Estimating the Size of the COVID-19 Population in British Columbia Using the Stratified Petersen Estimator

by

Viet Dao
B.A., Gustavus Adolphus College, 2017

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

© Viet Dao, 2023
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopying or other means, without the permission of the author.

We acknowledge and respect the lək̓ʷəŋən peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.

Estimating the Size of the COVID-19 Population in British Columbia Using the Stratified Petersen Estimator

by

Viet Dao
B.A., Gustavus Adolphus College, 2017

Supervisory Committee

Dr. Laura Cowen, Co-Supervisor
(Department of Mathematics and Statistics)

Dr. Junling Ma, Co-Supervisor
(Department of Mathematics and Statistics)

ABSTRACT

The presence of undetected COVID-19 cases is a known phenomenon. Mathematical modelling techniques, such as capture-recapture, provide a reliable method for estimating the true size of the infected population. Treating a positive SARS-CoV-2 diagnostic test result as the initial capture and a hospital admission with a COVID-19-related diagnosis code as the recapture, we developed a Lincoln-Petersen model with temporal stratification, taking into account factors that influence the occurrence of captures. Applying this model to repeated patient encounter data collected at the provincial level in British Columbia, we estimated the number of COVID-19 cases among males aged 35 or older during the first week of March 2021. Our analysis revealed that the true number of cases ranged from 4.94 to 9.18 times greater than the number of detected cases.

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisory committee, Laura and Junling, for their invaluable guidance and support throughout my degree. I would also like to extend my appreciation to my lab collaborators, Gracia and Kenny, for sharing their expertise and feedback, which have greatly contributed to the completion of this thesis. Lastly, I want to thank my partner, Kimmie, for being my unwavering source of motivation and encouragement.

This research was enabled in part by computing resources provided by the Digital Research Alliance of Canada (alliancecan.ca) and funding provided by the Visual and Automated Disease Analytics (VADA) Program.

Chapter 1
Introduction

Throughout human history, infectious diseases have been a significant threat to both public health and global stability. The recent COVID-19 pandemic has underlined the need for new ways to reduce the detrimental effects of such outbreaks, despite the knowledge learned from combating numerous viruses. Accurate estimation of case counts is crucial as it allows for effective resource allocation and informed decision-making to manage cases and understand disease patterns (Seddon et al., 2015). However, under-reporting of cases is common in public health studies (e.g., Toan et al., 2015; MacDougall et al., 2008). For COVID-19 in particular, Lau et al. (2021) revealed that the detection rate of COVID-19 cases in March 2020, across several global epicentres, was merely 1-2% based on case-fatality risk assessment. Similarly, in British Columbia, Skowronski et al. (2020) conducted a seroprevalence study and estimated that the actual number of COVID-19 cases in May 2020 varied between 2.25 and 20.5 (95% CI) times higher than the officially reported figures. In situations where it is challenging to mitigate under-reporting, mathematical models can be developed to provide reliable estimates of case counts using available data.

Recent studies have demonstrated the reliability of mathematical modelling methods in estimating the burden of the COVID-19 pandemic (Mehraeen et al., 2023). Among the methods examined is a novel N-mixtures-type model by Parker et al. (2021) which considers the observed count as under-counted and conditional on the hidden population size with a detection probability. Focusing on the Northern Health Authority region of British Columbia, the authors found the COVID-19 population in May 2020 to be 3.69-8.75 (95% CI) times greater than the reported numbers, aligning with the findings of the aforementioned serological study. This model possesses a unique advantage as it requires minimal data, relying solely on observed count data (i.e., publicly available COVID-19 case counts) for population size estimation. As the authors also noted, traditional methods such as capture-recapture are known to produce precise estimates and can be used to validate the results obtained, potentially paving the way for broader adoption of this model in disease modelling.

The Lincoln-Petersen estimator is a capture-recapture method to estimate the size

of a closed population, first used by Laplace in 1802 to estimate the population of France (Laplace, 1786). Subsequently, the concept underwent further development with notable contributions by two biologists, Petersen and Lincoln (White et al., 1982). The method has then been extensively applied to the study of both human and animal populations (e.g., see Chao et al., 2001; McCullough and Hirth, 1988; Xu et al., 2014).

The Lincoln-Petersen experiment involves two sampling events, one for tagging and one for recovery. In the first sampling, n_1 individuals out of the population size N are captured, uniquely tagged, and released back to the population. In the second sampling, n_2 out of N are captured which may include those previously tagged (m_2). Under conditions that allow the capture rate in the population and the capture rate in the second sample to be equal ($n_1/N = m_2/n_2$), the Lincoln-Petersen estimate of population size is then $\hat{N} = n_1 n_2/m_2$.

When we consider a population having COVID-19 as a closed population such as during a period of mandated travel restrictions, the sequential order of events for an individual in that population follows the pattern observed in the 2-sample experiment. Possessing a positive result to the diagnostic test for the SARS-CoV-2 virus constitutes the first capture (i.e. counting towards n_1). Being admitted to a hospital for treatment of COVID-19 constitutes the second capture (i.e. counting towards n_2). Let m_2 be the number of individuals captured on both occasions. Figure 1.1 illustrates the possible journeys of an individual within the COVID-19 population who remains susceptible during the two capture opportunities. As each event is random, it is possible for one to be in N without ever being captured.

Two key assumptions of the Petersen estimator are equal catchability within a sample and complete mixing of tagged and untagged individuals. However, there are cases where these assumptions are violated (e.g. varying catchability and differences in movements over time or space), which makes the simple Petersen estimator susceptible to substantial bias (Arnason et al., 1996). First explored by Schaefer (1951), Chapman and Junge Jr (1956), Darroch (1961) and more recently reviewed by Schwarz and Taylor (1998), stratification by time or space is one strategy to reduce this bias.

We develop a novel implementation of the stratified Petersen estimator using Bayesian Markov chain Monte Carlo (MCMC) fitting techniques to estimate the size of a population infected by the SARS-CoV-2 virus. We applied the model to repeated encounter data of COVID-19 patients in British Columbia, evaluated our findings, and discussed the potential implications on disease modelling strategies.

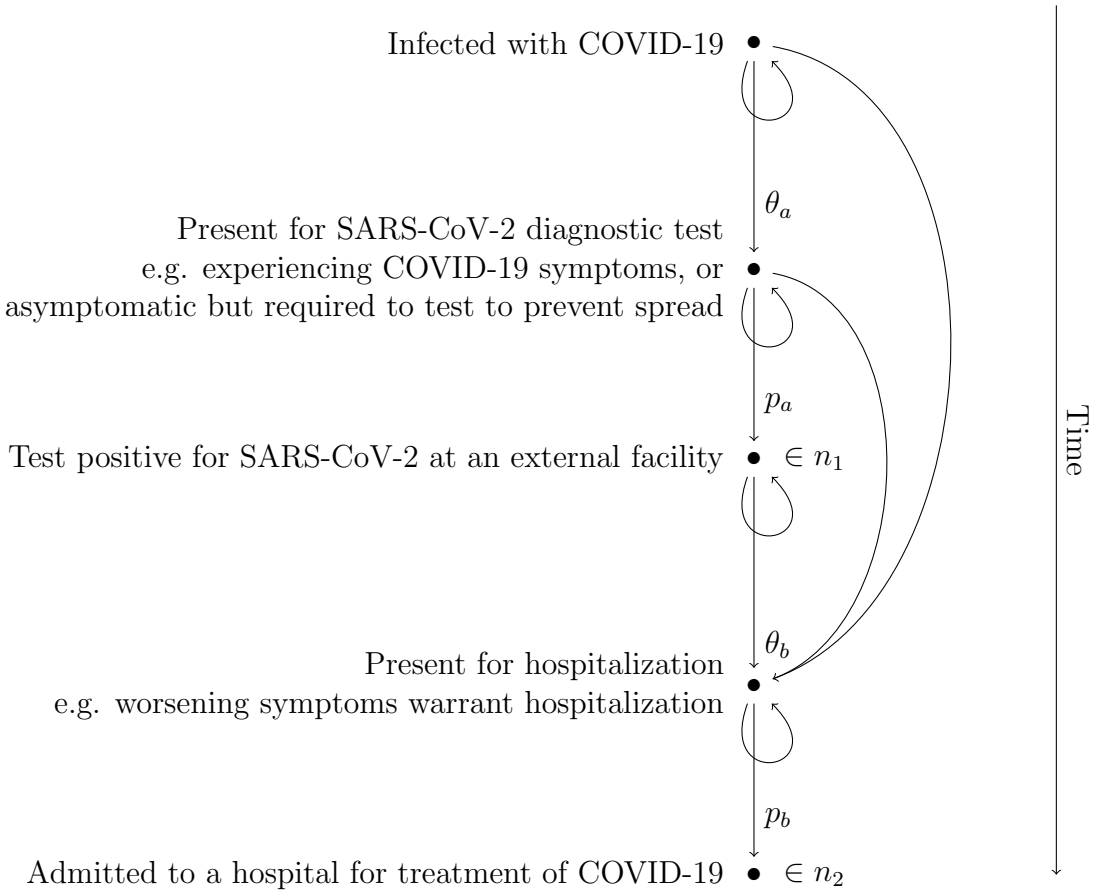

Figure 1.1: Possible journeys of an individual in the COVID-19 population, in chronological order.

Chapter 2
Methods

2.1 Model Development

A stratified Petersen experiment involves dividing the population into s nonoverlapping release strata and t nonoverlapping recovery strata, where capture probabilities are uniform with respect to the stratum. In temporal stratification, the release strata may be nonoverlapping time intervals in which separate groups of individuals become present for capture, and similarly for recovery strata. This type of stratification is suitable for modelling COVID-19 patient journeys, where it is reasonable to believe that differences in detection probabilities are attributable to specific factors such as demographic characteristics, movement restrictions, disease control measures, and testing policies over phases of the pandemic. On the other hand, spatial stratification may be more suitable when the patient groups have different capture probabilities due to spatial factors (e.g., dissimilar levels of access to tests or healthcare facilities). We will focus on temporal stratification in our simulation and case study due to the availability of data.

There are certain requirements applicable to the samplings as well as the strata in capture-recapture studies that the disease progression of COVID-19 does not necessarily abide by. We discuss these cases below and how they apply to the COVID-19 population. As for the sampling, conducting the two sampling events at different physical locations allow for movement and mixing within the population. In the study of the COVID-19 population, this can be satisfied when we consider testing to take place at external facilities (e.g. pop-up testing sites, private clinics) and hospitalization to take place at hospitals with the purpose of getting inpatient care. This requires data to contain location details. As for the strata, it is not necessary for the number of tagging strata and recovery strata to be equal, and it is also possible for them to overlap as long as movement from capture to recovery is ensured (e.g., in fisheries, recovery usually does not begin until several weeks after tagging has started) (Schwarz and Taylor, 1998). For mathematical and computational simplicity, however, we allowed the COVID-19 testing and hospitalization strata to completely overlap (i.e. the first

testing stratum is also the first hospitalization stratum). Consequently, we introduced an assumption that hospitalization might only occur after the first testing stratum (equivalently, from the second hospitalization stratum) to ensure individuals had time to become present for testing before being susceptible to hospitalization.

Let k denote the equivalent number of testing and hospitalization temporal strata. Whether or not an individual becomes present for COVID-19 testing or hospitalization may be influenced by various factors, such as the manifestation of COVID-19 symptoms and the accessibility of healthcare facilities (i.e. testing sites and hospitals). In order to incorporate these factors, we introduce θ_{aj} and θ_{bj} as the stratum-specific probabilities for an individual to become present for testing and for hospitalization, respectively, within stratum $j \in [1, k]$. For each individual i in the population ($i \in [1, N]$), we define $a_{ij} \sim Bernoulli(\theta_{aj})$ and $b_{ij} \sim Bernoulli(\theta_{bj})$ as two latent states which take a value of 1 if individual i is present for testing and hospitalization, respectively, within stratum j, and 0 otherwise. Let p_{aj} and p_{bj} represent the probabilities of actually obtaining a positive test result and being hospitalized (i.e., detected) in stratum j, respectively, provided that the individual is present for testing and hospitalization. We use $h_{ij} = 1$ to indicate individual i is captured (i.e. tests positive or gets hospitalized) in stratum j and $h_{ij} = 0$ otherwise. The observation model for each member of the population is then defined as $h_{ij}|a_{ij}, b_{ij} \sim Bernoulli(a_{ij}p_{aj} + b_{ij}p_{bj})$.

Data augmentation is an approach that may be used to estimate the unknown population size by parameterizing N by an occupancy rate ψ in a zero-inflated model with a known size M that is considerably larger than N (Kéry and Schaub, 2011). This superpopulation M consists of members of the population of interest N, and $M - N$ unobserved "pseudo-individuals". If we introduce a latent variable

$$z_i \sim Bernoulli(\psi) \tag{2.1}$$

indicating whether the individual i ($i \in [1, M]$) also belongs to N with probability ψ, the estimated population size can now be derived from

$$\hat{N} = \sum_{i=1}^{M} z_i \tag{2.2}$$

with an expected value of

$$\mathbf{E}(N) = M \times \psi. \tag{2.3}$$

Incorporating data augmentation, we have the following set of equations:
For the initial time stratum $j = 1$,

$$a_{i1}|z_i \sim Bernoulli(z_i\theta_{a1}) \tag{2.4}$$

Because an individual must be present for testing during at least one temporal

stratum before hospitalization,

$$b_{i1}|z_i = 0 \tag{2.5}$$

For time strata $j \geq 2$, a_{ij} and b_{ij} are conditioned on their histories.

$$a_{ij}|z_i, b_{ij} \sim Bernoulli(z_i \prod_{l=1}^{j-1}(1 - a_{il}) \prod_{l=1}^{j}(1 - b_{il})\theta_{aj}) \tag{2.6}$$

$$b_{ij}|z_i \sim Bernoulli(z_i \prod_{l=1}^{j-1}(1 - b_{il})\theta_{bj}) \tag{2.7}$$

The observation model is then

$$h_{ij}|z_i, a_{ij}, b_{ij} \sim Bernoulli(z_i a_{ij} p_{aj} + z_i b_{ij} p_{bj}) \tag{2.8}$$

Assuming that the observations are independent, the posterior distribution for our parameters is

$$f(\psi, \theta_{aj}, p_{aj}, \theta_{bj}, p_{bj}|\mathbf{h}, \mathbf{z}) \propto \prod_{i=1}^{M}\prod_{j=1}^{k}(z_i a_{ij} p_{aj} + z_i b_{ij} p_{bj})^{h_{ij}}(1 - z_i a_{ij} p_{aj} - z_i b_{ij} p_{bj})^{1-h_{ij}} \times f(\psi, \theta_{aj}, p_{aj}, \theta_{bj}, p_{bj}) \tag{2.9}$$

2.1.1 Prior Distributions

The model parameters ψ, θ_a, p_a, θ_b, p_b are all probabilities. Having no prior knowledge about their values, we follow a standard practice to set the prior distributions of these parameters to be non-informative Uniform(0,1) (Banner et al., 2020).

2.2 Bayesian Analysis Implementation

We conducted a Bayesian analysis of our model using MCMC techniques. The implementation was executed in the R language (R Core Team, 2022) using the NIMBLE package (de Valpine et al., 2017).

2.2.1 Simplifying the Model

As we were working in Population Data BC's secure server environment to access the research data, the model fitting process was computationally expensive, leading to time and memory allocation issues. To address these limitations, we selected a study period that was adequately short (e.g., 1-2 weeks), and we limited potential

factors causing temporal disparities in the presence and capture probabilities across strata (e.g., unchanged travel restrictions). This led to our assumption that $\theta_{aj} = \theta_a$, $p_{aj} = p_a$, $\theta_{bj} = \theta_b$, and $p_{bj} = p_b$ for $j = 1, 2, ..., k$, which effectively reduced the data structures for these parameters by an entire dimension.

2.2.2 Partially Observed Presence

The capture history can provide comprehensive information about the presence for testing and/or hospitalization. For instance, if an individual tested positive for SARS-CoV-2 in the third stratum, it implies they were present for testing but not present for hospitalization in that specific stratum.

Let a_{ij}^{obs} and b_{ij}^{obs} represent the observed presence histories where 0s and 1s are used to indicate the known presence status. At this stage, it became crucial to differentiate between a positive test and a hospitalization event in h_{ij}, as both were represented by 1s due to the Bernoulli distribution. To address this, a corresponding capture history $\tilde{h}_{ij}$ was generated, where each hospitalization occurrence was denoted by a 2, distinguishing it from a positive test. As a result, there were four possible scenarios for $\tilde{h}_{ij}$, based on which a_{ij}^{obs} and b_{ij}^{obs} were retroactively determined. Table 2.1 presents an example where $k = 5$ for each of the four scenarios.

Table 2.1: Mapping of capture in h_{ij} and $\tilde{h}_{ij}$ to observed presence in a_{ij}^{obs} and b_{ij}^{obs}. Latent values not deducible from the capture histories remained empty. Scenario descriptions are I) individual tested at time 2, hospitalized at time 3; II) individual tested at time 3 and not seen again; III) individual hospitalized at time 3; and IV) augmented individual.

	Scenario			
	I	II	III	IV
h_{ij}	0 1 1 0 0	0 0 1 0 0	0 0 1 0 0	0 0 0 0 0
$\tilde{h}_{ij}$	0 1 2 0 0	0 0 1 0 0	0 0 2 0 0	0 0 0 0 0
a_{ij}^{obs}	0 1 0 0 0	0 0 1 0 0	- - 0 0 0	- - - - -
b_{ij}^{obs}	0 0 1 0 0	0 0 0 - -	0 0 1 0 0	- - - - -

Chapter 3
Simulation Study

3.1 Simulation Configuration

In order to demonstrate the feasibility of our model, we conducted a simulation study using Digital Research Alliance of Canada's advanced computing resources with a set of parameter values commensurate with the COVID-19 population in our case study. We generated 30 data sets for each combination of true population size $N = 7250$, augmented population size $M = 20,000$, number of strata $k = 7$, test-presence probability $\theta_a = 0.025$, positive-test probability $p_a = 0.8$ or 0.9, hospitalization-presence probability $\theta_b = 0.003$ or 0.0075, and hospitalization probability $p_b = 0.35$, 0.55, or 0.75. This resulted in 12 parameter combinations (Table 3.1) and 360 simulated data sets in total. We imposed this limit on the number of data sets due to the significant computing resource allocation and run time required to complete the MCMC.

Table 3.1: Combinations of true parameter values for simulation study.

Combination	N	M	k	θ_a	p_a	θ_b	p_b
1	7250	20,000	7	0.025	0.8	0.003	0.35
2	7250	20,000	7	0.025	0.8	0.003	0.55
3	7250	20,000	7	0.025	0.8	0.003	0.75
4	7250	20,000	7	0.025	0.8	0.0075	0.35
5	7250	20,000	7	0.025	0.8	0.0075	0.55
6	7250	20,000	7	0.025	0.8	0.0075	0.75
7	7250	20,000	7	0.025	0.9	0.003	0.35
8	7250	20,000	7	0.025	0.9	0.003	0.55
9	7250	20,000	7	0.025	0.9	0.003	0.75
10	7250	20,000	7	0.025	0.9	0.0075	0.35
11	7250	20,000	7	0.025	0.9	0.0075	0.55
12	7250	20,000	7	0.025	0.9	0.0075	0.75

For each individual i in a data set, we simulated a capture history by first using the true values of θ_a and θ_b to generate a_{ij} and b_{ij} based on Equations 2.4, 2.5, 2.6, and 2.7. We then generated the capture history h_{ij} from a_{ij} and b_{ij} and the true values of p_a and p_b based on Equation 2.8. While we first generated a_{ij} and b_{ij} in order to construct h_{ij} in the simulation, in an actual experiment, they are non-observable latent variables and h_{ij} is the only observable data. As previously described in Section 2.2.2, these presence indicators might be partially or fully observed from h_{ij}, resulting in a_{ij}^{obs} and b_{ij}^{obs}. Following the data augmentation procedure, we appended $M-N$ empty rows to a_{ij}^{obs}, b_{ij}^{obs}, and h_{ij}.

For each data set, we fit the model to the augmented data using a `nimble` MCMC run with 3 chains, 30,000 iterations and 10,000 burn-in, resulting in 20,000 posterior samples. Model parameters were initialized from their prior distributions.

3.2 Results

For each true parameter value θ, let $\hat{\theta}$ be the corresponding posterior median representing the point estimate. We computed the bias, $\hat{\theta}-\theta$, for each probability estimate (ψ, θ_a, p_a, θ_b, p_b), and the relative bias, $(\hat{\theta}-\theta)/\theta$, for the N estimate.

Upon analyzing the results, it is evident that the estimation of the parameters had varying levels of success when examining each parameter individually, as well as across the different true parameter value combinations. For N and the corresponding ψ, the bias indicated consistent overestimation, with the lowest bias observed in combinations 6, 11, and 12 (Figures 3.1, 3.2). This suggests that the estimation of N was most accurate when θ_b and p_b had the highest respective values. For θ_a, there was a low level of bias across the combinations overall, except for combinations 3 and 10 which exhibited a higher level of bias (Figure 3.3). In the case of p_a, while the bias indicated consistent underestimation, combinations 1-6 ($p_a = 0.8$) exhibited a noticeably lower level of bias, with the lowest in combinations 1 and 6 (Figure 3.4). For θ_b, combinations 2, 4, 7, 8, and 10 had the lowest bias, while combinations 1, 5, and 6 had the highest bias (Figure 3.5). For p_b, there was a clear pattern indicating increasing bias as the true value of p_b increased. When grouped by this parameter's true value, combinations 1, 4, 7, and 10 (corresponding to $p_b = 0.35$) exhibited the lowest bias. This suggests that the estimation of p_b is most accurate when this probability is lower (Figure 3.6).

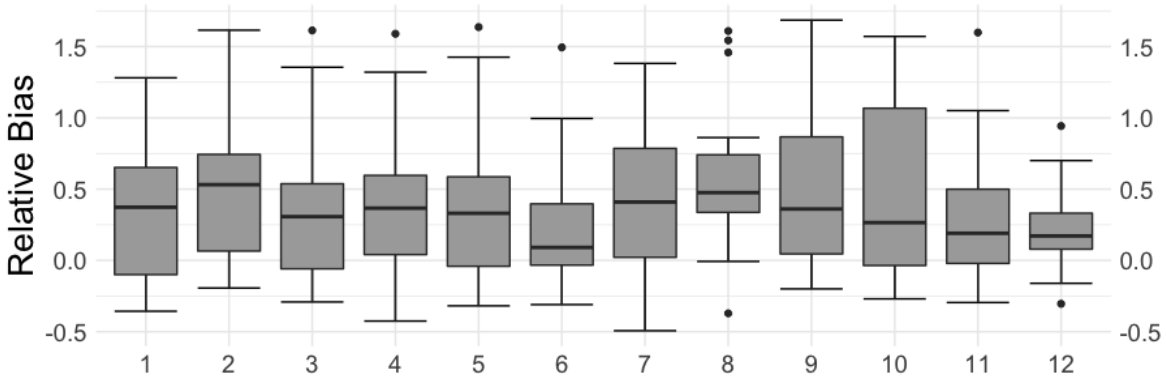

Figure 3.1: Relative bias of N estimate across all true parameter combinations.

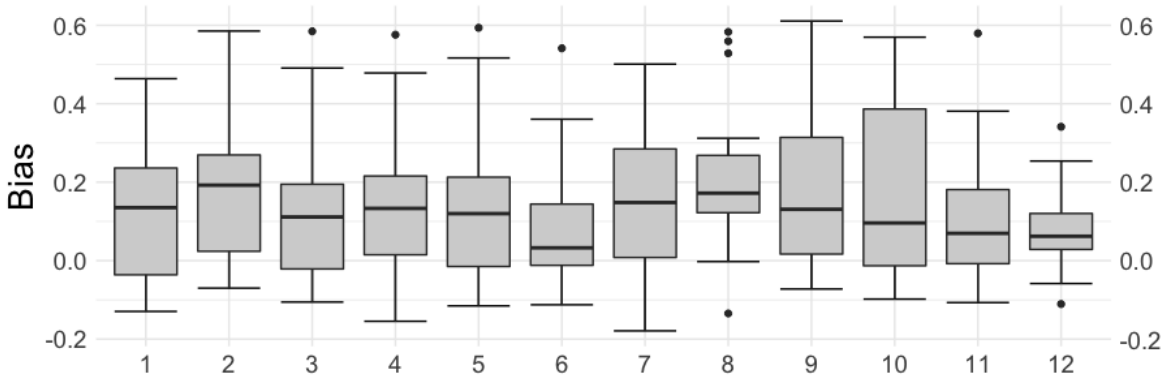

Figure 3.2: Bias of ψ estimate across all true parameter combinations.

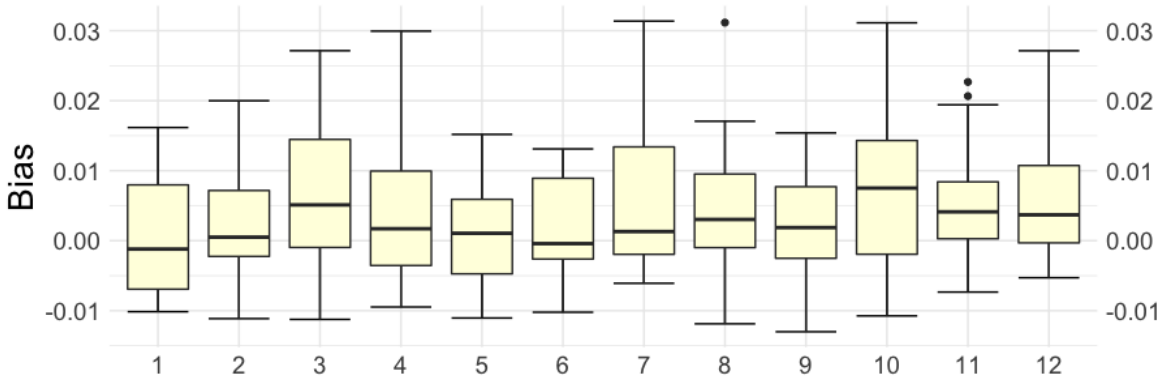

Figure 3.3: Bias of θ_a estimate across all true parameter combinations.

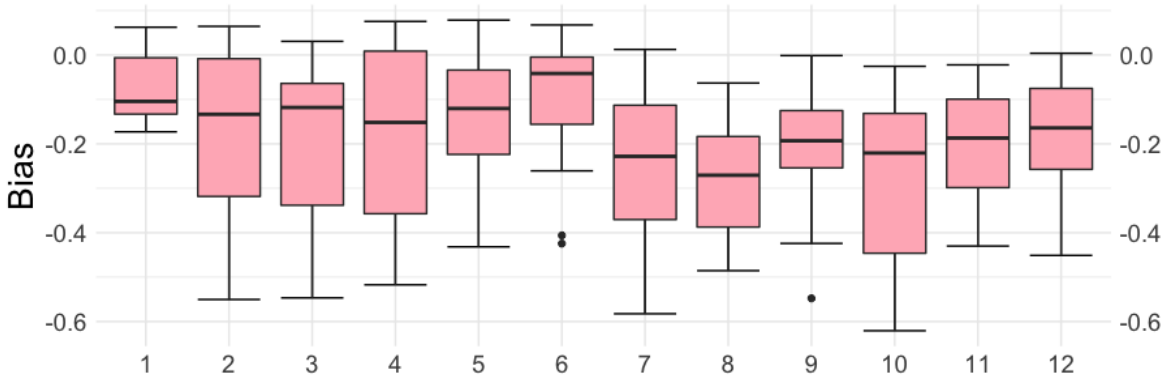

Figure 3.4: Bias of p_a estimate across all true parameter combinations.

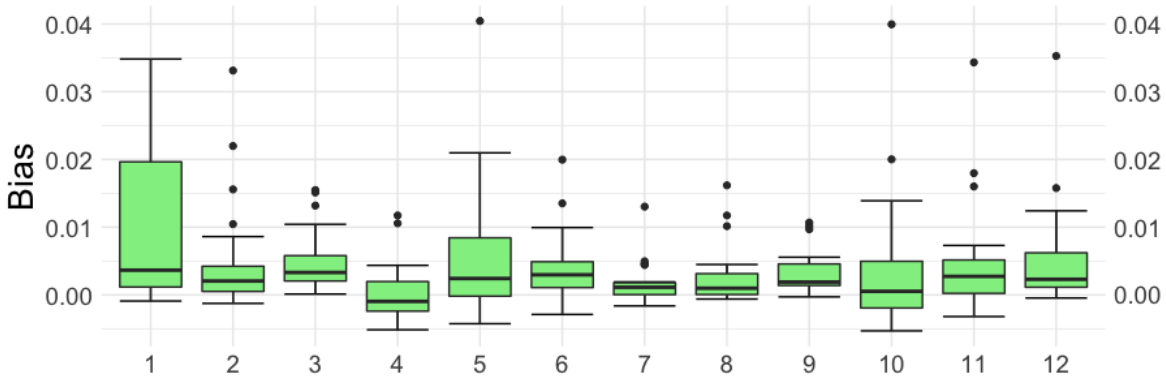

Figure 3.5: Bias of θ_b estimate across all true parameter combinations.

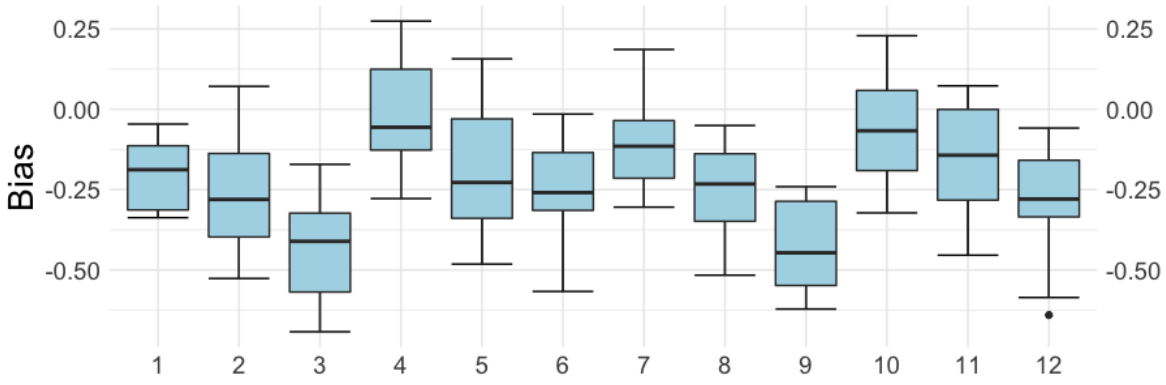

Figure 3.6: Bias of p_b estimate across all true parameter combinations.

In order to evaluate the precision of the estimates, we calculated the average posterior standard deviation (SD) and the root mean squared error (RMSE) of the estimates (Table 3.2). Overall, these evaluation metrics are relatively similar across the parameter combinations considered, with the exception of combinations 6, 11, and 12 where both measures for N and ψ are noticeably lower.

Table 3.2: Average posterior standard deviation (SD) and root-mean-square error (RMSE) of parameter estimates for simulation combinations 1-12.

	N		ψ		θ_a		p_a		θ_b		p_b	
	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE
1	2058	3085	0.103	0.190	0.010	0.009	0.179	0.076	0.009	0.011	0.174	0.108
2	2215	3379	0.111	0.169	0.010	0.007	0.168	0.186	0.007	0.008	0.230	0.155
3	2126	3638	0.106	0.182	0.011	0.009	0.156	0.181	0.007	0.004	0.236	0.154
4	2247	3522	0.112	0.176	0.011	0.011	0.174	0.183	0.007	0.004	0.231	0.147
5	2269	3757	0.113	0.188	0.010	0.007	0.172	0.125	0.008	0.009	0.212	0.171
6	1671	2837	0.084	0.142	0.008	0.007	0.158	0.125	0.007	0.005	0.218	0.136
7	2210	3596	0.115	0.180	0.010	0.010	0.166	0.155	0.006	0.003	0.251	0.138
8	2306	3398	0.115	0.170	0.010	0.010	0.175	0.119	0.006	0.004	0.244	0.140
9	2159	3672	0.108	0.183	0.009	0.007	0.166	0.115	0.006	0.003	0.236	0.131
10	2155	4524	0.108	0.226	0.010	0.010	0.171	0.172	0.008	0.010	0.211	0.156
11	1887	3007	0.094	0.150	0.009	0.007	0.163	0.125	0.007	0.007	0.226	0.164
12	1762	1923	0.088	0.096	0.009	0.007	0.154	0.128	0.007	0.007	0.213	0.155

Chapter 4
Case Study: COVID-19 Population in BC

4.1 Data

We were provided access to two main administrative health data sets, namely records of COVID-19 tests and hospital admissions taken place across British Columbia, through a Data Access Request to Population Data BC. The test records included unique patient identifiers (IDs), test sample collection timestamps, and test results. The hospitalization records included patient IDs linkable to the test data, along with admission timestamps and diagnosis codes. We specifically targeted the population of males who were 35 or older during the 7-day period between March 5th and March 11th, 2021. There were a total of 11,640 COVID-19 test records collected within this period for 11,041 individuals in the target demographic. Excluding non-positive results and records without provincial health numbers resulted in 1048 (9%) valid positive tests from 1033 unique individuals. As for the hospitalization data, aside from using the same filters applied to the test records, we selected only hospitalizations that contained at least one of the COVID-19-related diagnostic codes as defined by the National Center for Health Statistics (2020) (U071, U072, Z115, Z208, Z861, M358, J128) to ensure the two samples belonged to one infected population. There were 7931 hospitalizations involving 7661 individuals in the target demographic. Of these hospitalizations, 78 (1%), corresponding to 73 different patients, were related to COVID-19. Among these patients, 40 (54.8%) also had a positive COVID-19 test within the same time frame.

It was essential to structure the given data into capture histories (h_{ij}) and observed presence histories (a_{ij}^{obs}, b_{ij}^{obs}) then augment them. For each individual identified by the provided unique ID, we constructed a test history as a vector of length 7 (for 7 days) in which each positive-test-day was indicated by a 1 and a 0 otherwise. Each hospitalization history was similarly represented by a vector of length 7 with 1 and 0 values. As a result, each individual was mapped to two vectors, one for their test history and another for their hospitalization history.

Table 4.1: Summary statistics for 35-or-older males who tested positive and were hospitalized for COVID-19 from March 5th to March 11th, 2021.

Testing	Tested	Hospitalization stratum						
stratum	positive	03/05	03/06	03/07	03/08	03/09	03/10	03/11
03/05	160	0	1	0	1	1	1	0
03/06	110	0	0	0	0	0	2	0
03/07	115	0	0	0	0	2	0	0
03/08	138	0	0	0	0	0	0	0
03/09	152	0	0	0	0	0	0	1
03/10	159	0	0	0	0	0	0	1
03/11	177	0	0	0	0	0	0	0
	Untested	0	13	4	8	10	3	7
Total hospitalizations		0	14	4	9	13	6	9

Merging these two vectors by patient ID while keeping all the content from both even if they did not have a match resulted in a single capture history vector for each individual, which contained a number of tests and/or hospitalizations (it was possible to have multiple positive tests and hospitalizations over the 7 days). We cleaned the merged capture histories so that for each patient, their capture history contained a maximum of one earliest positive test and one earliest hospitalization following a test if the patient previously had a positive test. Hospitalizations dated in the first day were disregarded to follow the essential sequence of events in capture-recapture studies, where each individual must become present for capture at least once before being potentially recaptured. The two sampling sites also had to be distinct in order to minimize any potential bias resulting from population movements. The data did not provide explicit information about the locations where the tests were performed. However, advised by health research experts familiar with the data, we assumed that individuals who tested positive on the same day as their hospitalization likely underwent the test as part of the hospital admission process, and thus, those tests were not included in the capture history. On the other hand, tests conducted on different days from hospitalizations were assumed to have been performed at external facilities.

In the end, there were 1056 capture histories for further analysis, among which 1011 (95.7%) had a positive test, 55 (5.2%) were hospitalized, and 10 (0.9%) had a positive test followed by a hospital admission. Detailed capture counts by stratum are provided in Table 4.1.

The capture history and corresponding observed presence histories were each augmented by 18,944 rows (M = 20,000). We fit the model to the augmented data by running 3 MCMC chains with 30,000 iterations and 10,000 of burn-in for each chain. Parameters were independently initialized from their respective prior distributions. The Gelman-Rubin statistic $\hat{R}$ was used to assess convergence for each parameter (Gel-

man and Rubin, 1992). $\hat{R}$ was less than 1.1 for all of our model parameters, indicating strong evidence supporting the convergence of the samplers.

4.2 Results

The ψ estimate is not informative on its own, but when multiplied by M, it provides the expected value of N (Equation 2.3). As shown in Table 4.2, the median estimate of N as a sum of z_i's (Equation 2.2) is 7275 (95% CI = (5218, 9694)), which is 6.89 (95% CI = (4.94, 9.18)) times greater than the number of detected cases (n=1056). This estimate also suggests a detection rate of 14.52% (95% CI = (10.89%, 20.24%)). The estimates for θ_a and θ_b indicate very low chances of being present for testing and hospitalization, respectively. In contrast, the estimates of p_a and p_b indicate that individuals who were present for testing and hospitalization were very likely to test positive and be admitted to the hospital, respectively. The credible intervals of the hospitalization parameters are considerably wide, reflecting the sparse hospitalization counts (Table 4.1).

Table 4.2: Summary statistics for model parameters.

Parameter	Posterior Estimates			
	Mean	Median	Standard Deviation	95% Credible Interval
ψ	0.351	0.344	0.056	(0.246, 0.457)
N	7403	7275	1187.012	(5218, 9694)
θ_a	0.025	0.025	0.005	(0.017, 0.035)
p_a	0.868	0.894	0.103	(0.675, 0.999)
θ_b	0.003	0.003	0.002	(0.001, 0.008)
p_b	0.532	0.515	0.262	(0.142, 0.998)

Chapter 5
Discussion

We applied the stratified Petersen approach to model the population of males older than 35 infected with COVID-19 in British Columbia over a week of March 2021 using diagnostic test result and hospitalization data. This period marked the final week of restrictions on gatherings and events in the province, where transmission had stabilized after the second wave of COVID-19. The Gamma and Alpha variants were the dominating variants during this time (Hogan et al., 2021). Our estimate of the median size of this population is 6.89 (95% CI = (4.94, 9.18)) times greater than the number of detected cases (n=1056), indicating a detection rate of 14.52% (95% CI = (10.89%, 20.24%)). In other words, our model suggests that approximately 14.52% of all individuals who had COVID-19 during the study period tested positive and/or were hospitalized due to the disease. Our 95% credible interval aligns with the findings of previous studies conducted by Skowronski et al., 2020 and Parker et al., 2021 on COVID-19 populations in BC, which demonstrates the validity of our model. However, because our case study was conducted on a different time period and geographical boundary within the province, a comparative study of these methods using a shared data set would enable a direct comparison of their performance, allowing researchers to determine their suitability for specific applications. It is also worth noting the bias of the population size estimate, as demonstrated in our simulation study. While our observation suggests lower bias with larger second sample sizes, this behaviour can benefit from further examination. Prior research by Grimm et al. (2014), who assessed the levels of bias across various modifications of the Lincoln-Petersen estimator, or Evans and Bonett (1994), who introduced an adjusted estimator for small sample sizes, may provide valuable insights for determining strategies to limit this bias.

Apart from the population size, the other parameter estimates offer novel insights into the burden of COVID-19 in the province. Specifically, the probabilities of testing and hospitalization presence (θ_a, θ_b) can be utilized as non-clinical measures to quantify disease severity or testing policy, while the capture probabilities (p_a, p_b) may reflect the accessibility levels to healthcare services. To the best of our knowledge, these rates have not been extensively investigated by previous research.

Our study also demonstrates the potential applicability of our model to other COVID-19 populations, where demographic variables such sex and age could be used as covariates to refine the parameter estimates. We noted the computational challenges when running the MCMC on the secured computing environment. The runtime of our case study, configured on the secured research environment, lasted approximately 45 hours. Each run of our simulation study conducted on Digital Research Alliance of Canada's Cedar cluster required a varying amount of time, typically 36-48 hours. These limitations constrained our ability to apply the model to larger populations or use it for real-time forecasting. To mitigate this challenge, the data-efficient N-mixture model by Parker et al. (2021) holds promise as a possible alternative.

There were three categories of test results recorded in the data: positive, negative, and indeterminate. By focusing solely only the positive results, we effectively made the assumption that the proportion of false negatives and false positives was negligible. Future research may consider accounting for these potential errors. Additionally, while one could take multiple tests and attain different results, we assumed that their disease status did not change after testing positive (equivalent to the no-tag-loss condition). The Lincoln-Petersen estimate is designed for a closed population, and although this condition may not hold true in many cases for human populations, we believed our model was particularly applicable to the COVID-19 population during times of travel restrictions and social gathering limitations. Removals from the population (e.g., deaths) were also not considered a factor. Because our study period of 7 days was shorter than the typical duration from symptom onset to hospital discharge or death, we believed that the disease status remained constant and removals negligible (Byrne et al., 2020). Apart from constraining the length of the study period, we also limited the population of study to a single sex and age group to reduce variability and assumed constant presence and capture probabilities $(\theta_a, p_a, \theta_b, p_b)$ across strata. While this assumption reduced the computing cost, it may not be realistic in other settings such as during a longer period or a period of rapidly growing rates. Future studies covering another time span can consider examining population removals in more detail, as well as allowing the parameters to be time-dependent.

Bibliography

Arnason, A., Kirby, C., Schwarz, C., and Irvine, J. (1996). Computer analysis of marking data from stratified populations for estimation of salmonid escapements and the size of other populations. Canadian Technical Report of Fisheries and Aquatic Sciences No. 2106.

Banner, K. M., Irvine, K. M., and Rodhouse, T. J. (2020). The use of Bayesian priors in ecology: the good, the bad and the not great. *Methods in Ecology and Evolution*, 11(8):882–889.

Byrne, A. W., McEvoy, D., Collins, A. B., Hunt, K., Casey, M., Barber, A., Butler, F., Griffin, J., Lane, E. A., McAloon, C., O'Brien, K., Wall, P., Walsh, K. A., and More, S. J. (2020). Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ Open*, 10(8). Publisher: British Medical Journal Publishing Group _eprint: https://bmjopen.bmj.com/content/10/8/e039856.full.pdf.

Chao, A., Tsay, P., Lin, S.-H., Shau, W.-Y., and Chao, D.-Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in medicine*, 20(20):3123–3157.

Chapman, D. G. and Junge Jr, C. O. (1956). The estimation of the size of a stratified animal population. *The Annals of Mathematical Statistics*, pages 375–389.

Darroch, J. (1961). The two-sample capture-recapture census when tagging and sampling are stratified. *Biometrika*, 48(3/4):241–260.

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. *Journal of Computational and Graphical Statistics*, 26:403–417.

Evans, M. A. and Bonett, D. G. (1994). Bias reduction for multiple-recapture estimators of closed population size. *Biometrics*, pages 388–395.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Grimm, A., Gruber, B., and Henle, K. (2014). Reliability of different mark-recapture methods for population size estimation tested against reference population sizes constructed from field data. *PLoS One*, 9(6):e98840.

Hogan, C. A., Jassem, A. N., Sbihi, H., Joffres, Y., Tyson, J. R., Noftall, K., Taylor, M., Lee, T., Fjell, C., Wilmer, A., Galbraith, J., Romney, M. G., Henry, B., Krajden, M., Galanis, E., Prystajecky, N., and Hoang, L. M. N. (2021). Rapid increase in SARS-CoV-2 P.1 lineage leading to codominance with B.1.1.7 lineage, British Columbia, Canada, January-April 2021. *Emerg Infect Dis*, 27(11):2802–2809.

Kéry, M. and Schaub, M. (2011). *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic Press.

Laplace, P. S. (1786). Sur les naissances, les mariages et les mortsa Paris depuis 1771 jusqu'a 1784 et dans toute l'étendue de la France, pendant les années 1781 et 1782. *Mémoires de l'Académie Royale des Sciences présentés par divers savans*, pages 35–46.

Lau, H., Khosrawipour, T., Kocbach, P., Ichii, H., Bania, J., and Khosrawipour, V. (2021). Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters. *Pulmonology*, 27(2):110–115.

MacDougall, L., Majowicz, S., Doré, K., Flint, J., Thomas, K., Kovacs, S., and Sockett, P. (2008). Under-reporting of infectious gastrointestinal illness in British Columbia, Canada: who is counted in provincial communicable disease statistics? *Epidemiol Infect*, 136(2):248–56.

McCullough, D. R. and Hirth, D. H. (1988). Evaluation of the Petersen-Lincoln estimator for a white-tailed deer population. *The Journal of Wildlife Management*, pages 534–544.

Mehraeen, E., Pashaei, Z., Khajeh Akhtaran, F., Dashti, M., Afzalian, A., Ghasemzadeh, A., Asili, P., Kahrizi, M. S., Mirahmad, M., Rahimi, E., Matini, P., Afsahi, A. M., Dadras, O., and SeyedAlinaghi, S. (2023). Estimating methods of the undetected infections in the COVID-19 outbreak: a systematic review. *Infectious Disorders - Drug Targets*, 23:1–1.

National Center for Health Statistics (2020). New ICD-10-CM code for the 2019 novel coronavirus (COVID-19), december 3, 2020.

Parker, M. R. P., Li, Y. M., Elliott, L. T., Ma, J. L., and Cowen, L. L. E. (2021). Underreporting of COVID-19 in the Northern Health Authority region of British Columbia. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 49(4):1018–1038.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schaefer, M. B. (1951). Estimation of the size of animal populations by marking experiments. *Fishery Bulletin of the Fish and Wildlife Service*, page 189.

Schwarz, C. J. and Taylor, C. G. (1998). Use of the stratified-Petersen estimator in fisheries management: estimating the number of pink salmon (Oncorhynchus gorbuscha) spawners in the Fraser River. *Canadian Journal of Fisheries and Aquatic Sciences*, 55(2):281–296.

Seddon, J. A., Jenkins, H. E., Liu, L., Cohen, T., Black, R. E., Vos, T., Becerra, M. C., Graham, S. M., Sismanidis, C., and Dodd, P. J. (2015). Counting children with tuberculosis: why numbers matter. *The International Journal of Tuberculosis and Lung Disease*, 19(12):S9–S16.

Skowronski, D. M., Sekirov, I., Sabaiduc, S., Zou, M., Morshed, M., Lawrence, D., Smolina, K., Ahmed, M. A., Galanis, E., Fraser, M. N., Singal, M., Naus, M., Patrick, D. M., Kaweski, S. E., Mill, C., Reyes, R. C., Kelly, M. T., Levett, P. N., Petric, M., Henry, B., and Krajden, M. (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*.

Toan, N. T., Rossi, S., Prisco, G., Nante, N., and Viviani, S. (2015). Dengue epidemiology in selected endemic countries: factors influencing expansion factors as estimates of underreporting. *Tropical Medicine & International Health*, 20(7):840–863.

White, G. C., Anderson, D. R., Burnham, K. P., and Otis, D. L. (1982). *Capture-recapture and removal methods for sampling closed populations*. Los Alamos National Laboratory.

Xu, Y., Fyfe, M., Walker, L., and Cowen, L. L. (2014). Estimating the number of injection drug users in greater Victoria, Canada using capture-recapture methods. *Harm Reduction Journal*, 11:1–7.