WORKING PAPER · NO. 2020-54

Estimation of COVID-19 Prevalence from Serology Tests: A Partial Identification Approach

Panos Toulis

JUNE 2020

Becker Friedman Institute | Health Economics Initiative

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

Estimation of Covid-19 Prevalence from Serology Tests: A Partial Identification Approach

Panos Toulis*
University of Chicago,
Booth School of Business

Abstract

We propose a partial identification method for estimating disease prevalence from serology studies. Our data are results from antibody tests in some population sample, where the test parameters, such as the true/false positive rates, are unknown. Our method scans the entire parameter space, and rejects parameter values using the joint data density as the test statistic. The proposed method is conservative for marginal inference, in general, but its key advantage over more standard approaches is that it is valid in finite samples even when the underlying model is not point identified. Moreover, our method requires only independence of serology test results, and does not rely on asymptotic arguments, normality assumptions, or other approximations. We use recent Covid-19 serology studies in the US, and show that the parameter confidence set is generally wide, and cannot support definite conclusions. Specifically, recent serology studies from California suggest a prevalence anywhere in the range 0%-2% (at the time of study), and are therefore inconclusive. However, this range could be narrowed down to 0.7%-1.5% if the actual false positive rate of the antibody test was indeed near its empirical estimate ($\sim$0.5%). In another study from New York state, Covid-19 prevalence is confidently estimated in the range 13%-17% in mid-April of 2020, which also suggests significant geographic variation in Covid-19 exposure across the US. Combining all datasets yields a 5%-8% prevalence range. Our results overall suggest that serology testing on a massive scale can give crucial information for future policy design, even when such tests are imperfect and their parameters unknown.

Keywords: partial identification; disease prevalence; serology tests; Covid-19.
JEL classification codes: C12, C14, I10.

*Email: panos.toulis@chicagobooth.edu. Code is available at https://github.com/ptoulis/covid-19.

1 Introduction

Since December 2019 the world has been facing the Covid-19 pandemic, and its disastrous effects in human life and the economy. Responding to the pandemic, most countries have closed off their borders, and imposed unprecedented, universal lockdowns on their entire economies. The key reason for such drastic measures is uncertainty: we do not yet know the actual transmission rate, the lethality, or the prevalence of this new deadly disease. As governments and policy makers were caught by surprise, there is no doubt that these drastic measures were needed as a first line of defense. The data show that we would have to deal with a massive humanitarian disaster otherwise.

At the same time, as the economic pain mounts, especially for the most vulnerable and disadvantaged segments of the population, there is an urgent need to think of careful ways to safely reopen the economy. Estimating the true prevalence of Covid-19 has been identified as a key parameter to this effort (Alvarez et al., 2020). In the United States, the number of confirmed Covid-19 cases is 1,193,813 as of May 7 with 70,802 total deaths. This implies a (case) prevalence of 0.36% (assuming 328m as the US population), and a 5.9% mortality rate of Covid-19, which is even higher than the mortality rate reported at times by the World Health Organization.[1] However, the true prevalence, that is, the number of people who are currently infected or have been infected by Covid-19 over the entire population is likely much higher, and so the mortality rate should be significantly lower than 5.9%. A growing literature is attempting to estimate these numbers through epidemiological models (Li et al., 2020; Flaxman et al., 2020), or structural assumptions (Hortaçsu et al., 2020).

A more robust alternative seems to be possible through randomized serological studies that detect marker antibodies indicating exposure to Covid-19. In the US, there is currently a massive coordinated effort to evaluate the widespread application of these tests. The results are expected in late May of 2020.[2] The hope is that these tests will determine the true prevalence of the virus, and thus its lethality, and also determine whether someone is immune enough to return to work (the extent of immunity is still uncertain, however). Furthermore, seroprevalence studies can give information on risk factors for the disease, such as a patient's age, location, or underlying health conditions. They may also reveal important medical information on immune responses to the virus, such as how long antibodies last in people's bodies following infection, and could also identify those able to donate blood plasma, which is a possible treatment to seriously ill Covid-19 patients.[3] The development of serology tests is therefore essential to designing a careful strategy towards both effective medical treatments and a gradual reopening of the economy.

[1] The official mortality rate was revised from 2% in late January to 4% in early March; see also an official WHO situation report from early March: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_2.

[2] CDC page: https://www.cdc.gov/coronavirus/2019-ncov/lab/serology-testing.html.

[3] Food and Drug Administration (FDA) announcement on serology studies (04/07/2020): https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-serological-tests.

Until widespread serology testing is possible, however, we have to rely on a limited number of serology studies that have started to emerge in various areas of the globe, including the US. Table 1 presents a non-exhaustive summary of such studies around the world. For example, in Germany, serology tests in early April showed a 14% prevalence in a sample of 500 people.[4] In the Netherlands, a study in mid-April showed a lower prevalence at 3.5% in a small sample of blood donors.[5] [6] In the US, in a recent and relatively large study in Santa Clara, California, Bendavid et al. (2020) estimated an in-sample prevalence of 1.5% from 50 positive test results in a sample of 3330 patients. Using a reweighing technique, the authors extrapolated this estimate to 2-4% prevalence in the general population. A follow-up study in LA County found 35 positives out of 846 tests. What is unique about these last two studies is that data from a prior validation study are also available, where, say, 401 "true negatives" were tested with 2 positive results, implying a false positive rate of 0.5%. Upon publication, these studies received intense criticism because the false positive rate appears to be large enough compared to the underlying disease prevalence. For example, the Agresti-Coull and Clopper-Pearson 95% confidence intervals for the false positive rate are $[0.014\%, 1.92\%]$ and $[0.06\%, 1.79\%]$, respectively. These intervals for the false positive rate are not incompatible even with a 0% prevalence, since a 1.5% false positive rate achieves $0.015 \times 3330 \approx 50$ (false) positives on average, same as the observed value in the sample.

Such standard methods, however, are justified based on approximations, asymptotic arguments, prior specifications (for Bayesian methods), or normality assumptions, which are always suspect in small samples. In this paper, we develop a method that can assess finite-sample statistical significance in a robust way. The key idea is to treat all unknown quantities as parameters, and explore the entire parameter space to assess agreement with the observed data. Our method adopts the partial identification framework, where the goal is not to produce point estimates, but to identify sets of plausible parameter values (Wooldridge and Imbens, 2007; Tamer, 2010; Chernozhukov et al., 2007; Manski, 2003, 2010, 2007; Romano and Shaikh, 2008, 2010; Honoré and Tamer, 2006; Imbens and Manski, 2004; Beresteanu et al., 2012; Stoye, 2009; Kaido et al., 2019). Within that literature, our proposed method appears to be unique in the sense that it constructs a procedure that is valid in finite samples given the correct distribution of the test statistic. Importantly, the choice of the test statistic can affect only the power of our method, but not its validity. Such flexibility may be especially valuable in choosing a test statistic that is both powerful and easy to compute. Thus, the main benefit of our approach is that it is valid with *enough computation*, whereas more standard methods are only valid with *enough samples*.

[4] Report in German: https://www.land.nrw/sites/default/files/asset/document/zwischenergebnis_covid19_case_study_gangelt_0.pdf.

[5] Presentation slides in Dutch: https://www.tweedekamer.nl/sites/default/files/atoms/files/tb_jaap_van_dissel_1604_1.pdf.

[6] See also a summary of these projects in the journal "Science": https://www.sciencemag.org/news/2020/04/antibody-surveys-suggesting-vast-undercount-coronavirus-infections-may-be-unreliable.

Prevalence	Location	Time	Method	Notes
6.14%	China	01/21	PCR	Sample of 342 pneumonia patients (Liang et al., 2020).
2.6%	Italy	02/21	PCR	Used 80% of entire population in Vó, Italy (Lavezzo et al., 2020).
5.3%	USA	03/12	PCR	Used 131 patients with ILI symptoms (Spellberg et al., 2020).
13.7%	USA	03/22	PCR	Sample of 215 pregnant women in NYC (Sutton et al., 2020).
0.34%	USA	03/17	model	(Yadlowsky et al., 2020).
9.4%	Spain	03/28	serology	Sample of 578 healthcare workers (Garcia-Basteiro et al., 2020).
3%	Japan	03/31	serology	Random set of 1000 blood samples in Kobe Hospital (Doi et al., 2020).
36%	USA	04/02	PCR	Study in large homeless shelter in Boston (Baggett et al., 2020).
1.5%	USA	04/03	serology	Recruited 3330 people via Facebook (Bendavid et al., 2020).
2.5%	USA	04/04	model	Uses ILINet data; implies 96% unreported cases (Lu et al., 2020).
9.1%	Switzerland	04/06	serology	Sample of 1335 individuals in Geneva (Stringhini et al., 2020).
14%	Germany	04/09	serology	Self-reported 400 households in Gangelt. (Streeck et al., 2020).
3.1%	Netherlands	04/16	antigen	Used 3% of all blood donors (Jaap van Dissel, 2020).
0.4-40%	USA	04/24	model	Partial identification using #cases/deaths (Manski and Molinari, 2020).
Unreported				
90%	USA	03/16	model	Used airflight data to identify transmission rates (Hortaçsu et al., 2020).
85%	India	04/02	model	Extrapolated from respiratory-related cases (Venkatesan, 2020).

Table 1: Summary of recent Covid-19 prevalence studies crudely categorized as either statistical models or medical tests (PCR or serology). Most studies report intervals, but here we mainly report midpoints. *Top panel:* PCR stands for "polymerase chain reaction", and is a key test to detect presence of the virus' RNA; *serology* denotes a serology test to detect presence of antibodies (e.g., IgA, IgM, IgG). *ILI* stands for "influenza-like illness" and describes methods that use data recorded from patients with general influenza-like symptoms, including but not limited to Covid-19 cases. *Bottom panel*: Studies that aimed mainly to estimate percentage of Covid-19 cases that are not reported. This number can be used to calculate prevalence estimates, but needs to be adjusted for the exact timeframe of the study; e.g., the analysis of Hortaçsu et al. (2020) implies a 1.5-2.5% prevalence in Santa Clara county in mid-March.

The rest of this paper is structured as follows. In Section 2 we describe the problem formally. In Section 3.1 we describe the proposed method on a high level. A more detailed analysis along with a modicum of theory is given in Section 3.2. In Section 4 we apply the proposed method on data from the Santa Clara study, the LA County study, and a recent study from New York state.

2 Problem Setup

Here, we formalize the statistical problem of estimating disease prevalence through imperfect medical tests. Every individual i is associated with a binary status x_i: it is $x_i = 1$ if the individual has developed antibodies from exposure to the disease, and $x_i = 0$ if not. We will also refer to these cases as "positive" and "negative", respectively. Patient status is not directly observed, but can be estimated with a serology (antibody) test.

This medical antibody test can be represented by a function $t : \{0, 1\} \to \{0, 1\}$, and determines whether someone is positive or negative. As usual, the categorization of the test results can be described through the following table:

		true status, $x =$	
		0	1
$t(x) =$	0	true negative	false negative
	1	false positive	true positive

We will assume that each test result is an independent random outcome, such that the true positive rate and false positive rate, denoted respectively by q and p,[7] are constant:

$$P\{t(x) = 1 | x = 1\} = q, \tag{A1}$$

$$P\{t(x) = 1 | x = 0\} = p. \tag{A2}$$

This assumption may be untenable in practice. In general, patient characteristics, or test target and delivery conditions can affect the test results. For example, Bendavid et al. (2020) report slightly different test performance characteristics depending on which antibody (either IgM or IgG) was being detected. We note, however, that this assumption is not strictly necessary for the validity of our proposed inference procedure. It is only useful in order to obtain a precise calculation for the distribution of the test statistic (see Theorem 1 and remarks).

To determine test performance characteristics, and gain information about the true/false positive rates of the antibody test, there is usually a *validation study* where the underlying status of participating individuals is known. In the Covid-19 case, for example, such validation study could include pre-Covid-19 blood samples that have been preserved, and are thus "true negatives". To simplify, we assume that in the validation study there is a set $\mathcal{I}_{\text{c}}^{-}$ of participating individuals, where it is known that everyone is a true negative, and a set $\mathcal{I}_{\text{c}}^{+}$ where everyone is positive; i.e.,

$$x_i = 0, \text{ for all } i \in \mathcal{I}_{\text{c}}^{-}, \text{and } x_i = 1, \text{ for all } i \in \mathcal{I}_{\text{c}}^{+}.$$

There is also the *main study* with a set $\mathcal{I}_{\text{m}}$ of participating individuals, where the true status is not known. We assume no overlap between sets $\mathcal{I}_{\text{c}}^{-}, \mathcal{I}_{\text{c}}^{+}$ and $\mathcal{I}_{\text{m}}$, which is a realistic assumption. We define $N_{\text{c}}^{-} = |\mathcal{I}_{\text{c}}^{-}|$ and $N_{\text{c}}^{+} = |\mathcal{I}_{\text{c}}^{+}|$ as the respective number of participants in the validation study, and $N_{\text{m}} = |\mathcal{I}_{\text{m}}|$ as the number of participants in the main study. These numbers are observed, but the full patient sets or the patient characteristics, may not be observed.

[7] The terms "sensitivity" and "specificity" are frequently used in practice of medical testing. In our setting, sensitivity maps to the true positive rate (q), and specificity maps to one minus the false positive rate ($1 - p$). In this paper, we will only use the terms "true/false positive rate" as they are more precise and self-explanatory.

We also observe the positive test results in both studies:

$$
\begin{aligned}
S_{\mathrm{c}}^{-} &= \sum_{i \in \mathcal{I}_{\mathrm{c}}^{-}} t(x_i), \;\; S_{\mathrm{c}}^{+} = \sum_{i \in \mathcal{I}_{\mathrm{c}}^{+}} t(x_i), && \text{[Calibration study]} \\
S_{\mathrm{m}} &= \sum_{i \in \mathcal{I}_{\mathrm{m}}} t(x_i). && \text{[Main study]}
\end{aligned} \tag{1}
$$

Thus, S_{c}^{-} is the number of false positives in the validation study since we know that all individuals in $\mathcal{I}_{\mathrm{c}}^{-}$ are true negatives. Similarly, S_{c}^{+} is the number of true positives in the validation study since all individuals in $\mathcal{I}_{\mathrm{c}}^{+}$ are known to be positive. These numbers offer some simple estimates of the false positive rate and true positive rate of the medical test, respectively: $\hat{p} = S_{\mathrm{c}}^{-}/N_{\mathrm{c}}^{-}$ and $\hat{q} = S_{\mathrm{c}}^{+}/N_{\mathrm{c}}^{+}$. We use $(s_{\mathrm{c,obs}}^{-}, s_{\mathrm{c,obs}}^{+}, s_{\mathrm{m,obs}})$ to denote the observed values of test positives $(S_{\mathrm{c}}^{-}, S_{\mathrm{c}}^{+}, S_{\mathrm{m}})$, respectively, which are integer-valued random variables.

The statistical task is therefore to use observed data $\{(N_{\mathrm{c}}^{-}, N_{\mathrm{c}}^{+}, N_{\mathrm{m}}), (s_{\mathrm{c,obs}}^{-}, s_{\mathrm{c,obs}}^{+}, s_{\mathrm{m,obs}})\}$ and do inference on the quantity:

$$
\pi = \frac{\sum_{i \in \mathcal{I}_{\mathrm{m}}} x_i}{N_{\mathrm{m}}}, \tag{2}
$$

i.e., the unknown disease prevalence in the main study. We emphasize that π is a finite-population estimand — we discuss (briefly) the issue of extrapolation to the general population in Section 5. The challenge here is that S_{m} generally includes both false positives and true positives, which depends on the unknown test parameters, namely the true/false positive rates q and p. Since πN_{m} is the (unknown) number of infected individuals in the main study, we can use Assumptions (A1) and (A2) to write down this decomposition formally:

$$
\begin{aligned}
S_{\mathrm{c}}^{-} &\sim \mathrm{Binom}(N_{\mathrm{c}}^{-}, p), \;\; S_{\mathrm{c}}^{+} \sim \mathrm{Binom}(N_{\mathrm{c}}^{+}, q), \text{and} \\
S_{\mathrm{m}} &\sim \underbrace{\mathrm{Binom}(\pi N_{\mathrm{m}}, q)}_{\text{true positives}} + \underbrace{\mathrm{Binom}(N_{\mathrm{m}} - \pi N_{\mathrm{m}}, p)}_{\text{false positives}},
\end{aligned} \tag{3}
$$

where Binom denotes the binomial random variable. For brevity, we define $\boldsymbol{S} = (S_{\mathrm{c}}^{-}, S_{\mathrm{c}}^{+}, S_{\mathrm{m}})$ as our joint data statistic, and $\boldsymbol{\theta} = (p, q, \pi)$ as the joint parameter value. The independence of tests implies that the density of $\boldsymbol{S}$ can be computed exactly as follows.

$$
f(\boldsymbol{S} \mid \boldsymbol{\theta}) = \mathrm{d}(S_{\mathrm{c}}^{-}; N_{\mathrm{c}}^{-}, p) \cdot \mathrm{d}(S_{\mathrm{c}}^{+}; N_{\mathrm{c}}^{+}, q) \cdot \left[\sum_{j=0}^{j=S_{\mathrm{m}}} \mathrm{d}(j; \pi N_{\mathrm{m}}, q) \cdot \mathrm{d}(S_{\mathrm{m}} - j; N_{\mathrm{m}} - \pi N_{\mathrm{m}}, p) \right], \tag{4}
$$

where $\mathrm{d}(k; n, s)$ denotes the probability of k successes in a binomial experiment with n trials and s probability of success. There are several ways to implement Equation (4) efficiently — we defer discussions on computational issues to Section 5.

3 Method

We begin with an illustrative example to describe the proposed method on a high level. We give more details along with some theoretical guarantees in the section that follows.

3.1 Illustrative example

Let us consider the Santa Clara study (Bendavid et al., 2020) with observed data:

$$(N_{\mathrm{c}}^{-}, N_{\mathrm{c}}^{+}, N_{\mathrm{m}}) = (401, 197, 3330), \text{ and } (s_{\mathrm{c,obs}}^{-}, s_{\mathrm{c,obs}}^{+}, s_{\mathrm{m,obs}}) = (2, 178, 50).$$

The unknown quantities in our analysis are q, p and π: the true positive rate of the test, the false positive rate, and the unknown prevalence in the main study, respectively. Assume zero prevalence ($\pi = 0\%$), 90% true positive rate ($q = 0.9$), and 1.5% false positive rate ($p = 0.015$). We ask the question: "Is the combination $(p, q, \pi) = (0.015, 0.90, 0)$ compatible with the data?". Naturally, this can be framed in statistical terms as a null hypothesis:

$$H_0 : (p, q, \pi) = (0.015, 0.90, 0). \tag{5}$$

To test H_0 we have to compare the observed positive test results with the values that *could have been observed* if indeed the true model parameter values were $(p, q, \pi) = (0.015, 0.90, 0)$. Our model is simple enough that we can execute this hypothetical analysis exactly based on the density of $f(\boldsymbol{S}|\boldsymbol{\theta})$ in Equation (4), where $\boldsymbol{S} = (S_{\mathrm{c}}^{-}, S_{\mathrm{c}}^{+}, S_{\mathrm{m}})$ is the vector of all positive test results, and $\boldsymbol{\theta}$ is specified as in H_0; see Figure 1. To simplify visualization, in Figure 1 we fix the component S_{c}^{+} of $\boldsymbol{S}$ to its observed value ($S_{\mathrm{c}}^{+} = 178$), and only plot the density with respect to the other two components, $(S_{\mathrm{c}}^{-}, S_{\mathrm{m}})$; i.e., we plot $f(\boldsymbol{S} \mid H_0, S_{\mathrm{c}}^{+} = 178)$. One can visualize the full joint distribution $f(\boldsymbol{S} \mid H_0)$ as a collection of such conditional densities for all possible values of S_{c}^{+}.

The next step is to decide whether the observed value of $\boldsymbol{S}$, namely $\boldsymbol{s}_{\mathrm{obs}} = (2, 178, 50)$, is compatible with the distribution of Figure 1. We see that the mode of the distribution is around the point $(S_{\mathrm{c}}^{-}, S_{\mathrm{m}}) = (5, 45)$, whereas the point $(2, 50)$ is at the lower edge of the distribution. If the observed values were even further, say at $(S_{\mathrm{c}}^{-}, S_{\mathrm{m}}) = (2, 80)$, then we could confidently reject H_0 since the density at $(2, 80)$ is basically zero. Here, we have to be careful because the actual observed values are still somewhat likely under H_0. Our method essentially accepts H_0 when the density of this distribution at the observed value $\boldsymbol{s}_{\mathrm{obs}}$ of statistic $\boldsymbol{S}$ is above some threshold c_0, that is, we decide based on the following rule:

$$\text{Accept } H_0 \text{ if } f(\boldsymbol{s}_{\mathrm{obs}}|H_0) > c_0. \tag{6}$$

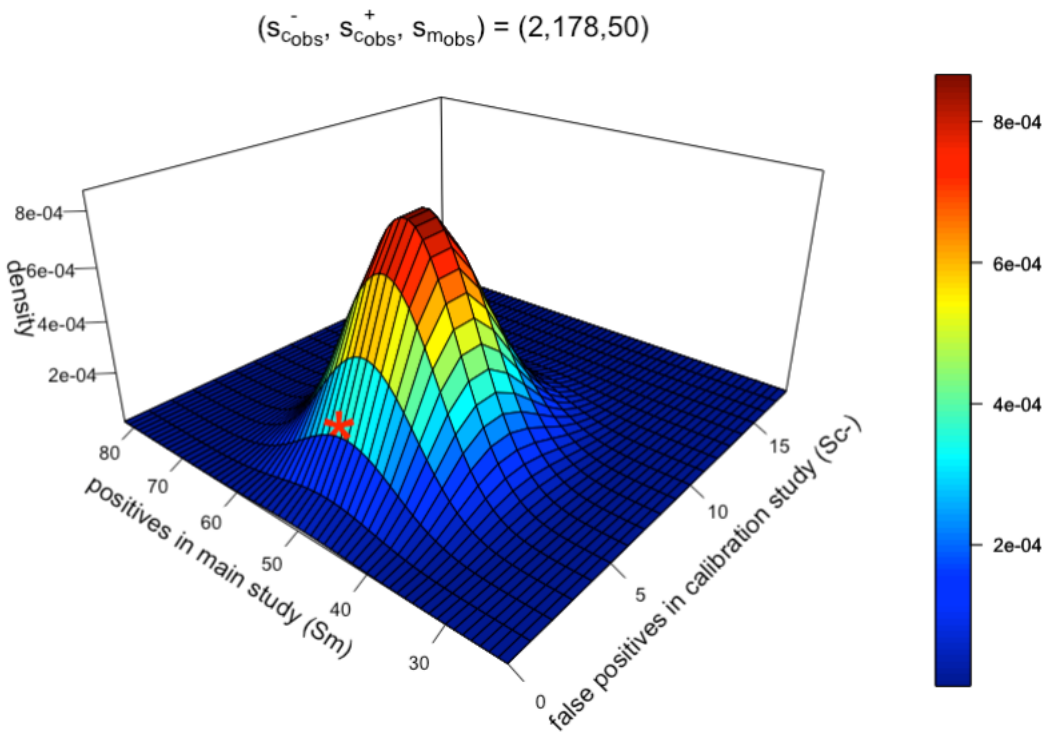

Figure 1: Density $f(\boldsymbol{S}|H_0)$ of test positives $\boldsymbol{S} = (S_{\mathrm{c}}^-, S_{\mathrm{c}}^+, S_{\mathrm{m}})$ conditional on H_0 of Equation (5), and with fixed $S_{\mathrm{c}}^+ = 178$ (its observed sample value) to ease visualization. The observed values $(s_{\mathrm{c,obs}}^-, s_{\mathrm{m,obs}}) = (2, 50)$ are marked with an asterisk. To test H_0 we need to calculate some kind of "p-value" for the observed point. In our construction, we simply test whether the density at the observed values exceeds an appropriately chosen threshold (see Section 3.2).

The test in Equation (6) is reminiscent of the likelihood ratio test, the key difference being that our test does not require maximizations of the likelihood function over the parameter space, which is computationally intensive, and frequently unstable numerically — we make a concrete comparison in the application of Section 4.3. Our test essentially uses the density of $\boldsymbol{S}$ as the test statistic for H_0, while threshold c_0 generally depends on the particular null values being tested. Assuming that the test of Equation (6) has been defined, we can then test for all possible combinations of our parameter values, $\boldsymbol{\theta} \in \Theta$, in some large enough parameter space Θ, and then invert this procedure in order to construct the confidence set. As usual, we would like this confidence set to cover the true parameters with some minimum probability (e.g., 95%). In the following section, we show that this is possible through an appropriate construction of the test in Equation (6), which takes into account the level sets of the density function depicted in Figure 1. The overall procedure is computationally intensive, but is valid in finite samples without the need of asymptotic or normality assumptions. The details of this construction, including the appropriate selection of the test threshold and the proof of validity, are presented in the following section.

3.2 Theoretical Details

Let $\boldsymbol{S} = (S_{\mathrm{c}}^{-}, S_{\mathrm{c}}^{+}, S_{\mathrm{m}}) \in \mathbb{S}$ denote the statistic, where $\mathbb{S} = \{0, \ldots, N_{\mathrm{c}}^{-}\} \times \{0, \ldots, N_{\mathrm{c}}^{+}\} \times \{0, \ldots, N_{\mathrm{m}}\}$, and let $\boldsymbol{\theta} = (p, q, \pi) \in \Theta$ be the model parameters. We take Θ to be finite and discrete; e.g., for probabilities we take a grid of values between 0 and 1. Let $\boldsymbol{s}_{\mathrm{obs}} = (s_{\mathrm{c,obs}}^{-}, s_{\mathrm{c,obs}}^{+}, s_{\mathrm{m,obs}})$ denote the observed value of $\boldsymbol{S}$ in the sample. Let $f(\boldsymbol{S}|\boldsymbol{\theta})$ denote the density of the joint statistic conditional on the model parameter value $\boldsymbol{\theta}$, as defined in Equation (4). Suppose that $\boldsymbol{\theta}_0$ is the true unknown parameter value, and assume that

$$P(\boldsymbol{\theta}_0 \in \Theta) = 1. \tag{A4}$$

Assumption (A4) basically posits that our discretization is fine enough to include the true parameter value with probability one. In our application, this assumption is rather mild as we are dealing with parameters that are either probabilities or integers, and so bounded within well-defined ranges. Moreover, this assumption is implicit essentially in all empirical work since computers operate with finite precision. Our goal is to construct a confidence set $\widehat{\Theta}_{1-\alpha} \subseteq \Theta$ such that $P(\boldsymbol{\theta}_0 \in \widehat{\Theta}_{1-\alpha}) \geq 1-\alpha$, where α is some desired level (e.g., $\alpha = 0.05$). Trivially, $\widehat{\Theta}_{1-\alpha} = \Theta$ satisfies this criterion, so we will aim to make $\widehat{\Theta}$ as narrow as possible. We will also need the following definition:

$$\nu(z, \boldsymbol{\theta}) = \big|\{\boldsymbol{s} \in \mathbb{S} : 0 < f(\boldsymbol{s}|\boldsymbol{\theta}) \leq z\}\big|. \tag{7}$$

Function ν depends on level sets of f, and counts the number of sample data points (over the sample space $\mathbb{S}$) with likelihood at $\boldsymbol{\theta}$ that is smaller than the observed likelihood at $\boldsymbol{\theta}$.

We can now prove the following theorem.

Theorem 1. *Suppose that Assumption (A4) holds. Consider the following construction for the confidence set:*

$$\widehat{\Theta}_{1-\alpha} = \big\{\boldsymbol{\theta} \in \Theta : f(\boldsymbol{s}_{\mathrm{obs}}|\boldsymbol{\theta}) \cdot \nu(f(\boldsymbol{s}_{\mathrm{obs}}|\boldsymbol{\theta}), \boldsymbol{\theta}) > \alpha\big\}. \tag{8}$$

Then, $P(\boldsymbol{\theta}_0 \in \widehat{\Theta}_{1-\alpha}) \geq 1 - \alpha$.

Proof. For any fixed $\boldsymbol{\theta} \in \Theta$ consider the function $g(z, \boldsymbol{\theta}) = z \cdot \nu(z, \boldsymbol{\theta})$, $z \in [0, 1]$. Note that, for fixed $\boldsymbol{\theta}$, function $g(z, \boldsymbol{\theta})$ is monotone increasing and generally not continuous with respect to z. Let $\mathbb{F}_{\boldsymbol{\theta}} = \{f(\boldsymbol{s}|\boldsymbol{\theta}) : \boldsymbol{s} \in \mathbb{S}\}$, and define as $z^*_{\boldsymbol{\theta}}$ the unique fixed point for which $g(z^*_{\boldsymbol{\theta}}, \boldsymbol{\theta}) - \alpha = 0$, if that point exists; if not, define the point as $z^*_{\boldsymbol{\theta}} = \max\{z \in \mathbb{F}_{\boldsymbol{\theta}} : g(z, \boldsymbol{\theta}) \leq \alpha\}$. It follows that $g(z^*_{\boldsymbol{\theta}}, \boldsymbol{\theta}) \leq \alpha$, for any $\boldsymbol{\theta}$, and so the event $\{f(\boldsymbol{s}_{\mathrm{obs}}|\boldsymbol{\theta}) \cdot \nu(f(\boldsymbol{s}_{\mathrm{obs}}|\boldsymbol{\theta}), \boldsymbol{\theta}) \leq \alpha\} = \{g(f(\boldsymbol{s}_{\mathrm{obs}}|\boldsymbol{\theta}), \boldsymbol{\theta}) \leq \alpha\}$ is the same as the event $\{f(\boldsymbol{s}_{\mathrm{obs}}|\boldsymbol{\theta}) \leq z^*_{\boldsymbol{\theta}}\}$. Now we can bound the coverage probability as follows.

$$
\begin{aligned}
P(\boldsymbol{\theta}_0 \notin \widehat{\Theta}_{1-\alpha}) &= P\left\{f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_0) \cdot \nu(f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_0), \boldsymbol{\theta}_0) \leq \alpha\right\} \\
&= P\left\{f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_0) \leq z^*_{\boldsymbol{\theta}_0}\right\} = \sum_{\boldsymbol{s}\in\mathbb{S}} \mathbb{I}\left\{f(\boldsymbol{s}|\boldsymbol{\theta}_0) \leq z^*_{\boldsymbol{\theta}_0}\right\} f(\boldsymbol{s}|\boldsymbol{\theta}_0) \qquad (9) \\
&\leq z^*_{\boldsymbol{\theta}_0} \sum_{\boldsymbol{s}\in\mathbb{S}: 0<f(\boldsymbol{s}|\boldsymbol{\theta}_0)\leq z^*_{\boldsymbol{\theta}_0}} 1 = z^*_{\boldsymbol{\theta}_0} \cdot \nu(z^*_{\boldsymbol{\theta}_0}, \boldsymbol{\theta}_0) = g(z^*_{\boldsymbol{\theta}_0}, \boldsymbol{\theta}_0) \leq \alpha.
\end{aligned}
$$

In the first line we used the test definition and Assumption (A4); in the second line, we used the monotonicity of g, and the fact that $\boldsymbol{\theta}_0$ is the true parameter value; in the last line, we used the definition of $z^*_{\boldsymbol{\theta}}$ and the uniform bound on g. ■

When $z^*_{\boldsymbol{\theta}}$ is not a discontinuity point of g, for all $\boldsymbol{\theta}$, then our test is exact in the sense that $P(\boldsymbol{\theta}_0 \in \widehat{\Theta}_{1-\alpha}) = 1 - \alpha$. In general, however, this condition will not hold for all Θ, and so the confidence set of Equation (8) may be conservative and lose power. We could potentially achieve more power if instead we define the confidence set as follows:

$$
\widehat{\Theta}^{\text{alt}}_{1-\alpha} = \left\{\boldsymbol{\theta} \in \Theta : \sum_{\boldsymbol{s}\in\mathbb{S}} \mathbb{I}\big\{f(\boldsymbol{s}|\boldsymbol{\theta}) \leq f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\big\} f(\boldsymbol{s}|\boldsymbol{\theta}) > \alpha\right\}. \qquad (10)
$$

Theorem 2. *Suppose that Assumption (A4) holds. Consider the construction of confidence set* $\widehat{\Theta}^{\text{alt}}_{1-\alpha}$ *as defined in Equation* (10)*. Then,* $P(\boldsymbol{\theta}_0 \in \widehat{\Theta}^{\text{alt}}_{1-\alpha}) \geq 1 - \alpha$.

Proof. The proof is almost identical to Theorem 1, if we replace the definition of g with $g(z, \boldsymbol{\theta}) = \sum_{\boldsymbol{s}\in\mathbb{S}} \mathbb{I}\left\{f(\boldsymbol{s}|\boldsymbol{\theta}) \leq z\right\} f(\boldsymbol{s}|\boldsymbol{\theta})$, $z \in [0, 1]$. With this definition g is a smoother function, which explains intuitively why this construction will generally lead to more power. ■

It is straightforward to see that $\widehat{\Theta}^{\text{alt}}_{1-\alpha} \subseteq \widehat{\Theta}_{1-\alpha}$ almost surely since

$$
\sum_{\boldsymbol{s}\in\mathbb{S}} \mathbb{I}\big\{f(\boldsymbol{s}|\boldsymbol{\theta}) \leq f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\big\} f(\boldsymbol{s}|\boldsymbol{\theta}) \leq f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}) \sum_{\boldsymbol{s}\in\mathbb{S}} \mathbb{I}\big\{f(\boldsymbol{s}|\boldsymbol{\theta}) \leq f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\big\} = f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}) \nu(f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}), \boldsymbol{\theta}).
$$

Since both constructions are valid in finite samples, the choice between $\widehat{\Theta}_{1-\alpha}$ or $\widehat{\Theta}^{\text{alt}}_{1-\alpha}$ should be mainly based on computational feasibility. The construction of $\widehat{\Theta}_{1-\alpha}$ may be easier to compute in practice as it depends on a summary of the distribution $f(\boldsymbol{s}|\boldsymbol{\theta})$ through the level set function ν, while the construction of $\widehat{\Theta}^{\text{alt}}_{1-\alpha}$ requires full knowledge of the entire distribution. If it is computationally feasible, however, $\widehat{\Theta}^{\text{alt}}_{1-\alpha}$ should be preferred because it is contained in $\widehat{\Theta}_{1-\alpha}$ with probability one, as argued above. This leads to sharper inference. See also the applications on serology studies in Section 4 for more details, where the construction of $\widehat{\Theta}^{\text{alt}}_{1-\alpha}$ is feasible.

3.3 Concrete Procedure and Remarks

Theorems 1 and 2 imply the following simple procedure to construct a 95% confidence set:

PROCEDURE 1

1. Observe the value $\boldsymbol{s}_{\text{obs}} = (s^-_{\text{c,obs}}, s^+_{\text{c,obs}}, s_{\text{m,obs}})$ of test positives in the validation study and the main study. Create a grid $\Theta \subset [0,1]^3$ for the unknown parameters $\boldsymbol{\theta} = (p, q, \pi)$.
2. For every $\boldsymbol{\theta}$ in Θ, calculate $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})$ as in Equation (4), and $\nu(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta})$ as in Equation (7).
3. Reject all values $\boldsymbol{\theta} \in \Theta$ for which $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}) \cdot \nu(f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}), \boldsymbol{\theta}) \leq 0.05$; alternatively, we can reject based on $\sum_{\boldsymbol{s}\in\mathbb{S}} \mathbb{I}\{f(\boldsymbol{s}|\boldsymbol{\theta}) \leq f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\} f(\boldsymbol{s}|\boldsymbol{\theta}) \leq 0.05$.
4. The remaining values in Θ form the 95% confidence sets $\widehat{\Theta}_{0.95}$ or $\widehat{\Theta}^{\text{alt}}_{0.95}$, respectively.

Remark 3.1 (Computation)**.** Procedure 1 is fully parallelizable over $\boldsymbol{\theta}$, and so the main computational difficultly is the need to sum over the sample space $\mathbb{S}$. Note, however, that Procedure 1 can work for any choice of $\boldsymbol{S}$ given its density $f(\boldsymbol{S}|\boldsymbol{\theta})$. Thus, our method offers valuable flexibility for inference; for instance, f could be simulated, or $\boldsymbol{S}$ could be a simple statistic (e.g., sample averages) and not necessarily an "arg max" estimator. See Section 5 for more discussion on computation. ■

Remark 3.2 (Identification)**.** Procedure 1 is not a typical partial identification method in the sense that there are settings under which the model of Equation (4) is point identified (i.e., when $N_{\text{m}}, N^-_{\text{c}}, N^+_{\text{c}} \to \infty$). However, we choose to describe Procedure 1 as a partial identification method for two main reasons. First, it is more plausible, in practice, that the calibration studies are small and finite ($N^-_{\text{c}}, N^+_{\text{c}} < \infty$), since a calibration study needs to have high-quality, ground-truth data. Second, it can happen that we don't have both kinds of calibration studies available (i.e., it could be that either $N^-_{\text{c}} = 0$ or $N^+_{\text{c}} = 0$). In both of these settings, the underlying model is no longer point identified, and so Procedure 1 is technically a partial identification method. ■

Remark 3.3 (Conservativeness)**.** Procedure 1 generally produces conservative confidence intervals. However, we can show that $P(\boldsymbol{\theta}_0 \notin \widehat{\Theta}^{\text{alt}}_{1-\alpha}) \geq \alpha - \epsilon$, where $\epsilon = \max_{s\in\mathbb{S}} P\{f(\boldsymbol{S}|\boldsymbol{\theta}_0) = f(s|\boldsymbol{\theta}_0)\}$. This value is very small, in general (e.g., $\epsilon \sim 10^{-3}$ in the Santa Clara study). In this case, the alternative construction is "approximately exact" in the sense that the coverage probability of $\widehat{\Theta}^{\text{alt}}_{1-\alpha}$ is almost equal to $(1-\alpha)$. ■

Remark 3.4 (Marginal inference)**.** The parameter of interest in our application could only be the disease prevalence, whereas the true/false positive rates of the antibody test may be considered "nuisance". In this paper, we directly project $\widehat{\Theta}_{1-\alpha}$ (or $\widehat{\Theta}^{\text{alt}}_{1-\alpha}$) on a single dimension to perform marginal inference (see Section 4), but this is generally conservative, especially at the boundary of the parameter space (Stoye, 2009; Kaido et al., 2019; Chen et al., 2018). A sharper way to do marginal inference with our procedures is an interesting direction for future work. ■

3.4 Comparison to other methods

How does our method compare to a more standard frequentist or Bayesian approach? Here, we discuss two key differences. First, as we have repeatedly emphasized in this paper, our method is valid in finite samples under only independence of test results, which is a mild assumption. In contrast, a standard frequentist approach, say based on the bootstrap, is inherently approximate and relies on asymptotics, while a Bayesian method requires the specification of priors and posterior sampling. Of course, our procedure requires more computation, mainly compared to the bootstrap, and can be conservative for marginal inference (see Remark 3.4), but this is arguably a small price to pay in a critical application such as the estimation of Covid-19 prevalence.

A second, more subtle, difference is the way our method performs inference. Specifically, we decide whether any $\boldsymbol{\theta} \in \Theta$ is in the confidence set based on the entire density $f(\boldsymbol{s}|\boldsymbol{\theta})$ over all $\boldsymbol{s} \in \mathbb{S}$, whereas both frequentist and Bayesian methods typically perform inference "around the mode" of the likelihood function $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})$ with fixed $\boldsymbol{s}_{\text{obs}} \in \mathbb{S}$ (we ignore how the prior specification affects Bayesian inference to simplify exposition). This can explain, on an intuitive level, how the inferences of the respective methods may differ. Figure 2 illustrates the difference. On the left panel, we plot the likelihood, $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})$, as a function of $\boldsymbol{\theta} \in \Theta$. Typically, in frequentist or Bayesian methods, the confidence set is around the mode, say $\hat{\boldsymbol{\theta}}$. We see that a parameter value, say $\boldsymbol{\theta}_1$, with a likelihood value, $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_1)$, that is low in absolute terms will generally not be included in the confidence set. However, in our approach, the value $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_1)$ is not important in absolute terms for doing inference, but is only important relative to all other values $\{f(\boldsymbol{s}|\boldsymbol{\theta}_1) : \boldsymbol{s} \in \mathbb{S}\}$ of the test statistic distribution $f(\boldsymbol{s}|\boldsymbol{\theta}_1)$. Such inference will typically include the mode, $\hat{\boldsymbol{\theta}}$, but will also include parameter values at the tails of the likelihood function, such as $\boldsymbol{\theta}_1$. As such, our method is expected to give more accurate inference in small-sample problems, or in settings with poor identifiability where the likelihood is non-smooth and multimodal. We argue that we actually see these effects in the application on Covid-19 serology studies analyzed in the following section — see also Section 4.3 and Appendix D for concrete numerical examples.

4 Application

In this section, we apply the inference procedure of Section 3.3 to several serology test datasets in the US. Moreover, we present results for combinations of these datasets, assuming that the tests have identical specifications. This is likely an untenable assumption, but it helps to illustrate how we can use our approach to flexibly combine all evidence. Before we present the analysis, we first discuss some data on serology test performance to inform our inference.

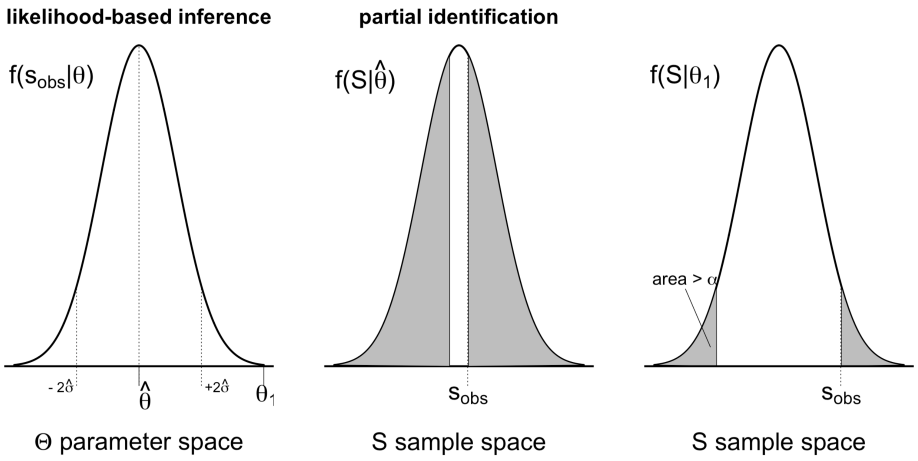

Figure 2: Illustration of main difference between standard methods of inference, and the partial identification method in this paper. *Left:* In standard methods, inference is typically based on the likelihood $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})$ as a function over Θ, and around some mode $\hat{\boldsymbol{\theta}}$. Parameter values with low likelihood, such as $\boldsymbol{\theta}_1$, are not included in the confidence set. *Middle & right:* In our method, inference is based on the entire distribution function $f(\boldsymbol{s}|\boldsymbol{\theta})$ over the statistic parameter space, $\mathbb{S}$. As such, $\hat{\boldsymbol{\theta}}$ will usually be in the confidence set (middle plot). Moreover, $\boldsymbol{\theta}_1$ will also be in the confidence set if $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_1)$ is high relative to the rest of $f(\boldsymbol{s}|\boldsymbol{\theta}_1)$ even when $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_1)$ is small relative to the mode $f(\boldsymbol{s}_{\text{obs}}|\hat{\boldsymbol{\theta}})$ (right plot).

4.1 Serology test performance

An important aspect of serology studies are the test performance characteristics. As of May 2020, there are perhaps more than a hundred commercial serology tests in the US, but they can differ substantially across manufacturers and technologies. In our application, we use data from Bendavid et al. (2020), who applied a serology testing kit distributed by Premier Biotech. Bendavid et al. (2020) used validation test results provided by the test manufacturer, and also performed a local validation study in the lab. The combined validation study estimated a true positive rate of 80.3% (95% CI: 72.1%-87%), and a false positive rate of 0.5% (95% CI: 0.1%-1.7%).

To get an idea about how these performance characteristics relate to other available serology tests we use a dataset published by the FDA based on benchmarking 12 other testing kits to grant emergency use authorization (EUA) status. The dataset is summarized in Table 2. We see that the characteristics of the testing kit used by Bendavid et al. (2020) are compatible with the FDA data shown in the table. For example, a true positive rate of 80% is below the mean and median of the point estimates in the FDA dataset. A false positive rate of 0.5% falls between the median and mean of the respective FDA point estimates. A reason for this skewness is likely the existence of one outlier testing kit that performs notably worse than the others (Chembio Diagnostic Systems). Removing this datapoint brings the mean false positive rate down to 0.6%, very close to the estimate provided by Bendavid et al. (2020).

Test characteristic		min	median	mean	max
true positive rate (q)	point estimate	77.4%	91.1%	90.6%	100%
	95% CI, low endpoint	60.2%	81%	80.1%	95.8%
	95% CI, high endpoint	88.5%	96.7%	95.3%	100%
false positive rate (p)	point estimate	0%	0.35%	1.07%	5.6%
	95% CI, low endpoint	0%	0.1%	0.52%	2.7%
	95% CI, high endpoint	0.3%	1.6%	3.18%	11.1%

Table 2: Performance characteristics of 12 different testing kits granted with emergency authorization status by the FDA. *Source:* Author calculations based on publicly available FDA dataset at https://www.fda.gov/medical-devices/emergency-situations-medical-devices/eua-authorized-serology-test-performance?mod=article_inline.

4.2 Santa Clara study

In the Santa Clara study, Bendavid et al. (2020) report a validation study and main study, with $(N_{\mathrm{c}}^{-}, N_{\mathrm{c}}^{+}, N_{\mathrm{m}}) = (401, 197, 3330)$ participants, respectively. The observed test positives are $\boldsymbol{s}_{\mathrm{obs}} = (s_{\mathrm{c,obs}}^{-}, s_{\mathrm{c,obs}}^{+}, s_{\mathrm{m,obs}}) = (2, 178, 50)$, respectively. Given these data, we produce the 95% confidence sets for (p, q, π) following both procedures in (8) and (10) described in Section 3.3. In Figure 11 of Appendix C, we jointly plot all triples in the 3-dimensional space $\widehat{\Theta}_{0.95}$ of Equation (8), with additional coloring based on prevalence values. We see that the confidence set is a convex space tilting to higher prevalence values as the false positive rate of the test decreases. The true positive rate does not affect prevalence, as long as it stays in the range 80%-95%.

To better visualize the pairwise relationships between the model parameters, we also provide Figure 3 that breaks down Figure 11 into two subplots, one visualizing the pairs (π, p) and another visualizing the pairs (π, q). The figure visualizes both $\widehat{\Theta}_{0.95}$ and $\widehat{\Theta}_{0.95}^{\mathrm{alt}}$ to illustrate the differences between the two constructions. From Figure 3, we see that the Santa Clara study is not conclusive about Covid-19 prevalence. A prevalence of 0% is plausible, given a high enough false positive rate. However, if the true false positive rate is near its empirical value of 0.5%, as estimated by Bendavid et al. (2020), then the identified prevalence rate is estimated in the range 0.4%-1.8% in $\widehat{\Theta}_{0.95}$. Under this assumption, we see that $\widehat{\Theta}_{0.95}^{\mathrm{alt}}$ offers a sharper inference, as expected, with an estimated prevalence in the range 0.7%-1.5%. Even though, strictly speaking, the statistical evidence is not sufficient here for definite inference on prevalence, we tend to favor the latter interval because (i) common sense precludes 0% prevalence in the Santa Clara county (total pop. of about 2 million); (ii) the interval generally agrees with the test performance data presented earlier, and (iii) it is still in the low end compared to prevalence estimates from other serology studies (see Table 1). Regardless, pinning down the false positive rate is important for estimating prevalence, especially when prevalence is as low as it appears to be in the Santa Clara study. Roughly speaking, a decrease of 1% in the false positive rate implies an increase of 1.3% in prevalence.

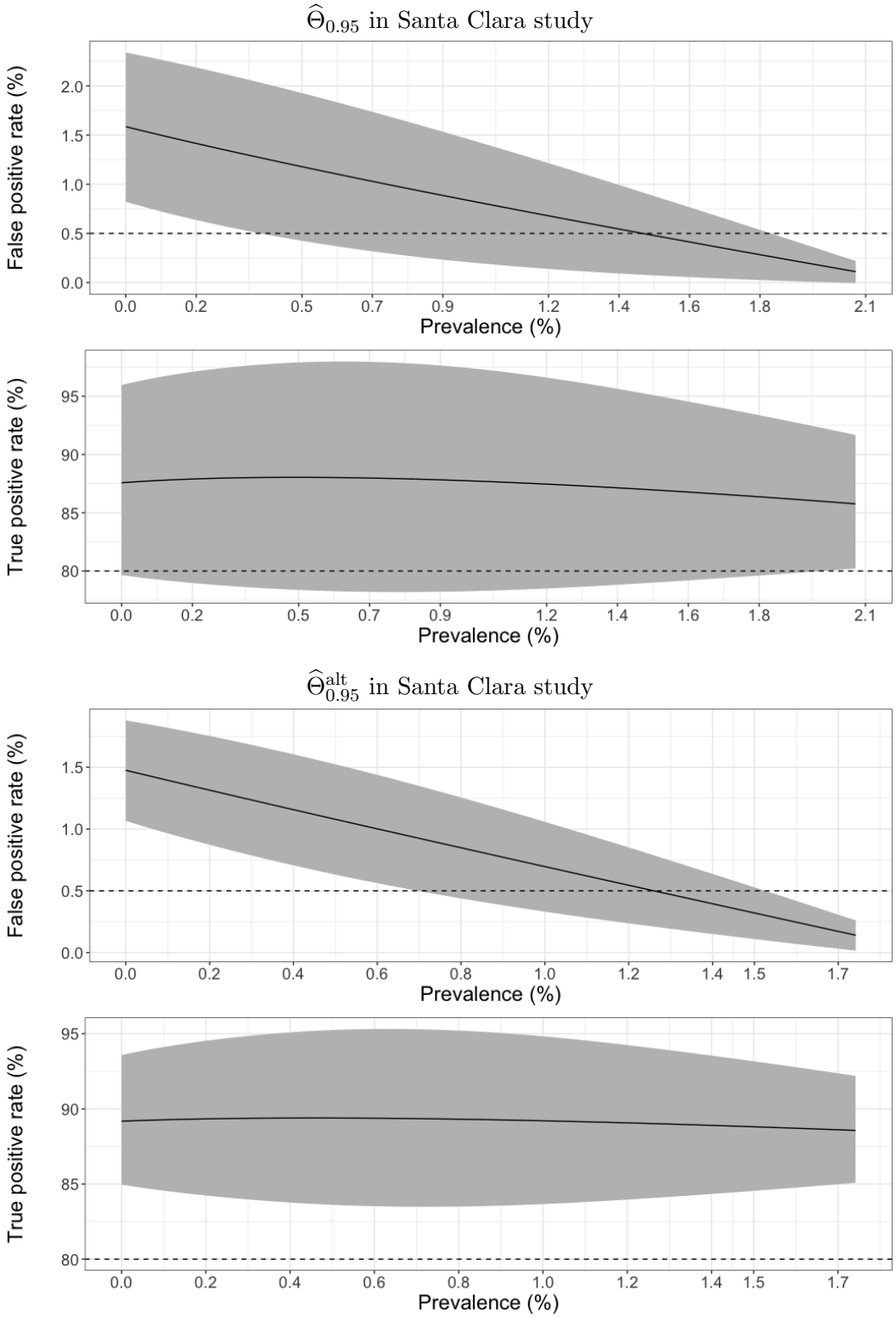

Figure 3: Confidence sets $\widehat{\Theta}_{0.95}$ and $\widehat{\Theta}^{\text{alt}}_{0.95}$ of the basic and the alternate 95% confidence set construction in Equations (8) and (10), respectively, for the Santa Clara study. Each confidence set is broken down to pairwise relationships. All values in a shaded area are in the corresponding confidence set. The horizontal dashed line corresponds to the empirical estimate of the false positive rate ($\hat{p} = 0.5\%$) of the serology test; the dashed line corresponds to the estimate of its true positive rate ($\hat{q} = 80\%$); see Section 4.1 for details.

4.3 Santa Clara study: Comparison to other methods

In this section, we aim to discuss how our method practically compares to more standard methods using data from the Santa Clara study. In our comparison we include Bayesian methods, a classical likelihood ratio-based test, and the Monte Carlo-based approach to partial identification proposed by Chen et al. (2018).[8]

4.3.1 Comparison to standard Bayesian and bootstrap-based methods

Due to initial criticism, the authors of the original Santa Clara study published a revision of their work, where they use a bootstrap procedure to calculate confidence intervals for prevalence in the range 0.7%-1.8%.[9] Some recent Bayesian analyses report wider prevalence intervals in the range 0.3%-2.1% (Gelman and Carpenter, 2020). In another Bayesian multi-level analysis, Levesque and Maybury (2020) report similar findings but mention that posterior summarization here may be subtle, since the posterior density of prevalence in their specification includes 0%. These results are in agreement with our analysis in the previous section only if we assume that the true false positive rate of the serology test was near its empirical estimate ($\sim$0.5%). We discussed intuitively the reason for such discrepancy in Section 3.4, where we argued that standard methods typically do inference "around the mode" of the likelihood, and may thus miscalculate the amount of statistical information hidden in the tails.

For a numerical illustration, consider two parameter values, namely $\boldsymbol{\theta}_1 = (0.5\%, 90\%, 1.2\%)$ and $\boldsymbol{\theta}_2 = (1.5\%, 80\%, 0\%)$, where the components denote the false positive rate, true positive rate, and prevalence, respectively. In the Santa Clara study, $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_1) = 2.2 \times 10^{-3}$ and $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_2) = 9.58 \times 10^{-8}$, that is, $\boldsymbol{\theta}_2$ (which implies 0% prevalence) maps to a likelihood value that is many orders of magnitude smaller than $\boldsymbol{\theta}_1$. In fact, $\boldsymbol{\theta}_1$ is close to the mode of the likelihood, and so frequentist or Bayesian inference is mostly based around that mode, ignoring the tails of the likelihood function, such as $\boldsymbol{\theta}_2$. For our method, however, the small value of $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_2)$ is more-or-less irrelevant — what matters is how this value compares to the entire distribution $f(\boldsymbol{s}|\boldsymbol{\theta}_2)$. It turns out that $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_2)\nu(f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_2), \boldsymbol{\theta}_2) = 0.137$, that is, 13.7% of the mass of $f(\boldsymbol{s}|\boldsymbol{\theta}_2)$ is below the observed level $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_2) = 9.58 \times 10^{-8}$. As such, $\boldsymbol{\theta}_2$ cannot be rejected at the 5% level (see also Appendix D). This highlights the key difference of our procedure compared to frequentist or Bayesian procedures. More generally, we expect to see such important differences between the inference from our method and the inference from other more standard methods in settings with small samples or poor identification (e.g., non-separable, multimodal likelihood).

[8] All these methods are fully implemented in the accompanying code at https://github.com/ptoulis/covid-19.
[9] Link: https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v2.full.pdf.

4.3.2 Comparison to likehood ratio test

As briefly described in Section 3.1, our test is related to the likelihood ratio test (Lehmann and Romano, 2006, Chapter 3). Here, we study the similarities and differences between the two tests, both theoretically and empirically through the Santa Clara study. Specifically, consider testing a null hypothesis that the true parameter is equal to some value $\boldsymbol{\theta}$ using the likelihood ratio statistic,

$$t(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}) = \frac{f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}' \neq \boldsymbol{\theta}} f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}')}. \tag{11}$$

Since f is known analytically from Equation (4), the null distribution of $t(\boldsymbol{S}|\boldsymbol{\theta})$ can be fully simulated. An exact p-value can then be obtained by comparing this null distribution with the observed value $t(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})$. We can see that this method is similar to ours in the sense that both methods use the full density $f(\boldsymbol{S}|\boldsymbol{\theta})$ in the test, and both are exact. The main difference, however, is that our method is using a summary of the density values $f(\boldsymbol{S}|\boldsymbol{\theta})$ that are below the observed value $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})$, which avoids the expensive (and sometimes numerically unstable) maximization in the denominator of the likelihood ratio test in (11). Our proposed method turns out to be orders of magnitude faster than the likelihood ratio approach as we get 50-fold to 200-fold speedups in our setup — see Section 5.1 for a more detailed comparison in computational efficiency.

To efficiently compare the inference between the two tests, we sampled 5,000 different parameter values from inside $\widehat{\Theta}_{0.95}$ — i.e., the 95% confidence set from the basic test in Equation (8) — and 5,000 parameter values from $\Theta \setminus \widehat{\Theta}_{0.95}$, and then calculated the overlap between the test decisions. The likelihood ratio test rejected 3% of the values from the first set, and 98% of the values from the second set, indicating a good amount of overlap between the two tests. The correlation between the p-value from the likelihood ratio test, and the values $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\nu(f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}), \boldsymbol{\theta})$, which our basic test uses to make a decision in Equation (8), was equal to 0.94. The correlation with the alternative confidence set construction is 0.90, using instead the values $\mathbb{I}\{f(\boldsymbol{s}|\boldsymbol{\theta}) \leq f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\}f(\boldsymbol{s}|\boldsymbol{\theta})$ in the above calculation. Since the likelihood ratio test is exact, these results suggest that our test procedures are generally high-powered.

In Figures 7 and 8 of Appendix A, we plot the 95% confidence sets from the likelihood ratio test described above for the Santa Clara study and the LA county study (of the following section). The estimated prevalence is 0%-1.9% for Santa Clara, which is shorter than $\widehat{\Theta}_{0.95}$ but wider than $\widehat{\Theta}^{\text{alt}}_{0.95}$, as reported earlier; the same holds for LA county. As with our method, prevalence here is estimated through direct projection of the confidence set, which may be conservative. It is also possible that with more samples the likelihood ratio test could achieve the same interval as $\widehat{\Theta}^{\text{alt}}_{0.95}$ (we used only 100 samples), but this would come at an increased computational cost. Overall, the likelihood ratio test produces very similar results to our method, but it is not as efficient computationally.

4.3.3 Comparison to Monte Carlo confidence set method of Chen et al. (2018)

In recent work, Chen et al. (2018) proposed a Monte Carlo-based method of inference in partially identified models. The idea is to sample from a quasi-posterior distribution, and then calculate q_n, the 95% percentile of $\{f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}^{(j)}),\ j = 1, \ldots\}$, where $\boldsymbol{\theta}^{(j)}$ denotes the j-th sample from the posterior. The 95% confidence set is then defined as:

$$\widehat{\Theta} = \{\boldsymbol{\theta} \in \Theta : f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}) \geq q_n\}. \tag{12}$$

We implemented this procedure with an MCMC chain that appears to be mixing well — see Appendix B and Figure 9 for details. The 95% confidence set, $\widehat{\Theta}$, is given in Figure 10 of Appendix B. Simple projection, yields a prevalence in the range 0.9%-1.43%. This suggests that our MCMC "spends more time" around the mode of the likelihood, which we back up with numerical evidence in Appendix B. Finally, we also tried Procedure 3 of Chen et al. (2018), which does not require MCMC simulations but is generally more conservative. Prevalence was estimated in the range 0.12%-1.65%, which is comparable to our method and the likelihood ratio test.

4.4 LA County study

Next, we analyze the results from a recent serology study in Los Angeles county, which estimated a prevalence of 4.1% over the entire county population.[10] We use the same validation study as before since this study was executed by the same team as the Santa Clara one. Here, the main study had $N_{\text{m}} = 846$ participants with $s_{\text{m,obs}} = 35$ positives.[11] For inference, we only use the alternative construction, $\widehat{\Theta}^{\text{alt}}_{0.95}$, of Equation (10) to simplify exposition. The results are shown in Figure 4.

In contrast to the Santa Clara study, we see that the results from this study are conclusive. The prevalence rate is estimated in the range 1.7%-5.2%. If the false positive rate is, for example, closer to its empirical estimate (0.5%) then the identified prevalence is relatively high, somewhere in the range 3%-5.2%. We also see that the true positive rate is estimated in the range 85%-95%, which is higher than the empirical point estimate of 80% provided by Bendavid et al. (2020). In fact, the empirical point estimate is not even in the 95% confidence set. Finally, as an illustration, we combine the data from the Santa Clara and LA county studies. The assumption is that the characteristics of the tests used in both studies were identical. The results are shown in Figure 12 of Appendix C. We see that 0% prevalence is consistent with the combined study as well. Furthermore, prevalence values higher than 2.5% do not seem plausible in the combined data.

[10] http://publichealth.lacounty.gov/phcommon/public/media/mediapubhpdetail.cfm?prid=2328

[11] This number was not reported in the official study announcement mentioned above. It was reported in a Science article referencing one of the authors of the study: https://www.sciencemag.org/news/2020/04/antibody-surveys-suggesting-vast-undercount-coronavirus-infections-may-be-unreliable.

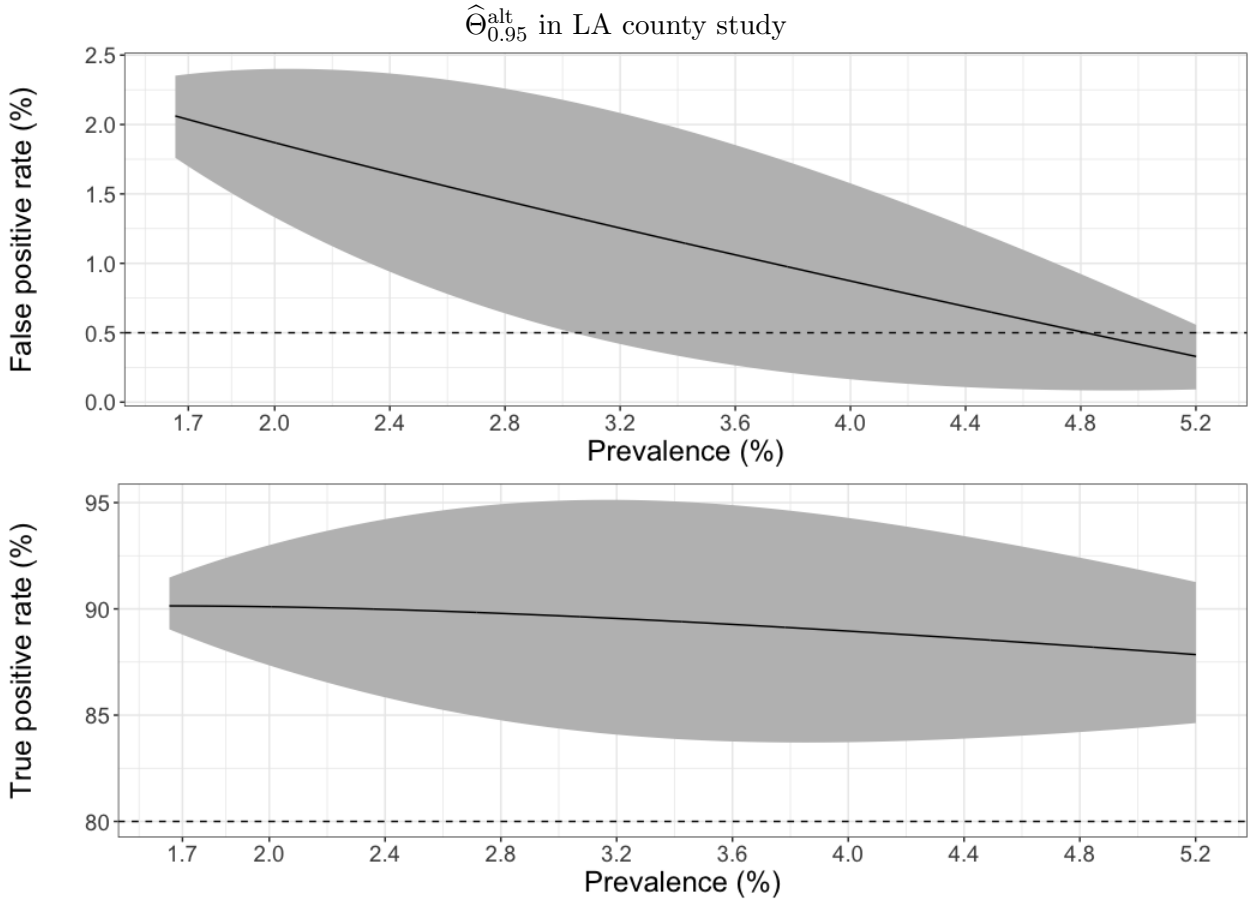

Figure 4: Visualization of $\widehat{\Theta}^{\text{alt}}_{0.95}$ for the LA county study. We see that the evidence for Covid-19 prevalence are stronger than the Santa Clara study. Prevalence is estimated in the range 1.7%-5.2%. This is shortened to 3%-5.2% if we assume a 0.5% false positive rate for the antibody test.

4.5 New York study

Recently, a quasi-randomized study was conducted in New York state, including NYC, which sampled individuals shopping in grocery stores. Details about this study were not made available. Here, we assume that the medical testing technology used was the same as in the Santa Clara and LA county studies, or at least similar enough that the comparison remains informative.

Under this assumption, we can use the same validation study as before, with $(N_{\text{c}}^-, N_{\text{c}}^+) = (401, 197)$ participants in the validation study, and $(s^-_{\text{c,obs}}, s^+_{\text{c,obs}}) = (2, 178)$ positives, respectively. The main study in New York had $N_{\text{m}} = 3000$ participants with $s_{\text{m,obs}} = 420$ observed test positives.[12] The $\widehat{\Theta}^{\text{alt}}_{0.95}$ confidence set on this dataset is shown in Figure 5. We see that the evidence in this study is much stronger than the Santa Clara/LA county studies with an estimated prevalence in the range 12.9%-16.6%. The true positive rate is now an important identifying parameter in the sense that knowing its true value could narrow down the confidence set even further.

[12] https://www.nytimes.com/2020/04/23/nyregion/coronavirus-antibodies-test-ny.html

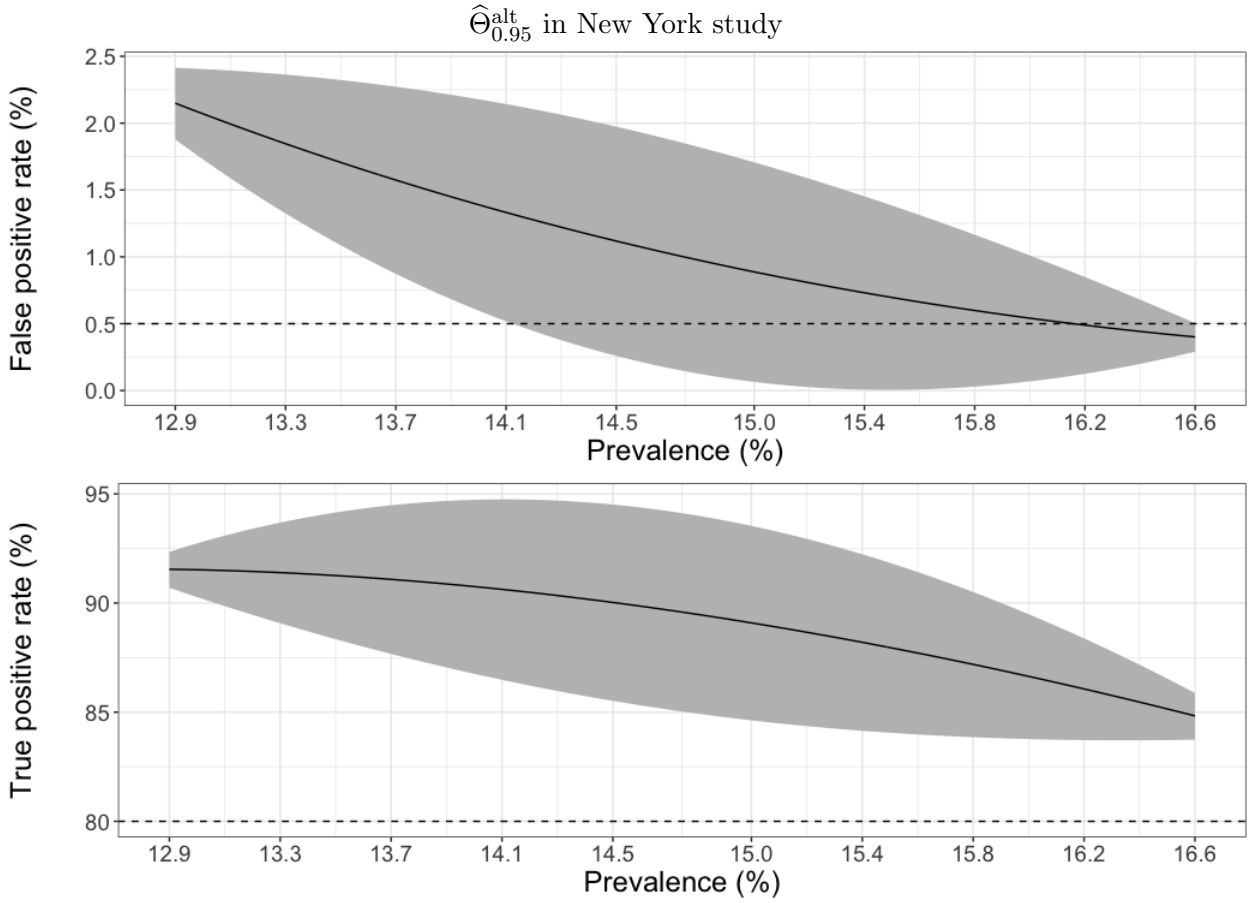

Figure 5: Visualization of confidence set, $\widehat{\Theta}^{\text{alt}}_{0.95}$, for New York State study. We see that this study gives strong evidence for prevalence in the range 13%-16.6%.

Finally, in Figure 13 of Appendix C we present prevalence estimates for a combination of all datasets presented so far, while using both constructions, $\widehat{\Theta}_{0.95}$ and $\widehat{\Theta}^{\text{alt}}_{0.95}$, to illustrate their differences. As mentioned earlier, this requires the assumption that the antibody testing kits used in all three studies had identical specifications, or at least very similar so that the comparison remains informative. This assumption is most likely untenable given the available knowledge. However, we present the results there for illustration and completeness. The general picture in the combined study is a juxtaposition of earlier findings. For example, both false and positive rates are now important for identification. The identified prevalence is in the range 5.2%-8.2% in $\widehat{\Theta}^{\text{alt}}_{0.95}$ (and 3.2%-8.9% in $\widehat{\Theta}_{0.95}$). These numbers are larger than the Santa Clara/LA county studies but smaller than the New York study.

5 Discussion

5.1 Computation

The procedure described in Section 3.3 is computationally intensive for two main reasons. First, we need to consider all values of $\boldsymbol{\theta} \in \Theta$, which is a three-dimensional grid. Second, given some $\boldsymbol{\theta}$, we need to calculate $f(\boldsymbol{s}|\boldsymbol{\theta})$ for each $\boldsymbol{s} \in \mathbb{S}$, which is also a three-dimensional grid.

To deal with the first problem we can use parallelization, since the test decisions in step 3 of our procedure are independent of each other. For instance, the results in Section 4 were obtained in a computing cluster (managed by Slurm) comprised of 500 nodes, each with x86 architecture, 64-bit processors, and 16GB of memory. The total wall clock time to produce all results of the previous section was about 1 hour. The results for, say, the Santa Clara study can be obtained in much shorter time (a few minutes) because they contain few positive test results. To address the second computational bottleneck we can exploit the independence property between $S_{\text{c}}^-, S_{\text{c}}^+$, and S_{m}, as shown in the product of Equation (4). Since any zero term in this product implies a zero value for f, we can ignore all individual term values that are very small. Through numerical experiments, we estimate that this computational trick prunes on average 97% of $\mathbb{S}$ leading to a significant computational speedup. For example, to test one single value $\boldsymbol{\theta} \in \Theta$ takes about 0.25 seconds in a typical high-end laptop, which is a 200-fold speedup compared to 50 seconds required by the likelihood ratio test of Section 4.3.2 — see Appendix A for more details.

5.2 Extrapolation to general population

As mentioned earlier, prevalence π in Equation (2) is a finite-population estimand, that is, it is a number that refers to the particular population in the study. Theorem 1 shows that our procedure is valid for π only under Assumption (A4). However, to extrapolate to the general population we generally need to assume that

$$\mathcal{I}_{\text{c}}^-, \mathcal{I}_{\text{c}}^+, \mathcal{I}_{\text{m}} \text{ are random samples from the population.} \tag{A3}$$

This is currently an untenable assumption. For example, in the Santa Clara study the population of middle-aged white women was overrepresented, while the population of Asian or Latino communities was underrepresented. The impact from such selection bias on the inferential task is very hard to ascertain in the available studies. Techniques such as post-stratification or reweighing can help, but at this early stage any extrapolation using distributional assumptions would be too speculative. However, selection bias is a well-known issue among researchers, and can be addressed as widespread and carefully designed antibody testing catches on. We leave this for future work.

6 Concluding Remarks

In this paper, we presented a partial identification method for estimating prevalence of Covid-19 from randomized serology studies. The benefit of our method is that it is valid in finite samples, as it does not rely on asymptotics, approximations or normality assumptions. We show that some recent serology studies in the US are not conclusive (0% prevalence is in the 95% confidence set). However, the New York study gives strong evidence for high prevalence in the range 12.9%-16.6%. A combination of all datasets shifts this range down to 5.2%-8.2%, under a test uniformity assumption. Looking ahead, we hope that the method developed here can contribute to a more robust analysis of future Covid-19 serology tests.

7 Acknowledgments

I would like to thank Guanglei Hong, Ali Hortascu, Chuck Manski, Casey Mulligan, Joerg Stoye, and Harald Uhlig for useful suggestions and feedback. Special thanks to Connor Dowd for his suggestion of the alternative construction (10), and to Elie Tamer for various important suggestions. Finally, I gratefully acknowledge support from the John E. Jeuck Fellowship at Booth School of Business.

References

Alvarez, F. E., Argente, D. and Lippi, F. (2020). A simple planning problem for covid-19 lockdown. Tech. rep., National Bureau of Economic Research.

Baggett, T. P., Keyes, H., Sporn, N. and Gaeta, J. M. (2020). Prevalence of sars-cov-2 infection in residents of a large homeless shelter in boston. *JAMA*.

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra, R., Tedrow, J. et al. (2020). Covid-19 antibody seroprevalence in santa clara county, california. *medRxiv*.

Beresteanu, A., Molchanov, I. and Molinari, F. (2012). Partial identification using random set theory. *Journal of Econometrics*, **166** 17–32.

Chen, X., Christensen, T. M. and Tamer, E. (2018). Monte carlo confidence sets for identified sets. *Econometrica*, **86** 1965–2018.

Chernozhukov, V., Hong, H. and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, **75** 1243–1284.

Doi, A., Iwata, K., Kuroda, H., Hasuike, T., Nasu, S., Kanda, A., Nagao, T., Nishioka, H., Tomii, K., Morimoto, T. et al. (2020). Estimation of seroprevalence of novel coronavirus disease (covid-19) using preserved serum at an outpatient setting in kobe, japan: A cross-sectional study. *medRxiv*.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H., Coupland, H., Mellan, T., Zhu, H., Berah, T., Eaton, J., Perez Guzman, P. et al. (2020). Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries.

Garcia-Basteiro, A. L. et al. (2020). Seroprevalence of antibodies against sars-cov-2 among health care workers in a large spanish reference hospital. *medRxiv*.

Gelman, A. and Carpenter, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *medRxiv*.

Honoré, B. E. and Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, **74** 611–629.

Hortaçsu, A., Liu, J. and Schwieg, T. (2020). Estimating the fraction of unreported infections in epidemics with a known epicenter: an application to covid-19. Tech. rep., National Bureau of Economic Research.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, **72** 1845–1857.

Kaido, H., Molinari, F. and Stoye, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica*, **87** 1397–1432.

Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Del Vecchio, C., Rossi, L., Manganelli, R., Loregian, A., Navarin, N. et al. (2020). Suppression of covid-19 outbreak in the municipality of vo, italy. *medRxiv*.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Levesque, J. and Maybury, D. W. (2020). A note on covid-19 seroprevalence studies: a meta-analysis using hierarchical modelling. *medRxiv*.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W. and Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*.

Liang, Y., Liang, J., Zhou, Q., Li, X., Lin, F., Deng, Z., Zhang, B., Li, L., Wang, X., Zhu, H. et al. (2020). Prevalence and clinical features of 2019 novel coronavirus disease (covid-19) in the fever clinic of a teaching hospital in beijing: a single-center, retrospective study. *medRxiv*.

Lu, F. S., Nguyen, A., Link, N. and Santillana, M. (2020). Estimating the prevalence of covid-19 in the united states: Three complementary approaches.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F. (2007). Partial identification of counterfactual choice probabilities. *International Economic Review*, **48** 1393–1410.

Manski, C. F. (2010). Partial identification in econometrics. In *Microeconometrics*. Springer, 178–188.

Manski, C. F. and Molinari, F. (2020). Estimating the covid-19 infection rate: Anatomy of an inference problem. Tech. rep., National Bureau of Economic Research.

Romano, J. P. and Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, **138** 2786–2807.

Romano, J. P. and Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica*, **78** 169–211.

Spellberg, B., Haddix, M., Lee, R., Butler-Wu, S., Holtom, P., Yee, H. and Gounder, P. (2020). Community prevalence of sars-cov-2 among patients with influenzalike illnesses presenting to a los angeles medical center in march 2020. *JAMA*.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, **77** 1299–1315.

Stringhini, S., Wisniak, A., Piumatti, G., Azman, A. S., Lauer, S. A., Baysson, H., De Ridder, D., Petrovic, D., Schrempft, S., Marcus, K. et al. (2020). Repeated seroprevalence of anti-sars-cov-2 igg antibodies in a population-based sample from geneva, switzerland. *medRxiv*.

Sutton, D., Fuchs, K., D'alton, M. and Goffman, D. (2020). Universal screening for sars-cov-2 in women admitted for delivery. *New England Journal of Medicine*.

Tamer, E. (2010). Partial identification in econometrics. *Annu. Rev. Econ.*, **2** 167–195.

Venkatesan, P. (2020). Estimate of covid-19 case prevalence in india based on surveillance data of patients with severe acute respiratory illness. *medRxiv*.

Wooldridge, J. and Imbens, G. (2007). What's new in econometrics? lecture 9: partial identification. *NBER Summer Institute*, **9** 2011.

Yadlowsky, S., Shah, N. and Steinhardt, J. (2020). Estimation of sars-cov-2 infection prevalence in santa clara county. *medRxiv*.

Appendix

A More details on the likelihood ratio test

The concrete testing procedure for the likelihood ratio test of Section 4.3.2 is as follows.

1. Define the test statistic:
$$t(\boldsymbol{s}|\boldsymbol{\theta}) = \frac{f(\boldsymbol{s}|\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}' \neq \boldsymbol{\theta}} f(\boldsymbol{s}|\boldsymbol{\theta}')}.$$
2. Calculate the observed value $t_{\text{obs}} = t(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})$.
3. Sample $\{\boldsymbol{s}^{(j)}, j = 1, \ldots, r\}$ from $f(\boldsymbol{S}|\boldsymbol{\theta})$ (we set $r = 1,000$ samples).
4. Calculate the one-sided p-value as $\frac{1}{r}\sum_{j=1}^{r} \mathbb{I}\{t(\boldsymbol{s}^{(j)}|\boldsymbol{\theta}) \geq t_{\text{obs}}\}$.

For the maximization in the denominator of $t(\boldsymbol{s}|\boldsymbol{\theta})$ we use the standard BFGS algorithm on a natural re-parameterization. Specifically, we define the natural parameters as follows:

$$\psi_0 = \text{logit}(p); \psi_1 = \text{logit}(q); \psi_2 = \text{logit}(\pi),$$

where $(p, q, \pi) \equiv \boldsymbol{\theta}$ are the original model parameters, i.e., false positive rate, true positive rate, and prevalence, respectively; and $\text{logit}(z) = \log(z/(1-z)), z \in (0, 1)$. To avoid numerical instabilities we define $\text{logit}(0) = \log(\epsilon/(1-\epsilon))$ and $\text{logit}(1) = \log((1-\epsilon)/\epsilon)$, where ϵ is a small constant, e.g., $\epsilon =$1e-8. Since the natural parameter $\boldsymbol{\psi} = (\psi_0, \psi_1, \psi_2)$ is unconstrained, the optimization routine becomes faster and easier; mapping back from $\boldsymbol{\psi}$ to $\boldsymbol{\theta}$ is also straightforward.

The maximization takes about 0.05 seconds of wall-clock time in a typical high-end laptop.[13] It therefore takes a total of 50 seconds to test one single hypothesis based on 1,000 samples of the likelihood ratio. In contrast, our partial identification method takes 0.25 seconds of wall-clock time to test the same single null hypothesis, a 200-fold speedup. As explained in Section 5.1 this is because the computation of $f(\boldsymbol{s}|\boldsymbol{\theta})$ can be done very efficiently due to the decomposition of f into three independent terms in Equation (4).

Since the likelihood ratio test cannot be fully implemented, we chose to sample randomly 5,000 parameter values from the basic confidence set, $\widehat{\Theta}_{0.95}$, and 5,000 values from $\Theta \setminus \widehat{\Theta}_{0.95}$, and then test each value using the likelihood ratio test. The idea is to explore the agreement of the two tests. The overlap between the likelihood ratio test decisions and the basic construction is 97.3% for the values from $\widehat{\Theta}_{0.95}$, and 97.7% for the values from $\Theta \setminus \widehat{\Theta}_{0.95}$. There was even more agreement with

[13] CPU: Intel(R) Core(TM) i7-8559U CPU at 2.70GHz; Memory: 16 GB 2133 MHz LPDDR3.

the alternative construction, specifically 97.3% and 99.6%, respectively. Figure 6 also shows some more detailed results. The x-axis represents the p-value generated in step 4 of the likelihood ratio test. The y-axis represents the "strength of evidence" calculated by our procedure. For the basic construction, $\widehat{\Theta}_{0.95}$, of Equation (8) the strength of evidence is defined as $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\nu(f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}), \boldsymbol{\theta})$ — the larger this value is, the stronger we reject. For the alternative construction, $\widehat{\Theta}^{\text{alt}}_{0.95}$, of Equation (10) the value is defined as $\sum_{\boldsymbol{s}\in\mathbb{S}} \mathbb{I}\{f(\boldsymbol{s}|\boldsymbol{\theta}) \leq f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\} f(\boldsymbol{s}|\boldsymbol{\theta})$. We see high correlation between the different tests (0.94 and 0.90, respectively).

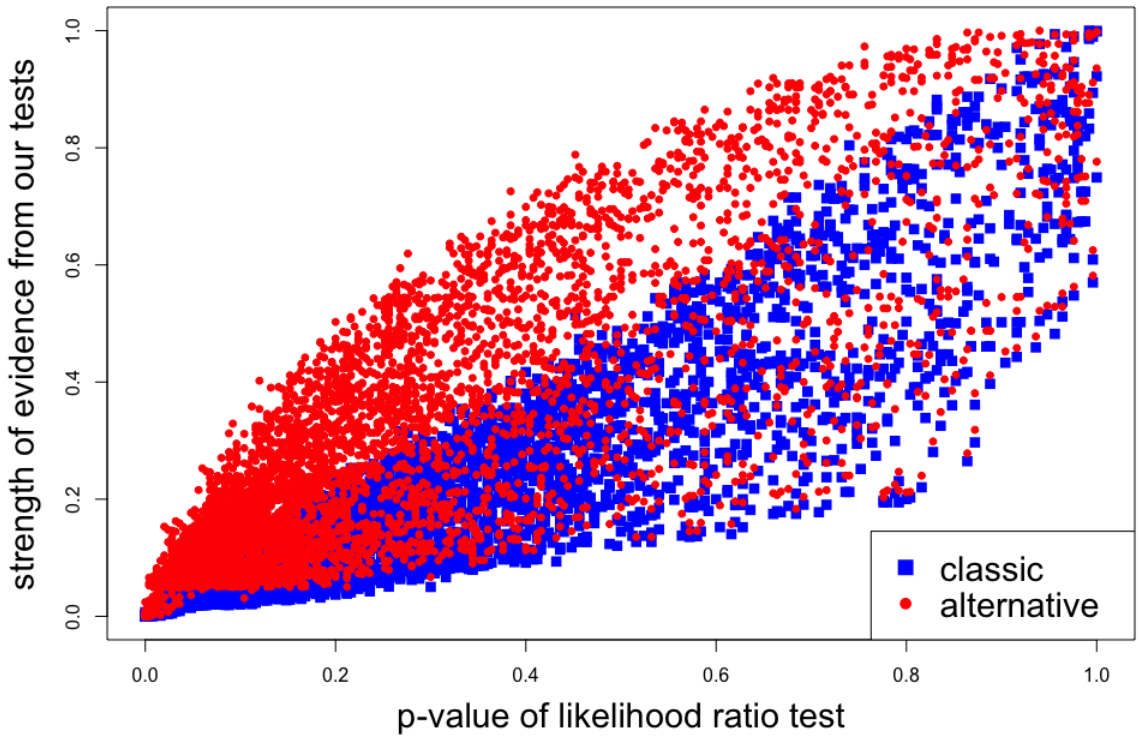

Figure 6: Relationship between likelihood ratio test and our testing procedures, namely the "classic" construction of Equation (8), and the alternative construction in Equation (10).

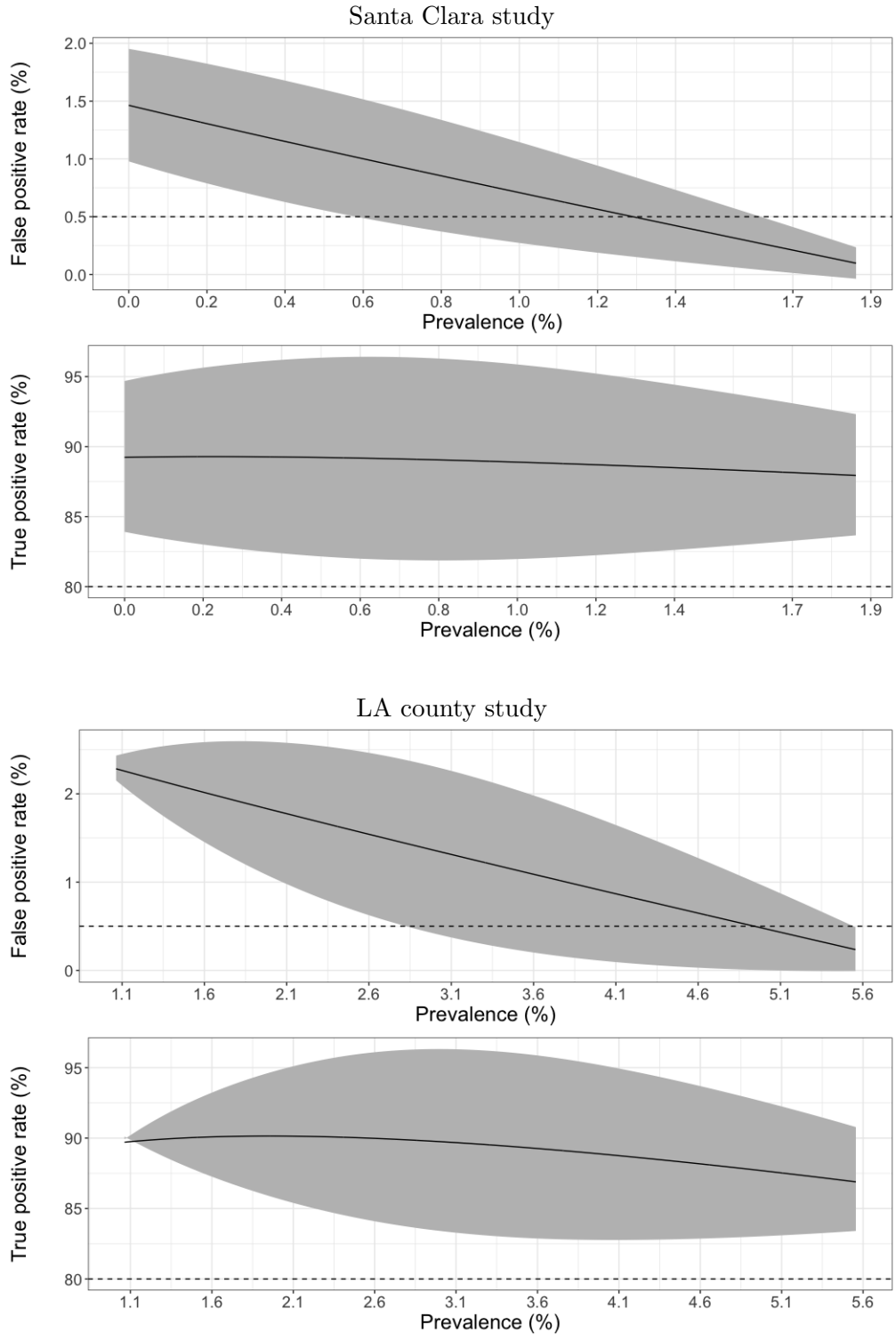

Figure 7: 95% confidence set from likelihood ratio test-based procedure of Section 4.3.2 for Santa Clara study (top) and LA county study (bottom).

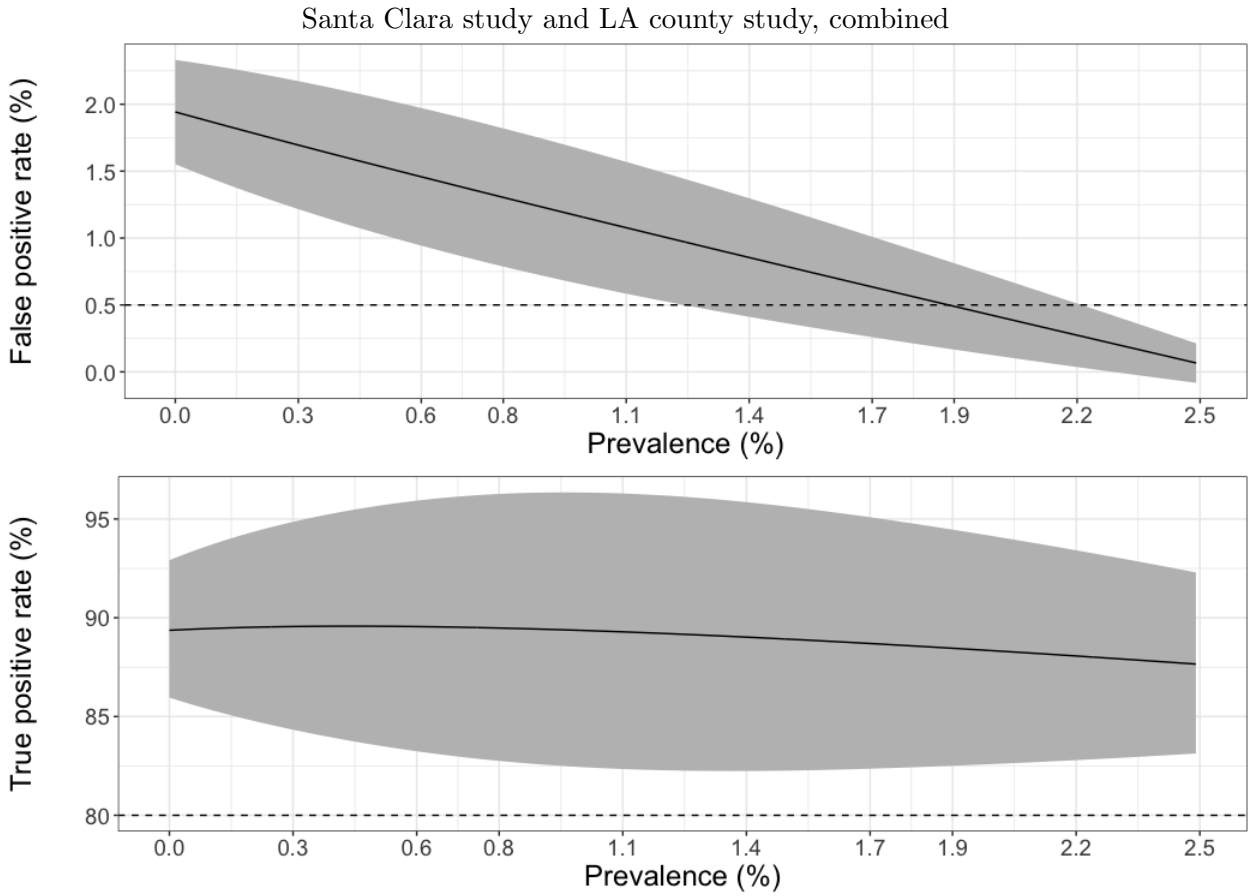

Figure 8: 95% confidence set from likelihood ratio test-based procedure of Section 4.3.2 for Santa Clara study and LA county study combined.

B More details on Monte Carlo confidence sets

To implement the method of Chen et al. (2018) we used the following procedure.

1. We defined the natural parameter, $\boldsymbol{\psi}$ as in the previous section, and imposed a uniform prior: $f(\psi) \propto 1$.

2. We implement the Metropolis-Hastings algorithm through a symmetric proposal distribution, $q(\boldsymbol{\psi}'|\boldsymbol{\psi}) \sim N(\widehat{\boldsymbol{\psi}}, \sigma^2 I)$, where $\widehat{\boldsymbol{\psi}}$ is the maximum-likelihood estimate.

3. We run the Metropolis-Hastings for 200,000 iterations and got samples from the posterior distribution $f(\boldsymbol{\psi}|\boldsymbol{s}_{\text{obs}}) \propto f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\psi})f(\boldsymbol{\psi})$. We discarded the first 20% of the posterior samples. Figure 9 shows that the MCMC chain appears to be mixing well.

4. We transformed the $\boldsymbol{\psi}$ samples back as $\boldsymbol{\theta}$ samples, and calculated q_n, the 95% percentile of $\{f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}^{(j)}),\ j = 1, \ldots\}$, where $\boldsymbol{\theta}^{(j)}$ denotes the j-th sample. The 95% confidence set is then defined as:
$$\widehat{\Theta} = \{\boldsymbol{\theta} \in \Theta : f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}) \geq q_n\}.$$

The 95% posterior credible interval (shown in the bottom panel of Figure 9) was 0.27%-1.56%, which is similar but slightly more narrow than the intervals from the Bayesian analyses described in Section 4.3. The 95% confidence set, $\widehat{\Theta}$, defined above is visualized in Figure 10. Simple projection, yields a prevalence in the range 0.9%-1.43%. We note that 0.9% corresponds to 32 true positives out of total 50 positives in the Santa Clara study. We can see from Figure 9 that this number corresponds to the mode of the posterior marginal distribution for prevalence. Given the symmetry of this posterior distribution around the mode, it is surprising that the low-end of the confidence set corresponds to 0.9% prevalence (i.e, 32 true positives out of 3,330 tests). We can explain this discrepancy numerically by checking that values higher than 32 (prevalence higher than 0.9%), that is, values on the right-end of the marginal posterior distribution, generally map to (much) higher likelihood values compared to values on the left-end. So, even though the marginal posterior looks symmetric, the likelihood values are not. Since, the confidence set $\widehat{\Theta}$ defined above is based on the likelihood values, it will be "skewed" towards higher values of $\boldsymbol{\theta}$.

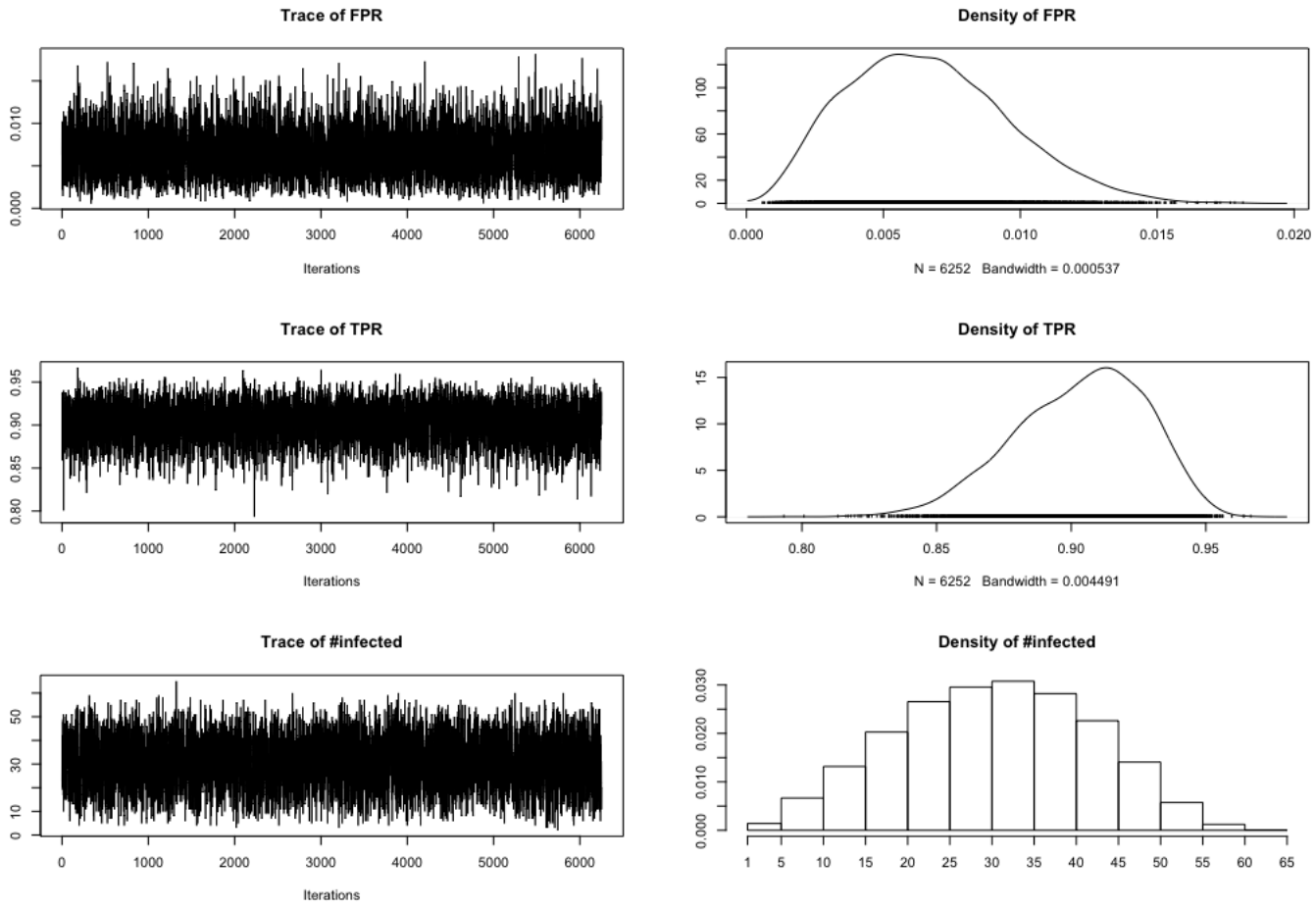

Figure 9: Mixing plots for the MCMC chain used to construct the 95% confidence set described in Section 4.3.3.

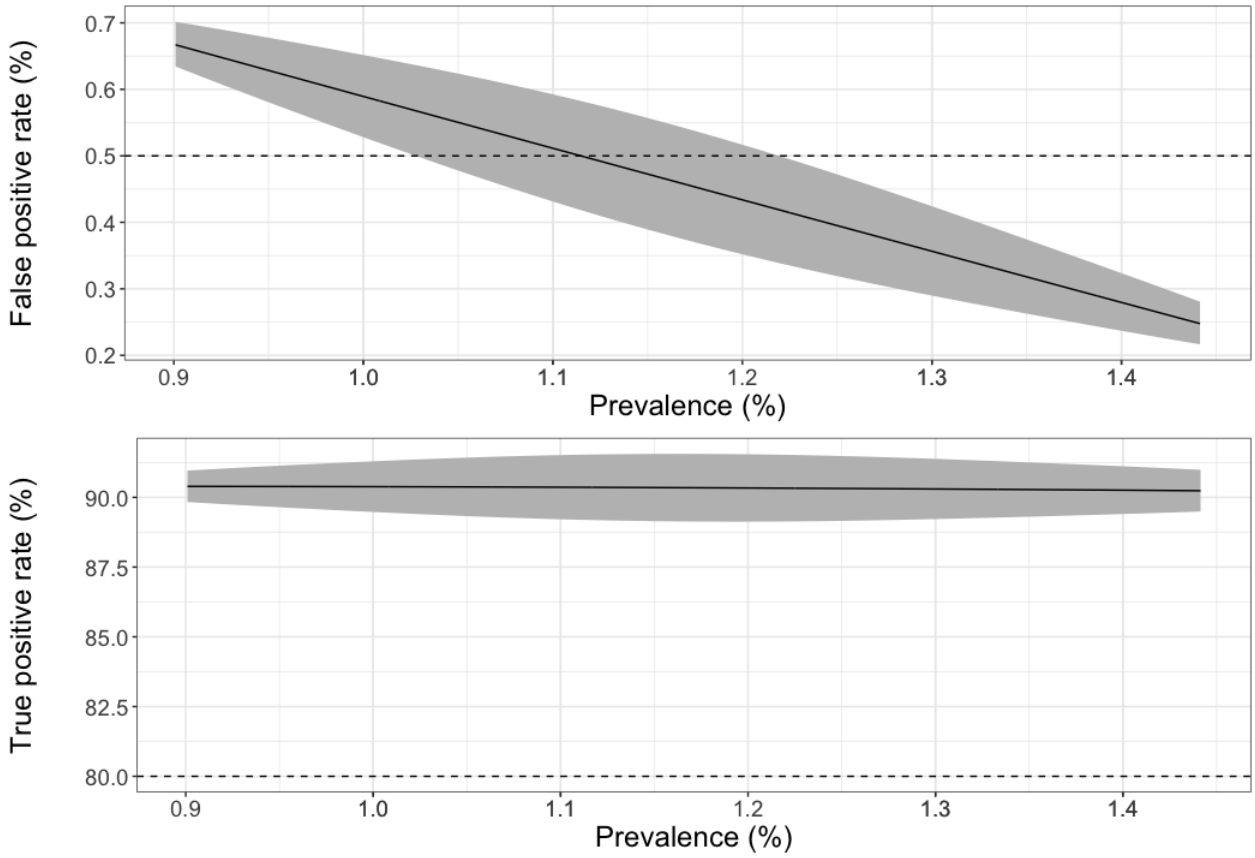

Figure 10: The MCMC-based 95% confidence set, $\widehat{\Theta}$, produced by Procedure 1 of Chen et al. (2018).

C Additional results from serology studies in the US

Here, we present additional results from our analysis on serology studies from the US. In particular, we combine datasets from various studies and analyze the results. This requires the assumption that the antibody testing kits used in all three studies had identical specifications, or at least very similar so that the comparison remains informative.

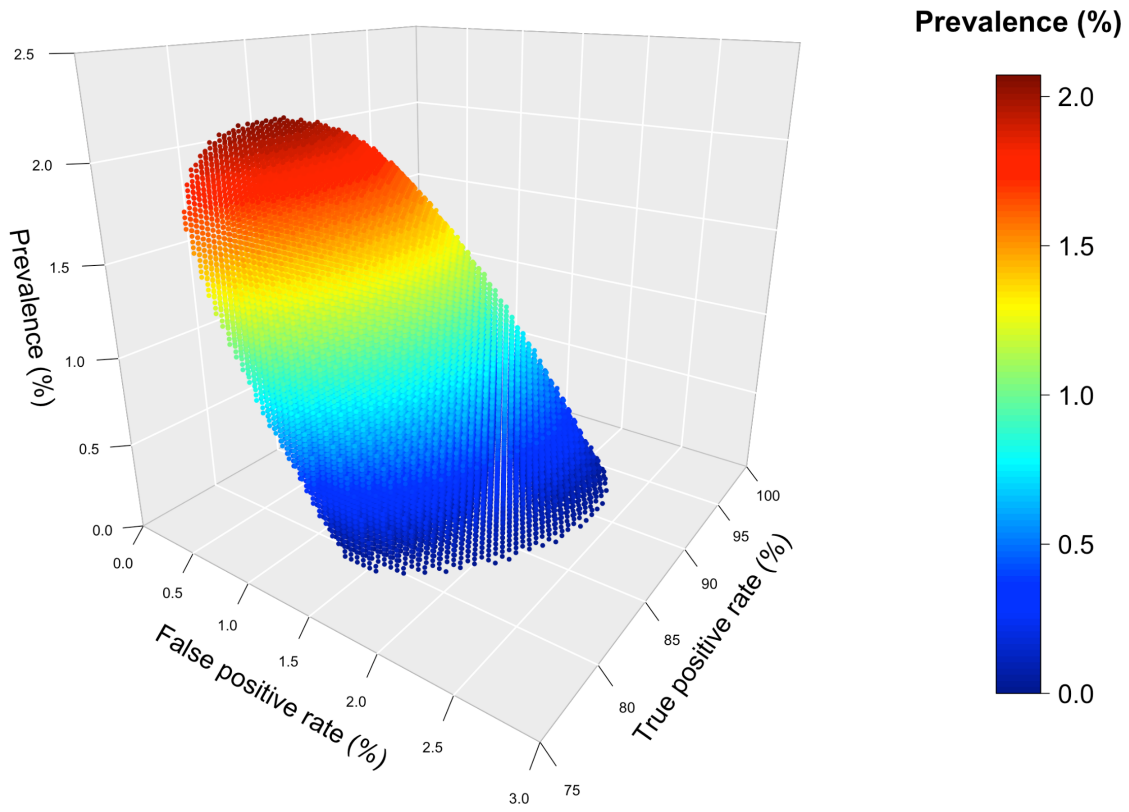

Figure 11: Confidence set, $\widehat{\Theta}_{0.95}$, for Santa Clara study, comprised of triples $\boldsymbol{\theta} = (p, q, \pi)$.

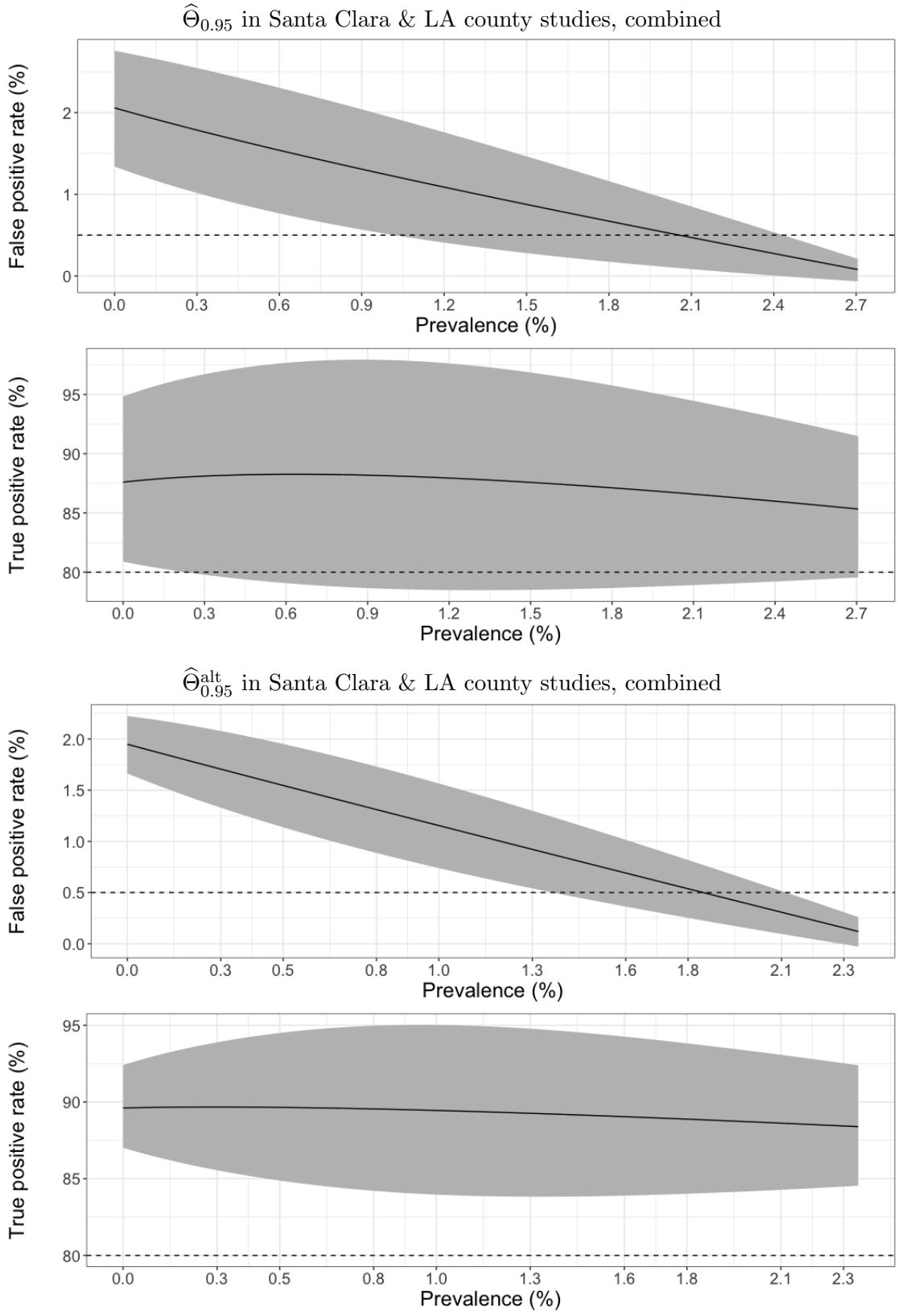

Figure 12: Visualization of confidence sets, $\widehat{\Theta}_{0.95}$ and $\widehat{\Theta}^{\text{alt}}_{0.95}$, for the Santa Clara and LA county studies, combined. The combined results are inconclusive (0% prevalence is in the confidence sets).

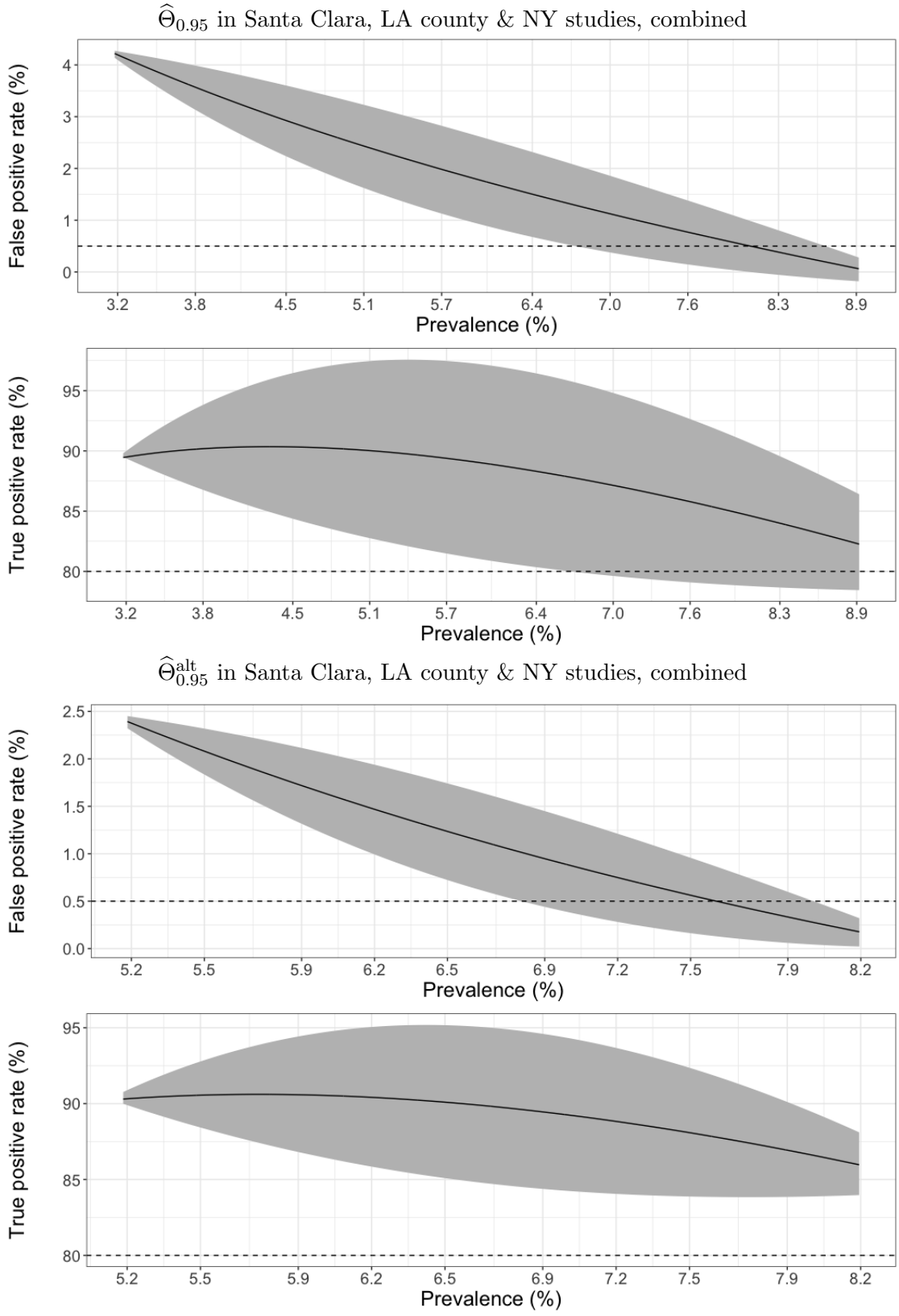

Figure 13: Visualization of confidence sets, $\widehat{\Theta}_{0.95}$ and $\widehat{\Theta}^{\text{alt}}_{0.95}$, for Santa Clara, LA county and New York studies, combined. Assuming the data combination is valid, the combined dataset estimates Covid-19 prevalence in the range 5.2%-8.2%.

D Numerical example illustrating differences with likelihood-based methods

Here, we illustrate the differences between our proposed inferential method and likelihood-based methods (both Bayesian and frequentist) through a simple numerical example. Suppose that $\mathbb{S}$ is such that $|\mathbb{S}| = N$, with N extremely large, and fix some parameter value $\boldsymbol{\theta}_1$ to test. Suppose also that the conditional density $f(\boldsymbol{S}|\boldsymbol{\theta}_1)$ is defined as:

$$f(\boldsymbol{s}_0|\boldsymbol{\theta}_1) = 0.95 - \epsilon_1, f(\boldsymbol{s}_1|\boldsymbol{\theta}_1) = \epsilon, \text{ and } f(\boldsymbol{s}|\boldsymbol{\theta}_1) = (0.05 - \epsilon + \epsilon_1)/(N-2), \ \forall \boldsymbol{s} \in \mathbb{S} \setminus \{\boldsymbol{s}_0, \boldsymbol{s}_1\}.$$

Set both ϵ and ϵ_1 to be infinitesimal values. As such, under $\boldsymbol{\theta}_1$ we observe $\boldsymbol{s}_0$ with probability roughly equal to 0.95, or $\boldsymbol{s}_1$ with some very small probability ϵ, or observe any other remaining value from $\mathbb{S}$ uniformly at random.

Suppose we observe $\boldsymbol{s}_{\text{obs}} = \boldsymbol{s}_1$ in the data. Should we reject or accept $\boldsymbol{\theta}_1$?

Since we can make ϵ arbitrarily small, an inferential method that focuses only on the likelihood function, would conclude that any $\boldsymbol{\theta} \in \Theta$ is more plausible than $\boldsymbol{\theta}_1$, as long as $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}) >> \epsilon$. Both frequentist and Bayesian methods would agree to such conclusion, and typically would perform inference around the mode of $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. However, our procedure makes a different conclusion, and actually accepts $\boldsymbol{\theta}_1$ (at the 5% level)! The reason is that

$$\sum_{\boldsymbol{s} \in \mathbb{S}} \mathbb{I}\{f(\boldsymbol{s}|\boldsymbol{\theta}_1) \leq f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_1)\} f(\boldsymbol{s}|\boldsymbol{\theta}_1) = \epsilon + \frac{0.05 - \epsilon + \epsilon_1}{N-2}(N-2) = 0.05 + \epsilon_1 > 0.05.$$

That is, even though $f(\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}_1)$ is equal to a tiny value, there is still 5% of the mass of $f(\boldsymbol{s}|\boldsymbol{\theta}_1)$ at or below that value.