Estimating Unobserved COVID-19 Infections in the United States

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

Preface

I am unsure as to what goes here.

Table of Contents

List of Tables

List of Figures

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.

```
knitr::opts_chunk$set(fig.width = 10, echo = FALSE)
```

Chapter 1

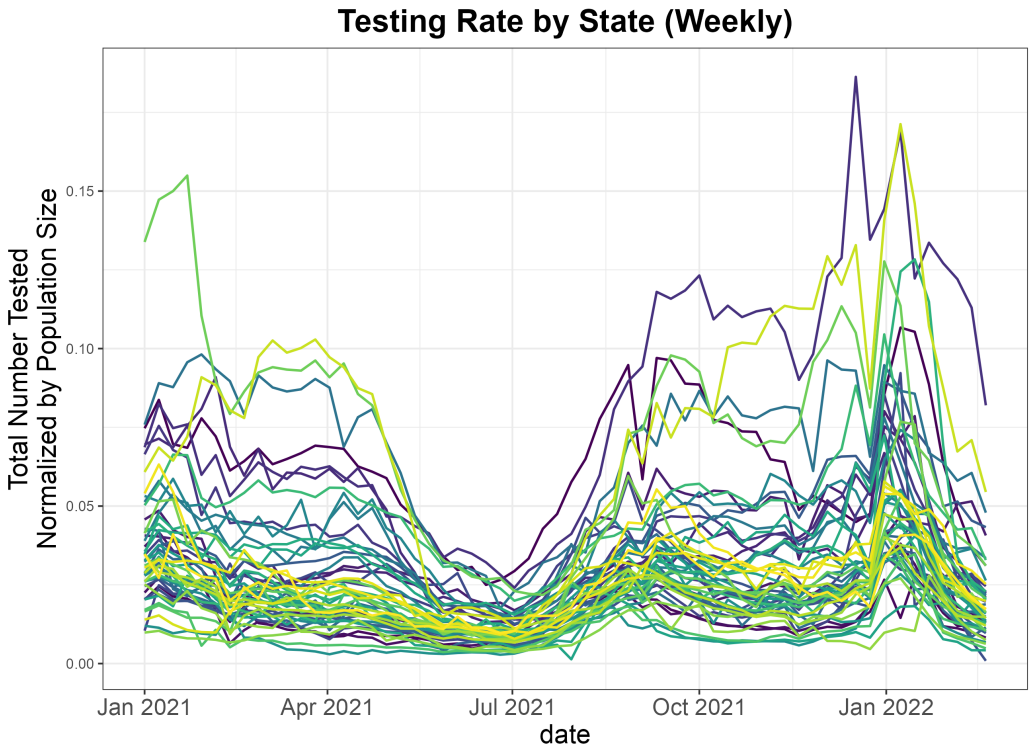
Motivation

Throughout the COVID-19 pandemic, observed infections have guided decisions at both the individual and government levels. At the state-level, policies on phased reopening, for example, often include criteria on COVID-19 cases (California Department of Public Health, 2021; Charles D. Baker, 2021; Tom Wolf, 2020).

To make this data accessible to the public, several organizations, including the CDC (Centers for Disease Control and Prevention, 2020), John Hopkins University (Dong, Du, & Gardner, 2020), and the New York Times (The New York Times, 2022), compiled comprehensive dashboards presenting key metrics such as positive cases and test positivity rates across states.

However, our interpretation of case counts as a measure of transmission is limited by the fact that testing rates impact these trends. The number of positive cases we observe in a county, for instance, will be a result of that county's testing capacity and testing behavior of its population. This means the relationship between observed infections and true total infections may not be monotonic.

The importance of considering testing rate led John Hopkins University to organize the most comprehensive testing database available in the United States (Dong et al., 2020), which enables us to see that testing rate varies substantially by state and time.



As we study the impact and transmission of SARS-CoV-2 as well as the efficacy of different interventions, we often turn to case counts for information. In this way, case counts form the basis for numerous types of analyses that inform our understanding of the pandemic. This means that bias in case counts due to unobserved infections can greatly impact our understanding of the pandemic.

One way testing rates can influence our understanding of COVID-19 is when we are seeking to make comparisons across different locations.

The government response to the pandemic has differed greatly by state, with a range of different policies and timelines as local governments weighted complex tradeoffs. The variability in state-level policies sparks several questions related to the consequences of these policies. Comparing case counts enables us to compare the impact of state-level management of the pandemic. For example, Kaufman *et al.* used cumulative case counts to study the effect of state-level social distancing policies (Kaufman et al., 2021). At the county scale, Jiang *et al.* evaluated the association between stay-at-home orders and daily incident cases (Jiang, Roy, Pollock, Shah, & McCoy, 2022), and Kao *et al.* looked at how the duration of multiple policy interventions – face mask mandates, stay-at-home orders, and gathering bans – affected monthly incidence (Kao et al., 2023).

The bias in case counts is particularly important for inference related to government interventions. With regard to government interventions, it is highly likely that lower testing resources may be related to less stringent policies in other respects. If this is the case, then lower cases may be observed in locations with less stringent policies as an artifact of inadequate testing rather than lower transmission. As a result, when we estimate the effect of a policy intervention based on observed cases, we may be underestimating the true impact.

Besides interventions, there has been substantial concern over the disparities in the impact of COVID-19. As a result, it is important to understand the relationship between various socioeconomic variables and case burden. Chen and Krieger showed a consistent monotonic relationship between the percent poverty and cumulative case burden at the zip-code tabulation area level in Illinois, with higher percent poverty associated with a higher case burden (J. T. Chen & Krieger, 2021). Similarly, Karmakar *et al.* showed in a cross-sectional analysis that for counties in the U.S., incident cases were associated with higher social vulnerability index (Karmakar, Lantz, & Tipirneni, 2021). This social vulnerability index is defined by the CDC, and includes information from a collection of census variables related to poverty, unemployment, and racial and ethnic minority status. Similar issues may arise when studying the effect of socioeconomic variables. Counties with higher social vulnerability (due to, for example, low economic resources) may also have lower testing resources, which may bias our comparisons to counties where testing is more adequate.

We also use cases to study the effect of vaccination at the population scale. Work in this area has been expansive. Harris showed an inverse relationship between cross-sectional COVID-19 incidence and county-level vaccination coverage during the Delta surge considering a sample of the counties with the largest population size (Harris, 2022), and Cuadros *et al.* found a similar trend in counties across the United States (Cuadros et al., 2022). Nevertheless, as the virus has evolved, the relationship between transmission and case counts has shifted, particularly with the evolution of the highly transmissible Omicron variant. Mclaughlin *et al.* found that there wasn't a relationship between the percentage of the population fully vaccinated and case counts, contrasting findings from other waves (McLaughlin, Wiemken, Khan, & Jodar, 2022). However, they did find that higher booster uptake rates were associated with meaningful decreases in case counts, and higher vaccination rates and booster rates were both associated with decreases in COVID-19 mortality.

Beyond the efficacy of vaccines at the individual level, these studies also demonstrate that we can use case data to quantify the impact of vaccination efforts as a public health intervention. Coupled with information about genetic variants that are circulating, they also can extend our knowledge about the effect of this intervention across different phases of the pandemic.

Looking to the future, infection counts also may be informative as we better understand the impacts of long COVID-19[1] on a population scale. There is increased concern over the poorly characterized but widespread phenomenon of lingering COVID-19 symptoms, which includes but is not limited to symptoms of fatigue, dyspnea, chest pain, and palpitation. The heterogeneity of presentations and definitions has complicated research on the syndrome, yet its impact has been pervasive. In light of this, the NIH has made the initiative XXX to better understand and treat long COVID-19 [SOURCE.]

[1]The syndrome goes by a number of names, including long-haul COVID-19, post-acute XXX, [SOURCE].

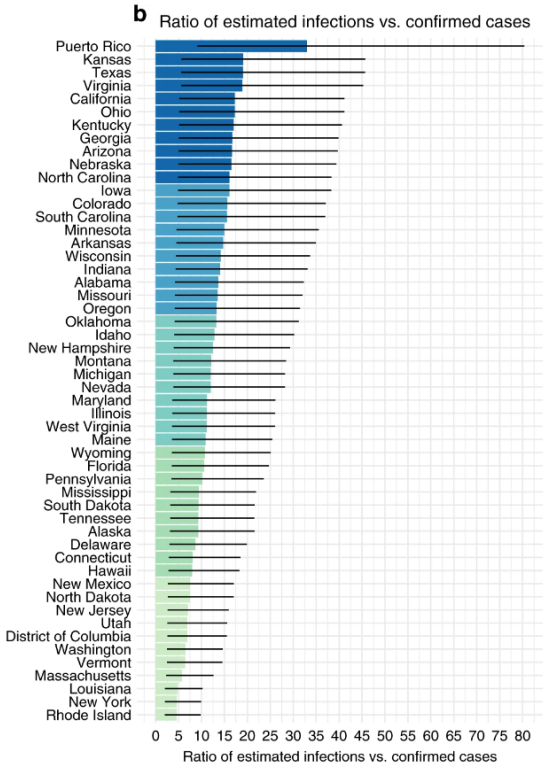
Infection counts are particularly relevant for the study of long COVID-19 at the population scale because, contrary to what we might expect, the severity of COVID-19 disease is not associated with the persistence of several symptoms, including anosmia, chest pain, cough, and palpitation (Dirican & Bal, 2022). Since lingering symptoms can be problematic even with mild cases, trying to characterize the cumulative burden of COVID-19 through a proxy such as hospitalization counts would not capture the full impact.

Ultimately, COVID-19 case counts are a key metric that informs our understanding of the pandemic. Case numbers are interesting in themselves to quantify the reach of the pandemic across different time periods, and they are also the inputs to an extensive array of analyses that aid our understanding of public health interventions, disparities in the impact of the virus, and differences in the dynamics among circulating genetic variants. This underlies the importance of quantifying the underestimation of COVID-19 infections and how the extent of underestimation differs across time and space.

Chapter 2

Overview of Approach

This work is based on the paper *Substantial underestimation of SARS-CoV-2 infection in the United States* by Wu *et al.* (Wu et al., 2020). The original implementation considered a single time interval early in the pandemic, with the objective to estimate the true number of cases as of April 18, 2020 at the state level. When we consider the estimates, we can look at both the estimates for total infections by state, but also the ratio of the estimated total cases to the observed cases. This enables us to think about the way case ascertainment varies by state, as we see below.



The core idea of the approach is to break up the unobserved infections into unobserved infections among those with no or mild symptoms or those with moderate to severe symptoms. We denote this symptom status by an indicator variable where S_1 represents having moderate to severe symptoms and S_0 represents hav-

ing no or mild symptoms. In what follows, $test+$ denotes the event that an individual *would* test positive if they were tested, not that they actually did. For example, $P(test+|S_1, untested)$ represents the probability a symptomatic individual would test positive if they were tested.

Then, our goal is to estimate the infections among the untested population by calculating the number of moderate to severe esymptomatic infections among the untested population as

$$N^+_{untested,S_1} = N_{untested}\ P(S_1|untested) \cdot P(test+|S_1, untested)$$

and the asymptomatic (or mild) infections among the untested population as

$$N^+_{untested,S_0} = N_{untested}(1 - P(S_1|untested))P(test+|S_0, untested).$$

Then we can estimate the total infections among the untested population as

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

which allows us to obtain the estimated number of true infections as

$$N^+ = N^+_{untested} + N^+_{tested}$$

where N^+_{tested} is the number of positive tests in a given location.

The uncertainty inherent in this estimation process is in the quantities $P(S_1|untested)$, $P(test+|S_1, untested)$, and $P(test+|S_0, untested)$.

It is particularly difficult to think about how we would estimate $P(test+|S_0, untested)$ or $P(test+|S_1, untested)$ directly because there is a lack of data on these quantities.

Instead, we define a random variable α that represents the ratio $\dfrac{P(test+|S_1, untested)}{P(test+|tested)}$, that is, $P(test+|S_1, untested) = \alpha P(test+|tested)$. We can think of α as the correction factor for estimating $P(+|S_1, untested)$ from the test positivity $P(test+|tested)$.

We can define β analogously for the asymptomatic case, where $\beta = \dfrac{P(test+|S_0, untested)}{P(test+|tested)}$, so we have $P(test+|S_0, untested) = \beta\ P(test+|tested)$.

This formulation enables us to estimate $P(test+|S_0, untested)$ and $P(test+|S_1, untested)$ with information from the observed test positivity rate among the tested population, which means it can reflect differences in transmission dynamics by the location and time interval considered.

We expect α to be higher than β to reflect that the test positivity rate among the asymptomatic untested population is lower than the symptomatic untested population. The specification of these distributions is discussed in greater detail in the [Definition of Prior Distributions for the Bias Parameters] section.

Because of the uncertainty around α and β, it is useful to relate these parameters to the asymptomatic rate of the virus, $P(S_0|test+, untested)$. Due to the importance of asymptomatic transmission to controlling the pandemic, the asymptomatic rate has been an area of substantial interest. This has led to extensive

studies on the topic, including multiple meta-analyses summarizing these results (Ma et al., 2021a; Sah et al., 2021a).

We can represent the relationship between $\theta = \{\alpha, \beta, P(S_1|untested)\}$ and $\phi = \{ P(S_0|test+, untested) \}$ by the deterministic function $M : \theta \rightarrow \phi$ for $\theta = \{P(S_1|untested), \alpha, \beta\}$ and $\phi = P(S_0|test+, untested)$ defined as:

$$P(S_0|test+, untested) = \frac{\beta(1 - P(S_1|untested))}{\beta(1 - P(S_1|untested)) + \alpha P(S_1|untested)}.$$

When we have prior knowledge about the distributions of the inputs and output of a deterministic function, we can use [Bayesian melding] to generate constrained distributions for the inputs and outputs that are in concordance with one another. In essence, this approach considers the distinct distributions we have for ϕ: the distribution informed by previous literature on the asymptomatic rate, and the distribution formed by evaluating M at values of θ. We can combine these distributions with logarithmic pooling to yield a constrained distribution for $\phi = P(S_0|test+, untested)$, and then can approximate the inverted distribution to obtain constrained distributions for the inputs $\theta = \{P(S_1|untested), \alpha, \beta\}$. These

We can summarize this process in the diagram that follows, where we repeat this process for every geographic unit (a state or county) and time interval (a 2 week interval). We divide the time period into 2-week intervals specifically due to the duration of test positivity, which is about two weeks on average (Kojima, Roshani, & Klausner, 2022; Mallett et al., 2020). This enables us to think of our estimates for each two-week period as incident infections.

Sample from $P(S_1|\text{untested}), \alpha, \beta$

$\downarrow$

Constrain priors with Bayesian melding to obtain constrained distributions for $P(S_1|\text{untested}), \alpha, \beta,$ and $P(S_0|test+, untested)$

$\downarrow$

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

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

$\downarrow$

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

$\downarrow$

Estimate unobserved positives as $N^+_{\text{untested}} = N^+_{\text{untested},S_0} + N^+_{\text{untested},S_1}$

$\downarrow$

Estimate total positives as $N^+ = N^+_{\text{untested}} + N^+_{\text{tested}}$

With the original implementation, $\alpha, \beta,$ and $P(S_1|untested)$ were assumed to be independent and identically distributed across states. However, because we are

considering a wider time interval over all of 2021 and into early 2022, it makes sense to vary these parameters by time and location. Due to the availability of data to inform β and $P(S_1|untested)$, we allow these parameters to vary by time and location, as discussed further in [Definition of Prior Distributions].

When we allow β and $P(S_1|untested)$ to vary over time and location, rather than implementing Bayesian melding once for each time interval, we must implement melding for each time interval and each location considered.

Chapter 3

Background

3.1 Probabalistic Bias Analysis

Often the focus of quantifying error about an effect estimate focuses on random error rather than the systematic error. For example, typical frequentist confidence intervals are frequent in medical and epidemiological literature, although they have faced rising criticism (Greenland et al., 2016). These confidence intervals quantify the fraction of the times we expect the true value to fall in this interval under the assumption that our model is correct. That is, if we ran an experiment 100 times and computed the effect size each time, we would expect the true value to fall within our 95% confidence interval in 95 of those times, on average. Neyman stressed this in his original publication formalizing the concept of a confidence interval in 1937 (Neyman, 1937). The nuance that the confidence interval is not the probability that the true value falls within this interval, however, is often lost in the discussion of results, in part because the true meaning of a confidence interval is less intuitive.

The aim of quantitative bias analysis is to estimate systematic error to give a range of possible values for the true quantity of interest. In this sense, it is a type of sensitivity analysis. It can be used to estimate various kinds of biases, from misclassification, as is implemented in this work, as well as selection bias and unmeasured confounding (Petersen, Ranker, Barnard-Mayers, MacLehose, & Fox, 2021). Often, the goal of performing such an analysis is to see how these sources of bias affect our estimates; in particular, under what situations of bias the observed effect would be null.

There are multiple different forms of bias analysis (Lash, Fox, & Fink, 2009). The most simple case, simple bias analysis, is correcting a point estimate for a single source of error. Multidimensional bias analysis extends this to consider sets of bias parameters, but still provides a corrected point estimate rather than a range of plausible estimates. Probabilistic bias analysis, meanwhile, defines probability distributions for bias parameters to generate a distribution of corrected estimates by repeatedly correcting estimates for bias under different combinations of the parameter values. Then, via Monte Carlo we obtain a distribution of corrected estimates that reflect the corrected values under different scenarios of bias, that is, under dif-

ferent combinations of the bias parameters. This can give us a better idea for the extent of uncertainty about the corrected estimates, although this uncertainty does depend on the specification of the bias parameter distributions. Inherent in bias analysis is the dependence of our results on the specification of bias parameters, which reflect what is known from available data, literature, or theory on the extent of bias that may occur. There is uncertainty about how we define these distributions or values; otherwise, if the precise values of the bias parameters were known, we could simply correct the estimates and probabilistic bias analysis would not be useful.

Although some forms of probabilistic bias analysis can be applied to summarized data, for example, frequencies in a contingency table, the methods are most often implemented with unsummarized data in its original form, as implemented here.

In choosing specific distributions for the bias parameters, different specifications may yield density functions where most of the density is within a similar interval (MAKE PLOT WITH EXAMPLE), which means the choice of the specific distribution will not be sensitive to the particular choice of density.

3.2 Bayesian Melding

3.2.1 Theoretical Background for the Approach

The Bayesian melding approach was proposed by Poole et al. (Poole & Raftery, 2000).

The Bayesian melding approach enables us to account for both uncertainty from inputs and outputs of a deterministic model. The initial motivation for the approach was to study the population dynamics of whales in the presence of substantial uncertainty around model inputs for population growth (Poole & Raftery, 2000). However, the framework provided by Poole et al. can applied in any circumstance where we have uncertainty around some quantities θ and ϕ where there is a deterministic function $M : \theta \to \phi$. Due the utility of Bayesian melding in various contexts, since this deterministic model M could take on a wide range of forms, the approach has since been applied in various fields, including urban simulations (Ševčíková, Raftery, & Waddell, 2007), ecology (Robson, 2014), and infectious disease (Powers et al., 2011).

At this point, we can define how Bayesian melding works more formally.

Let $M : \theta \to \phi$ be the deterministic model defined by the function relating a vector of input parameters θ to an output vector ϕ, and suppose we have a prior on θ denoted $q_1(\theta)$ and a prior on ϕ denoted $q_2(\phi)$.

However, note that we actually have two distinct priors on ϕ. There is the prior formed by the distribution induced on ϕ by the prior for θ and the function M, where we denote this induced prior $q_1^*(\phi)$. If M^{-1} exists, we can write this induced prior $q_1^*(\phi) = q_1(M^{-1}(\phi))|J(\phi)|$. This result follows from the fact $M(\theta) = \phi$, so we apply a change of variables to obtain the distribution of ϕ from the distribution of

θ. This is a generalization to the multivariate case of the change of variables result often covered in probability courses in the univariate case. That is, if we have a continuous random variable X with probability density function f_X and $Y = g(X)$ for a differentiable monotonic function, then the probability density function of Y is $f_Y(y) = f_X(g^{-1}(y))\left|\frac{dx}{dy}\right|$. In practice, M^{-1} rarely exists exists since θ is often of higher dimensionality then ϕ, in which cases M is not invertible. This means we generally approximate $q_1^*(\phi)$ without acquiring its analytical form.

In addition to this induced prior, we have the prior $q_2(\phi)$, which does not involve M nor the inputs θ. Since these priors are based on different sources of information and may reflect different uncertainties, often it useful to use both sources of information to inform our estimates. To do so, we need to combine the distributions for $q_1^*(\phi)$ and $q_2(\phi)$ to create a pooled distribution.

Multiple pooling strategies exist for distinct distributions, but one requirement for a Bayesian analysis is that the distribution should be independent of the order in which the prior is updated and the combining of the prior distributions. That is, updating the prior distributions using Bayes' theorem and then combining distributions should yield the same result as combining distributions and then updating this combined distribution; pooling methods that have this property are deemed externally Bayesian. Logarithmic pooling has been shown to be externally Bayesian under some conditions, which are likely to hold in most settings. Furthermore, logarithmic pooling has actually been shown to be the only pooling method where this holds (Genest, McConway, & Schervish, 1986).

The logarithmically pooled prior for ϕ by pooling $q_1^*(\phi)$ and $q_2(\phi)$ is proportional to

$$q_1^*(\phi)^\alpha q_2(\phi)^{1-\alpha}$$

where $\alpha \in [0, 1]$ is a pooling weight. Commonly, a choice of $\alpha = 0.5$ is used to give the priors equal weight. In this case, logarithmic pooling may be referred to as geometric pooling since it is equivalent to taking a geometric mean.

If M is invertible, we can obtain the contrained distributions for the model inputs by simply inverting. However, this is rare, so we have to think about how to proceed in the noninvertible case.

To get intuition for a valid strategy, consider a mapping $M : \theta \to \phi$ for $\theta \in \mathbb{R}$ and $\phi \in \mathbb{R}$ defined as follows. Note the choice of q_1, q_2 does not matter here as long as they are valid densities.

θ	$q_1(\theta)$	ϕ	$q_2(\phi)$
1	0.3	1	0.4
2	0.2	2	0.6
3	0.5	2	0.6

We see that M is not invertible since $\theta = 1$ and $\theta = 2$ both map to $\phi = 2$, which implies the inverse M^{-1} would not be well defined.

We can compute $q_1^*(\phi)$ using our function M and taking $q_1^*(\phi) = q_1(M^{-1}(\phi))$; in the continuous case we need to multiply by $|J(\phi)|$, but not in the discrete case (Blitzstein & Hwang, 2019).

So we have $q_1^*(1) = q_1(1) = 0.3$ since $M(1) = 1$, and $q_1^*(2) = q_1(2) + q_1(3) = 0.2 + 0.5 = 0.7$ since $M(2) = 2$ and $M(3) = 2$.

Then, we can compute the logarithmically pooled pooled prior with $\alpha = 0.5$ by taking $q_1^*(\phi)^\alpha q_2(\phi)^{1-\alpha}$.

For $\phi = 1$, we have $q_1^*(\phi)^\alpha q_2(\phi)^{1-\alpha} = (0.3)^{0.5}(0.4)^{0.5} = 0.3464$ For $\phi = 2$, we have $q_1^*(\phi)^\alpha q_2(\phi)^{1-\alpha} = (0.6)^{0.5}(0.7)^{0.5} = 0.6481$

To make this a valid density, however, these probabilities must sum to 1, so we renormalize by dividing by (0.3464 + 0.6481). This gives us the pooled prior $q^{\sim[\phi]}(\phi)$ as

$0.3464/(0.3464+0.6481) = 0.3483$ for $\phi = 1$ and $0.6481/(0.3464+0.6481) = 0.6517$ for $\phi = 2$.

Summarizing these results, we have

ϕ	$q_2(\phi)$	$q_1^*(\phi)$	$q^{\sim[\phi]}(\phi)$
1	0.4	0.3	0.3483
2	0.6	0.7	0.6517

However, we want the pooled prior on the inputs θ, that is, $q^{\sim[\theta]}(\theta)$.

Poole et al. reasoned as follows. Since M uniquely maps $\theta = 1$ to $\phi = 1$, the probability that $\theta = 1$ should be equal to the probability $\phi = 1$. That is, we should have $q^{\sim[\theta]}(1) = q^{\sim[\phi]}(1)$.

However, the relationship for $\theta = 2$ or $\theta = 3$ to ϕ is not one to one, but since $M(2) = 2$ and $M(3) = 2$, the sum of the probabilities for $\theta = 1$ and $\theta = 2$ should be equal to that for $\phi = 2$, that is, $q^{\sim[\theta]}(2) + q^{\sim[\theta]}(3) = q^{\sim[\phi]}(2) = 0.6517$.

The challenge here is how we divide the probability for $q^{\sim[\phi]}(2)$, which is defined, among $q^{\sim[\theta]}(2)$ and $q^{\sim[\theta]}(3)$. The prior for ϕ yields no information to assist in this choice, because knowing which value ϕ takes on does not give us any information about whether $\theta = 2$ or $\theta = 3$. Thus, the information we have about θ must be taken from $q_1(\theta)$.

That is, we can assign a probability for $q^{\sim[\theta]}(2)$ by considering the probability that $\theta = 2$ relative to the probability $\theta = 3$, computing

$$q^{\sim[\theta]}(2) = q^{\sim[\phi]}(2)\Big(\frac{q_1(2)}{q_1(2) + q_1(3)}\Big).$$

That is, if the probability θ takes on the value 2 is lower in this case than the probability $\theta = 3$ which we know from the prior on θ, $q_1(\theta)$, then the pooled prior on θ, $q^{\sim[\theta]}(2)$, should reflect this.

Using this reasoning, we have $q^{\sim[\theta]}(2) = (0.7)\frac{0.2}{0.2+0.5} = 0.1862$ and $q^{\sim[\theta]}(3) = (0.7)\frac{0.5}{0.2+0.5} = 0.4655$.

The result in this simple example, using $q_1(\theta)$ to determine how to distribute the probability for values of ϕ where multiple θ map to ϕ, can be used to derive general

formulas to compute $q^{\sim[\theta]}(\theta)$ for discrete and continuous distributions (Poole & Raftery, 2000).

3.2.2 Implementation through the Sampling-Importance-Resampling Algorithm

The steps are:

1. We draw θ from its prior distribution $q_1(\theta)$, where we note θ can be multidimensional.
2. For every θ_i we compute $\phi_i = M(\theta_i)$.
3. Since $q_1^*(\phi)$ is unlikely to have an analytical form, we can compute it via a density approximation by computing $M(\theta)$ for our sampled values of θ and then estimating the density from this sample using kernel density estimation.
4. Construct weights proportional to the ratio of the prior on ϕ evaluated at $M(\theta_i)$ to the induced prior q_1^* evaluated at $M(\theta_i)$. Note that this is applying the same logic as considering $q_1(\theta)/q_1^*(\theta)$, as discussed in the previous concrete example, but representing these probabilities in ϕ space.
5. Sample values from step (1) with probabilities proportional to the weights from (4).

3.2.3 Bayesian Melding Applied to COVID-19 Misclassification

In this work, we can relate the inputs $\theta = \{P(S_1|untested), \alpha, \beta\}$ and $\phi = P(S_0|test+, untested)$ by the deterministic model $M : \theta \rightarrow \phi$ given by

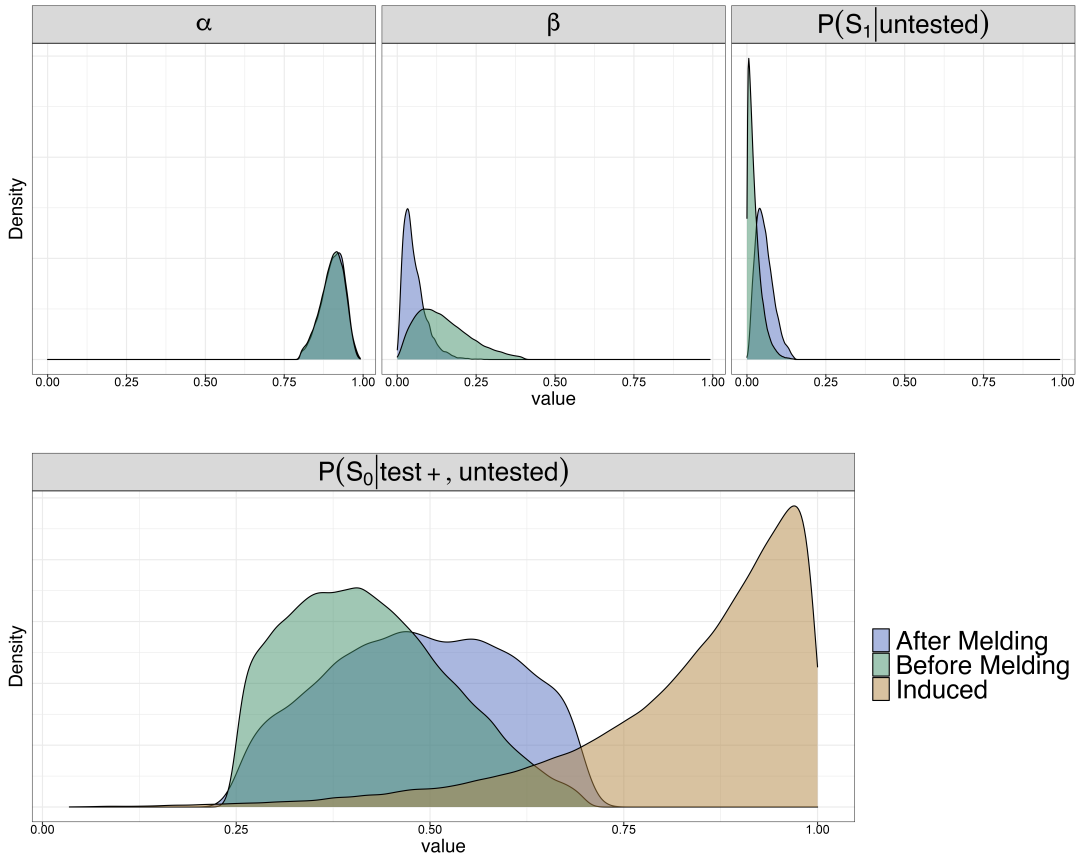
$$P(S_0|test+, untested) = \frac{\beta(1 - P(S_1|untested))}{\beta(1 - P(S_1|untested)) + \alpha P(S_1|untested)}.$$

The derivation of M is in the following section.

Now, we have two distributions on ϕ: the distribution based on data on the asymptomatic rate of infection of COVID-19, and the distribution formed by taking $M(\theta)$ where θ represents the values from the defined distributions of α, β, and $P(S_1|untested)$. With Bayesian melding, we pool these distributions using logarithmic pooling, and then implement the sampling-importance-resampling algorithm to obtain constrained distributions of the inputs θ that are in accordance with information about the asymptomatic rate of the virus.

Due to the uncertainty around our definitions of α and β, it is particularly useful to leverage the information we have about the asymptomatic rate of the virus $P(S_0|test+, untested)$ because a large collection of studies has been published in this area. In a meta-analysis pooling data from 95 studies, the pooled estimate among the confirmed population that was asymptomatic was 40.50% [95% CI, 33.50%-47.50%] (Ma et al., 2021b). Another meta-analysis including 350 studies estimated the asymptomatic percentage as 36.9% [95% CI: 31.8 to 42.4%] (Sah et al., 2021b).

To clarify the use of this method, we can look at the distributions before and after applying Bayesian melding.



Comparing these priors above, we see that although they have shared support, some values from the induced distribution we acquire by using M to generate values of ϕ from sampled values of θ are very unlikely to be in accordance with the information we know about SARS-CoV-2 asymptomatic infection. This is where Bayesian melding comes into play. Pooling these distributions enable us to take both the prior on $q_2(\phi)$ from published analyses on asymptomatic infection, and the induced prior, $q_1^*(\phi)$, into account to constrain the distributions of ϕ and θ to be in accordance.

3.2.4 Derivation of M

Define $M : \theta \to \phi$ for $\theta = (P(S_1|untested), \alpha, \beta)$ and $\phi = P(S_0|test+)$ as

$$P(S_0|test+, untested) = \frac{\beta(1 - P(S_1|untested))}{\beta(1 - P(S_1|untested)) + \alpha P(S_1|untested)}.$$

We define $test+$ to denote the event that an individual *would* test positive if they were tested, not that they actually did test positive.

Since we have $\alpha = \frac{P(test+|S_1,untested)}{P(test+|tested)}$ and $\beta = \frac{P(test+|S_0,untested)}{P(test+|tested)}$, we can write

$$= \frac{\frac{P(test+|S_0,untested)}{P(test+|tested)}(1 - P(S_1|untested))}{\frac{P(test+|S_0,untested)}{P(test+|tested)}(1 - P(S_1|untested)) + \frac{P(test+|S_1,untested)}{P(test+|tested)}P(S_1|untested)}$$

and cancelling out the term $P(test+|tested)$ we have

$$= \frac{P(test+|S_0, untested)(1 - P(S_1|untested))}{P(test+|S_0, untested)(1 - P(S_1|untested)) + P(test+|S_1, untested)P(S_1|untested)}.$$

Notice that $P(S_0|untested) = 1 - P(S_1|untested)$, so we have

$$P(S_0|test+) = \frac{P(test+|S_0, untested)P(S_0|untested)}{P(test+|S_0, untested)P(S_0|untested) + P(test+|S_1, untested)P(S_1|untested)}.$$

We can write the numerator

$$P(test+|S_0, untested)P(S_0|untested) = \Big(\frac{P(test+, S_0, untested)}{P(S_0, untested)}\Big)\Big(\frac{P(S_0, untested)}{P(untested)}\Big)$$

$$= \frac{P(test+, S_0, untested)}{P(untested)} = P(S_0, test+|untested).$$

With the same reasoning, we obtain $P(test+|S_1, untested)P(S_1|untested) = P(S_1, test+|untested)$.

Thus, using this result, we have

$$= \frac{P(S_0, test+|untested)}{P(S_0, test+|untested) + P(S_1, test+|untested)}.$$

We can write

$$P(S_0, test+|untested) = \frac{P(S_0, test+, untested)}{P(untested)} = \frac{P(test+, untested|S_0)P(S_0)}{P(untested)}$$

and similarly $P(S_1, test+|untested) = \frac{P(test+,untested|S_1)P(S_1)}{P(untested)}$.

Thus, we can write the denominator as

$$P(S_0, test+|untested) + P(S_2, test+|untested)$$

$$= \frac{P(test+, untested|S_1)P(S_1) + P(test+, untested|S_0)P(S_0)}{P(untested)}$$

$$= \frac{P(test+, untested)}{P(untested)} = P(test+|untested).$$

Putting these results together, we have

$$= \frac{P(S_0, test+|untested)}{P(test+|untested)} = \frac{\frac{P(S_0,test+,untested)}{P(untested)}}{\frac{P(test+,untested)}{P(untested)}} = \frac{P(S_0, test+, untested)}{P(test+, untested)} = P(S_0|test+, untested).$$

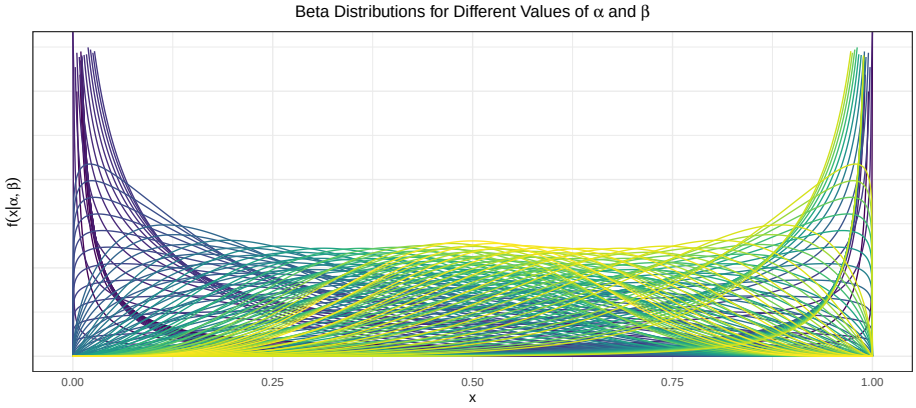
Chapter 4

Definition of Prior Distributions for Bias Parameters

4.1 Background on the Beta Distribution

The priors we are specifying here are for probabilities, with the exception of α and β, which represent ratios of probabilities. To define a prior for a parameter that takes on values in $[0, 1]$, a particularly useful distribution is the beta distribution, which is only defined on the interval $[0, 1]$. It is parameterized by two positive values α, β[1]. In this section, and only in this section, we refer to α and β as the parameters of the beta distribution, not the random variables that are inputs to the probabilistic bias analysis.

The β distribution is an extremely flexible distribution[2]; by altering the parameters α, β, we can get an extensive array of shapes, as seen below.



In defining a beta distribution to reflect knowledge about a parameter, we need to work backwards to find the parameters α and β that correspond to the desired mean and variance.

[1] In R, α = shape1 and β = shape2.

[2] The β distribution is also useful in Bayesian statistics, because it is the conjugate prior distribution for the binomial distribution and negative binomial distributions. That is, if we have a binomial likelihood with parameter p and p is distributed according to the β distribution, the resulting posterior follows a β distribution.

There are multiple parameterizations of the beta distribution, but R uses that where we define the probability density function as

$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}.$$

or equivalently as

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1},$$

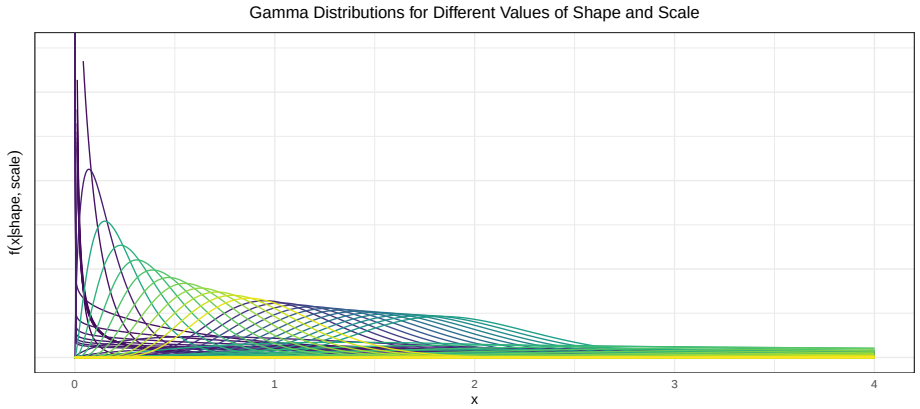
where the beta function $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

The expected value of the beta distribution is then $E(X) = \dfrac{\alpha}{\alpha+\beta}$ and the variance is given by $V(X) = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$; the derivation for both is given in the appendix. We can then solve for α and β to obtain $\alpha = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right)\mu^2$ and $\beta = \alpha\left(\frac{1}{\mu} - 1\right)$. At this point, we can easily write a function in R that generates the parameters of the beta distribution with the desired mean and variance.

4.2 Background on the Gamma Distribution

The gamma distribution is another very flexible distribution. However, the support of the gamma distribution is $(0,\infty)$ rather than $[0,1]$. Because some of the bias parameters are not probabilities (α and β are ratios of probabilities), we can instead use the gamma distribution to allow the random variable to take on values over 1.

As with the beta distribution, a variety of shapes are possible with the gamma distribution.



Let $k = shape$ and $\theta = scale$. The parameterization of the gamma distribution that R uses is

$$f(x|k,\theta) = \frac{1}{\Gamma(k)\ \theta^k}x^{k-1}e^{x/\theta}$$

where the mean $\mu = k\theta$ and the variance $\sigma^2 = k\theta^2$.

This allows us to obtain $\frac{\mu^2}{\sigma^2} = \frac{k^2\theta^2}{k\theta^2} = k = shape.$
Then, substituting this result in for k, we have

$$\sigma^2 = k\theta^2 = \frac{\mu^2}{\sigma^2}\theta^2$$

$$\frac{\sigma^4}{\mu^2} = \theta^2$$

$$\frac{\sigma^2}{\mu} = \theta = scale.$$

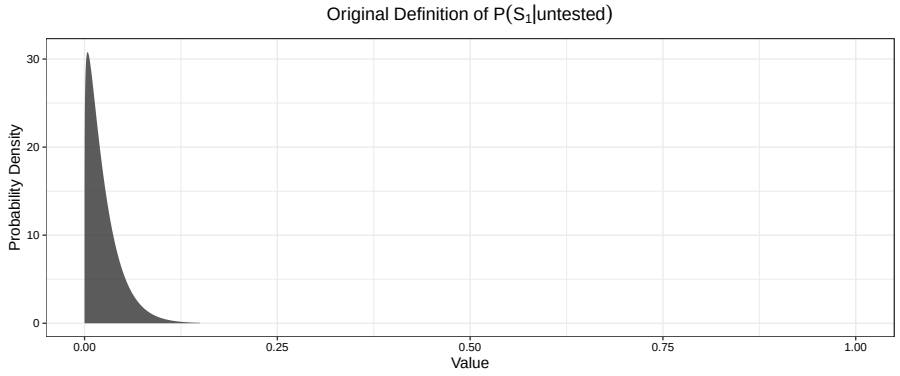
This allows us to calculate the shape and scale parameters of a gamma distribution with the desired mean and variance.

4.3 Definition of Prior Distributions for Incomplete Testing Correction

4.3.1 Defining $P(S_1|Untested)$

We recall that S_1 denotes the event that an individual has moderate to severe symptoms, so $P(S_1|Untested)$ is the probability of having moderate to severe symptoms among those who were not tested. We note that this would include people that have moderate to severe COVID-like symptoms that do indeed have COVID-19 as well as people that do not have COVID-19 and have some other respiratory illness.

The original distribution was defined such that $P(S_1|Untested) \sim TBeta(\alpha = 1.18, \beta = 45.97)$, bounded between 0 and 15%, as we see below.

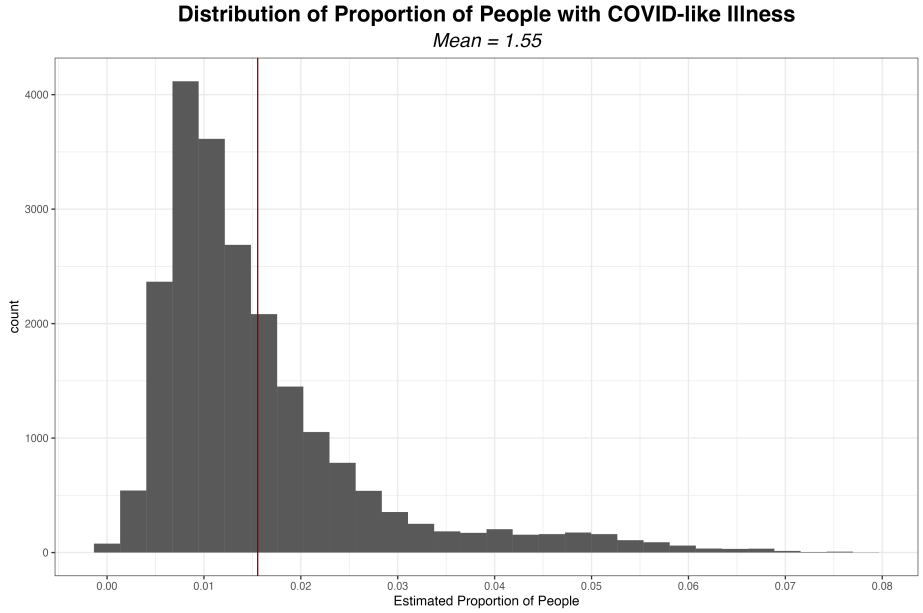


However, to implement this approach over a more extended time interval, we need to allow this parameter to vary by time. Due to state-specific differences in symptom prevalence, it also makes more sense to allow this parameter to vary by state.

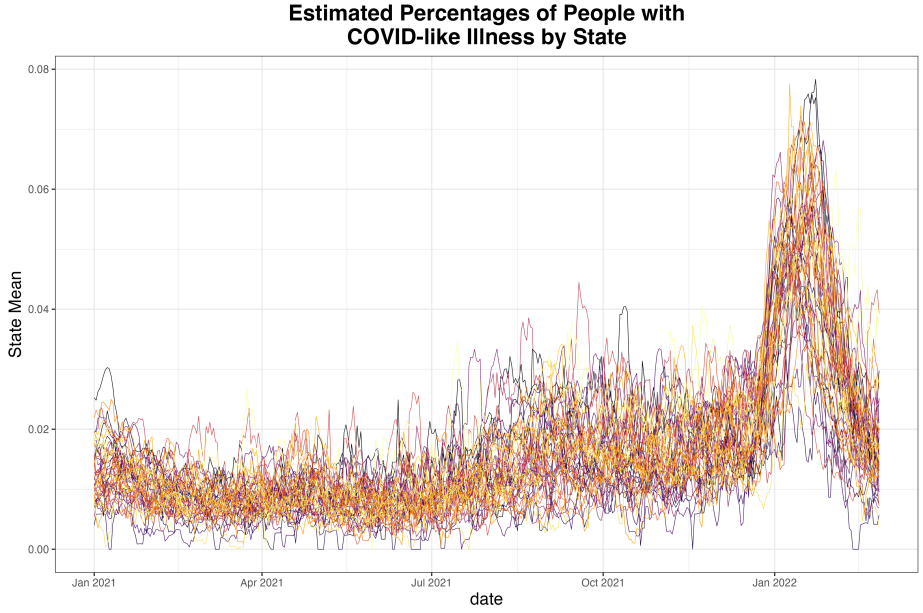
To do this, we can use the COVID-like illness indicator from the COVID-19 Trends and Impact Survey (Salomon et al., 2021). The COVID-19 Trends and Impact Survey (CTIS) is a large scale internet-based survey that invites a sample of

Facebook users to respond to questions on several topics of public health interest, including testing and symptom status. The survey effort selects participants using stratified random sampling by state, and responses are aggregated and made publicly available.

Below, we see that the distribution of the proportion of the population with COVID-like illness over all of 2021 is in a similar range as the original definition of $P(S_1|untested)$, with the bulk of the distribution between 0 and 15%.



We also see that although the general trend is similar between states, there is variability in the proportion experiencing COVID-19-like illness by state.

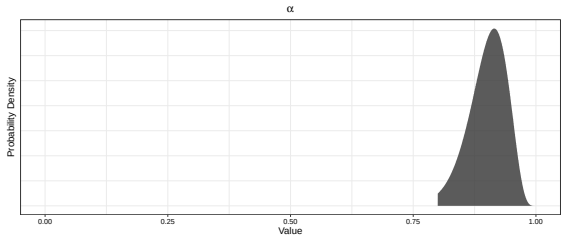


4.3.2 Defining α

α is defined as the ratio of $\frac{P(test+|S_1, untested)}{P(test+|tested)}$, applied to allow $P(test+|S_1, untested)$ to vary by state. $P(test+|tested)$ is the state-level empirical estimate, but α itself is not calculated using the state-level empirical estimate. Instead, $P(test+|S_1, untested)$ is calculated as $P(test+|S_1, untested) = \alpha P(test+|tested)$. So we can think about α as the adjustment to the test positivity rate as we estimate the probability of testing positive among symptomatic untested individuals. This is assumed to be high, that is, that the probability of testing positive among **symptomatic untested** individuals would be near 90% of the probability of testing positive among **tested individuals** (not all of whom would be symptomatic).

$\alpha \sim TBeta(\alpha = 49.73, \beta = 5.53)$, bounded between 80% to 100%, with the mean at $\frac{\alpha}{\alpha+\beta} = 0.90$.

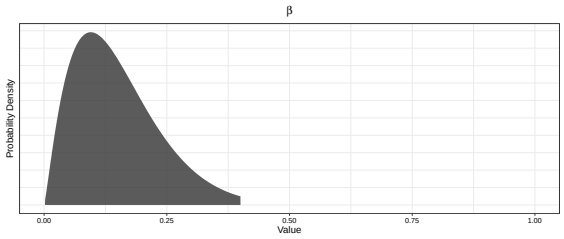
Due the the expansion of testing resources, it is plausible that $P(test+|untested, S_1)$ could exceed $P(test+|tested)$, so we will extend this distribution to be larger than one.



4.3.3 Defining β

Similar to the way we defined α, β is defined as the ratio of $\frac{P(test+|S_0, untested)}{P(test+|tested)}$, applied to allow $P(test+|S_0, untested)$ to vary by state. We use β to calculate $P(test+|S_1, untested)$ by the expression $P(test+|S_0, untested) = \beta P(test+|tested)$. We can think about β as the adjustment to the test positivity rate as we estimate the probability of testing positive among **aymptomatic untested** individuals (in contrast to α, which is symptomatic untested individuals). This is assumed to be substantially lower than α, reflecting we expect a much smaller proportion of asymptomatic untested individuals to test positive.

The original definition of β was $\beta \sim TBeta(\alpha = 2.21, \beta = 12.53)$ with the mean at $\frac{2.21}{2.21+12.53} = 0.15$ and bounded between 0.2% to 40%.

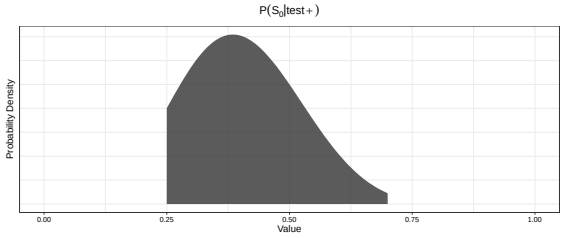


Because β is defined as the ratio of $\frac{P(test+|S_0, untested)}{P(test+|tested)}$, we can estimate β empirically by taking the screening test positivity rate as an estimate of $P(test+|S_0, untested)$ and then dividing by the overall test positivity rate $P(test+|tested)$. State-level estimates for screening test positivity and overall test positivity are available through the COVID-19 Trends and Impact Survey, enabling us to obtain a time and state-specific estimate of β.

4.3.4 Defining $P(S_0|test+, untested)$

$P(S_0|test+)$ is the probability of not having symptoms among those who test positive, that is, the percentage of asymptomatic infection among those with confirmed COVID-19. It is defined such that $P(S_0|Test+) \sim TBeta(\alpha = 6.00, \beta = 9.00)$, bounded between 25% and 70% with the mean at $\frac{\alpha}{\alpha+\beta} = 0.40$.

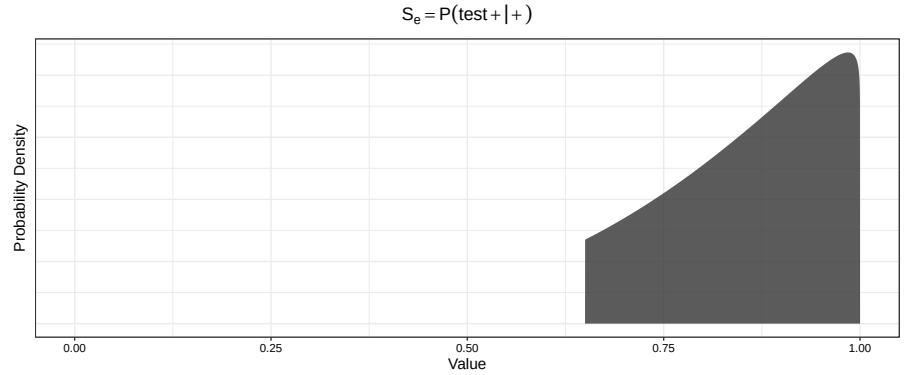
One large meta-analysis found $P(S_0|test+)$ to be 40.50% (95% CI: 33.50%-47.50%), although it did not restrict to screening studies (Ma et al., 2021a). Another meta-analysis, when restricting to screening studies, found $P(S_0|test+)$ to be 47.3% (95% CI: 34.0% -61.0%) (Sah et al., 2021a).



4.4 Definition of Priors for Test Inaccuracy Correction

4.4.1 Defining Test Sensitivity (S_e)

The test sensitivity $P(test+|+)$ is defined as $P(test+|+) \sim TBeta(\alpha = 4.20, \beta = 1.05)$, bounded between 0.65 and 1 and with mean $\frac{\alpha}{\alpha+\beta} = 0.80$.



Data available for informing this prior distribution:

In a population-based retrospective study including both inpatients and outpatients, the clinical sensitivity was estimated to be 89.9% (95% CI 88.2 – 92.1%) by considering repeat-tested patients who initially tested negative but later tested positive (Kortela et al., 2021). However, as Kortela *et al.* discussed, this approach is likely an overestimate of the true clinical sensitivity, because individuals will only be tested twice if there is high clinical suspicion that they do have COVID-19. To account for this, they produced an estimate of sensitivity including cases with high clinical suspicion in the denominator, which resulted in an estimate closer to 50%, yet this is likely an underestimate due to the fact that even those with a classical COVID-19 symptom presentation may not have COVID-19. They concluded that due to these biases, the true value most likely falls between the overestimate near 90% and the underestimate near 50%.

Another analysis of repeat-tested patients using data from a large sample of patients tested at the provincial Public Health Laboratory in Canada estimated the clinical sensitivity to be 90.7% (95% CI 82.6–98.9%) (Kanji et al., 2021). Green *et al.* found that the clinical sensitivity ranged from 58% to 96%: the estimate of 96% was dependent on the assumption that negative results, repeated or not, were true negatives, while the estimate of 58% assumed the rate of false negatives among the repeat-tested population would be the same as in the repeat-tested patients (Green et al., 2020). In a meta-analysis of 51 studies, Mustafa *et al.* found a pooled estimate of the clinical sensitivity as 0.96 (95% CI 93% - 98%) (Mustafa Hellou et al., 2021).

Because PCR tests are designed to target a highly conserved region of the viral genome, their sensitivity was expected to be relatively robust to the circulation of different variants of SARS-CoV-2. However, analytical sensitivity has shown some differences by genetic variants (Y. Chen et al., 2022). Viral shedding dynamics also have differed by genetic variant, but the variants dominant throughout most of the time period considered here, Delta and Omicron, have similar viral loads (Fall et al., 2022; Singanayagam et al., 2022).

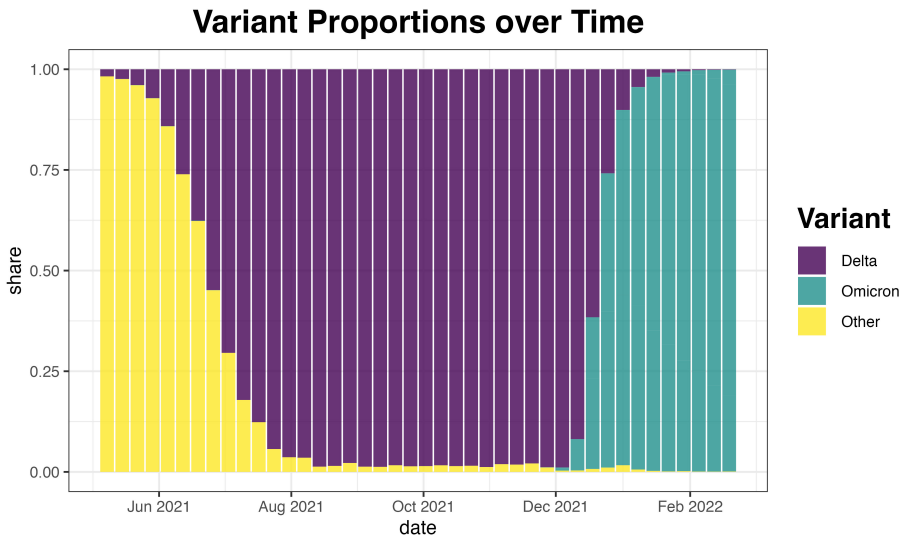
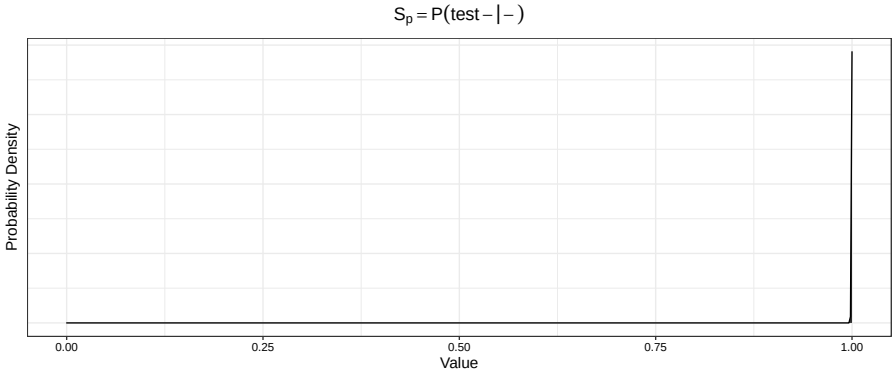


Figure 4.1: Variant proportions in the United States from genomic surveillance data collected by the CDC. Data is not available for time periods earlier than May 8, 2021.

Ultimately, although it is plausible that test sensitivity may vary by time due to differences in viral shedding dynamics over time as well as differences due to the mutations present in circulating variants, there is a lack of data to inform exactly how the sensitivity may vary over time. As a result, we assume the test sensitivity is independent and identically distributed across time periods.

4.5 Defining Test Specificity (S_p)

The test specificity $P(test-|-)$ is defined as $P(test-|-) \sim TBeta(\alpha = 4998.50, \beta = 0.25)$, bounded between 0.9998 and 1 and with mean $\frac{\alpha}{\alpha+\beta} = 0.99995$. The high certainty for this parameter is based on the CDC 2019-nCoV Real-Time RT-PCR Diagnostic Panel.



4.6 Summary Table of Bias Parameter Distributions

To add

4.7 Correction for Incomplete Testing

As discussed previously, once we have sampled values from the constrained distributions of $P(S_1|untested)$, α, β, we estimate the test positivity among the symptomatic untested population as $P(+|S_1, untested) = \alpha\ P(test+|tested)$ and we estimate the test positivity among the asymptomatic untested population as $P(+|S_0, untested) = \beta\ P(test+|tested)$. Then, we compute the positives among the symptomatic and mild/asymptomatic parts of the population respectively as

$$N^+_{untested,S_1} = N_{untested}\ P(S_1|untested) \cdot P(test+|S_1, untested) \quad \text{and}$$

$$N^+_{untested,S_0} = N_{untested}(1 - P(S_1|untested))P(test+|S_0, untested).$$

Then, we take the total positives among the untested population as

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

and finally we add the number of observed positives, N^+_{tested} to obtain the estimate for the positives among the total population, as

$$N^+ = N^+_{untested} + N^+_{tested}.$$

4.8 Correction for Diagnostic Test Inaccuracy

At this point, we have corrected for the incompleteness of testing. That is, we have an estimate of who would have tested positive if we tested the entire population. However, we also need to correct for imperfect test accuracy.

Test accuracy is broken up into two components, specificity and sensitivity.

We define test sensitivity and specificity as follows: * S_e = test sensitivity = the probability an individual tests positive if they have COVID-19 (probability of a true positive), that is, $P(test+|+)$. * S_p = test specificity = probability an individual tests negative if they do not have COVID-19 (probability of a true negative), that is, $P(test-|-)$.

Then, given that we have the number N^+ who tested positive (or, in the context of this work, would have tested positive), the specificity S_p, the sensitivity S_e, and the total population size N, we can calculate the true positives with the formula

$$\text{Number Truly Positive} = \frac{N^+ - (1 - S_p) \times N}{S_e + S_p - 1}$$

from Modern Epidemiology (Rothman, Greenland, & Lash, 2008).

4.8.1 Derivation of Formula for Correction for Diagnostic Test Inaccuracy

We define test sensitivity and specificity as follows: * S_e = test sensitivity = the probability an individual tests positive if they have COVID-19 (probability of a true positive), that is, $P(test+|+)$. * S_p = test specificity = probability an individual tests negative if they do not have COVID-19 (probability of a true negative), that is, $P(test-|-)$.

As defined previously, $S_e \sim TBeta(0.65, 1)$ and $S_p \sim TBeta(0.998, 1)$.

To correct case counts for diagnostic test inaccuracy, we use the formula

$$\text{Number Truly Positive} = \frac{N^+ - (1 - S_p) \times N}{S_e + S_p - 1}$$

as defined in Rothman et al. (2008).

To obtain this formula, we let:

- N denote the total population size
- N^+ denote the number *classified* as positive
- N^- denote the number *classified* as negative
- T^+ denote the number that is *truly* positive
- T^- denote the number that is *truly* positive

We also recall that

$$\text{Sensitivity} = S_e = P(test+|+)$$

$$\text{Specificity} = S_p = P(test-|-)$$

The quantity we want to estimate is the number of truly positive individuals when accounting for imperfect test accuracy, that is, T^+.

The number classified as positive, N^+ can be written as

$$N^+ = P(test+|+)T^+ + P(test+|-)T^-$$

where $P(test+|+)T^+$ is the number of true positives and $P(test+|-)N^-$ is the number of false positives. By the definitions of sensitivity S_e and specificity S_p we can write this more clearly as

$$N^+ = S_eT^+ + (1 - S_p)T^-.$$

Meanwhile, the number classified as negative, N^- can be written as

$$N^- = P(test-|-)T^- + P(test-|+)T^+$$

where $P(test-|-)T^-$ is the number of true negatives and $P(test-|+)T^+$ is the number of false negatives. Substituting in S_e and S_p we can express this as

$$N^- = S_pT^- + (1 - S_e)T^+.$$

At this point, we can solve the expression $N^- = S_pT^- + (1 - S_e)T^+$ for the number of people classified as positive for the number truly negative, T^-. This yields

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

Now, we can substitute this result into our expression for $N^+ = S_eT^+ + (1 - S_p)T^-$ and solve for the desired value, the number of truly positive individuals, T^+. This gives us

$$N^+ = S_eT^+ + (1 - S_p)\left(\frac{(N^- - (1 - S_e)T^+)}{S_p}\right)$$

$$S_pN^+ = S_pS_eT^+ + (1 - S_p)\left((N^- - (1 - S_e)T^+)\right)$$

$$S_pN^+ = S_pS_eT^+ + (1 - S_p)(N^-) - (1 - S_p)(1 - S_e)T^+$$

$$S_pN^+ - (1 - S_p)(N^-) = S_pS_eT^+ - (1 - S_p)(1 - S_e)T^+$$

$$S_pN^+ - (1 - S_p)(N^-) = (S_pS_e - (1 - S_p)(1 - S_e))T^+$$

$$S_pN^+ - (1 - S_p)(N^-) = (S_p + S_e - 1)T^+$$

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

At this point we can simplify the numerator as follows by using the fact $N = N^+ + N^-$. This gives us

$$\begin{aligned}
&= S_pN^+ - (1 - S_p)N^- \\
&= S_pN^+ + S_pN^- - N^- \\
&= S_p(N^+ + N^-) - N^- \\
&= S_pN - (N - N^+) \\
&= S_pN - (N - N^+) \\
&= (S_p - 1)N + N^+ \\
&= N^+ - (1 - S_p)N
\end{aligned}$$

so we have

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

Chapter 5

Comparison to the Covidestim Model

5.0.1 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.

The most 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 (https://academic.oup.com/aje/article/190/9/1908/6217341).

5.0.2 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 $\rightarrow$ 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.

5.0.3 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 (https://www.nature.com/articles/s41586-020-2918-0). 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 (FIND EXAMPLE) or different variants leading to less severe presentations (FIND PAPER ON

OMICRON SEVERITY).

One thing to consider is that infection fatality rate may vary over time, as treatments may vary, as well as the demographics of individuals being infected. For example, during the school year, more students may test positive but will be less likely to die on average than adults (PROVIDE SOURCE FOR THIS). However, these estimates are 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 Comparison to Other Indicators

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 (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html). 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 (https://www.cdc.gov/coronavirus/2019-ncov/covid-data/serology-surveillance/index.html).

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 \footnote{ 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 (https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient). This distinction is important here since we are interested in the strength of the monotonic relationship rather than the linear relationship between these values. }.

5.2 Seropositivity Data

To add

Chapter 6

Results

6.1 County-level

6.2 State-level

Appendix A

Appendix

A.1 Derivation of the Mean and Variance of the Beta Distribution

To add

References

Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability* (Second edition). Boca Raton: CRC Press.

California Department of Public Health. (2021, June 15). Blueprint for a Safer Economy. Retrieved December 16, 2022, from https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/COVID19CountyMonitoringOverview.aspx

Centers for Disease Control and Prevention. (2020, March 28). COVID Data Tracker. Retrieved December 16, 2022, from https://covid.cdc.gov/covid-data-tracker

Charles D. Baker. (2021). COVID-19 Order No. 65. Retrieved from https://www.mass.gov/doc/covid-19-order-65/download

Chen, J. T., & Krieger, N. (2021). Revealing the Unequal Burden of COVID-19 by Income, Race/Ethnicity, and Household Crowding: US County Versus Zip Code Analyses. *Journal of Public Health Management and Practice*, *27*(Supplement 1), S43–S56. http://doi.org/10.1097/PHH.0000000000001263

Chen, Y., Han, Y., Yang, J., Ma, Y., Li, J., & Zhang, R. (2022). Impact of SARS-CoV-2 Variants on the Analytical Sensitivity of rRT-PCR Assays. *Journal of Clinical Microbiology*, *60*(4), e02374–21. http://doi.org/10.1128/jcm.02374-21

Cuadros, D. F., Moreno, C. M., Musuka, G., Miller, F. D., Coule, P., & MacKinnon, N. J. (2022). Association Between Vaccination Coverage Disparity and the Dynamics of the COVID-19 Delta and Omicron Waves in the US. *Frontiers in Medicine*, *9*, 898101. http://doi.org/10.3389/fmed.2022.898101

Dirican, E., & Bal, T. (2022). COVID-19 disease severity to predict persistent symptoms: a systematic review and meta-analysis. *Primary Health Care Research & Development*, *23*, e69. http://doi.org/10.1017/S1463423622000585

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. http://doi.org/10.1016/S1473-3099(20)30120-1

Fall, A., Eldesouki, R. E., Sachithanandham, J., Morris, C. P., Norton, J. M., Gaston, D. C., … Mostafa, H. H. (2022). The displacement of the

SARS-CoV-2 variant Delta with Omicron: An investigation of hospital admissions and upper respiratory viral loads. *eBioMedicine*, *79*, 104008. http://doi.org/10.1016/j.ebiom.2022.104008

Genest, C., McConway, K. J., & Schervish, M. J. (1986). Characterization of Externally Bayesian Pooling Operators. *The Annals of Statistics*, *14*(2), 487–501. Retrieved from https://www.jstor.org/stable/2241231

Green, D. A., Zucker, J., Westblade, L. F., Whittier, S., Rennert, H., Velu, P., … Sepulveda, J. L. (2020). Clinical Performance of SARS-CoV-2 Molecular Tests. *Journal of Clinical Microbiology*, *58*(8), e00995–20. http://doi.org/10.1128/JCM.00995-20

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. http://doi.org/10.1007/s10654-016-0149-3

Harris, J. E. (2022). COVID-19 Incidence and hospitalization during the delta surge were inversely related to vaccination coverage among the most populous U.S. Counties. *Health Policy and Technology*, *11*(2), 100583. http://doi.org/10.1016/j.hlpt.2021.100583

Jiang, D. H., Roy, D. J., Pollock, B. D., Shah, N. D., & McCoy, R. G. (2022). Association of stay-at-home orders and COVID-19 incidence and mortality in rural and urban United States: a population-based study. *BMJ Open*, *12*(4), e055791. http://doi.org/10.1136/bmjopen-2021-055791

Kanji, J. N., Zelyas, N., MacDonald, C., Pabbaraju, K., Khan, M. N., Prasad, A., … Tipples, G. (2021). False negative rate of COVID-19 PCR testing: a discordant testing analysis. *Virology Journal*, *18*(1), 13. http://doi.org/10.1186/s12985-021-01489-0

Kao, S.-Y. Z., Sharpe, J. D., Lane, R. I., Njai, R., McCord, R. F., Ajiboye, A. S., … Ekwueme, D. U. (2023). Duration of Behavioral Policy Interventions and Incidence of COVID-19 by Social Vulnerability of US Counties, April–December 2020. *Public Health Reports*, *138*(1), 190–199. http://doi.org/10.1177/00333549221125202

Karmakar, M., Lantz, P. M., & Tipirneni, R. (2021). Association of Social and Demographic Factors With COVID-19 Incidence and Death Rates in the US. *JAMA Network Open*, *4*(1), e2036462. http://doi.org/10.1001/jamanetworkopen.2020.36462

Kaufman, B. G., Whitaker, R., Mahendraratnam, N., Hurewitz, S., Yi, J., Smith, V. A., & McClellan, M. (2021). State variation in effects of state social distancing policies on COVID-19 cases. *BMC Public Health*, *21*(1), 1239. http://doi.org/10.1186/s12889-021-11236-3

Kojima, N., Roshani, A., & Klausner, J. D. (2022). Duration of COVID-19 PCR positivity for Omicron vs earlier variants. *Journal of Clinical Virology Plus*, *2*(3), 100085. http://doi.org/10.1016/j.jcvp.2022.100085

Kortela, E., Kirjavainen, V., Ahava, M. J., Jokiranta, S. T., But, A., Lindahl, A., … Kekäläinen, E. (2021). Real-life clinical sensitivity of SARS-CoV-2 RT-PCR test in symptomatic patients. *PLOS ONE*, *16*(5), e0251661. http://doi.org/10.1371/journal.pone.0251661

Lash, T. L., Fox, M. P., & Fink, A. K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York, NY: Springer New York. http://doi.org/10.1007/978-0-387-87959-8

Ma, Q., Liu, J., Liu, Q., Kang, L., Liu, R., Jing, W., … Liu, M. (2021b). 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 Network Open*, *4*(12), e2137257. http://doi.org/10.1001/jamanetworkopen.2021.37257

Ma, Q., Liu, J., Liu, Q., Kang, L., Liu, R., Jing, W., … Liu, M. (2021a). 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 Network Open*, *4*(12), e2137257. http://doi.org/10.1001/jamanetworkopen.2021.37257

Mallett, S., Allen, A. J., Graziadio, S., Taylor, S. A., Sakai, N. S., Green, K., … Halligan, S. (2020). At what times during infection is SARS-CoV-2 detectable and no longer detectable using RT-PCR-based tests? A systematic review of individual participant data. *BMC Medicine*, *18*(1), 346. http://doi.org/10.1186/s12916-020-01810-8

McLaughlin, J. M., Wiemken, T. L., Khan, F., & Jodar, L. (2022). US County-Level COVID-19 Vaccine Uptake and Rates of Omicron Cases and Deaths. *Open Forum Infectious Diseases*, *9*(7), ofac299. http://doi.org/10.1093/ofid/ofac299

Mustafa Hellou, M., Górska, A., Mazzaferri, F., Cremonini, E., Gentilotti, E., De Nardo, P., … Paul, M. (2021). Nucleic acid amplification tests on respiratory samples for the diagnosis of coronavirus infections: a systematic review and meta-analysis. *Clinical Microbiology and Infection*, *27*(3), 341–351. http://doi.org/10.1016/j.cmi.2020.11.002

Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*(767), 333–380. http://doi.org/10.1098/rsta.1937.0005

Petersen, J. M., Ranker, L. R., Barnard-Mayers, R., MacLehose, R. F., & Fox, M. P. (2021). A systematic review of quantitative bias analysis applied to epidemiological research. *International Journal of Epidemiology*, *50*(5), 1708–

1730. http://doi.org/10.1093/ije/dyab061

Poole, D., & Raftery, A. E. (2000). Inference for Deterministic Simulation Models: The Bayesian Melding Approach. *Journal of the American Statistical Association*, *95*(452), 1244–1255. http://doi.org/10.1080/01621459.2000.10474324

Powers, K. A., Ghani, A. C., Miller, W. C., Hoffman, I. F., Pettifor, A. E., Kamanga, G., … Cohen, M. S. (2011). The role of acute and early HIV infection in the spread of HIV and implications for transmission prevention strategies in Lilongwe, Malawi: a modelling study. *The Lancet*, *378*(9787), 256–268. http://doi.org/10.1016/S0140-6736(11)60842-8

Robson, B. J. (2014). When do aquatic systems models provide useful predictions, what is changing, and what is next? *Environmental Modelling & Software*, *61*, 287–296. http://doi.org/10.1016/j.envsoft.2014.01.009

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (Third edition). Philadelphia Baltimore New York: Wolters Kluwer Health, Lippincott Williams & Wilkins.

Sah, P., Fitzpatrick, M. C., Zimmer, C. F., Abdollahi, E., Juden-Kelly, L., Moghadas, S. M., … Galvani, A. P. (2021b). Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences*, *118*(34), e2109229118. http://doi.org/10.1073/pnas.2109229118

Sah, P., Fitzpatrick, M. C., Zimmer, C. F., Abdollahi, E., Juden-Kelly, L., Moghadas, S. M., … Galvani, A. P. (2021a). Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences*, *118*(34), e2109229118. http://doi.org/10.1073/pnas.2109229118

Salomon, J. A., Reinhart, A., Bilinski, A., Chua, E. J., La Motte-Kerr, W., Rönn, M. M., … Tibshirani, R. J. (2021). The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences*, *118*(51), e2111454118. http://doi.org/10.1073/pnas.2111454118

Ševčíková, H., Raftery, A. E., & Waddell, P. A. (2007). Assessing uncertainty in urban simulations using Bayesian melding. *Transportation Research Part B: Methodological*, *41*(6), 652–669. http://doi.org/10.1016/j.trb.2006.11.001

Singanayagam, A., Hakki, S., Dunning, J., Madon, K. J., Crone, M. A., Koycheva, A., … Lackenby, A. (2022). Community transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated and unvaccinated individuals in the UK: a prospective, longitudinal, cohort study. *The Lancet Infectious Diseases*, *22*(2), 183–195. http://doi.org/10.1016/S1473-3099(21)00648-4

The New York Times. (2022, December 16). Coronavirus in the U.S.: Latest Map and Case Count. Retrieved December 16, 2022, from https://www.

nytimes.com/interactive/2021/us/covid-cases.html

Tom Wolf. (2020, November 19). Process to Reopen Pennsylvania. Retrieved December 16, 2022, from https://www.governor.pa.gov/process-to-reopen-pennsylvania/

Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., ... Benjamin-Chung, J. (2020). Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications*, *11*(1), 4507. http://doi.org/10.1038/s41467-020-18272-4