




# Estimation of SARS-CoV-2 antibody prevalence through serological uncertainty and daily incidence

Liangliang WANG<sup>1</sup> , Joosung MIN<sup>1</sup>, Renny DOIG<sup>1</sup>, Lloyd T. ELLIOTT<sup>1</sup> , and Caroline COLIJN<sup>2\*</sup> 

<sup>1</sup>Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

<sup>2</sup>Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada

**Key words and phrases:** Bayesian inference; COVID-19; epidemiology; MCMC; sero-prevalence; serology survey.

**MSC 2020:** Primary 62P10; secondary 62F15.

**Abstract:** Serology tests for SARS-CoV-2 provide a paradigm for estimating the number of individuals who have had an infection in the past (including cases that are not detected by routine testing, which has varied over the course of the pandemic and between jurisdictions). Such estimation is challenging in cases for which we only have limited serological data and do not take into account the uncertainty of the serology test. In this work, we provide a joint Bayesian model to improve the estimation of the sero-prevalence (the proportion of the population with SARS-CoV-2 antibodies) through integrating multiple sources of data, priors on the sensitivity and specificity of the serological test, and an effective epidemiological dynamics model. We apply our model to the Greater Vancouver area, British Columbia, Canada, with data acquired during the pandemic from the end of January to May 2020. Our estimated sero-prevalence is consistent with previous literature but with a tighter credible interval.

**Résumé:** Le dépistage sérologique du SRAS-CoV-2 permet d'estimer le nombre de personnes qui ont déjà été infectées (y compris les cas qui ne sont pas détectés au moyen de tests de dépistage réguliers, qui ont varié au cours de la pandémie et d'une province ou d'un territoire à l'autre). Une telle estimation est difficile lorsqu'il existe peu de données sérologiques et que l'incertitude du test sérologique n'est pas prise en compte. Nous proposons dans ce travail un modèle bayésien conjoint visant à améliorer l'estimation de la séroprévalence (la proportion de la population avec des anticorps SRAS-CoV-2) en intégrant de multiples sources de données, des lois a priori sur la sensibilité et la spécificité du test sérologique, et un modèle efficace des dynamiques épidémiologiques. Nous appliquons ce modèle à des données recueillies dans la région métropolitaine de Vancouver (Colombie-Britannique, Canada) pendant la pandémie de fin janvier à mai 2020. Notre estimation de la séroprévalence est cohérente avec la littérature antérieure tout en ayant un intervalle de crédibilité plus précis.

## 1. INTRODUCTION

SARS-CoV-2 has led to more than 6 million confirmed deaths globally and surpassed 500 million confirmed infections (Johns Hopkins University, 2021) as of April 2022. As COVID-19 involves asymptomatic carriers and cases with mild symptoms (Day, 2020), infection with SARS-CoV-2 may be much more widespread than indicated by the number of confirmed cases. Accurate

Additional Supporting Information may be found in the online version of this article at the publisher's website.

\* Corresponding author: [caroline\\_colijn@sfu.ca](mailto:caroline_colijn@sfu.ca)

estimates of sero-prevalence (the proportion of the population with SARS-CoV-2 antibodies) could inform both policy and non-pharmaceutical interventions (NPIs; Flaxman et al., 2020).

Serological studies for COVID-19 have led to estimates of sero-prevalence throughout the pandemic, including in Spain (Pollán et al., 2020), New York (Stadlbauer et al., 2020), and the United Kingdom (Steel & Donnarumma, 2021). However, serology tests are imperfect. Without considering the accuracy of the test, conclusions based on serology measurements may be misleading. For example, a study by the University of California in Santa Clara about serology measurements in Los Angeles has been criticized for their failure to incorporate accurate error rates in their results (McCormick, 2020; Sood et al., 2020).

The aim of this analysis is to provide a Bayesian method for estimating sero-prevalence by integrating data from different sources. Integration of multiple data modalities, such as the inclusion of confirmed case counts, ICU (intensive care unit) data, and death counts, can improve the accuracy of these estimates. In addition, any study of serology measurements must make reference to the sensitivity and specificity (the serological accuracy) of the test, thereby incorporating uncertainty. We consider a prior on these variables and integrate data from confirmed case counts in our Bayesian model. We also incorporate an epidemiological model for infection dynamics. Our epidemiological dynamics model, shown in Figure 1, is an exponential growth and decay model. This model is motivated by the fact that most epidemics grow approximately exponentially during their initial phases (Ma, 2020). Exponential growth and decay are related to commonly used compartmental models for epidemics, such as the susceptible–infectious–recovered (SIR) model and the susceptible–exposed–infectious–recovered (SEIR) model. In a SIR model, the fraction of infectious individuals grows exponentially about the disease-free equilibrium at a rate proportional to the difference between the transmission rate and the recovery rate (Ma, 2020). Similarly, in an SEIR model, the growth rate is also exponential but depends on the transmission rate, the rate at which symptoms develop, and the recovery rate in a more complicated form.

Our estimation approach has two main advantages. First, we use a general Bayesian hierarchical model that can easily incorporate multiple data sources and prior information. Although data from serology surveys are often noisy and limited, the estimation of the sero-prevalence can be improved by integrating multiple data sources (especially those of high quality) and prior knowledge about the sensitivity and specificity of serology tests. Second, our proposed epidemiological dynamics model is simpler than compartmental models.

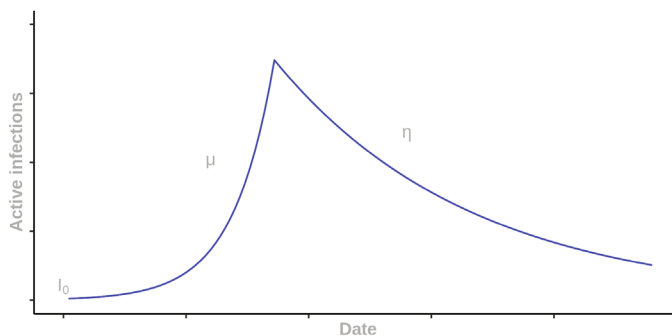


FIGURE 1: Epidemiological dynamics. We assume that the number of active cases increases exponentially with rate  $\mu$  and then decreases according to exponential decay with rate  $\eta$  following the introduction of NPI measures.  $I_0$  is the number of initial cases among the population. This schematic represents sero-prevalence during a rise and then a fall of the pandemic over a single phase. (For example, Phase 1 of the pandemic in B.C. from the end of January until the end of May 2020.) SIR models locally appear as exponential growth and decay.

It is straightforward to integrate our estimation approach with the serological data and case counts.

We apply our method to serology and case count data from the Greater Vancouver area, in British Columbia (B.C.), Canada, collected during Phase 1 (the end of January until the end of May 2020) of the pandemic (The British Columbia Centre for Disease Control broadly divides the pandemic in B.C. into phases covering periods with qualitatively similar epidemiological dynamics and measures). We combine serology measurements and confirmed case data using a Bayesian framework. We compare our results to those in Skowronski et al. (2020). Our results are similar to what is found in Skowronski et al. (2020), but our confidence intervals are tighter.

### 1.1. Serology and Vaccination

Sero-surveys that target the nucleocapsid (N) protein (to which antibodies are not induced by spike-based vaccination) are a valuable tool in estimating the proportion of the population previously infected with SARS-CoV-2 (Krutikov et al., 2022). By contrast, spike antibodies detect either vaccination or past infection. While antibodies and test sensitivity can wane with time, spike antibodies are detectable for months after infection (Krutikov et al., 2022). By building a profile of immunity in the population, sero-prevalence studies can improve our understanding about the future of the pandemic. Knowing the proportion of the population that has SARS-CoV-2 antibodies can help to determine whether a region or country is likely to require enhanced transmission control measures, including testing, contact tracing, self-quarantine, and non-pharmaceutical interventions (Subramanian, He & Pascual, 2021; Yang & Shaman, 2021). Furthermore, many SARS-CoV-2 infections are mild or asymptomatic, and are not detected by surveillance systems that focus on symptomatic individuals (Subramanian, He & Pascual, 2021). During and after the recent Omicron wave, many jurisdictions dramatically reduced testing, focusing only on symptomatic individuals at high risk of severe COVID-19. These factors mean that tracking sero-prevalence is of critical importance in characterizing immunity, and that sero-prevalence remains important after the deployment of vaccine programmes. Past infection and vaccination both induce cellular immune responses that protect individuals from severe disease (Xu, Dai & Gao, 2021; Hall et al., 2022). Our work does not focus on projecting forward and on the likely future health care impact of COVID-19 infections, but this immunity characterization (vaccination and infection combined) would be relevant for that task. The presence of either spike or nucleocapsid antibodies will indirectly convey information about the level and profile of cellular immune protection.

## 2. MATERIALS AND METHODS

Our model for antibody prevalence and the accuracy of serological assays is informed by two data modalities: confirmed cases of COVID-19 from testing; and serological data from a serology survey. In order to incorporate both types of data into our model, we define an underlying generative Bayesian process, and then we derive Bayesian inferences on the model parameters (including the sero-prevalence parameter and the serological accuracy, which will be detailed later in this section).

### 2.1. Epidemiological Dynamics Model

We denote the number of confirmed cases at time  $t$  by  $c_{cc}(t)$ . These case counts are typically reported daily. To specify our generative process, we first define the process underlying the spread of active cases (the case count) in the population. We consider a model with exponential growth followed by exponential decay for the dynamics for the true number of new cases in the population. This model reflects the assumption that the reproduction number could be above one before sufficient NPI measures are instituted, and then below one after such measures

are instituted. Specifically, we use a four-parameter exponential growth and decay model. We assume that at time  $t = 0$  there are  $I_0$  infected individuals in the population. The number of infections increases exponentially at rate  $\mu > 0$  until some time  $\tau$ . After the time  $t = \tau$ , the number of infected individuals is assumed to decrease exponentially with decay rate  $\eta > 0$ . This model is given by the following equation:

$$I(t) = \begin{cases} I_0 \exp(\mu t), & \text{if } t < \tau, \\ I_0 \exp(\mu\tau - \eta(t - \tau)), & \text{if } t \geq \tau. \end{cases} \quad (1)$$

Here,  $I(t)$  is the number of new infections at time  $t$  (both confirmed cases and cases that are unconfirmed because they are asymptomatic or untested). We assume that some proportion,  $p_{cc}$ , of active cases are reported in the confirmed cases. Of interest for serological assays is the cumulative number of infections at time  $t$ . We define this generally as  $C(t) = \int_{s=0}^t I(s) ds$ . However in practice, because we have counts at regular and discrete time points, it can be written as  $C(t) = \sum_{s=0}^t I(s)$ .

Immunoassays are used to determine how many people have been infected with COVID-19, by testing if an individual has SARS-CoV-2 antibodies. This allows us to measure the sero-prevalence of SARS-CoV-2 antibodies in a given population at time  $T$ . These data consist of two integers:  $m_p$ , the number of people who tested positive for SARS-CoV-2 antibodies and  $n$ , the total number of people tested.

In order to analyze the serological data, we must take into consideration the accuracy of the testing procedure. To do this, we consider the sensitivity and specificity of the immunoassays. The sensitivity of a test is the probability that an individual who has antibodies (D+) tests positive (T+): the true positive rate. Specificity is the probability that an individual who has not contracted the illness (D-) tests negative (T-): the true negative rate. We denote each of these respectively as  $S_{sens} = P(T+ | D+)$  and  $S_{spec} = P(T- | D-)$ . When modelling the observed data, we consider the overall probability that a given test comes back positive. The probability of a positive test is thus given by:

$$\begin{aligned} p &= P(T+) = P(T+, D+) + P(T+, D-) \\ &= P(D+)P(T+ | D+) + P(D-)P(T+ | D-) \\ &= p_s S_{sens} + (1 - p_s)(1 - S_{spec}). \end{aligned} \quad (2)$$

Here,  $p_s$  is the probability that a randomly selected individual from the population has the antibodies. Relating the serological data to the cumulative confirmed cases can then be done by writing  $p_s = C(T)/N$ , where  $N$  is the total population from which the  $n$  serology tests were randomly administered.

## 2.2. A Bayesian Model for Serology

The model we propose here is a fully Bayesian hierarchical model for serological data (including an estimate of antibody prevalence at a single point in time). Our model is informed by epidemiological data, in the form of regular confirmed case counts. The likelihood for this model considers both confirmed cases and results from a serological survey. The first component of the likelihood arises from the case counts. These are assumed to follow a binomial distribution in which the probability of observation is  $p_{cc}$ . We assume that there is some delay between the time when an individual is infected and the time when the test is reported. Accordingly, the confirmed cases on a day  $t$  are distributed according to  $c_{cc}(t) \sim \text{Binomial}(I(t - \tau_0), p_{cc})$ .

In the serological data,  $m_p$  is also assumed to follow a binomial distribution. Here the size parameter is the number of tests administered ( $n$ ) and the binomial proportion is the

probability of testing positive,  $p$ . These data are thus distributed according to  $m_p \sim \text{Binomial}(n, p)$ . Compressing notation, here we denote all observations as  $\mathbf{y} = \{(c_{cc}(t))_{t=1}^T, m_p\}$  and all parameters as  $\theta = \{\mu, \eta, \tau, \tau_0, I_0, p_{cc}, S_{sens}, \text{ and } S_{spec}\}$ . The full likelihood is thus:

$$L(\mathbf{y}|\theta) = \binom{n}{m_p} p^{m_p} (1-p)^{n-m_p} \prod_{t>\tau_0} \binom{I(t-\tau_0)}{c_{cc}} p_{cc}^{c_{cc}} (1-p_{cc})^{I(t-\tau)-c_{cc}}. \tag{3}$$

Here, we have defined  $p$  in terms of  $p_s, S_{sens}, S_{spec}$ , as in Equation (2). Note from Figure 2, case counts  $c_{cc}$  depend on  $I(t)$ , and the serology data  $m_p$  depend on  $p$ . Because  $p$  is a function of  $I(t)$ ,  $m_p$  also depends on  $I(t)$ . Consequently,  $m_p$  and  $c_{cc}$  are dependent and we consider the full likelihood in Equation (3) as a joint model rather than a product of two independent likelihood functions.

Three of our model parameters only take values in the (0, 1) interval:  $p_{cc}, S_{sens}$ , and  $S_{spec}$ . For these parameters we assume a beta prior distribution with shape parameters  $(a_{p_{cc}}, b_{p_{cc}})$ ,  $(a_{S_{sens}}, b_{S_{sens}})$ , and  $(a_{S_{spec}}, b_{S_{spec}})$  respectively. The remaining parameters can take values either in  $(0, \infty)$ : ( $\mu$  and  $\eta$ ), or in  $[1, \infty)$ : ( $I_0, \tau$ , and  $\tau_0$ ). For these parameters we assign a normal distribution truncated appropriately on the left, with hyperparameters for their means and standard deviations. These hyperparameters and priors are listed explicitly as follows:

$$p_{cc} \sim \text{Beta}(a_{p_{cc}}, b_{p_{cc}}),$$

$$S_{sens} \sim \text{Beta}(a_{S_{sens}}, b_{S_{sens}}),$$

$$S_{spec} \sim \text{Beta}(a_{S_{spec}}, b_{S_{spec}}),$$

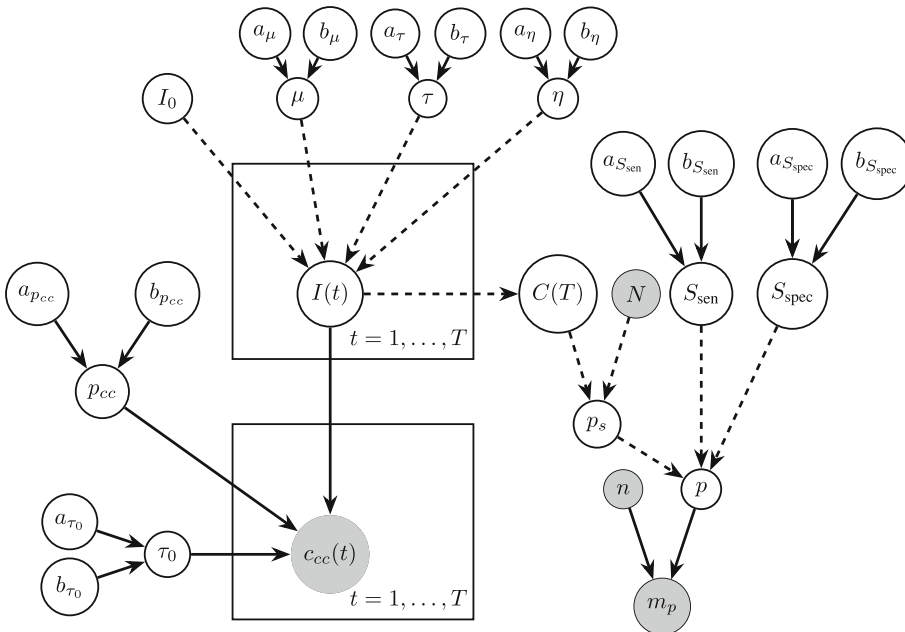


FIGURE 2: Plate diagram (Koller & Friedman, 2009) indicating the acyclic graph governing the relationships among variables in our model. Shaded circles denote constants, circles denote variables, solid arrows denote stochastic dependency, heavy dotted arrows denote deterministic dependency, and rectangles denote “plates” for indices.

$$\tau \sim \text{Truncated Normal}(1, \infty; a_\tau, b_\tau),$$

$$\tau_0 \sim \text{Truncated Normal}(1, \infty; a_{\tau_0}, b_{\tau_0}),$$

$$\mu \sim \text{Truncated Normal}(0, \infty; a_\mu, b_\mu),$$

$$\eta \sim \text{Truncated Normal}(0, \infty; a_\eta, b_\eta).$$

Bayes' rule can be used to compute the posterior  $p(\theta|\mathbf{y})$  from Equation (3) and prior distributions up to a constant of proportionality:

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto L(\mathbf{y}|\theta)\pi(\theta) \\ &= \binom{n}{m_p} p^{m_p}(1-p)^{n-m_p} \prod_{t>\tau_0} \binom{I(t-\tau_0)}{c_{cc}} p_{cc}^{c_{cc}}(1-p_{cc})^{I(t-\tau)-c_{cc}} \\ &\quad \times \pi(p_{cc})\pi(S_{\text{sens}})\pi(S_{\text{spec}})\pi(\mu)\pi(\eta)\pi(I_0)\pi(\tau)\pi(\tau_0), \end{aligned}$$

where  $\pi(\cdot)$ 's denote the prior distributions.

We perform posterior inference on this model using the *Stan* (The Stan Development Team, 2020) language. This is a probabilistic programming language that derives MCMC (Markov chain Monte Carlo) updates for estimating the posterior distribution of a generative process, conditioned on observations (*Stan* was also used in other COVID-19 work such as Flaxman et al., 2020, Manevski et al., 2020, and Unwin et al., 2020).

### 3. SIMULATION STUDIES

We explore the properties of our model through three simulation studies. In the first simulation study, we confirm that known parameter settings can be estimated and recovered by inference. In the second study, we examine the asymptotic behaviour of our model. In the final simulation study, we extend the model to simulated data with multiple breakpoints, showing how our model would perform on data spanning multiple phases of the pandemic.

#### 3.1. Simulation I: Parameter Recovery

To assess the extent to which parameters of our model can be estimated accurately, we simulate data according to the parameter settings displayed in Table 1 (these are the same parameters that we use for our experiment on real B.C. serological data). We simulate confirmed cases along a trajectory of 120 days ( $t = 1, \dots, 120$ ). We can compute the trajectory of infections  $I(t)$  through Equation (1) by using the parameter settings in Table 1. The sero-prevalence is computed as  $p_s = \sum_{t=1}^{120} I(t)/N$  (where  $N$  is the population size) and the probability of a positive serological test is computed from the relationship defined in Equation (2). We then simulate a sequence of case counts and the number of positive serology tests from the binomial distributions as defined in Section 2.2. We simulate 50 datasets using the parameter settings in Table 1 and aggregate our results over these 50 independent datasets in Figure 3.

Through this simulation study, we aim to answer two main questions: (i) How accurately do the posterior means estimate the true parameter values, and how well does the 95% credible interval capture the true values? (ii) To what extent does the choice of prior distribution affect the results? To this end, we fit our model under four different prior settings. We first distinguish the parameters into “positive-valued”,  $\mu, \eta, I_0, \tau_0, \tau$ , and “[0, 1]-valued”,  $p_{cc}, S_{\text{sens}}, S_{\text{spec}}$ . For the positive-valued parameters, we compare truncated normal and gamma priors, while keeping the 0–1-valued priors set according to a beta distribution. Similarly, for the [0, 1]-valued parameters, we compare truncated normal and beta priors, while keeping the positive-valued priors set to a

TABLE 1: Simulation I. Parameter settings for our simulation study, based on priors used for the B.C. data.

Parameter	Interpretation	Value
$\mu$	Growth rate	0.12
$\eta$	Decay rate	0.025
$I_0$	Initial active cases	5
$\tau_0$	Delay between infection and observation	7
$\tau$	Time at which daily cases begin to decrease	43
$p_{cc}$	Probability an active case is observed	0.05
$S_{sens}$	Sensitivity of immunoassay	0.96
$S_{spec}$	Specificity of immunoassay	0.99
$T$	Time at which serology tests are performed	120
$N$	Total population size	2,850,000
$n$	Number of serology tests administered	900

truncated normal. We select the prior parameters in such a way that the prior mean and variance for any given parameter is the same for each choice of prior distribution. A full list of prior parameter settings can be found in Table 5 of the Supplementary Material.

Simulation results for each prior setting are summarized in Figure 3. Although we purposely use prior distributions that are relatively far from the true parameter values in this simulation, the average 95% credible intervals for the posteriors in these simulations overlap with the true parameter values for all parameters. More specifically, the average posterior means for the growth rate and decay rate are close to the true values, although with some bias. The biases for the other parameter values are generally towards the prior distributions. Such results suggest that we can alleviate the biases by using more reasonable prior distributions and/or using a larger data set to reduce the effect of priors. Additionally, the similarity in performance across prior families suggests that our methods are not sensitive to the choice of prior distribution. A grid of pairwise scatterplots of the posterior samples of  $I_0$ ,  $p_{cc}$ , and  $\mu$  is shown in Figure 4. These three parameters were chosen to contrast with the scatterplots of the same parameters in the data analysis. The figure shows some weak negative correlation between  $I_0$  and  $p_{cc}$ . A full grid of pairwise scatterplots can be found in Figure 9 in the Supplemental Material.

### 3.2. Simulation II: Asymptotic Behaviour

The next simulation study we perform aims to confirm that: (i) our model estimates converge to the true values as more data become available, and (ii) that this convergence does not depend on the concentration of prior information. To assess the asymptotic behaviour of the posterior mean as an estimator, we perform an additional simulation wherein 50 datasets are generated, all of which represent a scenario where there are more sero-positive individuals in the population and more serology tests administered. In this model, we select values for  $\mu$ ,  $\eta$ , and  $\tau$  such that the maximum number of daily infections ( $I(t)$ ) is approximately 10,000.  $T$  and  $m$ , the time and number of serology tests administered, are set to 1200 and 90,000, respectively. To assess the performance of the posterior mean in this setting, we estimate three quantities: relative bias (bias divided by the true value); root mean square error (RMSE); and the coverage probability of the 95% posterior credible interval. We fit our model to each of the 50 datasets under three levels of prior variance in order to assess sensitivity to prior information; the prior settings used are summarized in Table 6 of the Supplementary Material. The results at the middle prior

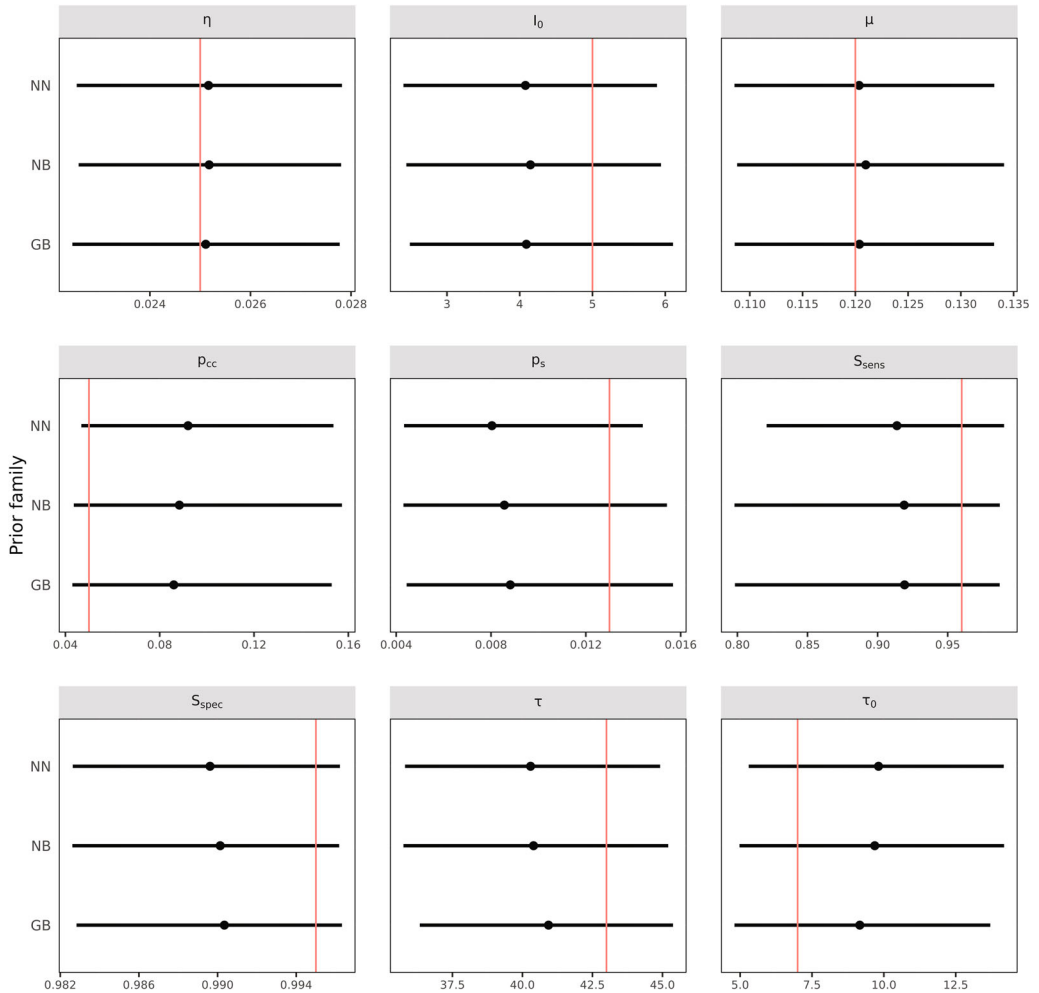


FIGURE 3: Simulation I. The average 95% credible intervals (solid lines) and posterior means (solid points) under each prior setting. Priors used were truncated Normal/Beta (NB), Gamma/Beta (GB), and truncated Normal/truncated Normal (NN) for the positive-valued/[0, 1]-valued prior distributions respectively. Red lines indicate true parameter values.

variance level are shown in Table 2. These results demonstrate that the posterior distributions concentrate on the true values as the amount of serology data available increases. It follows that these potential estimation issues, such as those for  $I_0$  and  $p_{cc}$  shown in Figure 4, are a result of an inadequate amount of data and not a fundamental issue with the estimability of our model. A full table of results at each of the prior variance levels can be found in the Supplemental Material in Table 6.

### 3.3. Simulation III: Multiple Breakpoints

Our final simulation study demonstrates that our model can be generalized to infections with multiple peaks. In this version of our model, we have two change points:  $\tau_1$  and  $\tau_2$ . These parameters delineate three distinct periods of exponential growth/decay: A first “wave” of cases prior to  $\tau_1$  at rate  $\mu_1$  followed by a period of decay at rate  $\eta$  until time  $\tau_2$ , followed by a period of growth at rate  $\mu_2$ . In this simulation, the time and number of serology tests administered are



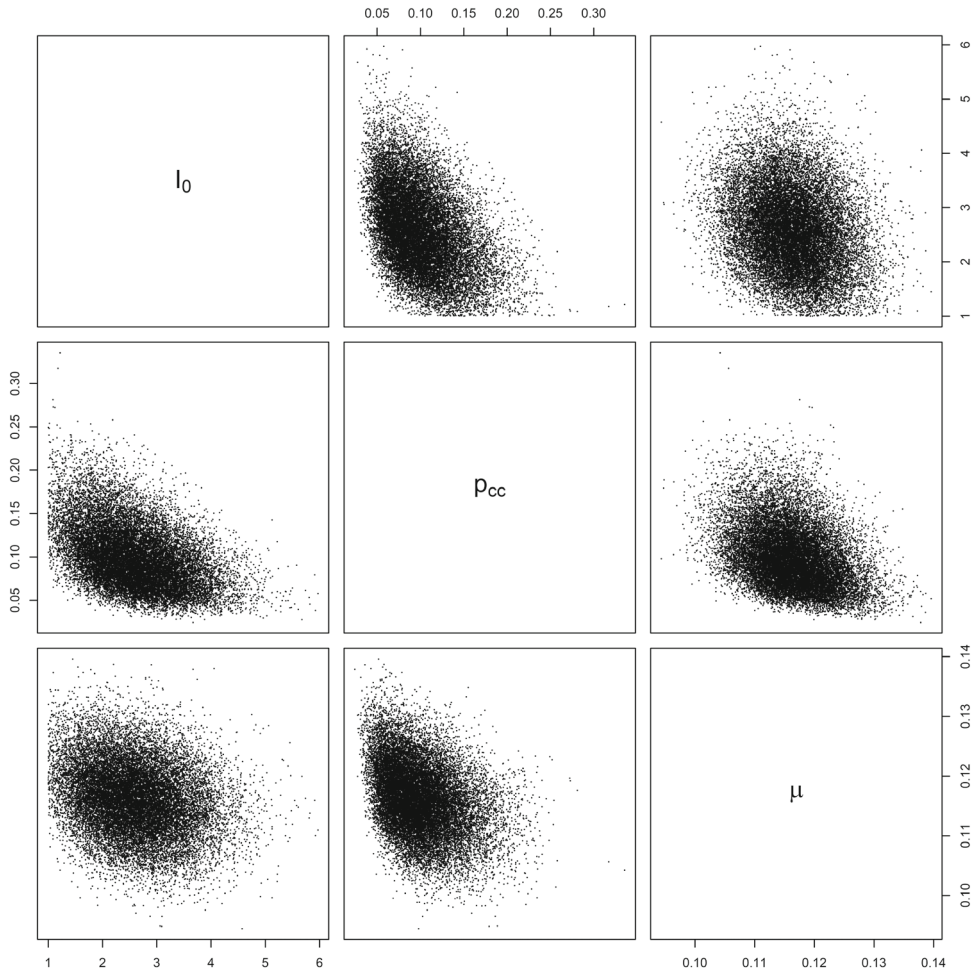


FIGURE 4: Simulation I. Pairwise scatterplots of the posterior samples from our first simulated dataset. A beta prior distribution was used for  $p_{cc}$  and a truncated normal prior was used for  $I_0$  and  $\mu$ .

unchanged from Simulation I, but case counts are generated up to  $T = 17,555$  time units past the time when serology tests are done. We examine multiple breakpoints in two experiments. In the first experiment, breakpoints were fixed to their known true values. In the second experiment, breakpoints were inferred from the data. The prior settings for both experiments can be found in Table 7 of the Supplemental Material.

For the estimated model associated with the estimated trajectory shown in Figure 5, we fix the change points at the true values, because these can usually be ascertained by visually inspecting case count data. The results suggest that our model can provide a reasonable estimate of the daily infections with multiple change points. Additionally, the 95% and 50% credible bands (shown by the light and dark grey regions, respectively) show that the true trajectory of daily infections is contained in the 95% band at all time points, but only by the 50% band when cases are low.

An additional model was fit wherein the change points were estimated. The resulting estimated trajectory is not shown, but is similar to that seen in Figure 5. Histograms of the posterior samples are shown in Figure 6. Solid black lines indicating the prior density are overlaid in Figure 6. Parameters that are deterministic functions of other parameters do not have their densities shown on the grid. Parameters whose prior density appear to be flat have posterior

TABLE 2: Simulation II. The relative bias and RMSE of the posterior mean and the coverage probability of the 95% posterior credible interval over 50 simulated datasets generated with more sero-positive individuals in the population.

Parameter	Relative bias	RMSE	Coverage probability
$I_0$	-0.0096	0.0101	1.00
$\tau_0$	-0.0020	0.0321	0.96
$p_{cc}$	0.0093	0.0101	1.00
$S_{sens}$	0.0097	0.0101	1.00
$S_{spec}$	0.0052	0.0052	1.00
$P(T+)$	-0.0004	0.0021	0.92
$P(T-)$	0.0011	0.0057	0.92
$p_s$	-0.0096	0.0101	1.00

Note: Results only shown for middle prior variance level.

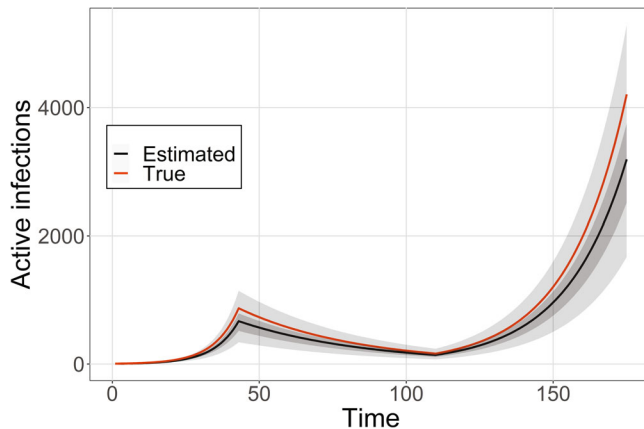


FIGURE 5: Simulation III. Estimated and true trajectories of the daily number of infected individuals. The orange line represents the true trajectory while the black line is the posterior mean.  $x$  axis indicates time in days since January 26. The 95% and 50% posterior credible bands are shown by the light and dark grey regions, respectively. Credible bands are found by interpolating the upper and lower bounds of the credible intervals using the R function `geom_ribbon`.

distributions sufficiently far out in the tail of the prior that they appear only as a horizontal line. We can see that the posterior distributions of  $\tau_1$  and  $\tau_2$  do not differ much from their prior distributions. In combination with the reasonable trajectory estimates, this observation suggests that exact knowledge of the change points is not necessary for good estimation of daily infections.

#### 4. B.C. PREVALENCE AND ACCURACY FROM SEROLOGY

We analyze data from the Greater Vancouver area in B.C., Canada for the Fraser and Vancouver Coastal Health Authorities. This analysis is based on serological data from Summer 2020 acquired by Skowronski et al. (2020). We use confirmed case data from the BCCDC (British Columbia Centre for Disease Control, 2020). Reporting of daily case counts starts on January 26, 2020 and is collected on weekdays. The Skowronski et al. (2020) study shows that in the

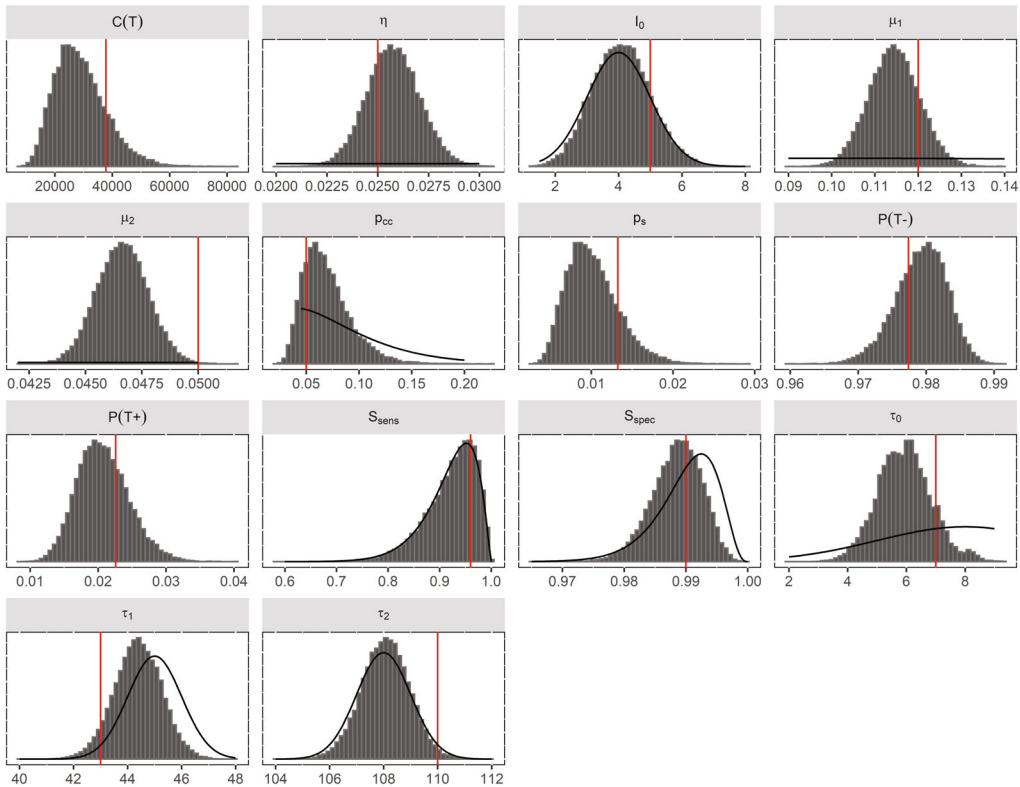


FIGURE 6: Simulation III. Posterior samples of model parameters for multiple change-point simulation study. Solid black lines indicate prior densities.

Greater Vancouver area, of 885 people tested between May 15 and May 27, four tested positive. For this analysis, we assume the population of the Greater Vancouver area around that time is 2.85 million (The Government of British Columbia, 2020). To determine settings for the exponential decay rates and growth rates, we referred to simulations we performed for which  $\mu = 0.1$  and  $\eta = 0.05$  (these are the growth rate and decay rate for a single phase of the pandemic in B.C.). These growth and decay rates are sourced from Anderson et al. (2020). To account for the uncertain nature of the prior data, we assign a standard deviation of 0.1 for the parameters  $\mu$  and  $\tau$ . The mean values for these estimates are also obtained from Anderson et al. (2020) and references therein. In that work, the estimate of the delay between infection and reporting was found to be approximately 1 week. Cases in B.C. first started to decrease around mid to late March 2020, giving us a reasonable mean prior estimate for  $\tau$  using 50 days since January 26, 2020. We set the prior mean for  $I_0$  (the initial number of infections) to reflect the estimate of eight active cases in B.C., as that number was reported on February 1, 2020 (Anderson et al., 2020).

In B.C., testing protocols were changed on April 14, leading to an increase in the testing rate (British Columbia Centre for Disease Control, 2020). For this reason we modified the likelihood function to incorporate two testing rates,  $p_{cc1}$  and  $p_{cc2}$ , which indicate the testing probability that a case is observed through a polymerase chain reaction (PCR) test before or after (respectively) April 14, 2020. Computing the binomial likelihood for a date  $t$  prior to April 14, 2020 ( $t < 76$ ) is done using  $p_{cc1}$ . After that date,  $p_{cc2}$  is used. The prior hyperparameters for these variables are set to (35, 65) and (65, 35) respectively, yielding a mean of 0.35 for  $p_{cc1}$  and 0.65 for  $p_{cc2}$ . These two values are found as the estimated sampling proportions for B.C. in Anderson et al. (2020). While our model differs from the model of Anderson et al. (2020), the interpretation

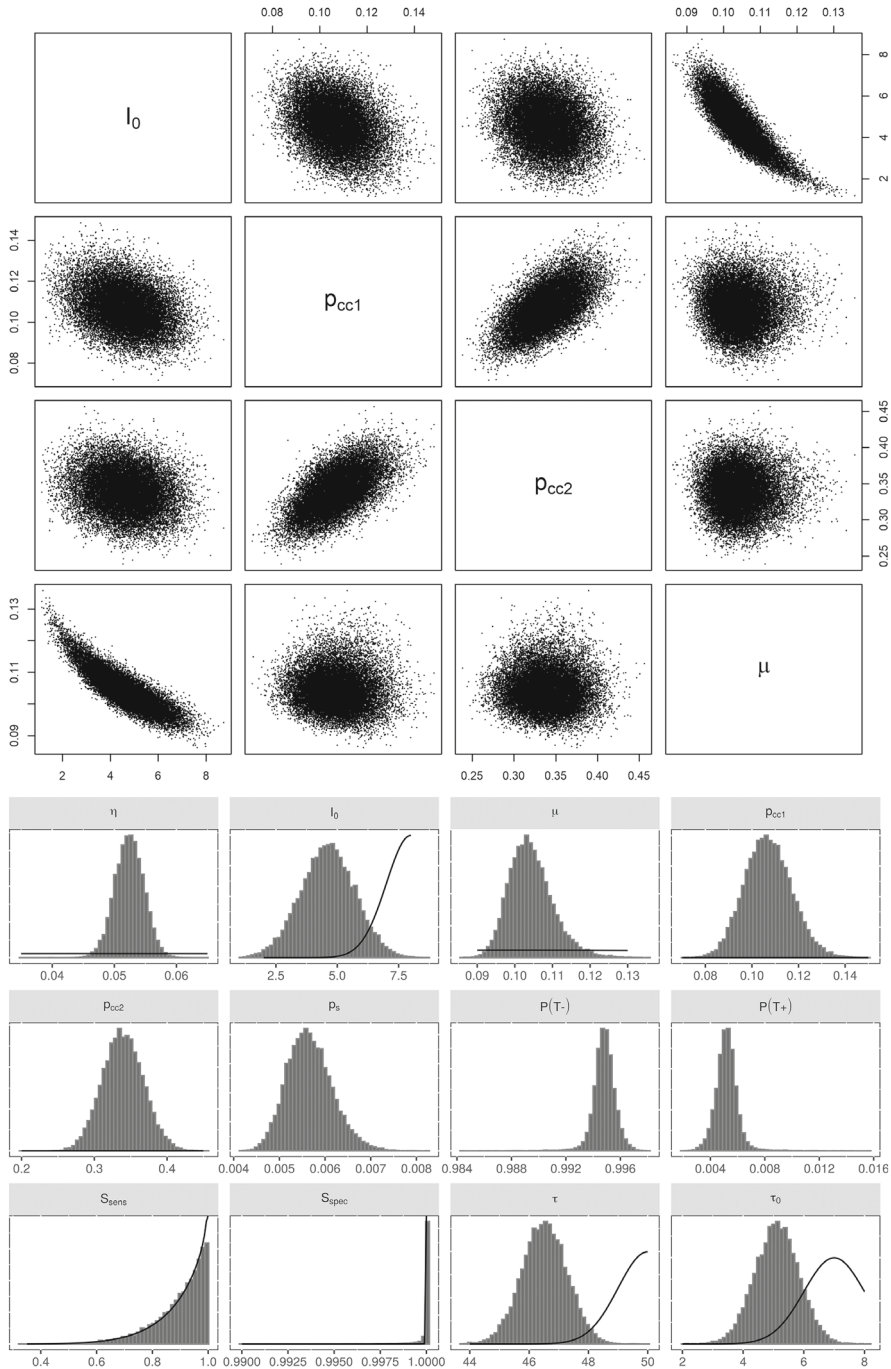


FIGURE 7: Top: Pairwise scatterplots of the posterior samples from our study on B.C. data. A beta prior distribution was used for  $p_{cc1}$  and  $p_{cc2}$ , and a truncated normal prior was used for  $I_0$  and  $\mu$ . Strong correlation is indicated between  $p_{cc1}$  and  $p_{cc2}$ . Bottom: Posterior samples for all model parameters plus derived parameters  $p_s$ ,  $P(T-)$ , and  $P(T+)$  for the experiment on B.C. serological data. All parameters have unimodal distributions, and there is evidence of MCMC mixing. The solid black line indicates the prior density for each parameter with an explicitly defined prior distribution.

TABLE 3: Posterior mean, median, and 95% credible interval for model parameters and derived parameters for B.C. serological data.

Parameter	Mean	Median	95% credible interval
$\mu$	0.104	0.104	(0.095, 0.118)
$\eta$	0.052	0.052	(0.048, 0.057)
$I_0$	4.612	4.620	(2.445, 6.741)
$\tau$	46.517	46.519	(45.029, 47.988)
$\tau_0$	5.111	5.117	(3.694, 6.514)
$p_{cc1}$	0.107	0.107	(0.088, 0.128)
$p_{cc2}$	0.339	0.339	(0.286, 0.395)
$S_{sens}$	0.898	0.924	(0.655, 0.998)
$S_{spec}$	1.000	1.000	(0.999, 1.000)
$p_s$	0.006	0.006	(0.005, 0.007)
$P(T-)$	0.995	0.995	(0.993, 0.996)
$P(T+)$	0.005	0.005	(0.004, 0.007)

Note:  $\mu$  and  $\eta$  indicate the exponential rise and fall of cases, respectively.  $I_0$  indicates the number of cases at the first time point.  $\tau$  indicates the time at which the number of cases changes from exponential growth to exponential decay.  $\tau_0$  indicates the number of days between infection of an individual and reporting.  $p_{cc1}$  and  $p_{cc2}$  indicate the probability of observation (before and after testing procedures were changed).  $S_{sens}$  and  $S_{spec}$  indicate the sensitivity and specificity of the serological assay.  $p_s$  indicates the sero-prevalence.

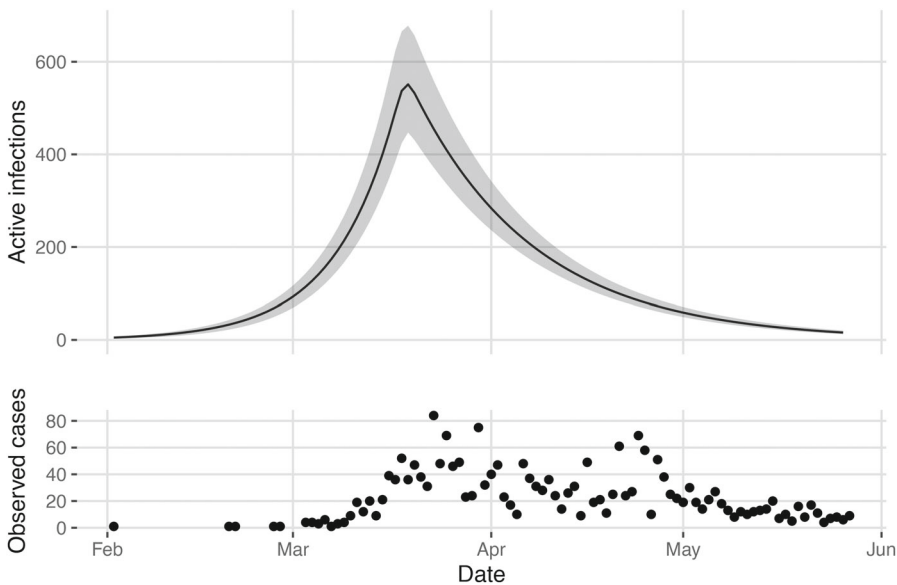


FIGURE 8: Top: Estimated number of infections in the Greater Vancouver area between January 26 and May 27. Posterior mean is shown by the solid line with 95% credible band in grey. Bottom: Observed cases in Vancouver over the same time frame.

TABLE 4: Posterior mean, median, and 95% credible interval for  $p_{cc1}$ ,  $p_{cc2}$ , and  $C$  (the cumulative count on May 27) for Greater Vancouver area serological data.

Parameter	Mean	Median	95% credible interval
$p_{cc1}$	0.107	0.107	(0.088, 0.128)
$p_{cc2}$	0.339	0.339	(0.286, 0.395)
$C$	16,150	16,060	(13,651, 19,193)

of our parameters for testing rates is similar. Based on the sensitivities and specificities reported in Skowronski et al. (2020), we selected prior parameters that yielded means of 0.9 and 0.99 respectively. The variance for both parameters was set to be 0.009: the corresponding shape parameters were computed from the mean and variance.

Our results suggest that during Phase 1 of the pandemic in B.C., approximately 0.57% of the B.C. population had been infected (95% confidence interval [0.48%, 0.68%]). In Figure 7, we provide histograms of the posterior samples from all of the model parameters. This includes the parameters for which we have defined priors, as well as  $p_s$ ,  $P(T-)$ , and  $P(T+)$ . The 95% credible intervals for the posteriors for all of the parameters are provided in Table 3.

A grid of pairwise scatterplots of the posterior samples for  $I_0$ ,  $p_{cc1}$ ,  $p_{cc2}$ , and  $\mu$  is shown in Figure 7. These scatterplots can be used to evaluate the relationships between the parameters of our model. The figure shows that  $p_{cc1}$  and  $p_{cc2}$  are highly correlated. There also appears to be some nearly linear relationship between  $I_0$  and  $\mu$ . This relationship can be understood by considering that by decreasing  $I_0$  and increasing  $\mu$ , we could leave the cumulative number of cases relatively unchanged. Note that a similar pattern is not present in the simulation scatterplots in Figure 4, indicating that the ability to estimate  $I_0$  and  $\mu$  improves as the sample size increases. This pattern suggests that issues in estimating  $I_0$  and  $\mu$  relate to the quantity of information about the infection curve.

From the model parameters, we can obtain samples from the posterior distribution of the number of active cases between January 26 and May 27. We provide our estimate of the number of active cases over the time period examined, according to our posterior samples. The posterior mean and 95% credible interval are shown in Figure 8 along with the observed case counts.

Table 4 summarizes the posterior distribution of the probability of case detection before and after April 14 ( $p_{cc1}$  and  $p_{cc2}$ ), as well as the cumulative number of cases in the Greater Vancouver area between January 26 and May 27. The cumulative number of cases (up to and including May 27) reflects the number of daily cases from our model resulting from the estimated model parameters. We find that the maximum number of simultaneous infections (including unobserved cases) during Phase 1 occurred on March 19 (552 infections, 95% confidence interval [446, 678]).

These results also suggest that on May 27 (the last day of the phase we investigate) the posterior mean percentage of people in the Greater Vancouver area who had COVID-19 antibodies was around 0.57% of the total population (95% confidence interval [0.48%, 0.68%]), approximately eight times the reported number of infections over the same time period. Previously, Skowronski et al. (2020) estimated a sero-prevalence of 0.55% in the Greater Vancouver area at the end of May 2020 (95% confidence interval [0.15%, 1.37%]) using conventional methods, without a connection to confirmed case counts from PCR tests or disease dynamics. These results show that our principled Bayesian methods, which incorporate case counts and disease dynamics, broadly match the conventional methods for estimation of sero-prevalence reported in Skowronski et al. (2020), while providing tighter confidence intervals.

## 5. CONCLUSION AND DISCUSSION

In this work, we have presented a Bayesian model for serological data, integrating test sensitivity and specificity with an epidemiological model for case counts. Our model improves upon previous work in serology measurements (McCormick, 2020; Sood et al., 2020) by incorporating uncertainty about testing accuracy. Bayesian methods allow higher accuracy than maximum likelihood methods in posterior estimation of parameters, through integration over uncertainty. Our prior includes an epidemiological model involving an increase followed by a decrease in case counts. This approximation (an increase, followed by a decrease) is consistent with Phase 1 of COVID-19 case counts in B.C. and in other locations; this model involves examination of COVID-19 data restricted to a single phase.

In B.C., testing policy was modified mid-April 2020 to widen testing to anyone with COVID-19-specific symptoms, overlapping with the phase we consider. We model this change by allowing the probability that an active case is observed to vary on either side of the change in testing procedures. Testing procedures are otherwise broadly constant (and focused on PCR testing of individuals with symptoms of COVID-19). Our procedure mitigates inaccuracies in estimation that could arise if the testing fraction were assumed to be constant.

We have demonstrated that our method can easily model multiple phases by coupling the parameters from one phase to the next (with change points constructed at dates reflecting inflection points identified by external model fits or by the times at which policies were changed). Alternatively, change points could be added as free variables in the *Stan* estimation code (The Stan Development Team, 2020), including the rise and fall of case counts independently over each learned range. In addition, the models we propose could be stratified by age, or extended to include hospitalization, ICU data, and deaths, through a more complex Bayesian model over these multiple modalities.

Our methods allow serological data to be combined with COVID-19 case counts from PCR tests, to better estimate the prevalence of antibody response, to provide insight on the fraction of infections that were detected, and to extrapolate future trends in the pandemic. We report a posterior mean antibody prevalence of 0.57% (95% confidence interval [0.48%, 0.68%]) in the Greater Vancouver area during Phase 1 of the pandemic, which broadly matches the previous report of 0.55% derived from the same serological data (Skowronski et al., 2020). The interval reported in Skowronski et al. (2020) was [0.15%, 1.37%]. Our confidence interval is tighter (varying by 0.19 percentage points, compared with 1.22 percentage points for Skowronski et al., 2020), and nested inside the confidence interval reported in Skowronski et al. (2020). Our posterior conditions on more data (case counts) and is constrained by disease dynamics, and so we expect less uncertainty in our estimates (this is reflected in our tighter confidence intervals). For our experiment on the Greater Vancouver area, our posterior mean matches classical methods (Skowronski et al., 2020), but our estimate is more precise.

We describe some ways in which our model could be improved to address various stages of pandemics, beyond the phase of the COVID-19 pandemic that we examine. If our model were to be applied in the absence of prior modelling studies (for example, during a time period closer to the beginning of an epidemic caused by a novel pathogen), then less informative priors should be used. Even in this setting with less prior information, some sensible parameter ranges can be taken from the physical context of the data. For example, if the initial data collection times occur near the beginning of the infection, we know the initial cases will be fairly low, even if we do not have reliable prior information about the exact number. Similarly, by visual inspection of case count plots, a reasonable prior on the change point,  $\tau$ , can be assigned. Some parameters, such as  $\tau_0$  and  $p_{cc}$ , would have less certainty; however, we can assign an entirely uninformative prior on  $p_{cc}$  if no prior intuition exists and an experienced epidemiologist would likely be able to set a sensible upper bound on the delay between infection and detection. Essentially, the prior

information here is not only previous studies of the same infection but also prior knowledge of infectious diseases in general.

With regards to deployment of our model to serological and case count data acquired after vaccination, note that none of the vaccines with widespread deployment in Canada (Moderna, Pfizer, Astrazeneca) target the nucleocapsid (N) protein. We noted in Section 1.1 that sero-surveys that target the nucleocapsid (N) protein will tend to yield negative for individuals who have been vaccinated but have not been infected. The protocol used by Skowronski et al. (2020) requires dual-assay positivity (spike and nucleocapsid antibodies) for determining seropositivity. Therefore, deployment of our model after vaccination does not require broad changes to the methodology. However, vaccination and changes in vaccine policy will modulate the rate parameters  $\mu$  and  $\eta$ , and so if our model is deployed to a range of data that spans the onset of vaccines or changes in vaccination policy, additional breakpoints should be used (as in Simulation III described in Section 3.3). Furthermore, if the time period is large enough to cover changes in vaccination policy, then waning of antibodies must be considered. This could be incorporated into the model by adding a decay term to the summation for  $C(t)$  yielding  $C(t) = \sum_{s=0}^t \exp(-\alpha(t-s))I(s)$ . Here  $\alpha$  is an additional latent variable indicating the waning rate of antibodies from infection. The prior on  $\alpha$  may be chosen with a mean of  $2.849 \times 10^{-3}$ , corresponding to an antibody half life of 8 months (Krutikov et al., 2022). For additional simplicity (which may improve inference), we could replace  $\exp(-\alpha(t-s))$  with 1 when  $t-s$  is less than 8 months, and with 0 otherwise.

Finally, we note that sensitivity and specificity of serology depend on time since infection (Abbasi, 2020). In this work, we assume that the sensitivity and specificity are not time-varying. This assumption could be relaxed by adding a convolution (as was done in the above discussion of vaccinations) to Equation (3).

## ACKNOWLEDGEMENTS

This work was funded by Genome B.C.'s COVID-19 Rapid Response Funding Initiative (project code COV-142). CC was supported by the federal government of Canada's Canada 150 research chair program. LW was supported by Discovery Grant RGPIN-2019-06131 from the Natural Sciences and Engineering Research Council of Canada (NSERC). LTE was supported by NSERC grants RGPIN-05484-2019 and DGEGR-00118-2019.

## REFERENCES

- Abbasi, J. (2020). The promise and peril of antibody testing for COVID-19. *Journal of the American Medical Association*, 323(19), 1881–1883.
- Anderson, S. C., Edwards, A. M., Yerlanov, M., Mulberry, N., Stockdale, J., Iyaniwura, S. A., Falcao, R. C. et al. (2020). Estimating the impact of COVID-19 control measures using a Bayesian model of physical distancing. medRxiv preprint <https://doi.org/10.1101/2020.04.17.20070086v1>.
- British Columbia Centre for Disease Control. (2020). *B.C. COVID-19 data*. <http://www.bccdc.ca/health-info/diseases-conditions/covid-19/data>
- Day, M. (2020). COVID-19: Identifying and isolating asymptomatic people helped eliminate virus in Italian village. *British Medical Journal (Online)*, 368, m1165.
- Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C. et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820), 257–261.
- Hall, V., Foulkes, S., Insalata, F., Kirwan, P., Saei, A., Atti, A., Wellington, E. et al. SIREN Study Group. (2022). Protection against SARS-CoV-2 after COVID-19 vaccination and previous infection. *The New England Journal of Medicine*, 386(13), 1207–1220.
- Johns Hopkins University. (2021). *COVID-19 Dashboard by the Center for Systems Science and Engineering*, Johns Hopkins University & Medicine, Baltimore. <https://coronavirus.jhu.edu/map.html>. Accessed Winter 2021.



- Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, Cambridge.
- Krutikov, M., Palmer, T., Tut, G., Fuller, C., Azmi, B., Giddings, R., Shrotri, M. et al. (2022). Prevalence and duration of detectable SARS-CoV-2 nucleocapsid antibodies in staff and residents of long-term care facilities over the first year of the pandemic (VIVALDI study): Prospective cohort study in England. *Lancet Healthy Longevity*, 3(1), e13–e21.
- Ma, J. (2020). Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 5, 129–141. <https://doi.org/10.1016/j.idm.2019.12.009>
- Manevski, D., Gorenjec, N. R., Kežar, N., Blagus, R. (2020). Modeling COVID-19 pandemic using Bayesian analysis with application to Slovene data. *Mathematical Biosciences*, 329, 108466.
- McCormick, E. (2020, April 23). Why experts are questioning two hyped antibody studies. *The Guardian*. <https://www.theguardian.com/world/2020/apr/23/coronavirus-antibody-studies-california-stanford>. Accessed Fall 2020.
- Pollán, M., Pérez-Gómez, B., Pastor-Barriuso, R., Oteo, J., Hernán, M. A., Pérez-Olmeda, M., Sanmartín, J. L. et al. (2020). Prevalence of SARS-CoV-2 in Spain: A nationwide, population-based seroepidemiological study. *The Lancet*, 396(10250), P535–544.
- Skowronski, D. M., Sekirov, I., Sabaiduc, S., Zou, M., Morshed, M., Lawrence, D., Smolina, K. et al. (2020). Low SARS-CoV-2 sero-prevalence based on anonymized residual sero-survey before and after first wave measures in British Columbia, Canada, March-May 2020. medRxiv preprint. <https://doi.org/10.1101/2020.07.13.20153148v1>
- Sood, N., Simon, P., Ebner, P., Eichner, D., Reynolds, J., Bendavid, E., & Bhattacharya, J. (2020). Seroprevalence of SARS-CoV-2-specific antibodies among adults in Los Angeles County, California. *Journal of the American Medical Association*, 323(23), 2425–2427.
- Stadlbauer, D., Tan, J., Jiang, K., Hernandez, M., Fabre, S., Amanat, F., Teo, C. et al. (2020). Seroconversion of a city: Longitudinal monitoring of SARS-CoV-2 seroprevalence in New York City. medRxiv preprint. <https://doi.org/10.1101/2020.06.28.20142190v1>
- Steel, K., & Donnarumma, H. (2021). Coronavirus (COVID-19) infection survey, UK: 26 February 2021. Office for National Statistics, UK.
- Subramanian, R., He, Q., & Pascual, M. (2021). Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proceedings of the National Academy of Sciences of the United States of America*, 118(9), e2019716118.
- The Government of British Columbia. (2020). *Census of Canada*, The Government of British Columbia, Victoria. <https://www2.gov.bc.ca/gov/content/data/statistics/people-population-community/census>. Accessed Fall 2020.
- The Stan Development Team. (2020). Stan modeling language users guide and reference manual. <https://mc-stan.org>
- Unwin, H. J. T., Mishra, S., Bradley, V. C., Gandy, A., Mellan, T. A., Coupland, H., Ish-Horowicz, J. et al. (2020). State-level tracking of COVID-19 in the United States. *Nature Communications*, 11(1), 1–9.
- Xu, K., Dai, L., & Gao, G. F. (2021). Humoral and cellular immunity and the safety of COVID-19 vaccines: A summary of data published by 21 May 2021. *International Immunology*, 33(10), 529–540.
- Yang, W. & Shaman, J. (2021). Development of a model-inference system for estimating epidemiological characteristics of SARS-CoV-2 variants of concern. *Nature Communications*, 12(1), 5573.

---

Received 23 March 2021

Accepted 11 April 2022