Qualifying Paper Report for Adjusting COVID-19 Seroprevalence Survey Results to Account for Test Sensitivity and Specificity

Naitong Chen

March 9, 2023

1 Summary

Over the past few years, tracking the spread of COVID-19 has been crucial to developing a scientific understanding the disease, which ultimately guided public health protocols aimed at controlling the spread of the disease across the world. As with any epidemic/pandemic, reported case counts wthin a defined geographic region is one of the most accessible statistics indicating the scale of the spread of a disease. However, due to testing availability, there may be many individuals in a given region that have been infected but not tested. This means that the total number of infection may be much greater than that reflected from reported case counts (cite).

To more accuratly estimate the cumulative number of infections over a period of time, an alternative approach is to conduct population-based seroprevalence studies. To carry out a seroprevalence study for a given region on a given disease, researchers begin by obtaining a sample representative of the population. Antibody tests are then performed for the disease of interest over each individual in the sample. A positive antibody test indicates a case of infection of the tested disease. Therefore, the proportion of positive tests in the sample can be used as an estimate of the porportion of population infected with the disease over some time interval, which we call the cumulative incidence. Given the population size of the corresponding region, one can estimate the total number of infected individuals in the population using the estimated cumulative incidence. It is worth noting, however, that seroprevalence studies cannot identify previous infections whose antibodies are no longer detectable or recent infections that have yet to produce detectable antibodies. At the same time, they also do not include individuals that have died after becoming infected. As a result, a seroprevalence study as we describe it here is only informative about cumulative incidence for the average period, prior to sample collection, over which antibodies are detectable, provided that the disease has a relatively low fatality rate.

We know that COVID-19 has a relatively low fatality rate (cite). We also know that an individual starts to produce detectable antibodies after an average of 25 days since infection, and that the antibodies stay detectable for months after infection (cite). Therefore, using data from a seroprevalence study conducted within the first few months of the COVID-19 pandemic, we can estimate cumulative incidence over the period from the beginning of the pandemic until roughly a month prior to when the samples were taken.

Since antibody tests are not 100% accurate, there may be positive cases that test negative and negative cases that test positive. Therefore, one would ideally also like to adjust cumulative incidence for test-kit performance. This is typically done as follows. We begin by defining test specificity sp as the proportion of noncases that test negative and test sensitivity se as the proportion of actual cases that test positive. Then with the true cumulative incidence being denoted as s, we can model the observed prevalence p, which is adjusted for test-kit performance, as

$$p = s \times se + (1 - s) \times (1 - sp).$$

To put in words, the observed prevalence can be decomposed into the proportion of actual cases that correctly test positive and noncases that incorrectly test positive. Given the total sample size n from a given region and the number of positive tests x from the sample, it is then reasonable to model the number of positive tests as the outcome of a Binomial distribution with the total sample size as the number of trials and observed prevalence as the probability of success, i.e.,

$$x \mid n, s, se, sp \sim \mathsf{Binom}\left(n, s \times se + (1 - s) \times (1 - sp)\right) = \mathsf{Binom}(n, p).$$

Using the above as the likelihood function of s, se and sp, we can construct a bayesian model by defining a set of prior distributions on each of s, se, and sp using distributions with a support on $[0, 1]$. (cite) applies this bayesian model to a dataset obtained from a seroprevalence study conducted in New York state between April 19 and April 28 in 2020. This dataset contains the number of positive antibody tests and the total number of tests from each of the 11 regions across New York state in the study. Full details of data can be found in (citation). With consideration of the average time between infection and when antibodies become detectable, this dataset can be used to estimate cumulative incidences from the beginning of the pandemic until Mar 29, 2020. This is because there are 25 days between Mar 29, 2020 and the seroprevalence study midpoint April 23, 2020.

Instead of directly applying the above Bayesian model where each region gets its own prior on cumulative incidence, (cite) remarks it is possible that regions close to each other geographically may share sociodemographic factors which are associated with the number of infections. As a result, (cite) groups the 11 regions into three super-regions (New York City, Westchester and Rockland Counties and Long Island, as well as rest of state), with regions from the same super-region sharing a common prior distribution on their cumulative incidences. Denoting s_{ij}, p_{ij}, n_{ij} and x_{ij} as the cumulative incidence, observed incidence, number of samples, and number of positive antibody tests from the i^{th} region in the j^{th} super-region, the final Bayesian model is defined as follows.

$$\begin{aligned}
s_{i1} &\overset{\text{i.i.d.}}{\sim} \mathsf{Beta}(2.1792, 9.8208) \quad \forall i \text{ in super-region1 (New York City)},\\
s_{i2} &\overset{\text{i.i.d.}}{\sim} \mathsf{Beta}(2.6641, 9.3359) \quad \forall i \text{ in super-region2 (Westchester, Rockland Counties and Long Island)},\\
s_{i3} &\overset{\text{i.i.d.}}{\sim} \mathsf{Beta}(1.1930, 10.8070) \quad \forall i \text{ in super-region3 (rest of state)},\\
se &\sim \mathsf{Beta}(205, 29)_{\{0.8, 0.95\}},\\
sp &\sim \mathsf{Beta}(288, 2)_{\{0.9, 1\}},\\
p_{ij} &= s_{ij} \times se + (1 - s_{ij}) \times (1 - sp),\\
x_{ij} \mid n_{ij}, p_{ij} &\overset{\text{indep}}{\sim} \mathsf{Binom}(n_{ij}, p_{ij}).
\end{aligned}$$

The priors for each region are chosen so that the mean of the prior matches the ratio between the cumulative reported case count up until March 29, 2020 and the total population of the corresponding super-region. On the other hand, the priors on test sensitivity and test specificity are based on validation studies: (cite) estimates the test specificity to be 0.9975 with a 95% confidence interval of $[0.961, 1]$, and the test sensitivity to be 0.879 with a 95% confidence interval of $[0.837, 0.921]$. The priors are then chosen so that means and variances of the priors on test specificity and sensitivity to match the results from the validation studies. Note that the subscripts denote truncation to the specified regions. With the model specified, we can use Markov Chain Monte Carlo to obtain samples from the posterior distribution (conditional distribution given observed data) of regional cumulative incidences as well as test specificity and sensitivity. Given a set of regional cumulative incidences from the posterior distribution, we can estimate the cumulative incidence for each super-region or the entire state using the average over the corresponding regional cumulative incidences weighted by the proportion of population living in each region. (cite) uses the median values over $100,000$ posterior samples as point estimates for each parameter of interest. At the same time, equal tailed 95% credible intervals are used to quantify the uncertainty about the estimates.

(cite) compares the results from the above Bayesian model to a non-Bayesian version of the same analysis. In the non-Bayesian version of the analysis, given a specific region (or super-region, or the entire state), with the sample proportion of positive tests being denoted pr, we can estimate the cumulative incidence of that region by rearranging the equation that adjusts for test specificity and sensitivity. Note that here the sample proportion of positive tests plays the same role as observed prevalence in the Bayesian version of the analysis. This leads to

$$s = (pr + sp - 1)/(se + sp - 1).$$

The point estimate of the cumulative incidence is obtained by plugging in the estimated test specificity and sensitivity values from the validation studies. To quantify the uncertainty around each point estimate, (cite) constructs an iterval using the 95% confidence interval endpoints from the validation studies. In particular, the lowerbound of this interval can be obtained by plugging in the lower endpoint of test specificity and upper endpoint of test sensitivity, and vice versa. It is worth noting, however, that this interval is not a 95% confidence interval around the true cumulative incidence of the corresponding region.

Comparing the results from both studies (cite), we can see that the point estimates from both the Bayesian and non-Bayesian analyses are relatively similar, but the Bayesian credible intervals are generally narrower than the intervals constructed using confidence interval endpoints. In addition, while some of the intervals constructed using confidence interval endpoints contain a negative lowerbound, this does not happen to any of the credible intervals constructed. We elaborate on advantages and disadvantages of the two analyses in the following sections.

2 Significance

no negative interval

interpretability of intervals

can incorporate prior knowledge

3 Limitations and challenges

talk about hierarchical model word misuse

talk about grouping of super-region being arbitrary

sensitivity

does not account for seroreversion

4 Paper-specific project

4.1 Data

The datasets used in Lewin et al. (2021) and Lewin et al. (2022) together can be viewed as one for a serial COVID-19 seroprevalence study in Quebec, Canada.

The first dataset contains numbers of antibody-positive samples as well as total number of samples for each region in Quebec (Montreal-Laval, surrounding Montreal-Laval, and other regions). These samples are collected relatively early on in the pandemic, from May 25 to July 9, 2020. The second dataset follows the same structure, but contains samples collected between January 25 and Mar 11, 2021.

In addition, Lewin et al. (2022) also contains results from a seroreversion substudy. Namely, we also have the number of antibody-positive samples from 2020 that remained positive in 2021.

Note that in the second study, the count of antibody-positive samples are available at a finer scale (each of Montreal-Laval, surrounding Montreal-Laval, and other regions are broken down into smaller regions), this is not the case for Lewin et al. (2021). Therefore the analysis will only be conducted at the bigger regional level.

4.2 Project Idea

We can apply the approach in Meyer et al. (2022) to the two datasets in Lewin et al. (2021) and Lewin et al. (2022) in one coherent Bayesian model that accounts for seroreversion.

Let S_{ij} denote the true seroprevalence in study i and region j ($i = 1$ and $i = 2$ correspond to first and second study, $j = 1, j = 2$ and $j = 3$ correspond to Montreal-Laval, surrounding Montreal-Laval, and other regions). Let P_{ij} denote the observed seroprevalances. Let Se, Sp, Sr denote sensitivity of test-kit, specificity of test-kit, as well as proportion of samples that has seroreverted (testing antibody-negative in 2021 but antibody-positive in 2020). Let x_{ij} denote the total number of antibody-positive samples in region j at study i, and n_{ij} denote the total number of samples in region j at study i. Finally, let x_r and n_r denote the total number of samples that seroreverted and the total number of samples in the seroreversion substudy.

The model can then be written as

$$\begin{aligned}
P_{11} &= S_{11} \times Se + (1 - S_{11}) \times (1 - Sp) \\
P_{12} &= S_{12} \times Se + (1 - S_{12}) \times (1 - Sp) \\
P_{13} &= S_{13} \times Se + (1 - S_{13}) \times (1 - Sp) \\
P_{21} &= (S_{21} - Sr \times S_{11}) \times Se + (1 - (S_{21} - Sr \times S_{11})) \times (1 - Sp) \\
P_{22} &= (S_{22} - Sr \times S_{12}) \times Se + (1 - (S_{22} - Sr \times S_{12})) \times (1 - Sp) \\
P_{23} &= (S_{23} - Sr \times S_{13}) \times Se + (1 - (S_{23} - Sr \times S_{13})) \times (1 - Sp) \\
S_{11} &\sim \mathsf{Beta}(\cdot, \cdot) \\
S_{12} &\sim \mathsf{Beta}(\cdot, \cdot) \\
S_{13} &\sim \mathsf{Beta}(\cdot, \cdot) \\
S_{21} &\sim \mathsf{Beta}(\cdot, \cdot)_{\{Sr \times S_{11}, 1\}} \\
S_{22} &\sim \mathsf{Beta}(\cdot, \cdot)_{\{Sr \times S_{12}, 1\}} \\
S_{23} &\sim \mathsf{Beta}(\cdot, \cdot)_{\{Sr \times S_{13}, 1\}} \\
Se &\sim \mathsf{Beta}(\cdot, \cdot) \\
Sp &\sim \mathsf{Beta}(\cdot, \cdot) \\
Sr &\sim \mathsf{Beta}(\cdot, \cdot)
\end{aligned}$$

$$L(P, S, Se, Sp, Sr \mid X, N) \propto \left(\prod_{i=1}^{2} \prod_{j=1}^{3} P_{ij}^{x_{ij}} (1 - P_{ij})^{n_{ij} - x_{ij}} \right) \left(Sr^{x_r} (1 - Sr^{n_r - x_r}) \right)$$

Note that for observed seroprevalences in the first study, we adjust for test sensivitity and test specificity (Following the setup in Meyer et al. (2022). We'll likely use the same prior on Se and Sp here). For observed seroprevalences in the second study, we adjust for both test sensivitity and test specificity as well as seroreversion. Namely, we also include the proportion that have seroreverted since the first study using information from the seroreversion substudy. This combines the setup in Lewin et al. (2022) and Meyer et al. (2022): we first substract the proportion that have seroreverted from the true seroprevalence and then adjust for test-kit performance. This is because the test-kit only has a chance at detecting antibodies if the subject has not seroreverted. To ensure we do not run into negative seroprevalence estimates, we truncate S_{21}, S_{22}, S_{23} accordingly.

Altogether, this can give us an estimate of the seroprevalance in Quebec, Canada in January to March 2021 adjusting for test-kit performance as well as seroreversion.

We can compare the results from this above Bayesian model to those from Lewin et al. (2022) (which is not Bayesian and does not account for test-kit performance) as well a frequentist equivalence of the above Bayesian model using the same data. We can use these results to check if the claims from Lewin et al. (2022) still hold under this different dataset, as well as to explore potential reasons as to why they do or do not hold.

It turns out that the intervals constructed from Rosenberg et al. (2020) (the non-Bayesian analysis that Meyer et al. (2022) compared to) is just by using the 95% confidence interval endpoints for test sensitivity and specificity to correct for the true seroprevalence using $P = S \times Se + (1 - S) \times (1 - Sp)$.

5 Future Directions

References

Antoine Lewin, Roseline Therrien, Gaston De Serres, Yves Grégoire, Josée Perreault, Mathieu Drouin, Marie-Josée Fournier, Tony Tremblay, Julie Beaudoin, Guillaume Beaudoin-Bussières, et al. Sars-cov-2 seroprevalence among blood donors in québec, and analysis of symptoms associated with seropositivity: a nested case-control study. *Canadian Journal of Public Health*, 112(4):576–586, 2021.

Antoine Lewin, Gaston De Serres, Yves Grégoire, Josée Perreault, Mathieu Drouin, Marie-Josée Fournier, Tony Tremblay, Julie Beaudoin, Amélie Boivin, Guillaume Goyette, et al. Seroprevalence of sars-cov-2 antibodies among blood donors in québec: an update from a serial cross-sectional study. *Canadian Journal of Public Health*, 113(3):385–393, 2022.

Mark J Meyer, Shuting Yan, Samantha Schlageter, John D Kraemer, Eli S Rosenberg, and Michael A Stoto. Adjusting covid-19 seroprevalence survey results to account for test sensitivity and specificity. *American Journal of Epidemiology*, 191(4):681–688, 2022.

Eli S Rosenberg, James M Tesoriero, Elizabeth M Rosenthal, Rakkoo Chung, Meredith A Barranco, Linda M Styer, Monica M Parker, Shu-Yin John Leung, Johanne E Morne, Danielle Greene, et al. Cumulative incidence and diagnosis of sars-cov-2 infection in new york. *Annals of epidemiology*, 48:23–29, 2020.

A Supplementary Material