Bayesian analysis of tests with unknown specificity and sensitivity*

Andrew Gelman† and Bob Carpenter‡

23 May 2020

Abstract

When testing for a rare disease, prevalence estimates can be highly sensitive to uncertainty in the specificity and sensitivity of the test. Bayesian inference is a natural way to propagate these uncertainties, with hierarchical modeling capturing variation in these parameters across experiments. Another concern is the people in the sample not being representative of the general population. Statistical adjustment cannot without strong assumptions correct for selection bias in an opt-in sample, but multilevel regression and poststratification can at least adjust for known differences between the sample and the population. We demonstrate these models with code in Stan and discuss their application to a controversial recent study of COVID-19 antibodies in a sample of people from the Stanford University area. Wide posterior intervals make it impossible to evaluate the quantitative claims of that study regarding the number of unreported infections. For future studies, the methods described here should facilitate more accurate estimates of disease prevalence from imperfect tests performed on non-representative samples.

1. Background

Correction of diagnostic tests for false positives and false negatives is a well-known probability problem. When the base rate is low, estimates become critically sensitive to misclassifications (Hemenway, 1997). This issue hit the news recently (Lee, 2020) with a recent study of coronavirus antibodies in a population with a low incidence rate.

This is a problem where not fully accounting for uncertainty can make a big difference in scientific conclusions and potential policy recommendations. In early April, 2020, Bendavid et al. (2020a) recruited 3330 residents of Santa Clara County, California and tested them for COVID-19 antibodies. 50 people tested positive, yielding a raw estimate of 1.5%. After corrections, Bendavid et al. (2020a) reported an uncertainty range of 2.5% to 4.2%, implying that the number of infections in the county was between 50 and 85 times the count of cases reported at the time. Using an estimate of the number of coronavirus deaths in the county up to that time, they computed an implied infection fatality rate (IFR) of 0.12–0.2%, much lower than IFRs in the range of 0.5%–1% that had been estimated from areas with outbreaks of the disease.

The estimates from Bendavid et al. (2020a) were controversial, and it turned out that they did not correctly account for uncertainty in the specificity (true negative rate) of the test. There was also concern about the adjustment they performed for non-representativeness of their sample. Thus, the controversy arose from statistical adjustment and assessment of uncertainty. In the present article we set up a Bayesian framework to clarify these issues, specifying and fitting models using the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2020). There is a long literature on Bayesian measurement-error models (see Gustafson, 2003) and their application to diagnostic testing (Greenland, 2009); our contribution here is to set up the model, supply code, and consider multilevel regression and poststratification, influence of

*We thank Julien Riou, Will Fithian, Sander Greenland, Blake McShane, and Joseph Candelora for helpful comments and the National Science Foundation, Office of Naval Research, National Institutes of Health, and Schmidt Foundation for financial support. R and Stan code for the computations in this paper are at `https://bob-carpenter.github.io/diagnostic-testing/`.

†Department of Statistics and Department of Political Science, Columbia University, New York.

‡Center for Computational Mathematics, Flatiron Institute, New York.

hyperpriors, and other challenges that arise in the problem of estimating population prevalence using test data from a sample of people.

2. Modeling a test with uncertain sensitivity and specificity

Testing for a rare disease is a standard textbook example of conditional probability, famous for the following counterintuitive result: Suppose a person tests positive for a disease, based on a test that has a 95% accuracy rate, and further suppose that this person is sampled at random from a population with a 1% prevalence rate. Then what is the probability that he or she actually has the disease? The usual intuition suggests that the conditional probability should be approximately 95%, but it is actually much lower, as can be seen from a simple calculation of base rates, as suggested by Gigerenzer et al. (2007). Imagine you test 1000 people. With a 1% prevalence rate, we can expect that 10 have the disease and 990 do not. Then, with a 95% accuracy rate (assuming this applies to both specificity and sensitivity of the test), we would expect $0.95 \times 10 = 9.5$ true positives and $0.05 \times 990 = 49.5$ false positives; thus, the proportion of positive tests that are true positives is $9.5/(9.5 + 49.5) = 0.16$, a number that is difficult to make sense of without visualizing the hypothetical populations of true positive and false positive tests.

A related problem is to take the rate of positive tests and use it to estimate the prevalence of the disease. If the population prevalence is π and the test has a specificity of γ and a sensitivity of δ, then the expected frequency of positive tests is $p = (1-\gamma)(1-\pi) + \delta\pi$. So, given known γ, δ and p, we can solve for the prevalence,

$$\pi = (p + \gamma - 1)/(\delta + \gamma - 1). \tag{1}$$

If the properties of the test are known, but p is estimated from a random sample, we can obtain a simple classical estimate by starting with a confidence interval for p and then propagating it through the formula. For example, Bendavid et al. (2020) report 50 positive tests out of 3330, which corresponds to an estimate $\hat{p} = 50/3000 = 0.015$ with standard error $\sqrt{0.015(1-0.015)/3330} = 0.002$. Supposing that their test had a specificity of $\gamma = 0.995$ and a sensitivity of $\delta = 0.80$, this yields an estimate of $(0.015 + 0.995 - 1)/(0.80 + 0.995 - 1) = 0.013$ with standard error $0.002/(0.80 + 0.995 - 1) = 0.003$.

Two immediate difficulties arise with the classical approach. First, if the observed rate $\hat{p}$ is less than $1 - \gamma$, the false positive rate of the test, then the estimate from (1) becomes meaninglessly negative. Second, if there is uncertainty in the specificity and sensitivity parameters, it becomes challenging to propagate uncertainty through the nonlinear expression (1).

We can resolve both these problems with a simple Bayesian analysis (Gelman, 2020).

First, suppose that estimates and uncertainties for sensitivity and specificity have been externally supplied. The model is then,

$$\begin{aligned} y &\sim \text{binomial}(n, p) \\ p &= (1-\gamma)(1-\pi) + \delta\pi \\ \gamma &\sim \text{normal}(\mu_\gamma, \sigma_\gamma) \\ \delta &\sim \text{normal}(\mu_\delta, \sigma_\delta), \end{aligned} \tag{2}$$

with π, γ, and δ constrained to be between 0 and 1. Stan code is given in Appendix A.1. For simplicity, we have specified independent normal prior distributions (with the constraint that both parameters are restricted to the unit interval), but it would be easy enough to use other distributions if prior information were available in that form.

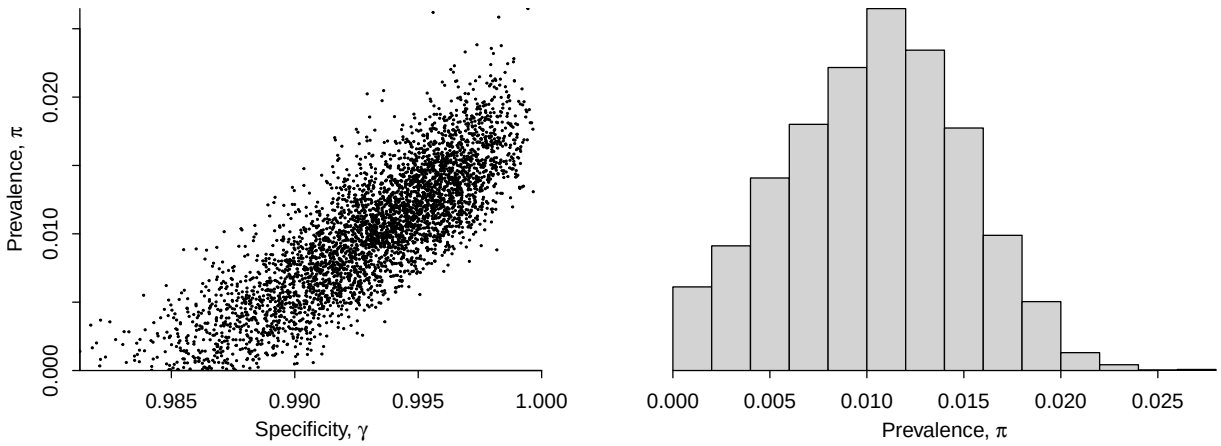

Figure 1: *Summary of inference from model with unknown specificity, sensitivity, and prevalence, based on data from Bendavid et al. (2020a): (a) scatterplot of posterior simulations of prevalence, π, and specificity, γ; (b) histogram of posterior simulations of γ. This model assumes the testing sites are identical and thus pools all data.*

In the example of Bendavid (2020a), prior information on specificity and sensitivity was given in the form of previous trials, specifically y_γ negative results in n_γ tests of known negative subjects and y_δ positive results from n_δ tests of known positive subjects. This yields the model,

$$\begin{aligned} y &\sim \text{Binomial}(n, p) \\ p &= (1-\gamma)(1-\pi) + \delta\pi \\ y_\gamma &\sim \text{Binomial}(n_\gamma, \gamma) \\ y_\delta &\sim \text{Binomial}(n_\delta, \delta). \end{aligned}$$

Fitting this model given the data in that report ($y_\gamma/n_\gamma = 399/401$ and $y_\delta/n_\delta = 103/122$) yields a wide uncertainty for p. Stan code is in Appendix A.2.

Figure 1a shows the joint posterior simulations for p and γ: uncertainty in the population prevalence is in large part driven by uncertainty in the specificity. Figure 1b shows the posterior distribution for π, which reveals that the data and model are consistent with prevalences as low as 0 and as high as nearly 2%.

The asymmetric posterior distribution with its hard bound at zero suggests that the usual central 95% interval will not be a good inferential summary. Instead we use highest posterior density or shortest posterior interval, for reasons discussed in Liu, Gelman, and Zheng (2015). The resulting 95% interval for π is (0, 1.8%), which is much different from the interval (1.1%, 2.0%) reported by Bendavid et al. (2020a). As a result, the substantive conclusion from that earlier report has been overturned. From the given data, the uncertainty in the specificity is large enough that the data do not supply strong evidence of a substantial prevalence.

3. Hierarchical model for varying testing conditions

The above analysis reveals that inference about specificity is key to precise estimation of low prevalence rates. In the second version of their report, Bendavid et al. (2020b) include data from 13 specificity studies and 3 sensitivity studies. Sensitivity and specificity can vary across experiments, so it is not appropriate to simply pool the data from these separate studies. Instead, we set up

	Posterior quantiles with weak prior			Posterior quantiles with stronger prior		
Parameter	2.5%	median	97.5%	2.5%	median	97.5%
Prevalence, π	0.003	0.017	0.354	0.003	0.012	0.021
Specificity, γ_1	0.986	0.998	1.000	0.988	0.995	0.998
Sensitivity, δ_1	0.031	0.757	0.973	0.642	0.818	0.910
μ_γ	4.63	5.69	7.20	4.71	5.26	5.87
μ_δ	−0.55	1.36	2.64	0.94	1.51	2.06
σ_γ	0.89	1.66	2.78	0.04	0.42	0.77
σ_δ	0.22	0.90	2.32	0.02	0.26	0.57

Figure 2: *Summary of inferences for the prevalence, specificity, and sensitivity of the Bendavid et al. (2020b) experiment, along with inferences for the hyperparameters characterizing the distribution of specificity and sensitivity on the logistic scale. (a) For the model with weak priors for σ_γ and σ_δ, the posterior inference for the prevalence, π, is highly uncertain. This is driven by the wide uncertainty for the sensitivity, which is driven by the large uncertainty in the hyperparameters for the sensitivity distribution. (b) Stronger priors on σ_γ and σ_δ have the effect of regularizing the specificity and sensitivity parameters, leading to narrower intervals for π, the parameter of interest in this study. The hyperparameters μ and σ are on the log odds scale and thus are difficult to interpret without transformation.*

a hierarchical model where, for any study j, the specificity γ_j and sensitivity δ_j are drawn from normal distributions on the logistic scale,

$$\begin{aligned}\text{logit}(\gamma_j) &\sim \text{normal}(\mu_\gamma, \sigma_\gamma)\\ \text{logit}(\delta_j) &\sim \text{normal}(\mu_\delta, \sigma_\delta).\end{aligned}$$

Stan code is given in Appendix A.3. This is different from model (2) in that, with data from multiple calibration studies, the hyperparameters μ and σ can be estimated from the data. In general it could make sense to allow correlation between γ_j and δ_j (Guo, Riebler, and Rue, 2017), but the way the data are currently available to us, specificity and sensitivity are estimated from separate studies and so there is no information about such a correlation. When coding the model, we use the convention that $j = 1$ corresponds to the study of interest, with other $j > 1$ representing studies of specificity or sensitivity given known samples. The parameters γ_1 and δ_1 represent the specificity and sensitivity for the study of interest.

One could also consider alternatives to the logistic transform, which allows the unbounded normal distribution to map to the unit interval but might not be appropriate for tests where the specificity can actually reach the value of 1.

We fit the above hierarchical model to the data from Bendavid et al. (2020b), assigning weak $\text{normal}^+(0, 1)$ priors to $\sigma_\gamma, \sigma_\delta$ (using the notation normal^+ for the truncated normal distribution constrained to be positive). The results are shown in Figure 2a. The 95% posterior interval for the prevalence is now $(0.00, 0.35)$. Where does that upper bound come from: how could an underlying prevalence of 35% be possible, given that only 1.5% of the people in the sample tested positive? The answer can be seen from the huge uncertainty in the sensitivity parameter, which in turn comes from the possibility that σ_δ is very large. The trouble is that the sensitivity information in these data comes from only three experiments, which is not enough to get a good estimate of the underlying distribution. This problem is discussed by Guo, Riebler, and Rue (2017).

The only way to make progress here is to constrain the sensitivity parameters in some way. One possible strong assumption is to assume that σ_δ is some small value. This could make sense

in the current context, as we can consider it as a relaxation of the assumption of Bendavid et al. (2020b) that $\sigma_\delta = 0$. We also have reason to believe that specificity will not vary much between experiments, so we will apply a soft constraint to the variation in specificities as well.

We replace the weakly informative normal$^+(0, 1)$ priors on $\sigma_\gamma, \sigma_\delta$ with something stronger, $\sigma_\gamma, \sigma_\delta \sim$ normal$^+(0, 0.2)$. To get a sense of what this means, start with the point estimate from Figure 2a of μ_δ, which is 1.36. Combining with this new prior implies that there's a roughly 2/3 chance that the sensitivity of the assay in a new experiment is in the range logit$^{-1}(1.36 \pm 0.2)$, which is $(0.76, 0.83)$. This seems reasonable.

Figure 2b shows the results. Our 95% interval for π is now $(0.003, 0.021)$; that is, the infection rate is estimated to be somewhere between 0.3% and 2.1%.

4. Prior sensitivity analysis

To assess the sensitivity of the above prevalence estimate to the priors placed on σ_γ and σ_δ, we consider the family of prior distributions,

$$\sigma_\gamma \sim \text{normal}^+(0, \tau_\gamma)$$
$$\sigma_\delta \sim \text{normal}^+(0, \tau_\delta),$$

where τ_δ and τ_γ are user-specified hyperparameters. Setting τ_δ and τ_γ to zero would force σ_δ and σ_γ to be zero and would enforce complete pooling, corresponding to Bendavid et al.'s (2020b) assumption that each test site has identical specificity and sensitivity. As the hyperparameters are increased, the scales of variation of σ_γ and σ_δ are allowed to vary more, and setting τ_γ and τ_δ to infinity would typically be considered noninformative in the sense of providing the least amount of constraint on the sensitivities and specificities. In practice, we often use normal$^+(0, 1)$ priors for hierarchical scale parameters, on the default assumption that the underlying parameters (in this case, the specificities) will probably vary by less than 1 on the logit scale.

For this problem, however, a weak prior does not work: as shown in the left panel of Figure 2, the resulting inferences for the sensitivities are ridiculously wide. No, we do not believe these tests could have specificities below 50%, yet such a possibility is included in the posterior distribution, and this in turn propagates to inappropriately wide intervals for the prevalence, π. As explained in the previous section, that is why we assigned a stronger prior, using hyperprior parameters $\tau_\gamma = \tau_\delta = 0.2$.

Figure 3 shows how these hyperprior parameters τ_γ and τ_δ affect inferences for the prevalence, π. The posterior median of π is not sensitive to the scales τ_γ and τ_δ of the hyperpriors, but the uncertainty in that estimate, as indicated by the central posterior 90% intervals, is influenced by these settings. In particular, in the graphs on the right, when the sensitivity hyperprior parameter τ_δ is given a high value, the upper end of the interval is barely constrained. The lower end of the interval is fairly stable, as long as the specificity hyperprior parameter τ_γ is not given an artificially low value.

When τ_γ and τ_δ are too low, the variation in specificity and sensitivity are constrained to be nearly zero, all values are pooled, and uncertainty is artificially deflated. As the hyperprior parameters are increased, the uncertainty in prevalence increases. This gets out of hand when the hyperprior for sensitivity is increased, because there are only three data points to inform the distribution it controls. This is an example of the general principle that wide hyperpriors on hierarchical scale parameters can pull most of the probability mass into areas of wide variation and dominate the data, leading to inflated uncertainty. Around the middle of these ranges, the posterior intervals are not as sensitive to variation in the hyperpriors. We would consider values

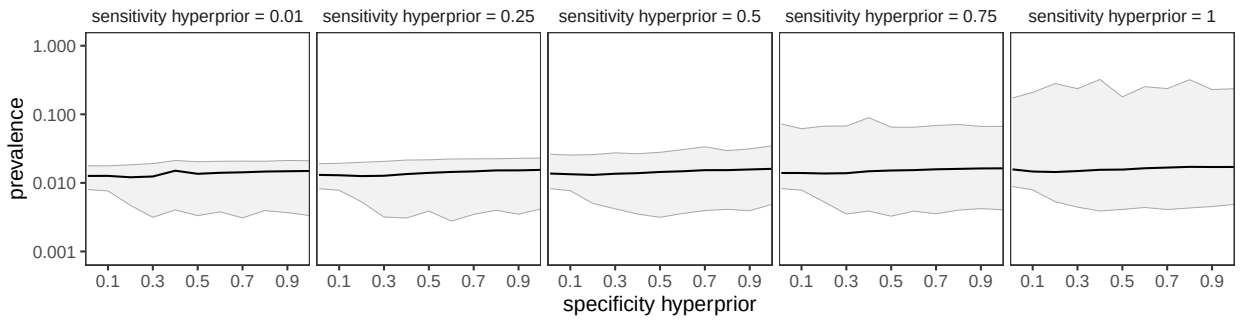

Figure 3: *Each panel shows a plot of the posterior median and central 90% posterior interval of the prevalence, π, as a function of τ_γ and τ_δ, the prior scales for the specificity and sensitivity hyperparameters, σ_γ and σ_δ. The posterior median of prevalence is not sensitive to τ_γ and τ_δ, but the endpoints of the 90% interval show some sensitivity. It is possible to use a weak hyperprior on the scale of the specificity distribution, σ_γ: this makes sense given that there are 13 prior specificity studies in the data. For the scale of the specificity distribution, σ_δ, it is necessary to use a prior scale of 0.5 or less to effectively rule out the possibility of extremely high prevalence corresponding to an unrealistically sensitivity parameter γ.*

$\tau_\gamma = \tau_\delta = 0.5$ to be weakly informative for this example, in that they are roughly consistent with inter-site variation in specificity in the range 73% to 99.3% and of specificity in the range 88% to 99.75%.

In addition we could consider priors on μ_γ and μ_δ. In this particular example, once we have constrained the variation in the specificities and sensitivities, enough data are available to estimate these population means without strong priors, but if this were a concern we could use moderately informative priors. For example, the normal$(4, 2)$ distribution puts 2/3 of the prior mass in the range 4 ± 2, which, after undoing the logistic transformation, corresponds to $(0.881, 0.995)$ on the probability scale. In general it is good practice to include informative priors for all parameters; we kept them out of this particular model just to keep the code a little bit cleaner.

The complexity of this sensitivity analysis might seem intimidating: if Bayesian inference is this difficult and this dependent on priors, maybe it's not a good idea?

We would argue that the problem is not as difficult as it might look. The steps taken in Sections 2 and 3 show the basic workflow: We start with a simple model, then add hierarchical structure. For the hierarchical model we started with weak priors on the hyperparameters and examined the inferences, which made us realize that we had prior information (that specificities and sensitivities of the tests were not so variable) which we then incorporated into the next iteration of the model. Performing the sensitivity analysis was fine—it helped us understand the inferences better—but it was not necessary for us to get reasonable inferences.

Conversely, non-Bayesian analyses would not be immune from this sensitivity to model choices, as is illustrated by the mistakes made by Bendavid et al. (2020b) to treat specificity and sensitivity as not varying at all, to set $\sigma_\gamma = \sigma_\delta = 0$ in our notation. An alternative could be to use the calibration studies to get point estimates of σ_γ and σ_δ, but then there would still be the problem of accounting for uncertainty in these estimates, which would return the researchers to the need for some sort of external constraint or bound on the distribution of the sensitivity parameters δ_j, given that only three calibration studies are available here to estimate these. This in turn suggests the need for more data or modeling of the factors that influence the test's specificity and sensitivity. In short, the analysis shown in Figure 3 formalizes a dependence on prior information that would arise, explicitly or implicitly, in any reasonable analysis of these data.

5. Extensions of the model

5.1. Multilevel regression and poststratification (MRP) to adjust for differences between sample and population

Bendavid et al. (2020a,b) compared demographics on 3330 people they tested, and they found differences in the distributions of sex, age, ethnicity, and zip code of residence compared to the general population of Santa Clara County. It would be impossible to poststratify the raw data on 2 sexes, 4 ethnicity categories, 4 age categories, and 58 zip codes, as the resulting 1856 cells would greatly outnumber the positive tests in the data. They obtained population estimates by adjusting for sex × ethnicity × zip code, but their analysis is questionable, first because they did not adjust for age, and second because of noisy weights arising from the variables they did adjust for. To obtain stable estimates while adjusting for all these variables, we would recommend applying a multilevel model to the exposure probability, thus replacing the constant π in the above models with something like this logistic regression:

$$\pi_i = \text{logit}^{-1}(\beta_1 + \beta_2 \cdot \mathit{male} + \beta_3 \cdot x^{\text{zip}}_{\mathit{zip}[i]} + \alpha^{\text{eth}}_{\mathit{eth}[i]} + \alpha^{\text{age}}_{\mathit{age}[i]} + \alpha^{\text{zip}}_{\mathit{zip}[i]}), \tag{3}$$

where *male* is a variable that takes on the value 0 for women and 1 for men; x^{zip} is a relevant predictor at the zip code level; *eth*[*i*], *age*[*i*], and *zip*[*i*] are index variables for survey respondent i; the β's are logistic regression coefficients; and the α's are vectors of varying intercepts. These varying intercepts have hierarchical priors

$$\alpha^{\mathit{name}} \sim \text{normal}(0, \sigma^{\mathit{name}}), \text{ for } \mathit{name} \in \{\text{eth, age, zip}\}.$$

In the regression model (3), it is important to include the predictor x^{zip}, which in this example might be percent Latino or average income in the zip code. Otherwise, with so many zip codes, the multilevel model will just partially pool most of the zip code adjustments to zero, and not much will be gained from the geographic poststratification. The importance of geographic predictors is well known in the MRP literature; see, for example, Caughey and Warshaw (2019).

The above model is a start; it could be improved by including interactions, following the general ideas of Ghitza and Gelman (2013). In any case, once this model has been fit, it can be used to make inferences for disease prevalence for all cells in the population, and these cell estimates can then be summed, weighting by known population totals (in this case, the number of people in each sex × ethnicity × age × zip code category in the population) to get inferences for the prevalence in the county. Here we are implicitly assuming that the data represent a random sample within poststratification cells.

In addition, priors are needed for $\sigma^{\text{eth}}, \sigma^{\text{age}}, \sigma^{\text{zip}}$, and β, along with the hierarchical specificity and sensitivity parameters from the earlier model.

We code the model in Stan; see Appendix A.4. Unfortunately the raw data from Bendavid et al. are not currently available, so we fit the model to simulated data to check the stability of the computation. We use a normal$(0, 2.5)$ prior for the centered intercept $\beta_1 + \beta_2\overline{\mathit{male}} + \beta_3\overline{x}^{\text{zip}}$ (corresponding to an approximate 95% prior interval of $(0.7\%, 99.3\%)$ for the probability that an average person in the sample has the antibody), a normal$(0, 0.5)$ prior for β_2, and normal$^+(0, 0.5)$ priors for σ^{eth}, σ^{age}, and σ^{zip}. These priors allow the prevalence to vary moderately by these poststratification variables.

Once the above model has been fit, it implies inferences for the prevalence in each of the $2 \times 4 \times 4 \times 58$ poststratification cells; as discussed by Johnson (2020), these can be averaged to get a population prevalence: $p_{avg} = \sum_j N_j p_j / \sum_j N_j$, where N_j is the number of people in cell j in the general population, and p_j is the prevalence as computed from the logistic model. We perform this summation in the generated quantities block of the Stan model in Appendix A.4.

5.2. Variation across location and over time

The aforementioned Santa Clara County study is just one of many recent COVID-19 antibody surveys. Other early studies were conducted in Boston, New York, Los Angeles, and Miami, and in various places outside the United States, and we can expect many more in the future. If the raw data from these studies were combined, it should be possible to estimate the underlying prevalences from all these studies using a hierarchical model, allowing specificity, sensitivity, and prevalence to vary by location, and adjusting for non-sampling error where possible. Such an analysis is performed by Levesque and Maybury (2020) using detailed information on the different tests used in different studies.

We will also be seeing more studies of changing infection rates over time. Stringhini et al. (2020) perform such an analysis of weekly surveys in Geneva, Switzerland, accounting for specificity and sensitivity and poststratifying by sex and age.

5.3. Including additional diagnostic data

We have so far assumed that test results are binary, but additional information can be gained from continuous measurements that make use of partial information when data are near detection limits (Gelman, Chew, and Shnaidman, 2004; Bouman, Bonhoeffer, and Regoes, 2020). Further progress can be made by performing different sorts of tests on study participants or retesting observed positive results.

Another promising direction is to include additional information on people in the study, for example from self-reported symptoms. Some such data are reported in Bendavid et al. (2020b), although not at the individual level. With individual-level symptom and test data, a model with multiple outcomes could yield substantial gains in efficiency compared to the existing analysis using only a single positive/negative test result on each participant.

6. Non-Bayesian approaches

As with any statistical analysis, alternative approaches are possible that would use the same information and give similar results.

In Section 2, it was necessary to account for uncertainty in all three parameters, while respecting the constraint that all three probabilities had to be between 0 and 1. We assume that both these aspects of the model could be incorporated into a non-Bayesian approach by working out the region in the space of (π, γ, δ) that is consistent with the data and then constructing a family of tests which could be inverted to create confidence regions.

This could be expanded into a multilevel model as in Section 3 by considering the specificities and sensitivities of the different experiments as missing data and averaging over their distribution, but still applying non-Bayesian inference to the resulting hyperparameters. The wide uncertainty intervals from the analysis in Section 3 suggest that some constraints or regularization or additional information on the hyperparameters would be necessary to get stable inferences here, no matter what statistical approach is used.

Fithian (2020) performs a non-Bayesian analysis of the data from Bendavid et al. (2020b), coming to the same basic conclusion that we do, demonstrating that the calibration data are incompatible with a model of constant specificity and that, once the specificity is allowed to vary, the observed rate of positive tests in the Santa Clara study does not allow rejection of the null hypothesis of zero infection rate. Had it been possible to reject zero, this would not be the end of

the story: at that point one could invert a family of tests to obtain a confidence region, as noted above.

Finally, some rough equivalent to the poststratification adjustment in Section 5.1 could be performed using a non-Bayesian weighting approach, using some smoothing to avoid the noisiness of raw poststratification weights. Similarly, non-Bayesian methods could be used to fit regressions allowing prevalence to vary over location and time.

7. Discussion

7.1. Limitations of the statistical analysis

Epidemiology in general, and disease testing in particular, features latent parameters with high levels of uncertainty, difficulty in measurement, and uncertainty about the measurement process as well. This is the sort of setting where it makes sense to combine information from multiple studies, using Bayesian inference and hierarchical models, and where inferences can be sensitive to assumptions.

The biggest assumptions in this analysis are, first, that the historical specificity and sensitivity data are relevant to the current experiment; and, second, that the people in the study are a representative sample of the general population. We addressed the first concern with a hierarchical model of varying sensitivities and specificities, and we addressed the second concern with multilevel regression and poststratification on demographics and geography. But this modeling can take us only so far. If there is hope or concern that the current experiment has unusual measurement properties, or that the sample is unrepresentative in ways not accounted for in the regression, then more information or assumptions need to be included in the model, as in Campbell et al. (2020).

The other issue is that there are choices of models, and tuning parameters within each model. Sensitivity to the model is apparent in Bayesian inference, but it would arise with any other statistical method as well. For example, Bendavid et al. (2020a) used an (incorrectly applied) delta method to propagate uncertainty, but this is problematic when sample size is low and probabilities are near 0 or 1. Bendavid et al. (2020b) completely pooled their specificity and sensitivity experiments, which is equivalent to setting σ_γ and σ_δ to zero. And their weighting adjustment has many arbitrary choices. We note these not to single out these particular authors but rather to emphasize that, at least for this problem, all statistical inferences involve user-defined settings.

For the models in the present article, the most important user choices are: (a) what data to include in the analysis, (b) prior distributions for the hyperparameters, and (c) the structure and interactions to include in the MRP model. For these reasons, it would be difficult to set up the model as a plug-and-play system where users can just enter their data, push a button, and get inferences. Some active participation in the modeling process is required, which makes sense given the sparseness of the data. When studying populations with higher prevalences and with data that are closer to random samples, more automatic approaches might be possible.

7.2. Santa Clara study

Section 3 shows our inferences given the summary data in Bendavid et al. (2020b). The inference depends strongly on the priors on the distributions of sensitivity and specificity, but that is unavoidable: the only way to avoid this influence of the prior would be to sweep it under the rug, for example by just assuming a zero variation in the test parameters.

What about the claims regarding the rate of coronavirus exposure and implications for the infection fatality rate? It's hard to say from this one study: the numbers in the data are consistent

with zero infection rate and a wide variation in specificity and sensitivity across tests, and the numbers are also consistent with the claims made in Bendavid et al. (2020a,b). That does not mean anyone thinks the true infection rate is zero. It just means that more data, assumptions, and subject-matter knowledge are required. That's ok—people usually make lots of assumptions in this sort of laboratory assay. It's common practice to use the manufacturer's numbers on specificity, sensitivity, detection limit, and so forth, and not worry about that level of variation. It's only when you are estimating a very low underlying rate that the statistical challenges become so severe.

For now, we do not think the data support the claim that the number of infections in Santa Clara County was between 50 and 85 times the count of cases reported at the time, or the implied interval for the IFR of 0.12–0.2%. These numbers are consistent with the data, but the data are also consistent with a near-zero infection rate in the county. The data of Bendavid et al. (2020a,b) do not provide strong evidence about the number of people infected or the infection fatality ratio; the number of positive tests in the data is just too small, given uncertainty in the specificity of the test.

Going forward, the analyses in this article suggest that future studies should be conducted with full awareness of the challenges of measuring specificity and sensitivity, that relevant variables be collected on study participants to facilitate inference for the general population, and that (de-identified) data be made accessible to external researchers.

References

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra-Walker, R., Tedrow, J., Tversky, D., Bogan, A., Kupiec, T., Eichner, D., Gupta, R., Ioannidis, J., and Bhattacharya, J. (2020a). COVID-19 antibody seroprevalence in Santa Clara County, California, version 1. https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v1.full.pdf

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra-Walker, R., Tedrow, J., Tversky, D., Bogan, A., Kupiec, T., Eichner, D., Gupta, R., Ioannidis, J., and Bhattacharya, J. (2020b). COVID-19 antibody seroprevalence in Santa Clara County, California, version 2. https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v2.full.pdf

Bouman, J. A., Bonhoeffer, S., and Regoes, R. R. (2020). Estimating seroprevalence with imperfect serological tests: a cutoff-free approach. https://www.biorxiv.org/content/10.1101/2020.04.29.068999v2

Campbell, H., de Valpine, P., Maxwell, L., de Jong, V. M. T., Debray, T., Jänisch, T., and Gustafson, P. (2020). Bayesian adjustment for preferential testing in estimating the COVID-19 infection fatality rate: Theory and methods. https://arxiv.org/abs/2005.08459

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B, Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* **76** (1), 1–32. https://www.jstatsoft.org/article/view/v076i01

Caughey, D., and Warshaw, C. (2019). Public opinion in subnational politics. *Journal of Politics* **81**, 352–363.

Fithian, W. (2020). Statistical comment on the revision of Bendavid et al. https://www.stat.berkeley.edu/~wfithian/overdispersionSimple.html

Gelman, A. (2020). Simple Bayesian analysis inference of coronavirus infection rate from the Stanford study in Santa Clara county. *Statistical Modeling, Causal Infer-*

ence, and Social Science, 1 May. https://statmodeling.stat.columbia.edu/2020/05/01/simple-bayesian-analysis-inference-of-coronavirus-infection-rate-from-the-stanford-study-in-santa-clara-county

Gelman, A., Chew, G., and Shnaidman, M. (2004). Bayesian analysis of serial dilution assays. *Biometrics* **60**, 407–417.

Ghitza, Y., and Gelman, A., (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science* **57**, 762–776.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest* **8**, 53–96.

Greenland, S. (2009). Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods. *International Journal of Epidemiology* **38**, 1662–1673.

Guo, J., Riebler, A., and Rue, H. (2017). Bayesian bivariate meta-analysis of diagnostic test studies with interpretable priors. *Statistics in Medicine* **36**, 3039–3058.

Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. London: CRC Press.

Hemenway, D. (1997). The myth of millions of annual self-defense gun uses: A case study of survey overestimates of rare events. *Chance* **10** (3), 6–10.

Johnson, D. (2020). Estimating seroprevalence with data from an imperfect test on a convenience sample. https://www.dougjohnson.in/post/estimating-seroprevalence-with-data-from-an-imperfect-test-on-a-convenience-sample/

Lee, S. M. (2020). Two antibody studies say coronavirus infections are more common than we think. Scientists are mad. *BuzzFeed News*, 22 Apr. https://www.buzzfeednews.com/article/stephaniemlee/coronavirus-antibody-test-santa-clara-los-angeles-stanford

Levesque, J., and Maybury, D. W. (2020). A note on COVID-19 seroprevalence studies: a meta-analysis using hierarchical modelling. https://www.medrxiv.org/content/10.1101/2020.05.03.20089201v1.full.pdf

Liu, Y., Gelman, A., and Zheng, T. (2015). Simulation-efficient shortest probability intervals. *Statistics and Computing* **25**, 809–819.

Stan Development Team (2020). *Stan Modeling Language User's Guide and Reference Manual*. https://mc-stan.org

Stringhini, S., Wisniak, A., Piumatti, G., Azman, A. S., Lauer, S. A., Baysson, H., De Ridder, D., Petrovic, D., Schrempft, S., Marcus, K., Yerly, S., Vernez, I. A., Keiser, O., Hurst, S., Posfay-Barbe, K. M., Trono, D., Pittet, D., Getaz, L., Chappuis, F., Eckerle, I., Vuilleumier, N., Meyer, B., Flahault, A., Kaiser, L., and Guessous, I. (2020). Repeated seroprevalence of anti-SARS-CoV-2 IgG antibodies in a population-based sample. https://www.medrxiv.org/content/10.1101/2020.05.02.20088898v1.full.pdf

A. Stan programs

A.1. Model with truncated normal priors on specificity and sensitivity

```
data {
  int y_sample;
  int n_sample;
  real mu_spec;
  real<lower=0> sigma_spec;
  real mu_sens;
  real<lower=0> sigma_sens;
}
parameters {
  real<lower=0, upper=1> p;
  real<lower=0, upper=1> spec;
  real<lower=0, upper=1> sens;
}
model {
  real p_sample;
  p_sample = p * sens + (1 - p) * (1 - spec);
  y_sample ~ binomial(n_sample, p_sample);
  spec ~ normal(mu_spec, sigma_spec);
  sens ~ normal(mu_sens, sigma_sens);
}
```

A.2. Model with binomial data on specificity and sensitivity

```
data {
  int<lower = 0> y_sample;
  int<lower = 0> n_sample;

  int<lower = 0> y_spec;
  int<lower = 0> n_spec;

  int<lower = 0> y_sens;
  int<lower = 0> n_sens;
}
parameters {
  real<lower=0, upper = 1> p;

  real<lower=0, upper = 1> spec;
  real<lower=0, upper = 1> sens;
}
model {
  real p_sample = p * sens + (1 - p) * (1 - spec);
  y_sample ~ binomial(n_sample, p_sample);
  y_spec ~ binomial(n_spec, spec);
  y_sens ~ binomial(n_sens, sens);
}
```

A.3. Hierarchical model for specificities and sensitivities

```
data {
```

```
  int<lower = 0> y_sample;
  int<lower = 0> n_sample;

  int<lower = 0> J_spec;
  int<lower = 0> y_spec[J_spec];
  int<lower = 0> n_spec[J_spec];

  int<lower = 0> J_sens;
  int<lower = 0> y_sens[J_sens];
  int<lower = 0> n_sens[J_sens];

  real<lower = 0> logit_spec_prior_scale;
  real<lower = 0> logit_sens_prior_scale;
}
parameters {
  real<lower = 0, upper = 1> p;

  real mu_logit_spec;
  real<lower = 0> sigma_logit_spec;
  vector<offset = mu_logit_spec, multiplier = sigma_logit_spec>[J_spec] logit_spec;

  real mu_logit_sens;
  real<lower = 0> sigma_logit_sens;
  vector<offset = mu_logit_sens, multiplier = sigma_logit_sens>[J_sens] logit_sens;
}
transformed parameters {
  vector[J_spec] spec = inv_logit(logit_spec);
  vector[J_sens] sens= inv_logit(logit_sens);
}
model {
  real p_sample = p * sens[1] + (1 - p) * (1 - spec[1]);
  y_sample ~ binomial(n_sample, p_sample);
  y_spec ~ binomial(n_spec, spec);
  y_sens ~ binomial(n_sens, sens);
  logit_spec ~ normal(mu_logit_spec, sigma_logit_spec);
  logit_sens ~ normal(mu_logit_sens, sigma_logit_sens);
  sigma_logit_spec ~ normal(0, logit_spec_prior_scale);
  sigma_logit_sens ~ normal(0, logit_sens_prior_scale);
}
```

A.4. Multilevel regression and poststratification

```
data {
  int<lower = 0> N;  // number of tests in the sample (3330 for Santa Clara)
  int<lower = 0, upper = 1> y[N];  // 1 if positive, 0 if negative
  vector<lower = 0, upper = 1>[N] male;  // 0 if female, 1 if male
  int<lower = 1, upper = 4> eth[N];  // 1=white, 2=asian, 3=hispanic, 4=other
  int<lower = 1, upper = 4> age[N];  // 1=0-4, 2=5-18, 3=19-64, 4=65+
  int<lower = 0> N_zip;  // number of zip codes (58 in this case)
  int<lower = 1, upper = N_zip> zip[N];  // zip codes 1 through 58
  vector[N_zip] x_zip;  // predictors at the zip code level
  int<lower = 0> J_spec;
  int<lower = 0> y_spec [J_spec];
  int<lower = 0> n_spec [J_spec];
```

```
  int<lower = 0> J_sens;
  int<lower = 0> y_sens [J_sens];
  int<lower = 0> n_sens [J_sens];
  int<lower = 0> J;  // number of population cells, J = 2*4*4*58
  vector<lower = 0>[J] N_pop;  // population sizes for poststratification
  real intercept_prior_mean;
  real<lower = 0> intercept_prior_scale;
  real<lower = 0> coef_prior_scale;
  real<lower = 0> logit_spec_prior_scale;
  real<lower = 0> logit_sens_prior_scale;
}
parameters {
  real mu_logit_spec;
  real mu_logit_sens;
  real<lower = 0> sigma_logit_spec;
  real<lower = 0> sigma_logit_sens;
  vector<offset = mu_logit_spec, multiplier = sigma_logit_spec>[J_spec] logit_spec;
  vector<offset = mu_logit_sens, multiplier = sigma_logit_sens>[J_sens] logit_sens;
  vector[3] b;  // intercept, coef for male, and coef for x_zip
  real<lower = 0> sigma_eth;
  real<lower = 0> sigma_age;
  real<lower = 0> sigma_zip;
  vector<multiplier = sigma_eth>[4] a_eth;  // varying intercepts for ethnicity
  vector<multiplier = sigma_age>[4] a_age;  // varying intercepts for age category
  vector<multiplier = sigma_zip>[N_zip] a_zip;  // varying intercepts for zip code
}
transformed parameters {
  vector[J_spec] spec = inv_logit(logit_spec);
  vector[J_sens] sens = inv_logit(logit_sens);
}
model {
  vector[N] p = inv_logit(b[1]
                          + b[2] * male
                          + b[3] * x_zip[zip]
                          + a_eth[eth]
                          + a_age[age]
                          + a_zip[zip]);
  vector[N] p_sample = p * sens[1] + (1 - p) * (1 - spec[1]);
  y ~ bernoulli(p_sample);
  y_spec ~ binomial(n_spec, spec);
  y_sens ~ binomial(n_sens, sens);
  logit_spec ~ normal(mu_logit_spec, sigma_logit_spec);
  logit_sens ~ normal(mu_logit_sens, sigma_logit_sens);
  sigma_logit_spec ~ normal(0, logit_spec_prior_scale);
  sigma_logit_sens ~ normal(0, logit_sens_prior_scale);
  a_eth ~ normal(0, sigma_eth);
  a_age ~ normal(0, sigma_age);
  a_zip ~ normal(0, sigma_zip);
  // prior on centered intercept
  b[1] + b[2] * mean(male) + b[3] * mean(x_zip[zip])
      ~ normal(intercept_prior_mean, intercept_prior_scale);
  b[2] ~ normal(0, coef_prior_scale);
  sigma_eth ~ normal(0, coef_prior_scale);
  sigma_age ~ normal(0, coef_prior_scale);
```

```
  sigma_zip ~ normal(0, coef_prior_scale);
  b[3] ~ normal(0, coef_prior_scale / sd(x_zip[zip]));  // prior on scaled coefficient
}
generated quantities {
  real p_avg;
  vector[J] p_pop;  // population prevalence in the J poststratification cells
  int count;
  count = 1;
  for (i_zip in 1:N_zip) {
    for (i_age in 1:4) {
      for (i_eth in 1:4) {
        for (i_male in 0:1) {
          p_pop[count] = inv_logit(b[1]
                                   + b[2] * i_male
                                   + b[3] * x_zip[i_zip]
                                   + a_eth[i_eth]
                                   + a_age[i_age]
                                   + a_zip[i_zip]);
          count += 1;
        }
      }
    }
  }
  p_avg = sum(N_pop .* p_pop) / sum(N_pop);
}
```