu_{X_{t+\tau}})(X_t - \mu_{X_t})]}{sd(X_{t+\tau})sd(X_t)}.$$

Assuming second order stationarity<sup>7</sup>, we have  $\mu_{X_{t+\tau}} = \mu_{X_t}$  and similarly  $\text{Var}(X_{t+\tau}) = \text{Var}(X_t)$ , so we can simplify the expression for  $\rho_{XX}(\tau)$  to yield

$$\rho_{XX}(\tau) = \frac{E[(X_{t+\tau} - \mu_X)(X_t - \mu_X)]}{\text{Var}(X)}.$$

The auto-correlation function  $\rho_{XX}(\tau)$  measures the linear dependence between  $X_{1+\tau}, \dots, X_n$  and  $X_1, \dots, X_{n-\tau}$ , that is, the difference between the original time series and the time series shifted forward by  $\tau$  time units.

We can extend this definition to quantify the linear relationship between distinct lagged time series  $X_1, X_2, \dots, X_t$  and  $Y_1, Y_2, \dots, Y_t$  by defining the cross correlation function. The function is only defined on two time series that are over the same time interval and sampled at the same frequency.

### Definition: Cross-Correlation

We compute the cross-correlation function (CCF) as

$$\rho_{XY}(s, t) = \frac{E[(X_s - \mu_{X_s})(Y_t - \mu_{Y_t})]}{\sqrt{\text{Var}(X_s)\text{Var}(Y_t)}}.$$

Again assuming the series satisfy second-order stationarity, we have

$$\rho_{XY}(s, t) = \frac{E[(X_s - \mu_X)(Y_t - \mu_Y)]}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

The implementation of the cross correlation in base R (`stats::ccf`) assumes second order stationarity (Venables & Ripley, 2002).

Looking at cross correlation can be useful in the sense that we can both consider the strength of correlation and the lag at which the correlation is maximized. Before

---

<sup>7</sup>Second order stationarity is also referred to as weak stationarity, and implies that the mean, variance are constant over time and the autocovariance function depends only on the difference between time points.

presenting the cross correlation results of the county level time series, we can consider a more concrete example, where the lag is known.

In Figure 5.20, we consider simulated data where  $(Z_t)$  is  $(Y_t)$  lagged by 3 time units with noise added. We can see that  $Z_t$  and  $Y_t$  are not second-order stationary since the mean clearly is not constant over time. However, to stabilize the mean, we can apply first order differencing, where we take the differences between consecutive observations.

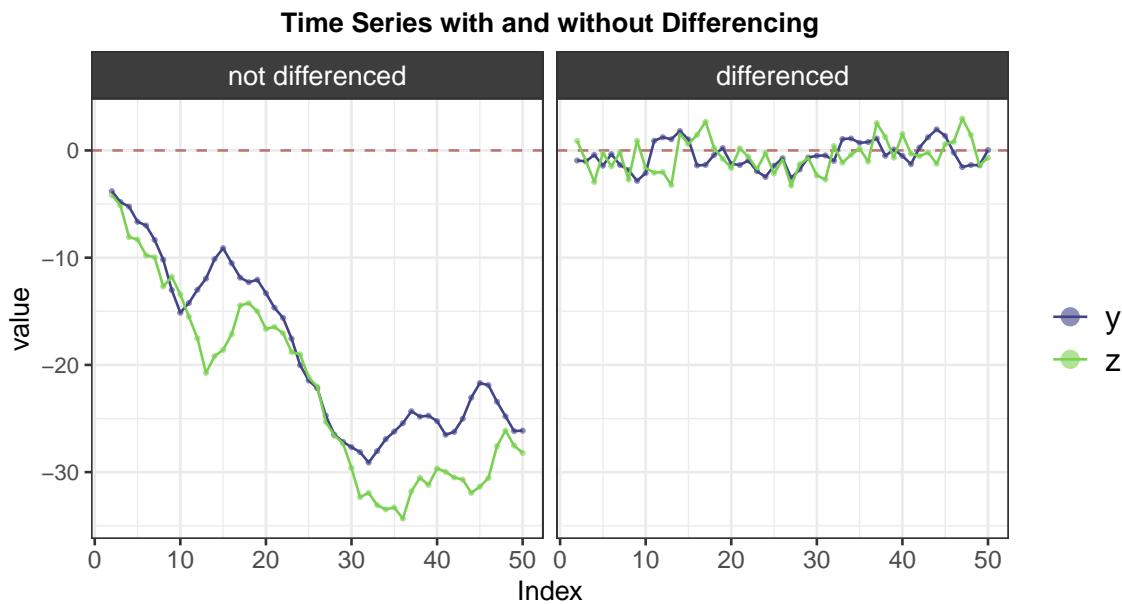


Figure 5.20

We can see the effect of applying differencing to the time series when we compute the cross correlations of  $(Z_t)$  and  $(Y_t)$ , as shown in Figure 5.21. The true lag of  $-3$  time units was recovered when considering the differenced time series, but not when we considered the original time series. In what follows, because the time series we are considering are not stationary, we consider the cross correlation between the differenced time series.

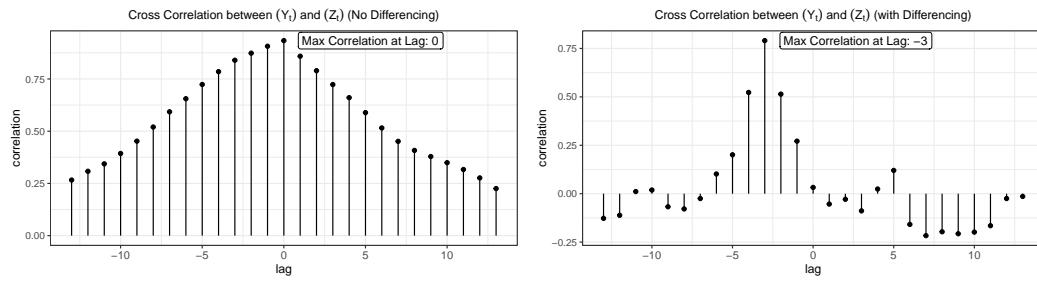


Figure 5.21

### 5.6.2 Cross Correlation Results Comparing Bias Corrected Counts, Covidestim Estimates, and Wastewater Concentrations

Because wastewater data is reported at the weekly time scale while the bias corrected estimates are at the 2-week time scale, we take a mean of the effective concentration for each 2 week interval, such that the time series are sampled at the same frequency.<sup>8</sup>

Since the effective concentration of SARS-CoV-2 in wastewater samples reported by Biobot is in genome copies per liter and is not directly comparable to estimates of infections, we place the wastewater concentration on a separate scale.

Looking at the counties in Figure 5.22, we see that, with the exception of Barnstable, MA, the wastewater trends are highly similar to trends captured by the bias corrected infection counts. We also see that the trends are similar both with regard to shape but also with regard to time, with little visible lag between the series. This is expected because although wastewater cases do in general lead cases, lead times generally are not on the order of 2 weeks. This means that since we are summarizing to 2-week intervals we would expect the lag to be very small, if present at all.

---

<sup>8</sup>We cannot interpret the cross correlation if the time steps are different.

## Comparing Bias Corrected Infections with Wastewater Concentrations Over Time

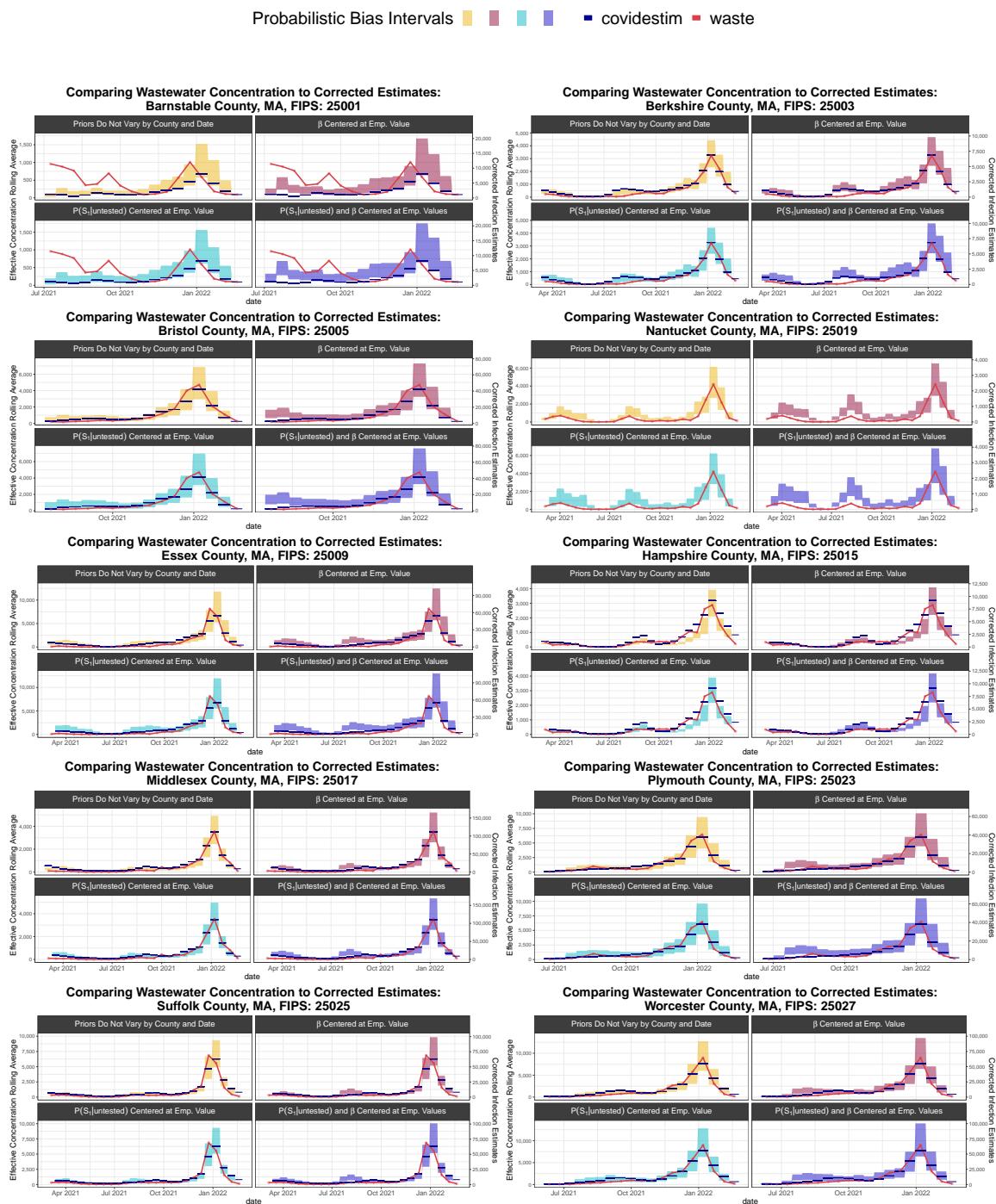


Figure 5.22

## Comparison Between Implementations of Probabilistic Bias Analysis

In Figure 5.23, we see that, in general, infections were highly correlated with the wastewater effective concentrations, which was true across all implementations of probabilistic bias analysis. In most cases, the implementation where priors did not vary by state or date were the most highly correlated with the wastewater concentrations. Exceptions to this were Barnstable County (25001), where the implementation with the prior for  $\beta$  centered at the empirical value was the most highly correlated, and Worcester County (25027), where the implementation with the prior for  $P(S_1|\text{untested})$  centered at empirical value was the most highly correlated. In all counties except for Barnstable, the lag at which the maximum correlation was obtained was 0 units, while for Barnstable it was -1, indicating that wastewater concentrations led infections by one two-week interval.

Given the small size of Barnstable relative to other counties and high variability in its early estimates in 2021 (as seen in Figure 5.22), it is possible that there were still aspects of the SARS-CoV-2 detection process that took time to refine. Another possibility is that the way Biobot aggregated wastewater concentrations by county failed to capture the infection dynamics in this county, since wastewater catchments are not contained within county lines. This is a central challenge in relating cases to wastewater concentrations, since these values are recorded for distinct geographic units.

Comparing the maximum correlations obtained the observed cases, only in Hampshire County were the observed cases more correlated with the wastewater concentrations than all implementations of probabilistic bias analysis. We also see again that in most cases the maximum correlation is obtained at zero lag in observed cases; however, for Barnstable, the correlation is highest when wastewater concentrations lead infections by two biweeks, and for Hampshire the correlation is highest when wastewater concentrations lead infections by 1 biweek.

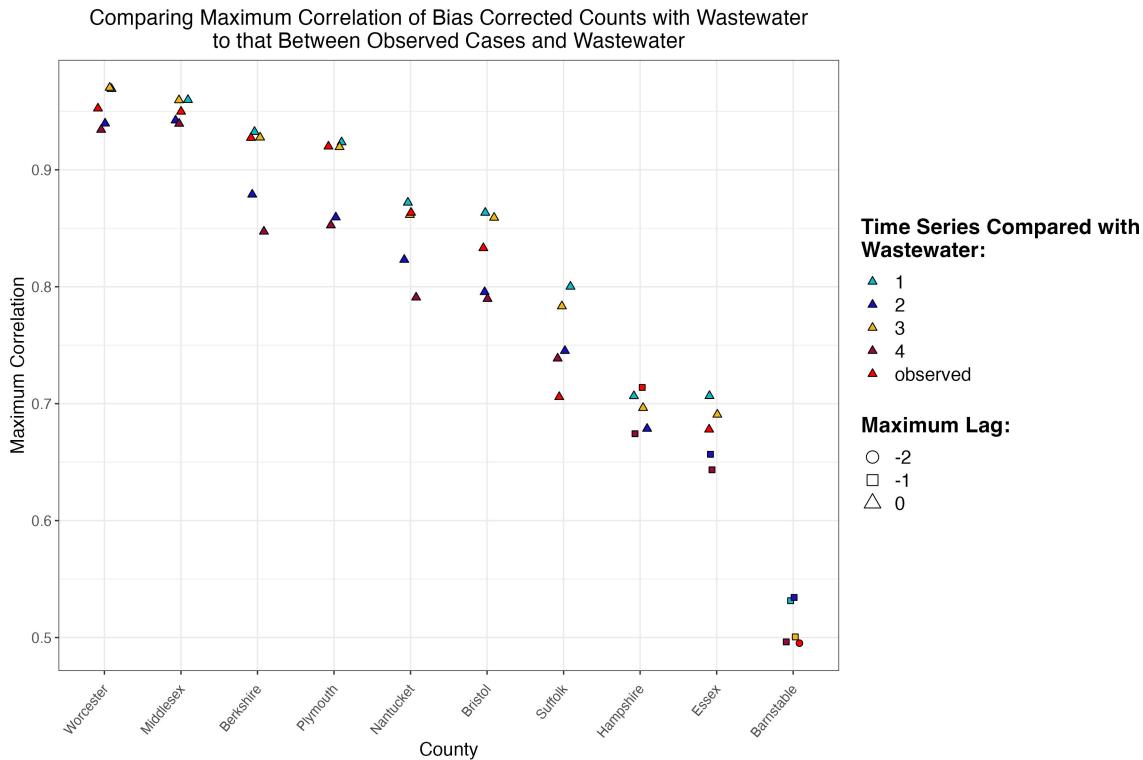


Figure 5.23

### Comparison Between Covidestim, Observed Cases, and Bias Corrected Counts

In Figure 5.24, we also compare the Covidestim estimates to the wastewater concentrations. In general, both Covidestim and bias-corrected counts are more correlated with wastewater concentrations than observed infections. Of note, Nantucket County (2019) is not included here because Covidestim does not report estimates are not reported for Nantucket.<sup>9</sup>

<sup>9</sup>In reporting of COVID-19 data, Nantucket values are grouped with Dukes County, which is likely why Covidestim does not try to estimate the grouped counts.

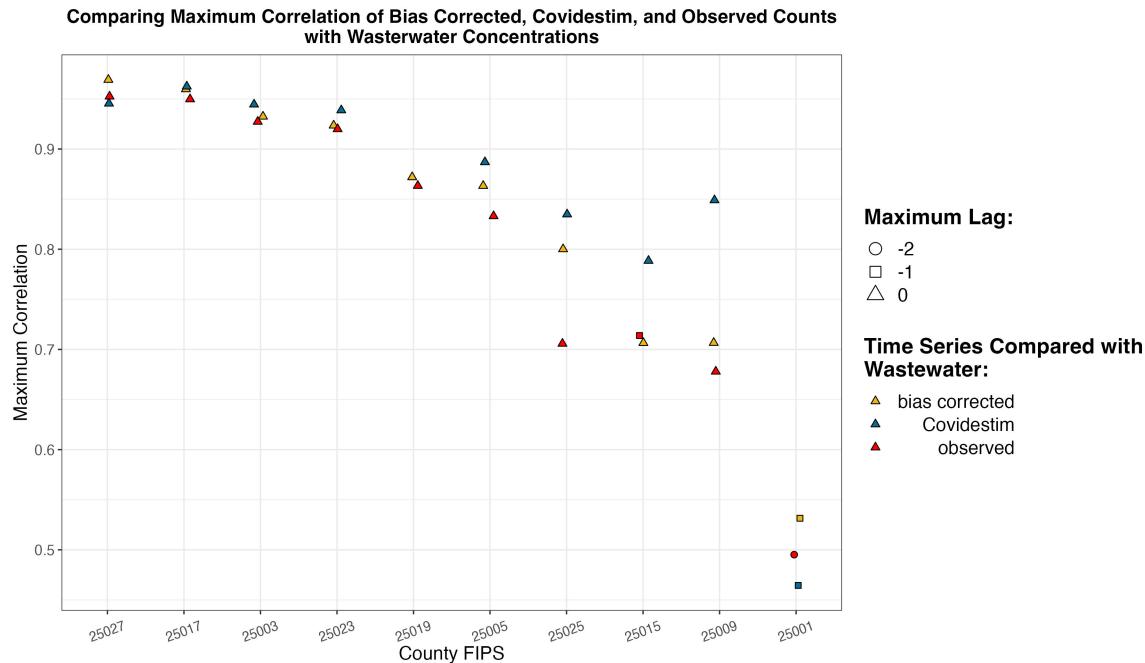


Figure 5.24

## Takeaways

The aim of the cross correlation analysis was to add another source of comparison for the county-level counts from an entirely different source of data – in particular, a source of data that is less impacted by access to testing or test behavior. We see that in most counties considered here, there is high agreement between the time series. An avenue for future exploration would be to consider this analysis among a broader set of counties to see which time series tends to be most highly correlated with wastewater concentrations, a question that we cannot confidently address here when looking only at counties in Massachusetts.

# Chapter 6

## Conclusion

The aim of this work is to consider possible scenarios for the extent of unobserved infections over an extended time during the COVID-19 pandemic, and to explore how we can present the uncertainty in the number of incident infections. Throughout the pandemic we often see line charts of observed cases or the test positivity rate. Advice has changed as has testing behavior, with warnings to not consider case counts in isolation, but rather to also look at trends in the test positivity rate (among other indicators). However, presentation of infections as intervals, their widths defined as a direct consequence of assumptions we make about the bias parameters, reflects genuine uncertainty about the number of true infections that may exist in a given area over time. Various models exist to try to get at this quantity of the number of true infections, incorporating a range of sources of data, including COVID-19 deaths, hospitalizations, seroprevalence data, and viral concentrations in wastewater, as well as estimates such as the infection fatality ratio. The strength of applying probabilistic bias analysis to consider possible values of true COVID-19 infections lies in its relative simplicity and transparency of assumptions, in addition to the ease of exploration of possible testing scenarios of the extent to which infections are going undetected. Although there is no ground truth to rigorously assess the accuracy of an estimate of the true number of COVID-19 infections, comparing approaches and understanding where and when they are non concordant provides useful insight into quantifying the range of true infections.



# Chapter 7

## Appendix

Placeholder

### 7.1 Smoothing Span

#### 7.1.1 Changing SPAN for LOESS Smoothing of $\beta$

### 7.2 Changing Mean and Variance for Prior Distribution Specifications

### 7.3 Relationship Between $(X + Y)_\alpha$ and $X_\alpha + Y_\alpha$ for Dependent Variables $X, Y$

#### 7.3.1 Simulation: Bivariate Normal

#### 7.3.2 Derivation of the Distribution of X+Y for Bivariate Normal

### 7.4 Total Tests and Positive Tests Across All States



# References

Placeholder

- CDC. (2020, February 11). Cases, data, and surveillance to track and analyze COVID-19. Retrieved April 12, 2023, from <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/serology-surveillance/index.html>
- Chandler, C. M., Bourassa, L., Mathias, P. C., & Greninger, A. L. (2021). Estimating the False-Positive Rate of Highly Automated SARS-CoV-2 Nucleic Acid Amplification Testing. *Journal of Clinical Microbiology*, 59(8), e01080–21. <http://doi.org/10.1128/JCM.01080-21>
- Commercial Laboratory Seroprevalence Surveys | Coronavirus | COVID-19 | CDC. (n.d.). Retrieved April 12, 2023, from <https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html>
- Duvallet, C., Wu, F., McElroy, K. A., Imakaev, M., Endo, N., Xiao, A., ... Matus, M. (2022). Nationwide Trends in COVID-19 Cases and SARS-CoV-2 RNA Wastewater Concentrations in the United States. *ACS ES&T Water*, 2(11), 1899–1909. <http://doi.org/10.1021/acsestwater.1c00434>
- Hopkins, L., Persse, D., Caton, K., Ensor, K., Schneider, R., McCall, C., & Stadler, L. B. (2023). 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. <http://doi.org/10.1016/j.scitotenv.2022.158967>
- O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L., Cummings, D. A. T., Azman, A. S., Paireau, J., ... Salje, H. (2021). Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, 590(7844), 140–145. <http://doi.org/10.1038/s41586-020-2918-0>
- Olesen, S. W., Imakaev, M., & Duvallet, C. (2021). Making waves: Defining the lead time of wastewater-based epidemiology for COVID-19. *Water Research*, 202, 117433. <http://doi.org/10.1016/j.watres.2021.117433>
- Pitzer, V. E., Chitwood, M., Havumaki, J., Menzies, N. A., Perniciaro, S., Warren, J. L., ... Cohen, T. (2021). The Impact of Changes in Diagnostic Testing Practices on Estimates of COVID-19 Transmission in the United States. *American Journal of Epidemiology*, 190(9), 1908–1917. <http://doi.org/10.1093/aje/kwab089>
- Shumway, R. H., & Stoffer, D. S. (2011). *Time series analysis and its applications: With R examples* (3rd ed). New York: Springer.

- Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York, NY: Springer New York. <http://doi.org/10.1007/978-0-387-21706-2>
- Xiao, A., Wu, F., Bushman, M., Zhang, J., Imakaev, M., Chai, P. R., ... Alm, E. J. (2022). Metrics to relate COVID-19 wastewater data to clinical testing dynamics. *Water Research*, 212, 118070. <http://doi.org/10.1016/j.watres.2022.118070>
- Zhan, Q., Babler, K. M., Sharkey, M. E., Amirali, A., Beaver, C. C., Boone, M. M., ... Solo-Gabriele, H. M. (2022). Relationships between SARS-CoV-2 in Wastewater and COVID-19 Clinical Cases and Hospitalizations, with and without Normalization against Indicators of Human Waste. *ACS ES&T Water*, 2(11), 1992–2003. <http://doi.org/10.1021/acsestwater.2c00045>