The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2928–2949
https://doi.org/10.1214/24-AOAS1915

MULTISITE DISEASE ANALYTICS WITH APPLICATIONS TO ESTIMATING COVID-19 UNDETECTED CASES IN CANADA

BY MATTHEW R. P. PARKER[1,a], JIGUO CAO[1,b], LAURA L. E. COWEN[2,d], LLOYD T. ELLIOTT[1,c] AND JUNLING MA[2,e]

[1]*Department of Statistics and Actuarial Sciences, Simon Fraser University,* [a]*mrparker909@gmail.com,* [b]*jiguo_cao@sfu.ca,* [c]*lloyde@sfu.ca*

[2]*Department of Mathematics and Statistics, University of Victoria,* [d]*lcowen@uvic.ca,* [e]*junlingm@uvic.ca*

Even with daily case counts, the true scope of the COVID-19 pandemic in Canada is unknown due to undetected cases. We develop a novel multisite disease analytics model which estimates undetected cases using discrete-valued multivariate time series in the framework of Bayesian hidden Markov modelling techniques. We apply our multisite model to estimate the pandemic scope using publicly available disease count data including detected cases, recoveries among detected cases, and total deaths. These counts are used to estimate the case detection probability, the infection fatality rate through time, the probability of recovery, and several important population parameters including the rate of spread and importation of external cases. We estimate the total number of active COVID-19 cases per region of Canada for each reporting interval. We applied this multisite model Canada-wide to all provinces and territories, providing an estimate of the total COVID-19 burden for the 90 weeks from 23 April 2020 to 10 February 2022. We also applied this model to the five health authority regions of British Columbia, Canada, describing the pandemic in B.C. over the 31 weeks from 2 April 2020 to 30 October 2020.

1. Introduction. The novel coronavirus pandemic began in December 2019, and spread rapidly to every corner of the globe. The impact of the pandemic on lives, societal and social norms, and on the world economy cannot be overstated (Béland et al. (2021), Cypress (2022), Tanaka (2022)). As of March 2022, the global total number of deaths attributed to the pandemic was over 6.1 million, from a recorded total of 474 million reported cases. And Canada has recorded a total of over 3.4 million confirmed cases and more than 37,000 confirmed deaths (Dong, Du and Gardner (2020)). It is well established that the number of confirmed cases of coronavirus are an underreporting of the true total (see, e.g., Feikin, Widdowson and Mulholland (2020), Buitrago-Garcia et al. (2020), Hasan et al. (2021), Chisale et al. (2022)). The reasons for under-reporting are varied and include the presence of asymptomatic, pauci-symptomatic, and presymptomatic cases, lack of testing for low severity cases, and periods with low testing volumes. Some of these causes can be controlled through testing protocols and wider availability/accessibility of testing. However, due to the impracticality of full census testing of large populations in Canada, under-reporting cannot be entirely mitigated. Undetected cases cause community infections and reduces the effectiveness of control measures such as contact tracing, quarantine, and isolation. They also cause underestimates of the social and economical impacts of the pandemic.

To properly understand the scope and impact of the pandemic, we must estimate the true total number of infections. Research conducted after the pandemic began has im-

Received July 2023; revised March 2024.
Key words and phrases. Infectious disease modelling, hidden Markov models, disease parameter estimation, infection fatality rate.

proved methodologies to produce such estimates. These methods include meta-analyses of asymptomatic prevalence (He et al. (2021), Alene et al. (2021)), extensions to susceptible-infectious-recovered (SIR) type modelling (Li et al. (2021), Subramanian, He and Pascual (2021), Huo, Chen and Ruan (2021)), and seroprevalence studies (Bendavid et al. (2021), Saeed et al. (2021), Halili et al. (2022)). Alternative models include integer-autoregressive models (Fernández-Fontelo et al. (2016), Fernández-Fontelo et al. (2020)) and case fatality rate (CFR) models (Dougherty et al. (2021)) and estimating underreporting using discrete count models (Parker et al. (2021)).

Much research has focused on understanding the course of the pandemic within Canada. Examples include seroprevalence of Montréal school age children (Zinszer et al. (2021)), CFR based models to estimate reporting rates across Canada (Dougherty et al. (2021)), intervention strategy analysis in Ontario (Tuite, Fisman and Greer (2020)), mental health and well-being of Canadians during the pandemic (Dozois (2021), Appleby et al. (2022)), models to forecast transmission and incidence (Chimmula and Zhang (2020), Mullah and Yan (2022)), and policy response analysis with comparisons between Canada, France, and Belgium (Desson et al. (2020)). State-space models have been used extensively in the literature to study disease populations, such as the disease N-mixtures work of DiRenzo et al. (2019) and the coronavirus modelling of Moriña et al. (2021). Among this backdrop of vital research, we provide up-to-date disease analytics for Canada as a whole, with results specific to each province and territory.

We propose a novel multisite modelling technique to better estimate disease dynamics such as domestic spread rate, recovery and death probabilities, and to estimate effectiveness of testing protocols and levels of underreporting of cases. By considering data from across the entire time span of the pandemic, estimates can be made throughout its course. Multisite modelling allows the total burden of the pandemic to be estimated across large regions by considering them as an amalgamation of smaller regions. This method of modelling increases precision compared to single site modelling, by pooling information across sites, which is an example of transfer learning (Weiss, Khoshgoftaar and Wang (2016)). The pooled information consists of all the observed data (including any covariate data) used in model fitting, informing estimates for parameters and covariate coefficients which do not vary over sites. Site heterogeneity is modelled by allowing parameters to vary across sites, such as the initial active cases in our case studies. To analyze the burden of COVID-19 as it has progressed through time in Canada, we model the pandemic period from 23 April 2020 to 10 February 2022.

Beyond the increased precision of estimates, our multisite hidden Markov model has several novel features in comparison with the single-site model, and related models (Parker et al. (2021)):

1. We provide a state-of-the-art framework for undercounted disease analytics across multiple sites, which can be easily adapted for study of future pandemics and disease outbreaks.

2. The multisite framework allows for both site- and time-varying covariates, which can improve both estimation accuracy and interpretability of results where data pooling across sites explicitly benefits the estimation of the parameters. In comparison, the single-site model can only account for time-varying covariates.

3. We model the disease spread rate separately for the detected and the undetected disease cases. This inclusion is nontrivial, and we show that the new parameters are identifiable. As seen in our Canada case study (Figure 5), the difference in rate of spread between the detected and undetected infections is substantial, and previ-

ous methodology would have only estimated the weighted average of the two rates. In contrast, the single-site model does not separately estimate the two modes of disease spread.

4. Alternative methods of estimating disease dynamics, which do not account for underreporting, are unable to provide estimates of infection fatality rate (IFR). Estimating IFR is necessary to understand the true mortality risk associated with contracting a disease, and our models provide ready estimates of IFR.

5. Under the assumption of a geometric distribution on recovery, we introduce a method of estimating the average recovery interval using a geometric series, which provides a principled measure for both patients and caregivers to understand the probable duration of illness (see Appendix A.3).

6. Unlike the single-site models, maximum likelihood estimation would be completely intractable due to the large number of latent states. For this reason we use Bayesian MCMC to improve computational efficiency and obtain model estimates and credible intervals.

We investigate the effect of vaccination coverage on virus spread rates, recovery probability, and death probability through inclusion of a vaccine coverage covariate. We also include indicators for time periods demarcated by the first confirmed cases in Canada of the variants of concern Delta (Mahumud et al. (2022)) and Omicron (Araf et al. (2022)), which have been shown to have very different vaccine efficacies (Kahn et al. (2022)). These indicators allow us to model changes in disease dynamics and interaction between vaccine efficacy and the dominant variant of concern.

The main contributions of this paper are: (1) a novel multisite disease analytics model, (2) estimates of the total burden of COVID-19 across Canada for 90 weeks of the pandemic, (3) estimates of reporting rates for each province and territory of Canada, (4) estimates of domestic spread rates among both detected and undetected cases, (5) estimates of IFR for COVID-19 in Canada, (6) estimates of average recovery period for active cases, and (7) comparisons between competing models.

2. Methods.

2.1. *Multisite model*. Throughout this work we use the following terminology. Currently infectious individuals present in the population are referred to as active cases. Initial active cases are the total active cases present in a region at the start of study, can be zero prior to the pandemic start date, or can be larger than one when the study start date is later than the pandemic start date. A new case is an active case which was not active during the previous sampling occasion. An active case can be either observed, meaning that the case has been detected/confirmed, or an active case can be unobserved, meaning that the case has not been detected/confirmed. Unobserved cases occur due to lack of testing or lack of reporting or (less commonly) due to false negative test results.

In our new multiple-site disease analytics model, each site is treated as statistically independent so that the likelihood function is a product of single-site likelihoods. The independent site assumptions are: a new case in site A is independent of a new case in site B; a new death in site A is independent of a new death in site B; a new recovery in site A is independent of a new recovery in site B; a newly observed case in site A is independent of a newly observed case in site B; a newly imported case in site A is independent of a newly imported case in site B.

TABLE 1
Definitions for each variable used in construction of the multisite disease analytics model

Statistics	
M	number of sampling sites
T	number of sampling occasions
a_{it}	detected cases still active at site i, time t $i \in \{1, 2, \ldots, M\}, t \in \{1, 2, \ldots, T\}$
n_{it}	new detected cases at site i, time t
d_{it}	new detected deaths at site i, time t
r_{it}	new detected recoveries at site i, time t
H_i	total population size at site i
Latent States	
N_{it}	total active cases at site i, time t
A_{it}	cases which remain active in site i from time t to $t+1$
D_{it}	cases which die in site i from time t to $t+1$
R_{it}	cases which recover in site i from time t to $t+1$
S_{it}	new cases from domestic spread in site i from time t to $t+1$
G_{it}	new cases from importation in site i from time t to $t+1$
Parameters	
λ_i	expected initial active cases per site
p_a	probability of a case remaining active
p_d	probability of a case dying
p_r	probability of a case recovering
ω_1	expected new domestic spread from unobserved cases
ω_2	expected new domestic spread from observed cases
γ_i	expected new imported cases per site
p	probability of detecting an active case
Derived Variables	
δ_i	fraction of population susceptible at site i
Ω_{it-1}	expected new domestic spread in site i time $t-1$ from all sources
α	inflation factor for proportion of observed deaths

Definitions for each of the variables used in constructing our model are listed in Table 1. The latent states (i.e., the unobserved states, unrelated to the latent stage in disease progression) in our model are identifiable due to parametric assumptions. We use the lowercase variables n_{it}, a_{it}, r_{it}, and d_{it} to denote observed data, while we use upper case variables N_{it}, A_{it}, R_{it}, and D_{it} to denote the unobserved latent states. For example, the latent states N_{it} are the total infectious individuals in the population, while the observed data n_{it} are the newly observed active cases, which are a subset of N_{it}. Let n_{it} denote the new detected case counts at sampling occasion $t = 1, \ldots, T$ and study site $i = 1, \ldots, M$. The new detected recoveries between t and $t+1$ among all detected active cases are r_{it}. The new detected deaths between t and $t+1$ are d_{it} (when deaths are fully detected, we let $d_{it} = D_{it}$ and D_{it} cease to be considered latent states).

The total number of active cases at time t are N_{it}, and a_{it} cases among them have been detected. The active cases at time t that will recover, die, or remain active between t and $t+1$ are, respectively, R_{it}, D_{it}, A_{it}. In the following we split the model into 10 components for

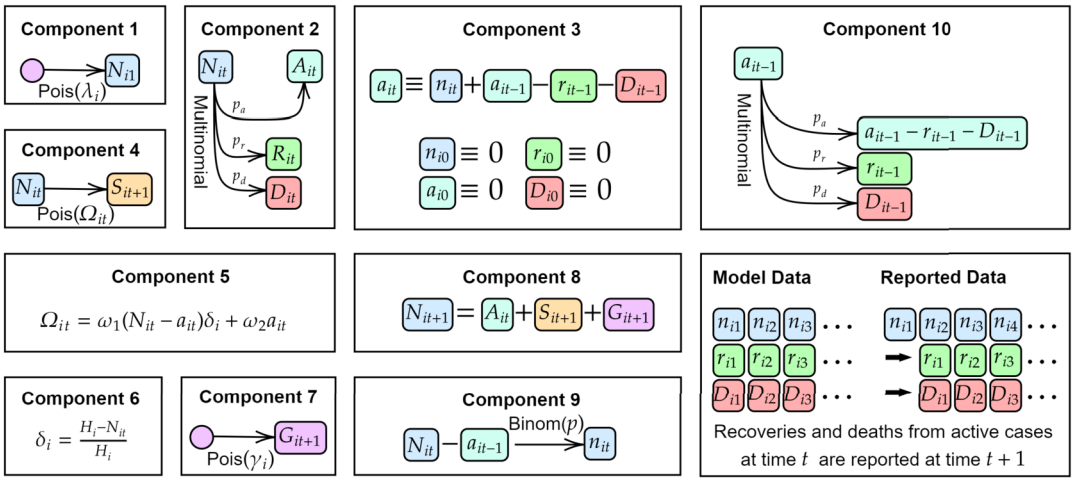

FIG. 1. *Diagram showing the* 10 *components of the multisite disease analytics model as well as the model data vs. the reported data matrices.*

ease of understanding:

(1) Initial Active Cases: $N_{i1} \sim \text{Poisson}(\lambda_i)$.

(2) Latent State Process: $\{A_{it}, D_{it}, R_{it}\} \sim \text{Multinomial}(N_{it}; p_a, p_d, p_r)$.

(3) Detected Active Cases: $a_{it} = n_{it} + a_{it-1} - r_{it-1} - d_{it-1}$, for $t > 0$.

(4) Domestic Spread: $S_{it} \sim \text{Poisson}(\Omega_{it-1})$, for $t > 1$.

(5) Ω_{it-1} : $\omega_1(N_{it-1} - a_{it-1}) \cdot \delta_i + \omega_2 a_{it-1}$ (mean domestic spread).

(6) δ_i : $(H_i - N_{it})/H_i$ (fraction of susceptible population, where H_i is the total population size).

(7) Imported Cases: $G_{it} \sim \text{Poisson}(\gamma_i)$, for $t > 1$.

(8) Active Cases Updates: $N_{it} = A_{it-1} + S_{it} + G_{it}$, for $t > 1$.

(9) Observation Process I: $n_{it} \sim \text{Binomial}(N_{it} - a_{it-1}, p)$.

(10) Observation Process II: $\{a_{it} - d_{it} - r_{it}, d_{it}, r_{it}\} \sim \text{Multinomial}(a_{it}; p_a^*, \alpha p_d, p_r)$.

Each model component is also summarized in Figure 1, and the full model diagram is shown in Figure 2. Components (1) through (8) describe the disease dynamics through time, while Components (9) and (10) describe the data detection mechanisms.

Component (1) models the initial active cases N_{i1}, using a Poisson distribution with mean λ_i. Here λ_i is the expected initial number of active cases per site i at the start of the study. N_{i1} is the total current active infections in the population of site i during the first sampling occasion. When the parameter λ_i is shared across multiple sites (e.g., $\lambda_1 = \lambda_2$) overdispersion may be an issue, and the Poisson distribution used in (1) may not be appropriate. When this is the case, a negative binomial distribution could be considered instead. Our two case studies allow λ_i to vary over sites, and so overdispersion is not an issue here.

The latent state process (Component 2) uses a multinomial distribution to partition each active case in N_{it} into one of the three categories A_{it} (remain active), D_{it} (die), and R_{it} (recover). The parameter p_a is the probability of remaining an active case between sampling occasions, while p_r and p_d are the probabilities of recovering or dying between sampling occasions. Here $p_a = 1 - p_r - p_d$. We assume that individuals are indistinguishable, which

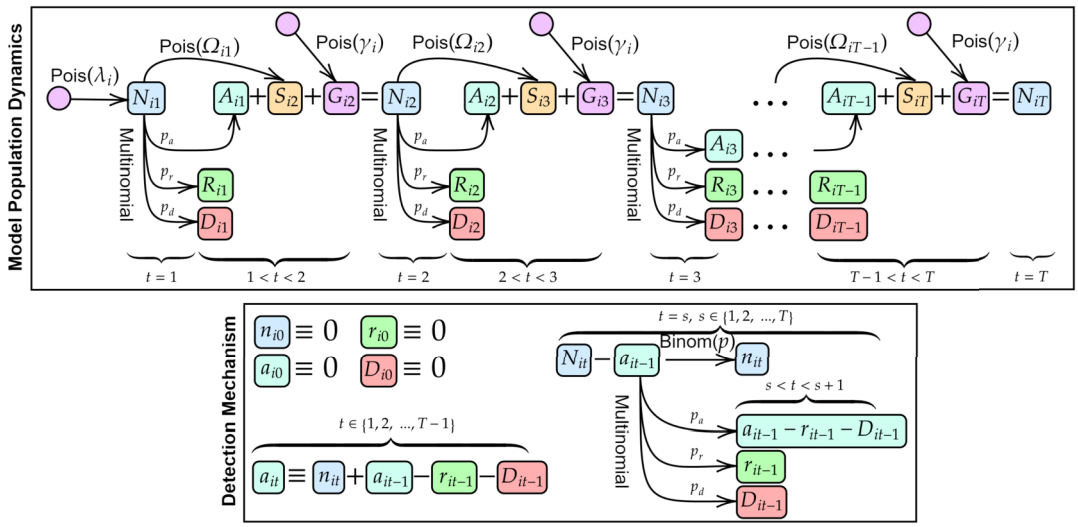

FIG. 2. *Diagram of the multisite model for sampling times* $t \in \{1, 2, \ldots, T\}$ *and sampling sites* $i \in \{1, 2, \ldots, M\}$. *Top*: *The disease dynamics and state process*. *Bottom*: *The detection mechanisms*.

is a requirement for these aggregate data models, and leads to p_r and p_d being constant across individuals, and is a justification for the choice of using the multinomial distribution.

Component (3) provides the calculation for the number of detected cases still active at time t: a_{it}. The value of a_{it} is entirely specified by the observed data. We calculate a_{it} recursively, with prestudy period data a_{i0}, r_{i0}, d_{i0}, n_{i0} usually chosen to be all zero by definition. However, when detected active cases are known from the time period prior to $t = 1$, then a_{i0} would be set as the number of known active cases (rather than being set to zero).

We are modelling an infectious disease, where infectious individuals are able to spread the disease to susceptible individuals with whom they come into close contact. We refer to this disease spread mechanism as domestic spread. We differentiate new infections as either coming from domestic spread or as being imported from an outside source (such as an infectious individual immigrating to the study region).

In Component (4), S_{it} models the domestic spread within the population during each time interval. Because we assumed that individuals are indistinguishable in each site, each susceptible individual has the same probability of being infected, giving binomially distributed case counts. In the large population limit, this binomial distribution converges to Poisson. S_{it} allows for exponential growth of cases but does not allow a spontaneous outbreak within a population having zero active cases. Ω_{it} can be thought of as the expected total new infections per time interval, which is calculated using the two parameters ω_1 and ω_2 (Component 5). The parameters ω_1 and ω_2 are the average new infections per undetected $(N_{it-1} - a_{it-1})$ and per detected (a_{it-1}) active case, respectively. We make the simplifying assumption that the transmission rates ω_1 and ω_2 remain constant unless mediated by additional covariates, such as vaccination rates in our Canada case study. There is strong evidence that the asymptomatic and symptomatic patients have different transmission rate (Wu et al. (2021)), while asymptomatic patients may be a major source of unobserved infections. By modelling ω_1 and ω_2 separately, we are able to accurately estimate the disease dynamics. Alternative epidemiological models for disease spread, such as SIR, often make use of the basic reproductive number. For comparison to our model, the basic reproductive number R_0 is the product of the mean transmission rate $(1 - p)\omega_1 + p\omega_2$ and the mean infectious period $(1/p_r)$, where p is the probability of detecting a case (see Component 9).

Component (6) describes the value δ_i that modulates the growth of cases as the population becomes saturated with infection. Here H_i is the total population size of site i, which is the

maximum number of infected individuals that are possible in that region. The susceptible population for site i and time t is the total population of the region less the total active cases: $H_i - N_{it}$. The fraction of susceptible population δ_i decreases to zero linearly as N_{it} approaches H_i; that is, the spread rate goes to zero as the susceptible population is depleted.

Imported cases, G_{it}, are new active cases entering the population (Component 7), for example, from travel. The parameter γ is the average new number of imported cases per time interval. Imported cases allow for disease to occur even when $N_{i1} = 0$, allowing for spontaneous outbreaks in regions with no previous active cases. G_{it} allows for linear growth of cases over time.

We have now introduced all of the latent states necessary to calculate disease dynamics updates through time. The initial active cases at $t = 1$ is N_{i1}, described in Component (1). The number of active cases N_{it} is updated over time using Component (8). The active cases which remained active from the previous time step are A_{it-1}. The new active cases from domestic spread are S_{it}. The new active cases from importation are G_{it}. From these we calculate the active cases for $t > 1$: $N_{it} = A_{it-1} + S_{it} + G_{it}$.

Component (9) describes the reporting of case counts. With p being the probability of detecting a case, $1 - p$ gives the underreporting rate. We would like to have $n_{it} \sim \text{Binomial}(N_{it}, p)$, which is the traditional N-mixtures parameterization of detection probability (Royle (2004)). However, since N_{it} comprises all active cases, including those which have already been detected, this would allow double counting (because detected cases n_{it} are tracked until recovery or death). Instead, we subtract the already detected active cases prior to the binomial thinning: $N_{it} - a_{it-1}$. Individuals are indistinguishable for these aggregate count models, and so we choose the binomial distribution to model detection.

Component (10) models the reporting of recoveries and deaths by partitioning a_{it} into cases that remain active, cases that die, and cases that recover. We use $\alpha = 1$ when deaths are underreported. In the situation where deaths are considered to be fully reported (when detected deaths are D_{it} rather than d_{it}), we set $\alpha = 1/p$, which increases the proportion of deaths among detected cases compared to proportion of deaths among all cases. Note that $p_a^* = 1 - \alpha p_d - p_r$.

The proportion of deaths among all cases, p_d, is modelled through the multinomial in Component (2). The proportion of deaths among the observed cases, αp_d, is modelled by the multinomial in Component (10). When $\alpha = 1$, p is both the probability of reported cases as well as the probability of reported death. In this case the two proportions are the same (we would expect a similar number of deaths per capita in the observed sample as in the overall population). When $\alpha = 1/p$ and $p < 1$, the two proportions are different. If $p = 1$, then the proportions are the same because the sample is equal to the total active cases. However, if $p < 1$, then the proportion of deaths in the sample will be larger than the proportion in the population by the factor $\alpha = 1/p$.

We have built on the Parker et al. (2021) model to allow for multiple sites, which in our B.C. case study corresponds to the five health authority regions of B.C. and in our Canada case study corresponds to the provinces and territories of Canada. Deaths and recoveries are occasionally recorded directly during the first observation period for an individual (such as when a recovery or death occurs in the same week as the initial positive test result). This leads to a small simplification of the Parker et al. (2021) detection process: $n_{it} \sim \text{Binomial}(p, N_{it} - a_{it-1})$ rather than $n_{it} \sim \text{Binomial}(p, N_{it} - a_{it-1} + d_{it-1} + r_{it-1})$. With this improvement we not only continue to avoid double counting cases but also allow deaths and recoveries to occur in the same reporting period as the first observation for an individual. We have split the spread rate ω into two new parameters, ω_1 and ω_2. This allows us to model different domestic spread rates due to detected and undetected cases. To account for local cluster saturation and population saturation effects, we incorporated a penalty term

to the model which allows ω_1 to diminish as the number of active cases increases. By using the known population size of a region as an upper bound on total possible infections, we linearly reduce the domestic spread to zero when the population is saturated: replacing $\omega_1(N_{it} - a_{it})$ with $\omega_1(N_{it} - a_{it})(H_i - N_{it})/H_i$, where H_i is the total population of region i. An exponential decay could be considered rather than linear decay if case cluster saturation is expected to be a dominant effect, which can be explored in future work. The full joint distribution for our multisite model is $f(\{\lambda_i\}, \{\gamma_i\}, \omega_1, \omega_2, p, p_r, p_d, \{n_{it}\}, \{r_{it}\}, \{d_{it}\}) = \mathscr{L} \cdot (\prod_{i=1}^{M} \pi_{\lambda_i} \cdot \pi_{\gamma_i}) \cdot \pi_{\omega_1} \cdot \pi_{\omega_2} \cdot \pi_p \cdot \pi_{p_d} \cdot \pi_{p_r}$ with likelihood function $\mathscr{L}$ given by

$$
\begin{aligned}
\mathscr{L} = \prod_{i=1}^{M} \Bigg(& \text{Poisson}(N_{i1}; \lambda_i) \cdot \Bigg(\prod_{t=1}^{T} \text{Binomial}(n_{it}; N_{it} - a_{it-1}, p) \\
& \cdot \text{Mult}(A_{it}, D_{it}, R_{it}; N_{it}, p_a, p_d, p_r) \Bigg) \\
& \cdot \Bigg(\prod_{t=1}^{T-1} \text{Mult}(a_{it} - d_{it} - r_{it}, d_{it}, r_{it}; a_{it}, p_a, p_d, p_r) \Bigg) \\
& \cdot \Bigg(\prod_{t=2}^{T} \text{Poisson}(G_{it}; \gamma_i) \cdot \text{Poisson}\big(S_{it}; \omega_1(N_{it-1} - a_{it-1}) \\
& \cdot (H_i - N_{it})/H_i + \omega_2 a_{it-1}\big) \Bigg) \Bigg).
\end{aligned} \tag{1}
$$

Here π_x is the prior distribution chosen for parameter x. Note that, while the sites are independent conditioned on their parameters, the likelihood shares parameters across sites. Parameters can be separated across sites or pooled by grouping sites using categorical site covariates. A single-site model can be obtained by dropping the site subscript i from equation (1) (see Supplementary Material A.1). We conducted a simulation study to verify parameter identifiability for our multisite model (see Supplementary Material A.2.1), finding all parameters to be identifiable. We also performed a simulation study assessing model robustness to misspecification (see Supplementary Material A.2.2), where we found the model to be robust against overdispersed domestic spread.

2.2. *Parameter covariates*. The model described in Section 2.1 assumes that the parameters ω_1, ω_2, p, p_r, and p_d are shared across sites and that they are constant over time. These assumptions are very strict. However, it is straightforward to relax or even remove these assumptions, as we do in both of our case studies. To allow a parameter to vary over sites, such as p, we may replace p with p_i. In this way each site has a detection probability separate to that of the other sites. An alternative would be to assign sites into categories, such as "small population" and "large population" sites, each category c having it's own p_c parameter. To add time dependence to a parameter, we add time varying covariates to the model. As an example, we use the testing volume vol_{it} per site i and per week t as a covariate for probability of detection p by replacing p in the model: $p \to p_i + p_{\text{vol}} \cdot \text{vol}_{it}$, where p_i is the site varying baseline detection and the coefficient p_{vol} is the testing volume effect size. Whether a particular parameter should be constant, vary with sites, or vary with covariates is dependent on individual case studies and on the available data. As such, we include Section 3.1.2 and Section 3.2.2 to discuss the specific modelling choices made for each of our case studies. Many more potentially important covariates could be included than are considered in our case studies. For example, availability of at-home test kits would almost certainly correlate with decreasing detection rates. However, data for our case studies are limited by availability, quality, and duration of data reporting.

2.3. *Model fitting*. We used Bayesian Markov-chain Monte-Carlo (MCMC) methods for parameter estimation and implemented model fitting using the statistical computing software R (R Core Team (2022)), together with the R package `Nimble` (de Valpine et al. (2017), de Valpine et al. (2021)). Bayesian MCMC is orders of magnitude more computationally efficient than traditional maximum likelihood estimation methodology for this type of model, due to the presence of large numbers of latent states. Bayesian MCMC also provides posterior estimates for each of the latent states, such as G_{it} and S_{it}, which are directly useful for estimating IFR.

Throughout this work we use noninformative priors to illustrate the identifiability of each model parameter. In the case that prior information is available for a particular case study, the prior information can be included through the judicious choice of prior distributions in order to improve estimation. Our simulations (see Appendix A.2.1) showed each model parameter to be identifiable, even with noninformative priors. Regarding the wide posterior distribution for p in our simulation study, the inclusion of reasonable covariates (such as testing volume in our case studies) substantially reduces the width of the posterior, as evidenced in Figure 4. We mainly used uninformative uniform prior distributions with reasonable upper and lower bounds for our case studies. For example, ω_1 was given a prior of Uniform(0, 5), as $\omega_1 \geq 0$, and an ω_1 of 5 is far larger than would be expected. Our investigations indicate that, with the exception of extreme priors, parameter estimates were not sensitive to the prior. We found that a burn-in of 100,000 was usually enough for the MCMC chains to converge. However, for some combinations of data and covariates a burn-in of 1,000,000 was necessary. We used 200,000 iterations, which was more than sufficient to produce posterior estimates.

3. Applications.

3.1. *Case study*: *Canada*.

3.1.1. *Data*. We used three publicly available sources of data for our Canada case study. The majority of the data came from the Government of Canada's COVID-19 daily epidemiology update (PHAC: Government of Canada (2022a)), which contains testing volumes, case counts, recoveries, and deaths due to COVID-19 for each province and territory from 23 April 2020 to 10 February 2022. However, the testing volumes for the Yukon Territory were largely missing from 3 June 2021 onward. For this reason we used the testing volume data reported by Yukon Health (Government of Yukon (2022)) to supplement the PHAC data. Our third set of data was the provincial vaccination status reports, also from the PHAC (Government of Canada (2022b)). These reports include total counts per province of individuals who had received one vaccination shot and who had received at least two vaccination shots. All count data were aggregated by week to alleviate the issue of lump sum data reporting (such as when COVID-19 cases detected over a weekend are only reported after the weekend). After aggregation we were left with 90 weeks of data for model fitting.

3.1.2. *Parameter covariates*. We chose several parameter covariates for the Canada-wide model, which are summarized in Table 2. The parameters λ_i, γ_i, p, ω_1, and ω_2 were allowed to vary by site (province/territory), for example, using p_i in the covariate structure for p. The parameters γ_i, p, p_r, p_d, ω_1, and ω_2 were allowed to vary over site and time using site- and time-varying covariates, such as the covariate vol_{it}. Data pooling was incorporated by keeping all covariate coefficients constant across sites, for example, the coefficient p_{vol}.

We allowed each site to have a unique baseline detection p_i and used testing volume vol_{it} per site and per week as a covariate for detection probability, since more testing should lead to higher detection rates. Thus, the probability of detection is modelled as: $p \to p_i + p_{\text{vol}} \cdot \text{vol}_{it}$. Figure A5 in the Supplementary Material shows the testing volume data for each site.

TABLE 2

Parameter covariates for the Canada-wide case study. Subscript i indicates site dependence; subscript t indicates time dependence. Testing volume is represented by vol, *while vaccination rates, one dose or two doses, are represented by* vac_1 *and* vac_2, *respectively. The indicator variables denoting the Omicron and Delta time periods are o and Δ. Data pooling across sites explicitly benefits the estimation of the coefficients γ_o, γ_Δ, p_{r0}, p_{ro}, $p_{r\Delta}$, p_{d0}, p_{do}, $p_{d\Delta}$, ω_{1o}, $\omega_{1\Delta}$, ω_{2o}, and $\omega_{2\Delta}$*

Parameter	Covariate Structure
λ_i	$\lambda_i \to \lambda_i$
γ_i	$\gamma_i \to \gamma_i + \gamma_{\text{vac}_1} \cdot \text{vac}_{1it} + \gamma_{\text{vac}_2} \cdot \text{vac}_{2it} + \gamma_o \cdot o_t + \gamma_\Delta \cdot \Delta_t$
p	$p \to p_i + p_{\text{vol}} \cdot \text{vol}_{it}$
p_r	$p_r \to p_{r0} + p_{r\,\text{vac}_1} \cdot \text{vac}_{1it} + p_{r\,\text{vac}_2} \cdot \text{vac}_{2it} + p_{ro} \cdot o_t + p_{r\Delta} \cdot \Delta_t$
p_d	$p_d \to p_{d0} + p_{d\,\text{vac}_1} \cdot \text{vac}_{1it} + p_{d\,\text{vac}_2} \cdot \text{vac}_{2it} + p_{do} \cdot o_t + p_{d\Delta} \cdot \Delta_t$
ω_1	$\omega_1 \to \omega_{1i} + \omega_{1\,\text{vac}_1} \cdot \text{vac}_{1it} + \omega_{1\,\text{vac}_2} \cdot \text{vac}_{2it} + \omega_{1o} \cdot o_t + \omega_{1\Delta} \cdot \Delta_t$
ω_2	$\omega_2 \to \omega_{2i} + \omega_{2\,\text{vac}_1} \cdot \text{vac}_{1it} + \omega_{2\,\text{vac}_2} \cdot \text{vac}_{2it} + \omega_{2o} \cdot o_t + \omega_{2\Delta} \cdot \Delta_t$

We chose to keep the baseline recovery and death rates p_{ro} and p_{do} constant over sites under the assumption that the disease dynamics would be mostly impacted by the dominant virus strain rather than geographic location when considering aggregation at the provincial level.

Vaccination rates (at least one dose, vac_{1it}, and two doses, vac_{2it}) were used as covariates for the parameters p_r, p_d, ω_1, ω_2, and γ. Figure A6 in the Supplementary Material shows the vaccination rates as a portion of the total population. Single dose rates are shown in red, while second dose rates are shown in blue. The vaccination rates for the Northwest Territories indicate a data anomaly in June 2021, where both single and double dose rates were reduced. This may be due to a data correction which occurred in June 2021.

The dates of emergence in Canada for the two variants of concern—Delta and Omicron—were used as covariates for time period change points for the variables p_r, p_d, γ, ω_1, and ω_2. As start date for the time periods, we used the first week of confirmed cases in Canada, 4 April 2021 (week 54) for Delta, and 28 November 2021 (week 84) for Omicron.

3.1.3. *Results*. Figure 3 shows both the detected case counts n_{it} and the estimated total active cases $\hat{N}_{it}$ for each site (province or territory). Due to the exponential growth and decay in case numbers, the active cases are plotted on a log scale. Every site saw a large increase in cases during the Omicron time period, whereas the Delta time period saw an initial drop in cases from May to August 2021 in most sites.

The detection probability $\hat{p}$ changes over time due to testing volumes (Figure 4). Detection rates in Newfoundland and Labrador, Nunavut, and Yukon were much lower than other sites for the majority of the pandemic, while the Prince Edward Island detection rate was consistently higher than 35%. Several provinces (such as B.C., Ontario, and Quebec) showed consistent growth in detection rates over the course of the pandemic, with a shared noticeable slump during the first four months of the Delta time period.

The probability of death, $\hat{p}_d$, was seen to hold steady at 0.26% in all sites for the first period of the pandemic, before decreasing (Figure A7 in the Supplementary Material). The effect of vaccinations on p_d is seen as a dramatic decrease throughout early 2021. The Delta period saw an increase in p_d, which was ameliorated by increasing vaccination coverage in each site. The Omicron period saw a sharp drop in mortality, with p_d levelling out around 0.06% across Canada. Nunavut saw a larger probability of death during the Delta time period than the other sites, with a 95% credible interval of $(0.19, 0.20)\%$, which does not overlap with any of the other site credible intervals. The three territories saw probability of death drop earlier than the provinces, likely due to earlier vaccination drives within the territories;

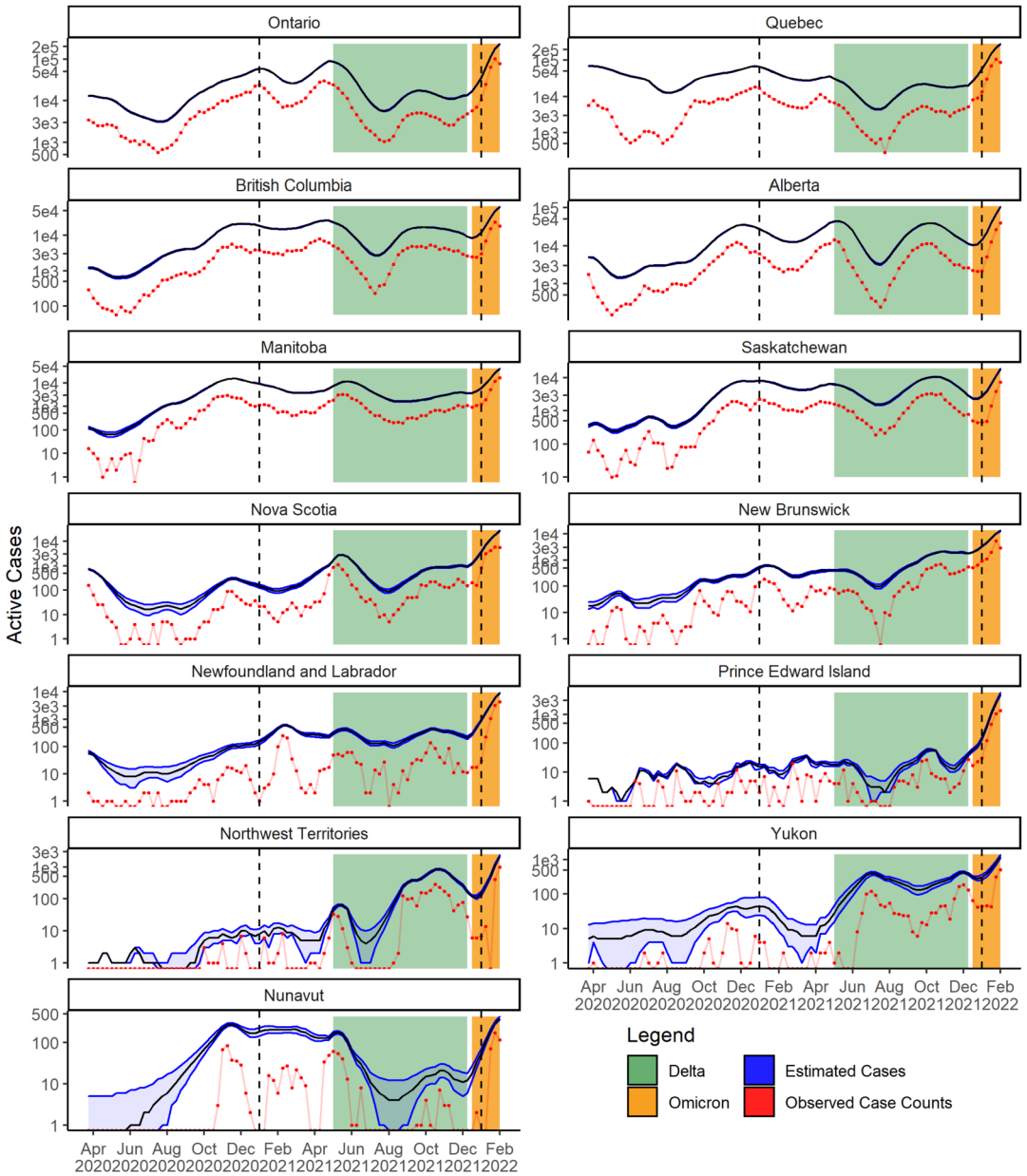

FIG. 3. *Detected case counts* (*lower lines*) *and estimated total active cases* (*upper lines*) *for each province/territory from* 23 *April* 2020 *to* 10 *February* 2022. *The time periods for the two variants of concern Delta and Omicron are depicted with shaded regions. The two vertical dashed lines indicate* 1 *January* 2021 *and* 1 *January* 2022. *Active cases are plotted on a log scale. Bands indicate* 95% *credible intervals.*

Yukon reached 60% vaccination by 15 May 2021, while Saskatchewan reached 60% by 26 June 2021. The median death probabilities are shown in Table A1 in the Supplementary Material.

Probability of recovery was steady at 35.0% during the early stages of the pandemic for all sites until January 2021 when the recovery probability began to increase along with the increase in vaccination coverage across Canada (Figure A8 in the Supplementary Material). The Delta period saw a small decrease in recovery probability for all sites, while the Omicron period saw a large decrease. Probability of recovery peaked much earlier for the three territo-

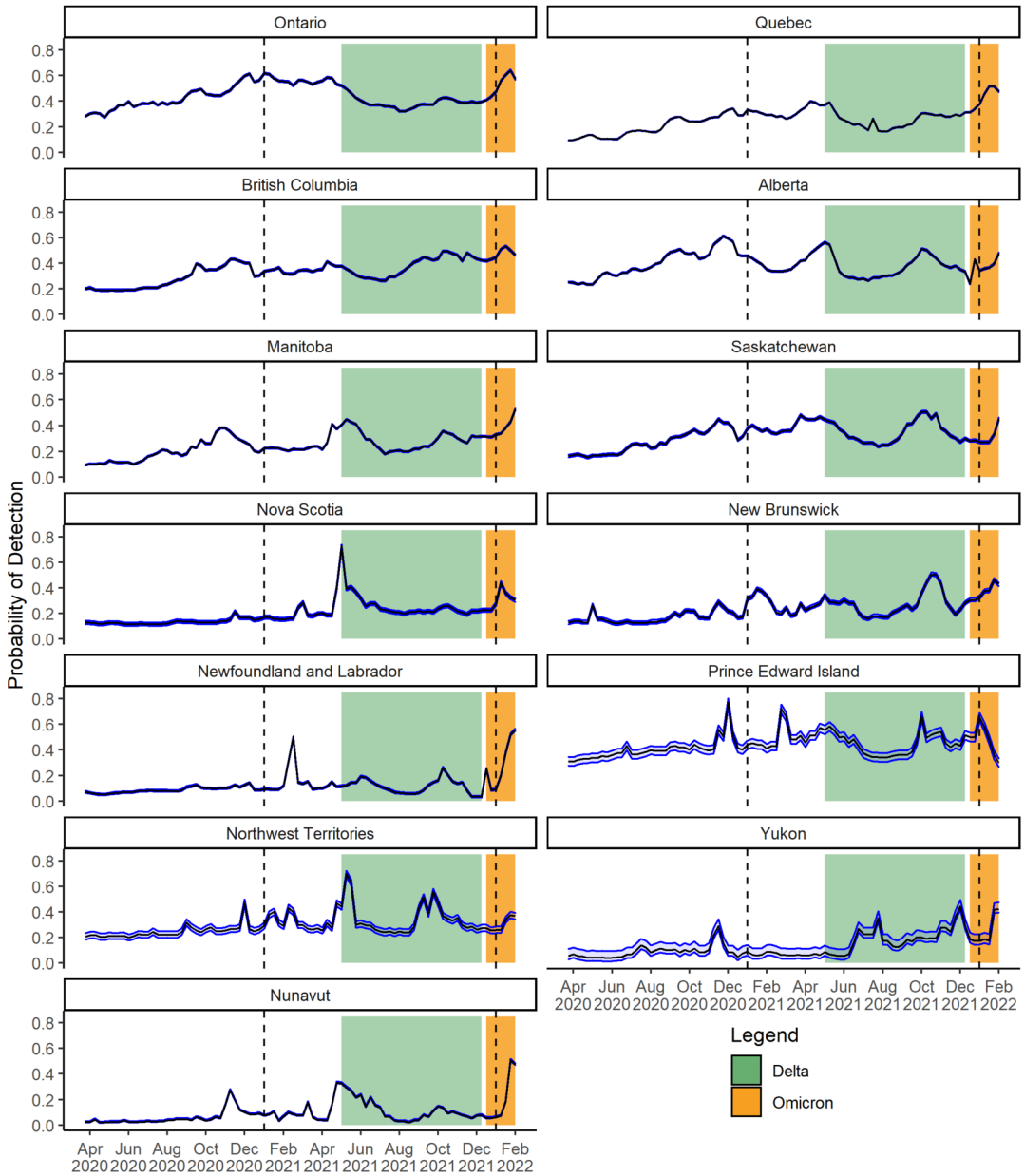

FIG. 4. *Estimated weekly detection probability for active COVID-19 cases for each province or territory from* 23 *April* 2020 *to* 10 *February* 2022. *The time periods for the two variants of concern—Delta and Omicron—are depicted with shaded regions. The two vertical dashed lines indicate* 1 *January* 2021 *and* 1 *January* 2022. *Bands show* 95% *credible intervals.*

ries than for any of the provinces, likely due to the earlier vaccination drives in the territories. The median recovery probabilities are shown in Table A2 in the Supplementary Material.

The domestic spread for undetected cases, $\hat{\omega}_1$, accounted for most cases in the pre-2021 period (Figure 5). As vaccination coverage increased across Canada, the spread rate for detected cases, $\hat{\omega}_2$, increased, and in all sites became larger than the spread rate for undetected cases during the Delta time period. During the Omicron time period, the two spread rates converged to similar levels. During the early pandemic, Prince Edward Island had the largest spread rate among undetected cases, with a 95% credible interval of $(1.11, 1.26)$, much larger than the Canada-wide average of $(0.55, 0.59)$. During this early pandemic period, Yukon

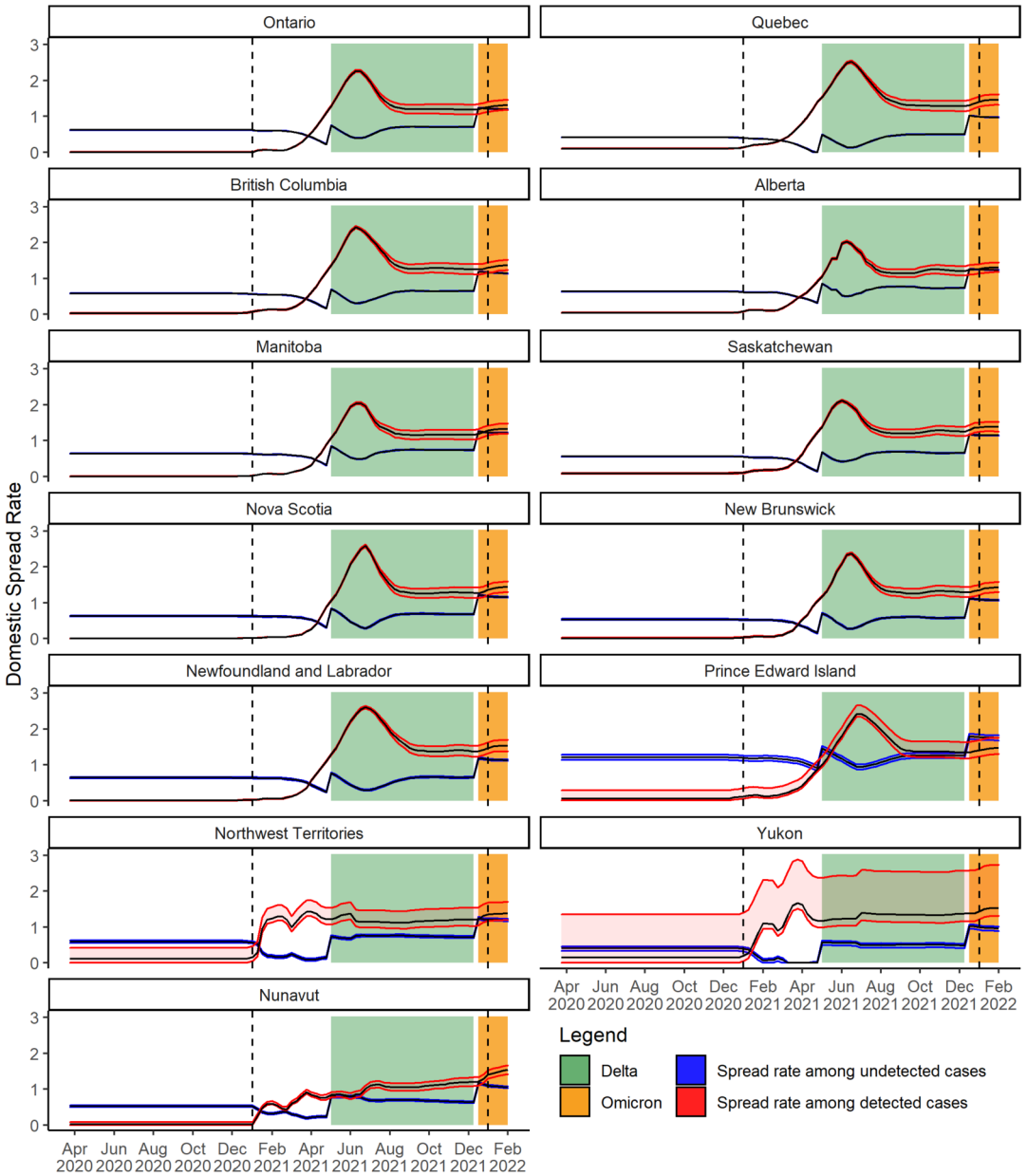

FIG. 5. *Estimated weekly domestic spread rates for detected* ($\hat{\omega}_2$, *lower line at April* 2020, *except for the three territories*) *and undetected* ($\hat{\omega}_1$, *upper line at April* 2020, *except for the three territories*) *active COVID-19 cases for each province/territory from* 23 *April* 2020 *to* 10 *February* 2022. *The time periods for the two variants of concern Delta and Omicron are depicted with shaded regions. The two vertical dashed lines indicate* 1 *January* 2021 *and* 1 *January* 2022. *Bands show* 95% *credible intervals.*

had the lowest rate, with a 95% credible interval of $(0.25, 0.36)$. Conversely, the spread rate among detected cases was much smaller, with the Canada-wide average having a 95% credible interval of $(0.14, 0.32)$. In the Omicron period, the spread rate for undetected cases was less than the spread rate for detected cases, with the Canada-wide 95% credible intervals of $(1.17, 1.21)$ and $(1.22, 1.62)$, respectively. The median spread rates are shown in Table A3 for undetected and in Table A4 for detected cases in the Supplementary Material.

Weekly importation rates, $\hat{\gamma}_i$, are generally low across Canada, with the exceptions of Quebec and British Columbia (Figure A9 in the Supplementary Material). British Columbia

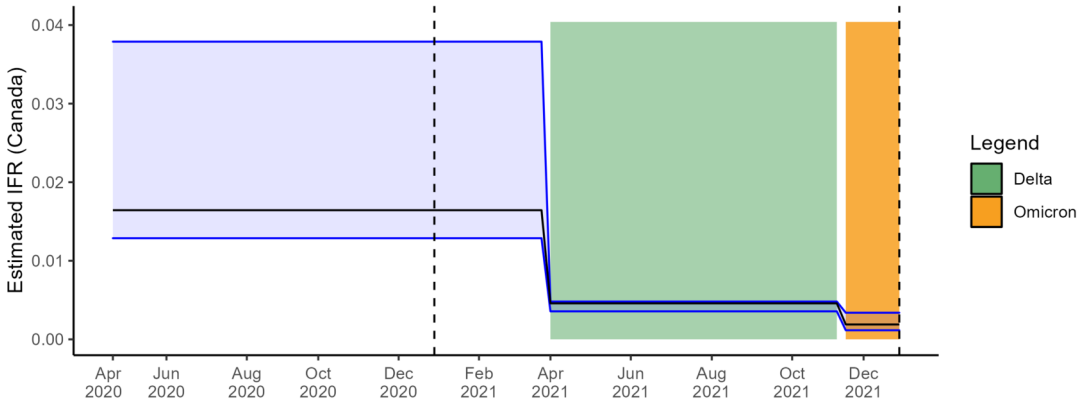

FIG. 6. *Estimated infection fatality rate (IFR) for COVID-19 infections within Canada. The time periods for the two variants of concern Delta and Omicron are depicted with coloured bands. The estimated IFR for each time period is* 0.015 (*early pandemic period*), 0.005 (*Delta period*), *and* 0.002 (*Omicron period*). *The two vertical dashed lines indicate* 1 *January* 2021 *and* 1 *January* 2022. *Bands show* 95% *credible intervals*.

had the second largest importation rate, at around 80 imported cases per week. Quebec had the largest, at around 1400 imported cases per week. It is important to note that, while γ_i are identifiable parameters in the model, γ_i measures average linear growth in cases and so is not necessarily the true importation rate. Rather, the very large γ_i estimate for Quebec could indicate better control of the exponential spread rate due to quarantines and other public health measures. Evidence for this can be seen in Figure 5, where the domestic spread rate among undetected cases for Quebec is much lower than for the other provinces. In contrast to the large importation rate of Quebec, Prince Edward Island had an importation rate of just 0.03 cases per week. This implies that most new cases in Prince Edward Island were due to the exponential growth of domestic spread.

To estimate the infection fatality rate (IFR) for Canada as a whole, we used the estimated new cases, $\hat{S}_{it} + \hat{G}_{it}$ per time period P, then we estimated the IFR as $\widehat{\text{IFR}}_P = \sum_{i,t\in P} D_{it} / \sum_{i,t\in P}(\hat{S}_{it} + \hat{G}_{it})$. The estimated IFRs for Canada as a whole with 95% credible intervals are: 0.015 (0.012, 0.038) for the early pandemic period, 0.005 (0.004, 0.006) for the Delta period, and 0.002 (0.001, 0.003) for the Omicron period (Figure 6).

Figures comparing the prior to the posterior distributions for each model parameter are included in the Supplementary Material (Figures A10 to A15). All parameters were identifiable. However, ω_{2o} shows signs of potential instability. This is likely due to the small number of sampling occasions for the Omicron time period. To evaluate how well our model fits the data, we compared the observed data against the posterior prediction data and found that they match well. Figure A16 in the Supplementary Material shows the actual observed new active cases n_{it} as well as the posterior median $\hat{n}_{it}$. The model reproduces the observed data closely, even in the Omicron tail of the data.

3.2. *Case study*: *British Columbia*.

3.2.1. *Data*. We used several publicly available sources to compile the data for our B.C. case study. The B.C. Surveillance Reports (BC Centre for Disease Control (2020)) were used to gather counts of cases, recoveries, and deaths. These data are shown for each health authority region in Figure A17 in the Supplementary Material. Province of B.C. laboratory data (Province of British Columbia (2020)) was used as a source for COVID-19 testing volumes per region. Start dates for the Phases of the B.C. Recovery plan were obtained from the Government of B.C. emergency preparedness response web pages (Government of British Columbia (2020)).

Data for this B.C. case study was limited to the date range 2 April 2020 (Week 1) to 30 October 2020 (Week 31). After this period the BC Surveillance Reports began to exclude the data necessary for fitting these models (case counts, case recoveries, and deaths split by health authority region). These methods could be easily applied to more recent pandemic data if the required aggregate data were made publicly available.

3.2.2. *Parameter covariates*. In this case study, we explored many combinations of parameter covariates in order to better understand our model and to investigate the relationships between the covariates and the data. Health authority region (denoted *reg*) was used as a covariate for both λ and γ in all of our considered models. The covariate *reg* was also used for ω_1 and ω_2 in several models to indicate region dependency. Phase of the B.C. Recovery Plan (denoted *pha*) was also considered as a potential covariate for γ, ω_1 and ω_2, to indicate that the parameters change with time at the boundary of the recovery plan phases. For detection probability p, we considered the parameter covariates: *reg*, *pha*, COVID-19 testing volume (*vol*), and a baseline offset which was constant across regions (B).

3.2.3. *Model selection*. We used the widely applicable information criterion (WAIC: Watanabe (2010)) to aid in model selection and study the impact of covariates. Gelman, Hwang and Vehtari (2014) discuss the limitations of relying solely on log-predictive density methods (such as WAIC) for model selection. In particular, the authors note that such model selection procedures can overfit the model to the data, providing suboptimal predictive performance. To that end, they caution to view information criteria as an approach to understand fitted models rather than to choose from among them.

While we do use WAIC to aid in model selection, it is not the sole criteria. For example, we select the second best performing model according to WAIC as our preferred model. We did this for three reasons: (1) the top two ranked models perform very similarly in terms of estimated latent variables, (2) the number of model parameters is far less for the second ranked model, and (3) the covariates used in the second model carry far more explanatory value than those in the first ranked model.

3.2.4. *Results*. Our results for fitting 22 models with different organizations of the covariates and conditions on ω to the B.C. data are summarized in Table A5 in the Supplementary Material. Model M1 performed best in terms of WAIC, and model M22 performed worst (we label the models in descending order of WAIC). Excluding the four models which used no covariates for ω_1 and ω_2, every model for which ω_1 was allowed to differ from ω_2 performed better than the models which forced $\omega_1 = \omega_2$. This is strong evidence in favour of the models which allow the domestic spread rate to be affected by detection status, since those models perform uniformly better in terms of WAIC. In addition, every model that allowed ω_1 and ω_2 to vary with Phase performed better than models not varying with Phase. This suggests spread rates were not constant over the course of the early pandemic.

We identify model M2 as the best model, despite M1 having the lower WAIC score. Figure A18 in the Supplementary Material compares the estimated active cases for each health authority region between model M1 and model M2. The two models perform similarly for all sites, except Fraser Health Authority Region, for which the model M2 predicts a substantially larger initial number of active cases. However, model M1 uses 18 more parameters than model M2 and so is at greater risk of overfitting. Model M2 uses testing volume as a covariate for detection probability, which is far more interpretable than B.C. Recovery Plan Phase (which is used instead for model M1 and is correlated with many confounding factors, such as season, health mandates, and human mobility). Figure A18 also compares the results for the five single-site models (shown in green) against the two best fitted multisite models

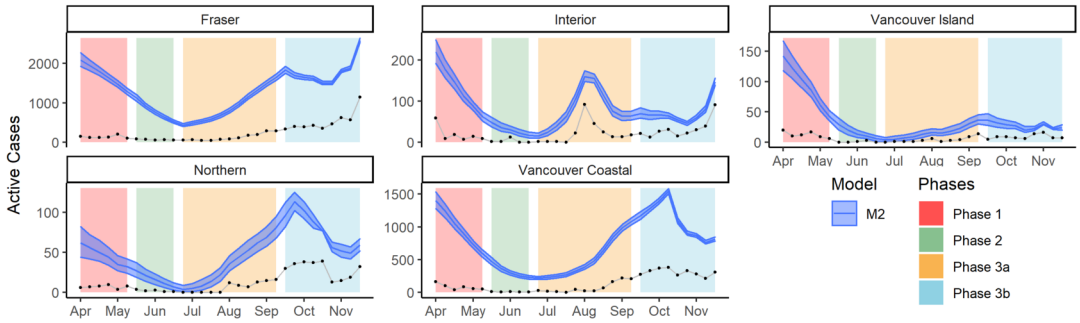

FIG. 7. *Plots of active cases split by health authority region. Data starts with* 2 *April* 2020 (*Week* 1) *and ends with* 30 *October* 2020 (*Week* 31). *The upper bands show the* 95% *credible intervals for total active cases as estimated using model M*2. *The lower dotted line shows the newly detected active cases each week.*

(M1 in red and M2 in blue). The single-site models show larger uncertainty in their estimates through wide credible intervals, illustrating an advantage of using multisite modelling.

We show the results for our chosen model M2 in Figure 7. The trend for all regions over this period of study was for a large initial number of active cases, which falls off leading up to week 10 and builds leading into the later stages of the 31-week period. All five regions showed signs of peaking at different times between weeks 17 and 26, with a second peak beginning around week 30, and likely continuing after this period of study.

3.3. *Comparing results.* We compare results between three case studies overlapping in time intervals and regions to illustrate several important points. Model tails (time periods near the beginning and the end of the study) are less accurate, due to a lack of information outside of the study time boundaries (Section 3.3.1). This is particularly important to note when interpreting our Canada case study results. Since the Omicron time period lies on the boundary, there is less certainty in our estimates for that time period. Future applications of these methods should be aware of these boundary limitations and should not put excessive weight on estimates near the study time boundaries. Another important point is the dependence of these models on data quality. The difference in case counts between the B.C. CDC Surveillance reports and the PHAC situation reports impacts the estimates of total active cases (see Figure A19 in the Supplementary Material). However, when the data sources are the same, the multisite models produce more precise estimates than the single-site models (Section 3.3.2). Therefore, when using these models, it is important to use the most accurate and highest quality case count data available.

3.3.1. *British Columbia comparison.* Figure A19 in the Supplementary Material illustrates the estimated total cases in B.C., as estimated by the single-site health authority region model (green) and as estimated by the multisite Canada case study model (blue). The two models diverge considerably in estimated total cases in both tails, while agreeing in the middle period. There are several reasons for this. First, the data used to fit the models is not the same. The detected case counts are a form of administrative data and have been updated and corrected over time, and the counts used in the Canada model are likely to be more accurate than the counts used in the B.C. model (Figure A19). Second, the Canada model has no covariate equivalent to the B.C. Recovery Plan phases covariate of the B.C. model; thus, the B.C. model allows for more granular changes in dynamics. Third, the Canada model pools information from the pandemic across Canada, allowing the province and territory estimates to be informed by national trends. Fourth, the Canada model is trained on 90 weeks of data rather than 31, and over 13 sampling sites rather than five. The larger amount of data used

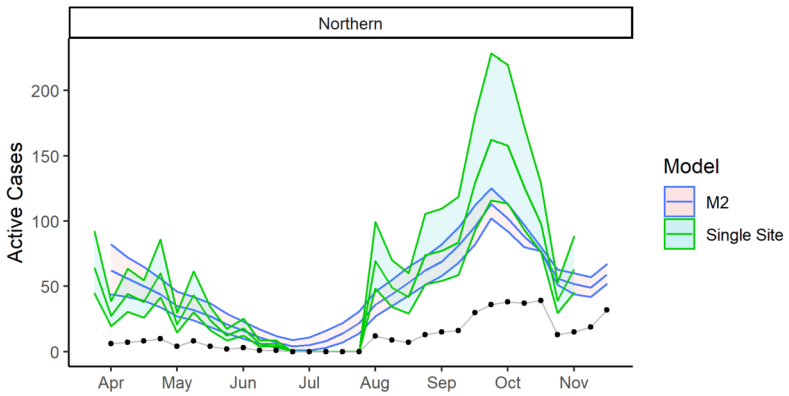

FIG. 8. *Plot of active cases for the Northern Health Authority Region. Data starts with* 2 *April* 2020 (*Week* 1) *and ends with* 30 *October* 2020 (*Week* 31). *The band extending through November* 2020 *shows the* 95% *credible interval for total active cases as estimated using model M*2. *The second band shows the* 95% *confidence intervals for total active cases, as estimated using the single site model from Parker et al.* (2021), *denoted Single Site. The lower dotted line shows the newly detected active cases each week.*

for the Canada-wide model means its results are likely to be more accurate. Fifth, the B.C. model begins with three earlier weeks of data so that the beginning tail of the model is likely more accurate than the Canada model. Conversely, the Canada model contains 62 weeks of more recent data than the B.C. model, meaning the end tail of the Canada model is likely to be more accurate.

3.3.2. *Northern health authority comparison.* We compare the Northern Health Authority Region results from the single-site model of Parker et al. (2021) with results from our multisite model M2 (Figure 8). The single-site results were obtained using maximum likelihood methods, and so the variability shown indicates the 95% confidence interval, whereas the multisite model uses Bayesian MCMC to obtain 95% credible intervals. The two models show excellent agreement through overlapping credible/confidence bands and share a common maximum during week 24, but the multisite model has increased precision over the single-site model. An important advantage to the Bayesian MCMC method of model fitting is the ability to estimate active cases even during periods when no new cases are detected, as was the case for the Northern Health Region during weeks 13 through 16.

4. Discussion. Throughout the pandemic, case counts have been only a lower bound on the total number of active infections. Our Canada case study estimates the levels of underreporting for each province and territory. Estimates of total active infections vary widely in magnitude across Canada and over time, with over 200,000 concurrent active cases at times in Ontario and Quebec, and under 400 concurrent active cases at any time in Nunavut. Our results show that, although the first confirmed infection in Nunavut occurred in October 2020, it is likely that there were over 100 active infections in Nunavut at that time and that the first infection likely would have occurred between June and August 2020.

Our estimates for detection probability show that there were similar detection levels across Canada. In the early pandemic, most provinces/territories had very low testing volumes (less than 10 tests per 1000 population size), which led to low detection rates (under 20% detection for most provinces/territories). Among the provinces, Prince Edward Island had the highest minimum detection rate (30%), while Newfoundland and Labrador had the lowest (3%). Among the territories, Northwest Territories had the highest minimum detection rate (20%), while Nunavut had the lowest (2%). Testing volume is not the only factor in determining detection, as can be seen by comparing the results from Ontario with those from Quebec

(Figure 4). Both provinces had very similar testing volumes throughout the pandemic; however, the detection rate was substantially higher in Ontario than in Quebec. Ontario had a low of 27% and a high of 64%, compared with Quebec's low of 10% and high of 52%. The reason for this is unclear. It could be due to differences in testing protocols, differences in access to testing, geographic or political differences, or many other possible factors. Future research into the effects of public messaging and health policies on detection rates would be beneficial and could improve our understanding of this phenomenon. Underreporting becomes a larger issue late in a pandemic, as fewer cases are recorded (availability of at-home test kits whose results are not reported, less urgent testing protocols, less severe symptoms due to vaccinations, etc). Decisions regarding targeted vaccination roll-outs are critical to the well-being of at-risk persons, such as the elderly and those with compromised immune systems, and knowledge of the levels of underreporting can inform those policy decisions as well as help individuals and their care-givers to make better informed health decisions.

In Canada the case-fatality rate for COVID-19 was estimated to be 4.9% in April 2020 (Abdollahi et al. (2020)) and 3.36% by the end of 2020 (Shim (2021)). However, CFR is larger than IFR when there are undetected cases. Understanding the IFR during recent periods of the pandemic allows us to understand the personal risk associated with contracting SARS-CoV-2. We estimated the weekly probability of death p_d over the course of the pandemic and found that for all provinces and territories of Canada the mortality rates have decreased over time. In general, mortality rates decreased with increased first dose of vaccination (Figure A7 in the Supplementary Material), increased during the Delta time period, and decreased again in the Omicron time period. Weekly probability of death is not to be confused with overall probability of death, which is better measured as IFR. The weekly probability of death is the average probability of an active case dying during a particular week (if they neither die nor recover during that week, then they will have a further chance of dying in the following week and so on until they either recover or die). IFR is a more easily interpreted measure of mortality rates than p_d, since IFR is the overall mortality rate for infected individuals. Previous findings regarding IFR using data across 15 countries determined it likely that IFR for COVID-19 was $< 0.2\%$ (Ioannidis, Axfors and Contopoulos-Ioannidis (2020)). Fisman et al. (2020) estimated the overall IFR in Ontario to be 0.8% (0.75%, 0.85%), with a large range based on age (from 0.01% up to 12.7%). Our IFR estimate for the early pandemic time period of 1.5% is much higher than the 0.2% from Ioannidis, Axfors and Contopoulos-Ioannidis (2020). However, our estimates for the Delta and Omicron periods are more moderate, respectively, 0.5% and 0.2%.

Estimates of recovery probability are useful for understanding the length of an average infection. During the early pandemic period, the weekly probability of recovery was 35.0%, and the weekly probability of remaining an active case was 64.7%. This implies that the average recovery time and 95% recovery interval was 11.3 (6.1, 15.4) days (obtained by solving the simple geometric series in p_a, see Supplementary Material A.3). For the Delta and Omicron periods, the average recovery time was 8.5 (6.0, 12.8) days and 14.0 (6.1, 17.0) days, respectively. We see the average recovery time increased during the Omicron period, while the lower bound on recovery time has remained around six days throughout the pandemic.

According to National Collaborating Centre for Infectious Diseases (2022), both the Delta and Omicron variants of concern have been found to be more transmissible than the earlier variants. Thus, we would expect to find an increased ω_1 and ω_2 during those time periods. Our Canada model agrees with this expectation (Figure 5), where domestic spread rates are seen to increase across Canada at the Delta boundaries as well as at the Omicron boundaries. A single vaccine dose is correlated with decreased spread rate for the undetected cases (corresponding loosely to asymptomatic, presymptomatic, and low severity cases) and is correlated with increased spread rate for the detected cases (corresponding loosely to symptomatic and

medium to high severity cases). The increased spread rate could be explained by decreases in severity for vaccinated individuals (Lauring et al. (2022)) so that the detected cases are more inclined to mobility and interaction events as well as relaxed health measures correlated with increasing vaccination rates. Receiving a second vaccine dose is correlated with an increase in spread rate for undetected cases, which may be explained by a decrease in perceived personal danger from contracting the disease when vaccinated with two doses. The second vaccine dose is correlated with a large decrease in spread rate for the detected cases in all provinces but not in the territories.

In this manuscript we have introduced a multisite disease dynamics model for underreported disease counts. We have used a discrete time modelling approach in order to more accurately match the data, which is widely available during a pandemic (periodic case, recovery, and death counts). An alternative approach would be to modify a classical multicompartment model, such as SIR, to incorporate underreporting. Comparing these two approaches would be of interest and is considered for future work.

Our British Columbia case study was limited in scope by a lack of available public data after 30 October 2020. For future pandemics and disease outbreaks, we would urge all public health authorities (in B.C., in Canada, and abroad) to make weekly aggregate counts of cases, recoveries, and deaths publicly accessible as early as possible to promote greater knowledge and more expedient research. The decision to do so can in turn lead to better informed, more timely health policy decision making, which can save lives and reduce the burden on our health care systems.

Acknowledgments. The authors are grateful to the Editor, the Associate Editor, and the anonymous reviewers for their time and effort. Their valuable feedback helped us to improve the quality of our manuscript. The authors gratefully acknowledge the Micheal Smith Foundation for Health Research and the Victoria Hospitals Foundation for support through a COVID-19 Research Response grant as well as a Canadian Statistical Sciences Institute Rapid Response Program—COVID-19 grant to LC that supported this research.

Funding. LC was supported by a Michael Smith Foundation for Health Research and Victoria Hospitals Foundation Grant # COV-2020-1061 and a Canadian Statistical Sciences Institute Rapid Response Program- COVID-19.

We would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) for providing PGS-D support for MP [funding reference number 569754].

LTE was supported by a Michael Smith Health Research BC Scholar Award and NSERC grant numbers RGPIN/05484-2019 and DGECR/00118-2019.

JC was supported by an NSERC discovery grant (RGPIN-2023-04057) and the Canada Research Chairs program.

SUPPLEMENTARY MATERIAL

Code supplemental (DOI: 10.1214/24-AOAS1915SUPPA; .zip). The Code Supplemental includes data for the BC and Canada case studies, R codes necessary for running both case studies and the simulation study, as well as our simulation results (Parker et al. (2024a)).

Online supplementary material (DOI: 10.1214/24-AOAS1915SUPPB; .pdf). The Supplementary Material contains additional results including Figures, Tables, simulations, and supporting information (Parker et al. (2024b)).

REFERENCES

ABDOLLAHI, E., CHAMPREDON, D., LANGLEY, J. M., GALVANI, A. P. and MOGHADAS, S. M. (2020). Temporal estimates of case-fatality rate for COVID-19 outbreaks in Canada and the United States. *CMAJ*, *Can. Med. Assoc. J.* **192** E666–E670. https://doi.org/10.1503/cmaj.200711

ALENE, M., YISMAW, L., ASSEMIE, M. A., KETEMA, D. B., MENGIST, B., KASSIE, B. and BIRHAN, T. Y. (2021). Magnitude of asymptomatic COVID-19 cases throughout the course of infection: A systematic review and meta-analysis. *PLoS ONE* **16** e0249090. https://doi.org/10.1371/journal.pone.0249090

APPLEBY, J. A., KING, N., SAUNDERS, K. E., BAST, A., RIVERA, D., BYUN, J., CUNNINGHAM, S., KHERA, C. and DUFFY, A. C. (2022). Impact of the COVID-19 pandemic on the experience and mental health of university students studying in Canada and the UK: A cross-sectional study. *BMJ Open* **12** e050187.

ARAF, Y., AKTER, F., TANG, Y.-D., FATEMI, R., PARVEZ, M. S. A., ZHENG, C. and HOSSAIN, M. G. (2022). Omicron variant of SARS-CoV-2: Genomics, transmissibility, and responses to current COVID-19 vaccines. *J. Med. Virol.* **94** 1825–1832.

BC CENTRE FOR DISEASE CONTROL (2020). BC COVID-19 data [surveillance reports]. Retrieved from http://www.bccdc.ca/health-info/diseases-conditions/covid-19/data. Accessed: 2022-02-25.

BÉLAND, D., DINAN, S., ROCCO, P. and WADDAN, A. (2021). Social policy responses to COVID-19 in Canada and the United States: Explaining policy variations between two liberal welfare state regimes. *Soc. Policy Adm.* **55** 280–294.

BENDAVID, E., MULANEY, B., SOOD, N., SHAH, S., BROMLEY-DULFANO, R., LAI, C. et al. (2021). COVID-19 antibody seroprevalence in Santa Clara County, California. *Int. J. Epidemiol.* **50** 410–419.

BUITRAGO-GARCIA, D., EGLI-GANY, D., COUNOTTE, M. J., HOSSMANN, S., IMERI, H., IPEKCI, A. M., SALANTI, G. and LOW, N. (2020). Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLoS Med.* **17** e1003346.

CHIMMULA, V. K. R. and ZHANG, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals* **135** 109864.

CHISALE, M. R. O., RAMAZANU, S., MWALE, S. E., KUMWENDA, P., CHIPETA, M., KAMINGA, A. C., NKHATA, O., NYAMBALO, B., CHAVURA, E. et al. (2022). Seroprevalence of anti-SARS-CoV-2 antibodies in Africa: A systematic review and meta-analysis. *Rev. Med. Virol.* **32** e2271.

CYPRESS, B. S. (2022). COVID-19: The economic impact of a pandemic on the healthcare delivery system in the United States. *Nurs. Forum* **57** 323–327. https://doi.org/10.1111/nuf.12677

DE VALPINE, P., PACIOREK, C., TUREK, D., MICHAUD, N., ANDERSON-BERGMAN, C., OBERMEYER, F., WEHRHAHN CORTES, C., RODRÌGUEZ, A., TEMPLE LANG, D. et al. (2021). NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling. R package version 0.11.1.

DE VALPINE, P., TUREK, D., PACIOREK, C., ANDERSON-BERGMAN, C., TEMPLE LANG, D. and BODIK, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Statist.* **26** 403–413.

DESSON, Z., WELLER, E., MCMEEKIN, P. and AMMI, M. (2020). An analysis of the policy responses to the COVID-19 pandemic in France, Belgium, and Canada. *Health Public Technol.* **9** 430–446.

DIRENZO, G. V., CHE-CASTALDO, C., SAUNDERS, S. P., GRANT, E. H. C. and ZIPKIN, E. F. (2019). Disease-structured N-mixture models: A practical guide to model disease dynamics using count data. *Ecol. Evol.* **9** 899–909.

DONG, E., DU, H. and GARDNER, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20** 533–534.

DOUGHERTY, B. P., SMITH, B. A., CARSON, C. A. and OGDEN, N. H. (2021). Exploring the percentage of COVID-19 cases reported in the community in Canada and associated case fatality ratios. *Infect. Dis. Model.* **6** 123–132. https://doi.org/10.1016/j.idm.2020.11.008

DOZOIS, D. J. A. (2021). Anxiety and depression in Canada during the COVID-19 pandemic: A national survey. *Can. Psychol.* **62** 136–142.

FEIKIN, D. R., WIDDOWSON, M.-A. and MULHOLLAND, K. (2020). Estimating the percentage of a population infected with SARS-CoV-2 using the number of reported deaths: A policy planning tool. *Pathogens* **9** 838.

FERNÁNDEZ-FONTELO, A., CABAÑA, A., PUIG, P. and MORIÑA, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Stat. Med.* **35** 4875–4890. https://doi.org/10.1002/sim.7026

FERNÁNDEZ-FONTELO, A., MORIÑA, D., CABAÑA, A., ARRATIA, A. and PUIG, P. (2020). Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. *PLoS ONE* **15** e0242956.

FISMAN, D. N., DREWS, S. J., TUITE, A. R. and O'BRIEN, S. F. (2020). Age-specific SARS-CoV-2 infection fatality and case identification fraction in Ontario, Canada Technical Report medRxiv.

GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016.

GOVERNMENT OF BRITISH COLUMBIA (2020). Phase 1: BC's restart plan. Retrieved from https://www2.gov.bc.ca/gov/content/safety/emergency-preparedness-response-recovery/covid-19-provincial-support/phase-1. Accessed: 2020-12-08.

GOVERNMENT OF CANADA (2022a). COVID-19 daily epidemiology update. Web Archive: /web/20220114205328/ https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html?redir=1. Accessed: 2022-01-14.

GOVERNMENT OF CANADA (2022b). COVID-19 vaccination in Canada. Web Archive: /web/20220120212246/ https://health-infobase.canada.ca/covid-19/vaccination-coverage/. Accessed 2022-01-20.

GOVERNMENT OF YUKON (2022). COVID-19 data dashboard. Web: https://covid-19-data-dashboard.service.yukon.ca/. Accessed 2022-02-18.

HALILI, R., BUNJAKU, J., GASHI, B., HOXHA, T., KAMBERI, A., HOTI, N., AGAHI, R., BASHA, V., BERISHA, V. et al. (2022). Seroprevalence of anti-SARS-CoV-2 antibodies among staff at primary healthcare institutions in Prishtina. *BMC Infect. Dis.* **22** 57.

HASAN, T., PHAM, T. N., NGUYEN, T. A., HIEN THI, T. L., DUYET, V. L., THUY, T. D. et al. (2021). Seroprevalence of SARS-CoV-2 antibodies in high-risk populations in Vietnam. *Int. J. Environ. Res. Public Health* **18** 6353.

HE, J., GUO, Y., MAO, R. and ZHANG, J. (2021). Proportion of asymptomatic coronavirus disease 2019: A systematic review and meta-analysis. *J. Med. Virol.* **93** 820–830.

HUO, X., CHEN, J. and RUAN, S. (2021). Estimating asymptomatic, undetected and total cases for the COVID-19 outbreak in Wuhan: A mathematical modeling study. *BMC Infect. Dis.* **21** 476.

IOANNIDIS, J. P. A., AXFORS, C. and CONTOPOULOS-IOANNIDIS, D. G. (2020). Population-level COVID-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters. *Environ. Res.* **188** 109890.

KAHN, F., BONANDER, C., MOGHADDASSI, M., RASMUSSEN, M., MALMQVIST, U., INGHAMMAR, M. and BJÖRK, J. (2022). Risk of severe COVID-19 from the Delta and Omicron variants in relation to vaccination status, sex, age and comorbidities - surveillance results from southern Sweden, July 2021 to January 2022. *Euro Surveill.* **27**. https://doi.org/10.2807/1560-7917.ES.2022.27.9.2200121

LAURING, A. S., TENFORDE, M. W., CHAPPELL, J. D., GAGLANI, M., GINDE, A. A. and SELF, W. H. (2022). Clinical severity of, and effectiveness of mRNA vaccines against, covid-19 from omicron, delta, and alpha SARS-CoV-2 variants in the United States: Prospective observational study. *BMJ* **376** e069761.

LI, C., ZHU, Y., QI, C., LIU, L., ZHANG, D., WANG, X., SHE, K., JIA, Y., LIU, T. et al. (2021). Estimating the prevalence of asymptomatic COVID-19 cases and their contribution in transmission—using Henan province, China, as an example. *Front. Med.* **8**.

MAHUMUD, R. A., ALI, M. A., KUNDU, S., RAHMAN, M. A., KAMARA, J. K. and RENZAHO, A. M. N. (2022). Effectiveness of COVID-19 vaccines against delta variant (B.1.617.2): A meta-analysis. *Vaccines* **10** 277.

MORIÑA, D., FERNÁNDEZ-FONTELO, A., CABAÑA, A., ARRATIA, A., ÁVALOS, G. and PUIG, P. (2021). Cumulated burden of Covid-19 in Spain from a Bayesian perspective. *Eur. J. Public Health* **31** 917–920.

MULLAH, M. A. S. and YAN, P. (2022). A semi-parametric mixed model for short-term projection of daily COVID-19 incidence in Canada. *Epidemics* **38** 100537. https://doi.org/10.1016/j.epidem.2022.100537

NATIONAL COLLABORATING CENTRE FOR INFECTIOUS DISEASES (2022). Updates on COVID-19 Variants of Concern (VOC). Retrieved from https://nccid.ca/covid-19-variants/. Accessed: 2022-03-09.

PARKER, M. R. P., CAO, J., COWEN, L. L. E., ELLIOTT, L. T. and MA, J. (2024a). Code supplement to "Multi-site disease analytics with applications to estimating COVID-19 undetected cases in Canada." https://doi.org/10.1214/24-AOAS1915SUPPA

PARKER, M. R. P., CAO, J., COWEN, L. L. E., ELLIOTT, L. T. and MA, J. (2024b). Appendix supplement to "Multi-site disease analytics with applications to estimating COVID-19 undetected cases in Canada." https://doi.org/10.1214/24-AOAS1915SUPPB

PARKER, M. R. P., LI, Y., ELLIOTT, L. T., MA, J. and COWEN, L. L. E. (2021). Under-reporting of COVID-19 in the northern health authority region of British Columbia. *Canad. J. Statist.* **49** 1018–1038.

PROVINCE OF BRITISH COLUMBIA (2020). BC COVID-19—laboratory information. Retrieved from https://governmentofbc.maps.arcgis.com/home/item.html?id=ba047e4a9bd24beb9ca6e94c05eddef9. Accessed: 2021-04-19.

R CORE TEAM (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

ROYLE, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60** 108–115. https://doi.org/10.1111/j.0006-341X.2004.00142.x

SAEED, S., DREWS, S. J., PAMBRUN, C., YI, Q.-L., OSMOND, L. and O'BRIEN, S. F. (2021). SARS-CoV-2 seroprevalence among blood donors after the first COVID-19 wave in Canada. *Transfusion* **61** 862–872. https://doi.org/10.1111/trf.16296

SHIM, E. (2021). Regional variability in COVID-19 case fatality rate in Canada, February-December 2020. *Int. J. Environ. Res. Public Health* **18**. https://doi.org/10.3390/ijerph18041839

SUBRAMANIAN, R., HE, Q. and PASCUAL, M. (2021). Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proc. Natl. Acad. Sci. USA* **118** e2019716118.

TANAKA, S. (2022). Economic impacts of SARS/MERS/COVID-19 in Asian countries. *Asian Econ. Policy Rev.* **17** 41–61.

TUITE, A. R., FISMAN, D. N. and GREER, A. L. (2020). Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. *CMAJ*, *Can. Med. Assoc. J.* **192** E497–E505. https://doi.org/10.1503/cmaj.200476

WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594.

WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *J. Big Data* **3** 9.

WU, P., LIU, F., CHANG, Z., LIN, Y., REN, M., ZHENG, C., LI, Y., PENG, Z., QIN, Y. et al. (2021). Assessing asymptomatic, presymptomatic, and symptomatic transmission risk of severe acute respiratory syndrome coronavirus 2. *Clin. Infect. Dis.* **73** e1314–e1320.

ZINSZER, K., MCKINNON, B., BOURQUE, N., PIERCE, L., SAUCIER, A., OTIS, A., CHERIET, I., PAPENBURG, J., HAMELIN, M. et al. (2021). Seroprevalence of SARS-CoV-2 antibodies among children in school and day care in Montreal, Canada. *JAMA Netw. Open* **4** e2135975. https://doi.org/10.1001/jamanetworkopen.2021.35975