THE ROLE OF AGE AND INFECTION ASCERTAINMENT IN THE COVID-19 EPIDEMIC IN BRITISH COLUMBIA

by

Harvir Singh Bhattal

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE

in

HONOURS CELLULAR, ANATOMICAL AND PHYSIOLOGICAL SCIENCES

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2021

Table of Contents

Abstract

The COVID-19 pandemic has presented a global health crisis of unprecedented scale – with a grave cost of life, and substantial social and economic ramifications. Mathematical models of COVID-19 have been indispensable in informing policy to minimize this cost. In order to explore the role of age and the suspected underreporting of infections in B.C.'s epidemic, we constructed an age-stratified model of COVID-19 that incorporates infection ascertainment. We found that fitting this model to reported case data, without additional constraint, is insufficient to precisely identify age group-specific infection dynamics and the extent of infection ascertainment. Therefore, we constrained the transmission parameters with external contact data – after which our model was able to simulate the observed epidemic with a high degree of accuracy. Furthermore, the current report developed a novel method to estimate infection ascertainment using only case and hospitalization data (along with a single time-point of seroprevalence). This method was used to back-calculate the true epidemic in B.C. from its outset to date. We confirm suspicions of significant underreporting of infections as the epidemic in March 2020 is estimated to be 14.4 (95% CI, 9.6-17.1) times larger than reported. Finally, we utilized our model to evaluate the effectiveness of different vaccination strategies in preventing hospitalization events. Based on the utility in blocking chains of transmission, the present report recommends vaccinating high-contact essential workers earlier than currently scheduled in B.C.

Acknowledgements

I would like to thank my supervisor Dr. Daniel Coombs – for whose patient guidance during this project I am greatly appreciative. I would also like to thank my family for their unconditional support.

List of Figures

List of Tables

Chapter 1: Introduction

1.1 COVID-19: Outbreak to Pandemic

On December 31, 2019, the Wuhan Municipal Health Commission issued its first public report on a cluster of 27 cases of pneumonia of unidentified etiology in Wuhan, Hubei Province, China (1). Due to the threat of a possible disease outbreak, the Chinese Centre for Disease Control and Prevention (CDC) collected and analyzed isolates from positive patients in order to the identify the causal agent. Preliminary screens for known respiratory pathogens – including the Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) which ravaged Asia during the SARS epidemic of 2003, the Middle Eastern Respiratory Syndrome coronavirus (MERS-CoV), and influenza viruses – were unsuccessful. Gene sequencing of the isolates revealed that the causal agent for these atypical pneumonias was in fact a novel coronavirus (2,3). This virus was provisionally termed the 2019-nCoV by the World Health Organization (WHO), and later renamed to SARS-CoV-2 by the International Committee on Taxonomy of Viruses (based on its phylogenetic similarity to SARS-CoV). The disease caused by SARS-CoV-2 was named the Coronavirus Disease 2019 (COVID-19) by the WHO (4).

Despite public health regulations, including restricting travel into and out of Wuhan, the virus spread across China and to neighbouring nations. On January 30, 2020, the WHO formally declared the outbreak a "public health emergency of international concern" (PHEIC) – its highest level of alarm (5). In the coming weeks and months, COVID-19 proved to be both highly transmissible and highly deadly as epidemics overwhelmed healthcare systems around the world. By March 11, there were more than 118,000 reported cases and 4,000 reported deaths in 114 countries and the WHO had declared COVID-19 a pandemic (6).

British Columbia, Canada reported its first case of COVID-19 on January 26, 2020. As of April 2021, the epidemic has grown to a cumulative total of more than 100,000 reported cases and 1,400 reported deaths (7). The B.C. government has taken several public health measures, at both a province-wide and region-specific level, in an effort to curtail the epidemic. These include restrictions on public and private gatherings, a mandatory mask mandate, and most recently a centralized immunization program (8,9).

1.2 Mathematical Epidemiology

Our understanding of infectious disease dynamics is limited by our inability to perform controlled experiments with pathogenic agents in host populations – which would be unethical if not impossible with humans. Therefore, we must rely on mathematical models of disease transmission to answer scientific questions. These models are constructed on a theoretical basis and then validated against epidemiological data (reported infections, hospitalizations, deaths, etc.). Since these data are historical in nature, our analysis will necessarily be reactive. However, with sufficient data, collected and reported using consistent methods, our models may be sufficiently determined to provide real-time predictive power (10). Predictive models, such as those constructed by the British Columbia Centre for Disease Control (BC CDC), have been used to inform decisions taken by public health authorities throughout the course of the COVID-19 epidemic in B.C. (7,8).

1.2.1 SIR Model

The simplest model for understanding disease transmission is the SIR compartmental model based on the 1927 seminal work of Kermack and McKendrick (11). In this model, individuals of a population transition between different compartments at specified rates (Fig. 1). An individual begins in the *Susceptible* (S) compartment and is initially susceptible to infection. If infected, the

individual transitions into the *Infective* (I) compartment and is infectious to other susceptibles. Once the individual recovers (or dies) from the disease, they transition into the *Recovered* (R) compartment. The state variables $\langle S(t),\ I(t),\ R(t)\rangle$ describe the dynamical system at any point in time $t > t_0$ and are determined as the solution (for a given initial value $\langle S(t_0),\ I(t_0),\ R(t_0)\rangle$) to the following system of non-linear ODEs:

$$\frac{dS}{dt} = -\beta\left(\frac{S(t)}{N}\right) I(t)$$
$$\frac{dI}{dt} = \beta\left(\frac{S(t)}{N}\right) I(t) - \gamma I(t)$$
$$\frac{dR}{dt} = \gamma I(t)$$

In this model, we assume that contact mixing is homogenous between compartments. Thus, for a contact rate of c, an infective makes contact with $c \cdot (S(t)/N)$ susceptibles per unit time. For a probability p of transmitting infection, this infective will produce $c \cdot p \cdot (S(t)/N)$ secondary infections per unit time; and the $I(t)$ infectives in the population will produce a total of $c \cdot p \cdot (S(t)/N) \cdot I(t)$ new infections per unit time. By defining the transmission parameter as $\beta := c \cdot p$, we have derived the transition rate between the S and I compartments: $\beta(S(t)/N)I(t)$. Infectives leave the I compartment at a rate of $\gamma I(t)$ per unit time. Thus, the length of stay in the I compartment is exponentially distributed and has mean value of $1/\gamma$. It is important to consider that this model ignores birth, death, and immigration; so, the size of the population $N = S(t) + I(t) + R(t)$ is constant and we have that $\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$. This model also ignores the possibility of reinfection.

The epidemic will grow if and only if $\frac{dI}{dt} = (\beta(S(t)/N) - \gamma)I(t) > 0$. Since at the outset of an epidemic $S(t) \approx N$, this inequality is satisfied if and only if β/γ, which is defined as the

basic reproduction number R_0, is greater than 1. Conversely, the epidemic will become extinct when $R_0 < 1$, as is the desired outcome from public health interventions such as mass immunization. The R_0 value can alternatively be conceptualized as the number of secondary infections a single infective produces in a completely susceptible population – and is thus a common metric used to quantify the transmissibility of a pathogen (12).

1.3 Challenges with Modelling COVID-19

The paradox with modelling disease dynamics is that our models are least developed when they are needed most. At the outset of an epidemic, when (often underprepared) public health agencies must conduct risk assessments, our models are data-starved and thus provide limited utility. This often leads to a lagged public health response (and an associated cost of life). Since the public health authority was unable to adequately assess the threat, COVID-19 overwhelmed hospitals in Wuhan by January 2020 (6). The world stood witness to a brutal epidemic. The first estimate for the R_0 of COVID-19 (2.0 to 3.1), published to preprint[1] on January 23, further underscored the threat of this virus (12). Despite this warning, several countries suffered severe epidemics in the following months (6). This is likely because governments are unable to justify lockdowns without direct evidence of a threat to their own populations – which is only realized post-hoc.

B.C. experienced its first wave of COVID-19 infections (and associated hospitalizations and deaths) from March-May 2020 (7). However, the full extent of the epidemic at this stage is unknown as B.C. primarily tested high-risk groups and individuals connected to travel (13). Therefore, reported case counts necessarily deviate from the true epidemic by some unknown

[1] Due to the urgent demand posed by the pandemic, many of the articles cited in this report were published to preprint and are yet to be peer-reviewed.

(possibly substantial) multiplicative factor. Even after broad testing of all symptomatic individual was implemented (April 21 onwards), reported case counts are suspected to ascertain only a fraction of true infections – herein termed the ascertainment fraction – due to asymptomatic infection and the testing behaviours of symptomatic individuals (which may vary throughout time and with age, disease severity, socioeconomic status, etc.) (13,14). Therefore, analyzing true epidemiological trends from reported case counts is confounded by the ascertainment fraction, which is difficult to accurately estimate (15,16).

The risk for severe disease (hospitalization or death) due to COVID-19 is well known to skew towards older populations (17). Older individuals have, accordingly, been more willing to restrict contacts than younger individuals. Due to data-limitations (caused by low infection counts and poor testing capacity), early models of COVID-19 were not structured by age and therefore masked age-specific transmission dynamics – many of which may be key to characterizing the true epidemic (10). It is important to note that age exacerbates uncertainty with the ascertainment fraction since older individuals are more likely to show severe symptoms and thus more likely to get tested.

In the current project, we seek to construct a model of the COVID-19 epidemic in B.C. that is structured by age and incorporates testing behaviour. To this end, we will attempt to computationally constrain this model using reported B.C. case and hospitalization data. Once sufficiently developed, we will apply this model to simulate the epidemic under various immunization programs – so as to advise on the ideal vaccination schedule to halt this epidemic as rapidly as possible. In addition, we develop a novel method to estimate infection ascertainment using case and hospitalization data (along with a single time-point of seroprevalence data).

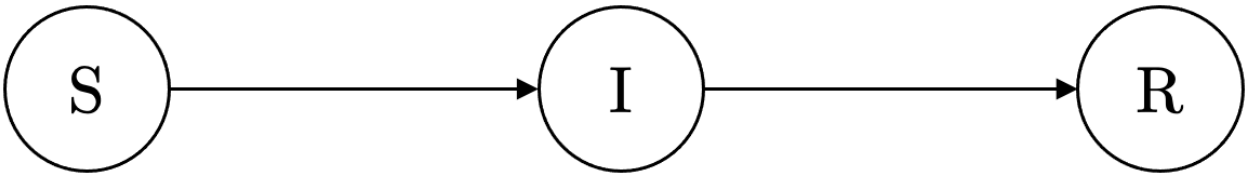

Figure 1: Schematic of the SIR compartmental model of disease transmission first introduced by Kermack and McKendrick (11). The S, I, and R compartments represent the susceptibles, infectives, and recovered individuals in the population, respectively.

Chapter 2: Methods

2.1 Data and Code

COVID-19 case and hospitalization data was extracted from the BC CDC linelist in (7). Analysis in the current project was primarily conducted using MATLAB and R. The code required to reproduce our results is available at this project's public GitHub repository (https://github.com/hbhattal/covid-19-model).

2.2 Analysis of Age

Decisions on the age-cutoffs for both the two- and three-age group models were informed by unbiased k-means clustering of age cohorts based on infection and hospitalization dynamics (see Appendix A). Practical considerations on data limitations were also taken when deciding these cutoffs.

2.2.1 k-means Clustering

k-means clustering is an unbiased clustering algorithm that groups observations so as to minimize the Euclidian distance between observations of a cluster to their centroid (18). Since the final four months of 2020 showed consistent growth and consistent testing behaviour (13), age cohorts were clustered on either two or three centroids (depending on the number of age groups) for either the per-capita case counts, per-capita hospitalization counts, or case-hospitalization fractions. k-means clustering was conducted using the built-in `kmeans` function in R.

2.3 Model Fitting

The classic Kermack-McKendrick SIR model of disease transmission (11) was structured by age and subcompartmentalized to incorporate infection ascertainment to develop the generalized n-age group $SI_RI_{UR}R$ model of COVID-19 transmission (see in section 3.1). We sought to identify

the transmission and ascertainment fraction parameters by fitting our two- and three-age group model to the cumulative reported case count data over different time periods and under various constraints. All model fitting was conducted in MATLAB.

2.3.1 Simulating Epidemic

In order to simulate the epidemic for a given parameter set, we solved the system of non-linear ordinary differential equations describing the two- or three-age group model using the built-in `ode45` function. The timespan (`tspan`) argument is calculated as the sequence array describing the duration over which the system of ODEs is solved. The initial value (`y1`) argument is the array describing the susceptibles (S), reported infectives (I_R), unreported infectives (I_{UR}), and recovered individuals (R) for each age group. The recovered individuals in each group are computed as the cumulative number of cases 5 days before the start date (since the mean infectious period is 5.0 days) (19). The active infectives in each group are computed as the cumulative number of cases at the start date minus the number of recovered individuals. The proportion of active infectives that are reported is the ascertainment fraction, θ; while the proportion unreported is $1-\theta$. The susceptibles in each age group are computed as the population of the age group minus the sum of all active infectives and recovered individuals.

2.3.2 Optimization

The sum of squared residuals (SSR) is calculated at each time-point between the cumulative case count curves corresponding to the observed epidemic and the epidemic simulated by the model over the designated time period. The normalized SSR (NSSR) is calculated by normalizing all the residuals as a proportion of the observed epidemic at each time-point. The optimal parameters were identified by minimizing either the SSR or NSSR under various parameter constraints using

the built-in `fmincon` function. In order to confirm that the optima returned were indeed stable global optima, this procedure was repeated with multiple different initial guesses.

2.3.3 Parameter Constraint: Contact Behaviour

The contact diary surveys in (20) provided cross-sectional data on the contact behaviour of B.C. residents at four time-points (May, July, Sept, Dec 2020) during the COVID-19 pandemic. The contact matrices reported in this study were broken down into age cohorts of 10 years (18-29, 30-39, 40-49, etc.). These contact matrices were adapted for the age groups studied in the current analysis by combining age cohorts using a simple weighted average, where the weight is the number of participants surveyed in each age cohort. Note that since participants under the age of 18 were not surveyed, this group's contact behaviour is assumed to be consistent with the contact behaviour of the youngest group.

2.4 Estimating Infection Ascertainment

In the present analysis, we introduced a novel method to estimate infection ascertainment as the ratio of the infection-hospitalization fraction and the case-hospitalization fraction ($\theta = \phi_I/\phi_C$; see section 3.3).

2.4.1 Infection-hospitalization Fraction (ϕ_I)

In order to calculate the infection-hospitalization fraction for each age group, we used the seroprevalence study in (21) to estimate the cumulative number of infections in the B.C. population. This study collected blood samples from May 9 to July 21, 2020 and found 0.56% (95% CI, 0.42-0.69%) of the population was positive for COVID-19 IgG antibodies. Although the data for each age cohort in B.C. was not reported, the data from the whole of Canada showed no statistically significant difference between cohorts. Therefore, we took this value as the seroprevalence for all age groups analyzed. Since the median time from symptom onset to

seroconversion for IgG antibodies is 14 days, this study shows the seroprevalence from two-weeks prior to sample collection. Therefore, we calculated the infection-hospitalization fraction as the ratio of the cumulative reported hospitalizations in each age groups and the cumulative infection count (as estimated from this seroprevalence) for each day April 25 to July 7. However, since sample collection was not uniform throughout the collection period, each of these estimates was weighted by the number of samples collected that week, and the weighted average for ϕ_I was reported.

2.4.2 Case-hospitalization Fraction (ϕ_C)

The case-hospitalization fraction, for each age group, was calculated as the ratio between total number of new hospitalizations and the total number of new cases for each month.

2.4.3 True Epidemic

The ascertainment fraction in each age group for each month was calculated as the ratio of the infection-hospitalization fraction and the monthly case-hospitalization fraction. The true epidemic was computed by dividing the daily case count by this ascertainment fraction for each age group.

2.5 Vaccination

The generalized n-age group $SI_RI_{UR}R$ model was further developed to include hospitalization and vaccination (n-age group $V{\sim}SI_RI_{UR}HR$ model; see section 3.1). The three-age group model was used to simulate two contrasting vaccination strategies, vaccinating the oldest population first vs. the youngest population first, to determine which minimizes the cumulative number of hospitalizations in the population after the epidemic has run its course.

2.5.1 Simulating Vaccination Schedules

The epidemic was simulated under a given parameter set using the built-in `ode45` function as described in subsection 2.3.1. Sensitivity analysis was conducted by quantifying the effects of

varying different parameters relevant to the vaccine's effectiveness on the resultant cumulative hospitalizations under each strategy.

Chapter 3: Results

3.1 Epidemiological Model

In order to model the effects of age and testing behaviour on COVID-19 transmission in B.C. we begin by developing the classic Kermack-McKendrick SIR model (11). Since our objective is to constrain this model using reported case data, we will construct it with ease of computation in mind.

3.1.1 Infection Ascertainment

The classic SIR model implicitly assumes that all infections are ascertained – which is simply not practical for COVID-19 (14). We will control for this by specifying the I compartment into subcompartments of reported (I_R) and unreported cases (I_{UR}) (Fig. 2). For an instantaneous ascertainment fraction of $\theta(t)$, we have that $I_R(t) = I(t)\theta(t)$ and that $I_{UR}(t) = I(t)\big(1-\theta(t)\big)$. The differentials for these subcompartments are:

$$\frac{dI_R}{dt} = \frac{dI}{dt}\theta(t) + I(t)\frac{d\theta}{dt}$$
$$\frac{dI_{UR}}{dt} = \frac{dI}{dt}\big(1-\theta(t)\big) - I(t)\frac{d\theta}{dt}$$

For computational simplicity, we will assume a constant ascertainment fraction which removes the $\frac{d\theta}{dt}$ terms. Thus, this model is described by the following dynamical system:

$$\frac{dS}{dt} = -\beta\left(\frac{S(t)}{N}\right)I(t)$$
$$\frac{dI_R}{dt} = \left[\beta\left(\frac{S(t)}{N}\right)I(t) - \gamma I(t)\right]\theta$$
$$\frac{dI_{UR}}{dt} = \left[\beta\left(\frac{S(t)}{N}\right)I(t) - \gamma I(t)\right](1-\theta)$$
$$\frac{dR}{dt} = \gamma I(t)$$

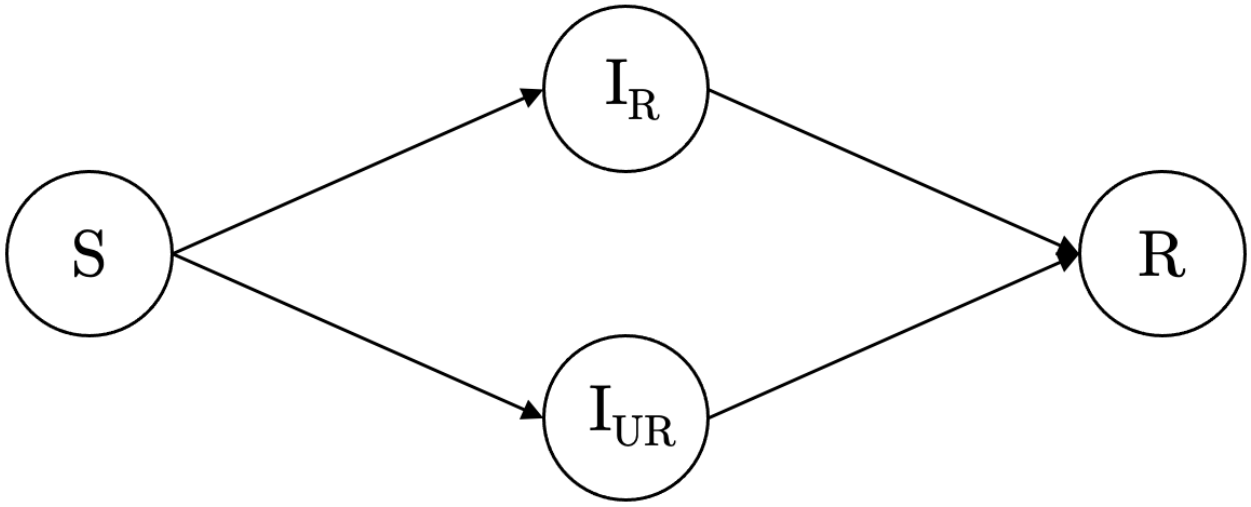

Figure 2: Schematic of the n-age group $SI_RI_{UR}R$ model of COVID-19 transmission, shown for the j-th group only. The S, I_R, I_{UR} and R compartments represent the susceptibles, reported infectives, unreported infectives, and recovered individuals in the population, respectively.

Since we are constraining this model using reported case data, it will be advantageous to parametrize the functions in terms of $I_R(t)$. Using the trivial identity $I(t) = I_R(t)/\theta$, the dynamical system describing this model is as follows:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta \left(\frac{S(t)}{N}\right) \left(\frac{I_R(t)}{\theta}\right) \\
\frac{dI_R}{dt} &= \left[\beta \left(\frac{S(t)}{N}\right) \left(\frac{I_R(t)}{\theta}\right) - \gamma \left(\frac{I_R(t)}{\theta}\right)\right] \theta \\
\frac{dI_{UR}}{dt} &= \left[\beta \left(\frac{S(t)}{N}\right) \left(\frac{I_R(t)}{\theta}\right) - \gamma \left(\frac{I_R(t)}{\theta}\right)\right] (1-\theta) \\
\frac{dR}{dt} &= \gamma \left(\frac{I_R(t)}{\theta}\right)
\end{aligned}$$

We will term this model the $SI_RI_{UR}R$ model of COVID-19 transmission. This model makes the assumption that unreported cases are equally as infectious as reported cases. This is likely not true as unreported cases are more likely to include asymptomatic individual, who present a lower relative risk for transmission (22). And although we could have parametrized this model to also delineate symptomatic and asymptomatic transmission, this analysis would have been

muddled by the fact that both reported and unreported cases include both symptomatic and asymptomatic individuals (in unknown proportions). Therefore, such an analysis would significantly increase the number of parameters, most of which would likely be underdetermined by the data. Note that we could have simplified the expression for $\frac{dI_R}{dt}$ but did not so as to prime our model to be structured by age in the following subsection.

3.1.2 Age-Structure

We will further structure the $SI_RI_{UR}R$ model into n different age groups, so as to allow us to delineate intra- and inter-group transmission dynamics for COVID-19.

For ease of analysis with multiple groups, we will introduce the idea of force of infection. The force of infection that the susceptibles in the j-th group of a population experience from the infectives in the i-th group is defined to be $\mathcal{F}_{i,j}(t) = \beta_{i,j} I_i(t)$. $I_i(t)$ describes the number of infectives in group i and $\beta_{i,j} = c_{i,j} \cdot p$ is the transmission parameter between these infectives and individuals in group j. We see that the total number of new infections that infectives in group i produce in group j per unit time is $\mathcal{F}_{i,j}(t)(S_j/N_j) = \beta_{i,j}(S_j/N_j)I_i(t)$. This is consistent with the previous whole population analysis (for $i = j = n = 1$), where the number of new infections per unit time is $\beta(S(t)/N)I(t)$.

The total number of new infections the susceptibles in group j experience is the sum of those produced by the forces of infection ($\mathcal{F}_{i,j}(t)$) applied by each group i of the population. This is computed as:

$$\sum_i^n \left\{ \mathcal{F}_{i,j}(t) \left(\frac{S_j(t)}{N_j} \right) \right\} = \left(\frac{S_j(t)}{N_j} \right) \sum_i^n \mathcal{F}_{i,j}(t)$$
$$= \left(\frac{S_j(t)}{N_j} \right) \sum_i^n \beta_{i,j} I_i(t).$$

Therefore, the SIR model structured by age (shown for the j-th group only) is as follows:

$$\frac{dS_j}{dt} = -\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \left(\beta_{i,j} I_i(t)\right)$$
$$\frac{dI_j}{dt} = \left(\frac{S_j(t)}{N_j}\right)\sum_i^n \left(\beta_{i,j} I_i(t)\right) - \gamma_j I_j(t)$$
$$\frac{dR_j}{dt} = \gamma_j I_j(t)$$

We seek to structure the $SI_RI_{UR}R$ model by age; however, this derivation is tedious. A simpler derivation (which yields the same result) is to incorporate infection ascertainment into the above age-structured SIR model (using the methods described in subsection 3.1.1.). Namely, we specify the I_j compartment into reported (I_{jR}) and unreported (I_{jUR}) cases and replace $I_i(t)$ with $I_{iR}(t)/\theta_i$ for each group i and j in the population. These methods yield the generalized n-age group $SI_RI_{UR}R$ model (Fig. 2):

$$\frac{dS_j}{dt} = -\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i}\right)$$
$$\frac{dI_{jR}}{dt} = \left[\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i}\right) - \gamma_j\left(\frac{I_{jR}(t)}{\theta_j}\right)\right]\theta_j$$
$$\frac{dI_{jUR}}{dt} = \left[\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i}\right) - \gamma_j\left(\frac{I_{jR}(t)}{\theta_j}\right)\right](1-\theta_j)$$
$$\frac{dR_j}{dt} = \gamma_j\left(\frac{I_{jR}(t)}{\theta_j}\right)$$

Choosing the number of age groups (n) with which to analyze the epidemic is a non-trivial problem. Analyses with greater numbers of age groups are more granular and thus able to acutely identify age-specific trends. On the other hand, such analyses necessarily decrease the size of data available to constrain our model in each age group, which contributes to noise. Furthermore, since this model includes n^2 β-parameters and n θ-parameters, computationally constraining these parameters grows in run-time proportional to n^2.

We will begin our analysis with $n = 2$ age groups (and later conduct sensitivity analysis with $n = 3$ age groups). The solution to this system of non-linear ODEs (for some initial value describing the epidemic at $t = t_0$) describes the theoretical epidemic simulated under the parameter set $\{\beta_{i,j},\ \theta_j \mid i, j = 1\ or\ 2\}$. We will attempt to constrain these parameters by fitting this model to the reported COVID-19 case data in order to gain insight into the dynamics of B.C.'s epidemic.

3.1.2.1 Estimating R_0

Structuring our model by age complicates the estimation of the basic reproduction number (R_0). As with for the simple *SIR* model, the epidemic will grow when $\frac{dI}{dt} > 0$. For the generalized n-age group $SI_RI_{UR}R$ model, this occurs if and only if the sum of the epidemics in each age group is positive:

$$\sum_j^n \left\{\frac{dI_j}{dt}\right\} = \sum_j^n \left\{ \left(\frac{S_j(t)}{N_j}\right) \sum_i^n \left(\beta_{i,j} I_i(t)\right) - \gamma_j I_j(t) \right\} > 0$$

Since at the outset of an epidemic $S_j(t) \approx N_j$ for each age group, this condition is satisfied if and only if

$$\sum_j^n \sum_i^n \left(\beta_{i,j} I_i(t)\right) - \sum_j^n \left(\gamma_j I_j(t)\right) > 0$$

With some algebraic manipulation, we have that

$$\sum_i^n \sum_j^n \left(\beta_{i,j} I_i(t)\right) - \sum_i^n \left(\gamma_i I_i(t)\right) > 0$$

$$\sum_i^n \left\{ I_i(t) \left(\sum_j^n \beta_{i,j} - \gamma_i \right) \right\} > 0$$

In order to analytically solve this inequality, we must assume that at the outset of the epidemic each age group has a comparable (non-zero) number of infectives $I_i(t) = I_0$. It follows that

$$I_0 \left\{ \sum_i^n \sum_j^n \beta_{i,j} - \sum_i^n \gamma_i \right\} > 0$$

Thus, we define our estimate of R_0 as

$$R_0 := \frac{\sum_{i,j}^n \beta_{i,j}}{\sum_i^n \gamma_i}$$

This satisfies the condition that $R_0 > 1$ when the epidemic is growing ($\frac{dI}{dt} > 0$) and that $R_0 < 1$ when the epidemic is going to extinction ($\frac{dI}{dt} < 0$). For the two-age group $SI_RI_{UR}R$ model, the basic reproduction number can be estimated as

$$R_0 = \frac{\beta_{1,1} + \beta_{1,2} + \beta_{2,1} + \beta_{2,2}}{\gamma_1 + \gamma_2}$$

3.1.3 Hospitalization

Due to its relevance to public health, we will incorporate hospitalization after infection into our generalized n-age group $SI_RI_{UR}R$ model. If an infective experiences disease of sufficient severity to require hospitalization, they will most certainly have been tested after admission (if they had not already tested positive before). Therefore, we will assume that all hospitalized cases are derived from ascertained infections. The fraction of reported infectives (I_{jR}) in each group j that develop into hospitalizations is termed the case-hospitalization fraction (ϕ_{Cj}). For simplicity, we will assume that this fraction (ϕ_{Cj}) is equally infectious as the fraction of infectives who are not hospitalized ($1 - \phi_{Cj}$) and will therefore remain in the I_{jR} compartment for an equivalent duration. After this duration, whereas other infectives will transition to the R_j compartment,

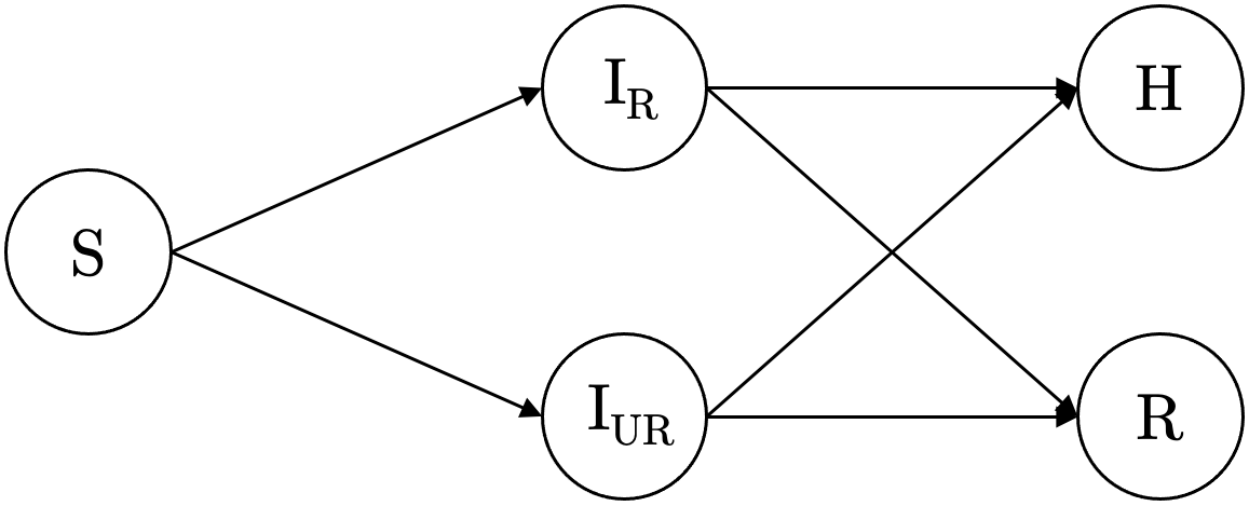

Figure 3: Schematic of the n-age group $SI_RI_{UR}HR$ model of COVID-19 transmission, shown for the j-th group only. The S, I_R, I_{UR}, H, and R compartments represent the susceptibles, reported infectives, unreported infectives, cumulative hospitalizations, and recovered individuals in the population, respectively.

hospitalized cases will transition to the H_j compartment. We will define the H_j compartment as the cumulative number of hospitalized cases for each group j – not the number of patients currently in hospital. Although the alternate definition may be useful for some analyses, the current report will make no attempt to fit to this metric. This is because, unlike the infectious period, the period of hospitalization varies considerably from patient to patient (23) – contributing to noise while fitting. This model is termed the generalized n-age group $SI_RI_{UR}HR$ model (Fig. 3) and described by the following dynamical system:

$$\frac{dS_j}{dt} = -\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i}\right)$$
$$\frac{dI_{jR}}{dt} = \left[\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i}\right) - \gamma_j\left(\frac{I_{jR}(t)}{\theta_j}\right)\right]\theta_j$$
$$\frac{dI_{jUR}}{dt} = \left[\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i}\right) - \gamma_j\left(\frac{I_{jR}(t)}{\theta_j}\right)\right](1-\theta_j)$$
$$\frac{dH_j}{dt} = \phi_{Cj}\left(\gamma_j I_{jR}(t)\right)$$

$$\frac{dR_j}{dt} = \gamma_j \left(\frac{I_{jR}(t)}{\theta_j}\right) - \phi_{Cj}\left(\gamma_j I_{jR}(t)\right)$$

3.1.4 Vaccination

We will further develop our generalized n-age group $SI_RI_{UR}HR$ model to incorporate the effect of immunization. In order to simplify our analysis, we will assume that an individual will experience the full utility of a vaccine immediately after administration of a single dose. To a first approximation, this assumption will only deviate from reality, in which immunity builds progressively, by a shift in the timescale – which is of little concern to the current analysis. We will assume that that susceptibles (S_j) in group j are vaccinated and transition into the VS_j compartment at a rate of $\mathcal{V}_j(t)$ individuals per unit time. Further, we will assume that vaccinated individuals (VS_j) are still susceptible to infection, though protected to an extent of δ_j. If infected, vaccinated individuals will transition into the VI_j compartment, which includes both reported (VI_{jR}) and unreported (VI_{jUR}) infections. We will assume that these vaccinated infectives (VI_j) carry lower viral loads and thus that transmission is attenuated by a factor of σ_j. Finally, we will assume that vaccinated infectives are protected from severe disease (i.e., hospitalization) to an extent of τ_j. We term this model the generalized n-age group $V{\sim}SI_RI_{UR}HR$ model (Fig. 4), which is described by the following dynamical system:

$$\begin{aligned}
\frac{dS_j}{dt} &= -\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i} + (1-\sigma_i)\frac{VI_{iR}(t)}{\theta_i}\right) - \mathcal{V}_j(t) \\
\frac{dVS_j}{dt} &= \mathcal{V}_j(t) - (1-\delta_j)\left(\frac{VS_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i} + (1-\sigma_i)\frac{VI_{iR}(t)}{\theta_i}\right) \\
\frac{dI_{jR}}{dt} &= \left[\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i} + (1-\sigma_i)\frac{VI_{iR}(t)}{\theta_i}\right)\right. \\
&\qquad \left. - \gamma_j\left(\frac{I_{jR}(t)}{\theta_j}\right)\right]\theta_j
\end{aligned}$$

$$\frac{dI_{jUR}}{dt} = \left[\left(\frac{S_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i} + (1-\sigma_i)\frac{VI_{iR}(t)}{\theta_i}\right) - \gamma_j\left(\frac{I_{jR}(t)}{\theta_j}\right)\right](1-\theta_j)$$

$$\frac{dVI_{jR}}{dt} = \left[(1-\delta_j)\left(\frac{VS_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i} + (1-\sigma_i)\frac{VI_{iR}(t)}{\theta_i}\right) - \gamma_j\left(\frac{VI_{jR}(t)}{\theta_j}\right)\right]\theta_j$$

$$\frac{dVI_{jUR}}{dt} = \left[(1-\delta_j)\left(\frac{VS_j(t)}{N_j}\right)\sum_i^n \beta_{i,j}\left(\frac{I_{iR}(t)}{\theta_i} + (1-\sigma_i)\frac{VI_{iR}(t)}{\theta_i}\right) - \gamma_j\left(\frac{VI_{jR}(t)}{\theta_j}\right)\right](1-\theta_j)$$

$$\frac{dH_j}{dt} = \phi_{Cj}\left(\gamma_j I_{jR}(t) + (1-\tau_j)\gamma_j VI_{jR}(t)\right)$$

$$\frac{dR_j}{dt} = \gamma_j\left(\frac{I_{jR}(t)}{\theta_j} + \frac{VI_{jR}(t)}{\theta_j}\right) - \phi_{Cj}\left(\gamma_j I_{jR}(t) + (1-\tau_j)\gamma_j VI_{jR}(t)\right)$$

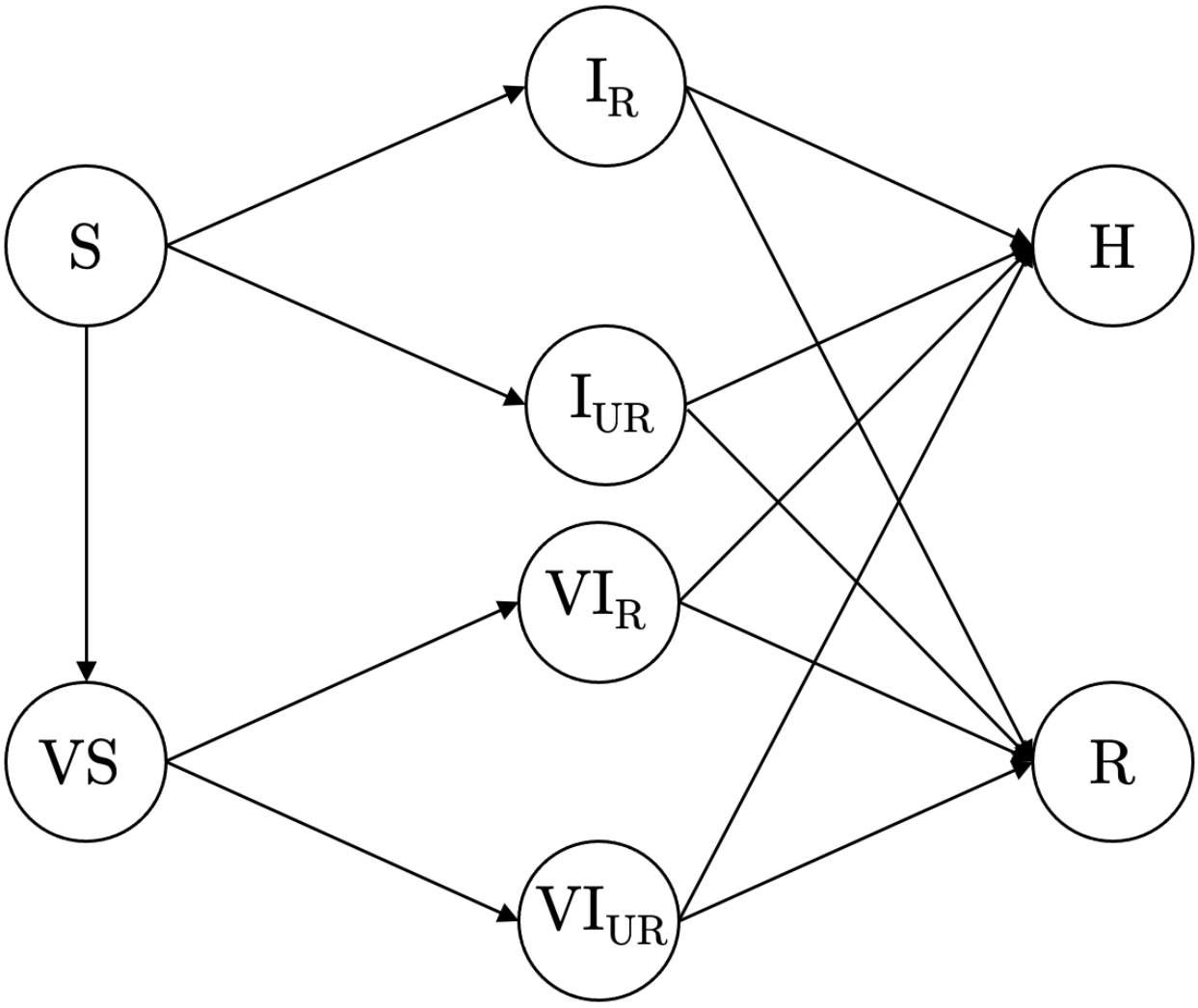

Figure 4: Schematic of the n-age group $V{\sim}SI_RI_{UR}HR$ model of COVID-19 transmission, shown for the j-th group only. The S, I_R, I_{UR}, H, and R compartments represent the susceptibles, reported infectives, unreported infectives, cumulative hospitalizations, and recovered individuals in the population, respectively. The VS, VI_R, and VI_{UR} compartments represent the vaccinated versions thereof.

3.2 Model Fitting

Appendix A includes an extensive, technical analysis on the role of age in reported COVID-19 infection and hospitalization dynamics in B.C. This analysis employed the k-means clustering algorithm to justify an age-cutoff of 60 with which to analyze the two-age group $SI_RI_{UR}R$ model. Therefore, we will begin our analysis with this cutoff, such that Group 1 (74.8% of the population) represents ages 0-59 and Group 2 (25.2%) represents ages 60+. Since the γ_i parameters are fixed as the inverse of the mean infectious period (5.0 days for all age groups) (19), this model has a six-dimensional parameter space:

$$\beta = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} \end{bmatrix}, \qquad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

We will attempt to constrain the transmission and ascertainment fraction parameters by fitting our model to the reported case count data for B.C. throughout its epidemic. In particular, we will fit the cumulative reported cases predicted by the model for each age group to those observed from the data by minimizing the sum of squared residuals (see Methods) – either normalized (NSSR) or non-normalized (SSR).

Optimizing our model by minimizing the SSR weights each group by the magnitude of observed cumulative cases, whereas minimizing the NSSR weights each group equally. Occasionally, the first (but never the latter) method may entirely fail to fit age groups with relatively low case counts. However, weighting by cumulative case counts (as with SSR) guarantees that our simulation will provide the closest fit to the whole population. Therefore, unless the fit fails for any age group, the current analysis will employ the SSR method by default. Further, when we wish to compare the fit of our model at different points in the epidemic, we will

employ the NSSR method (which will allow us to control for differences in the cumulative case count).

3.2.1 Transmission (β)

In order to analyze the transmission parameters, we first assumed an ascertainment fraction (θ_i) of 100% for both age groups. This allowed us to study reported infection dynamics as though they represent true infection dynamics. To begin our analysis, we fit our model (under various constraints of interest) to the months of August-November 2020, during which time the epidemic saw consistent growth (Fig. A.1).

We first analyzed our model under the hypothesis of no mixing between age groups. Thus, an infective in group i will only make contact with other individuals in group i, with transmission parameters B_i, such that:

$$\beta = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}, \qquad B_i \geq 0$$

Optimization under this constraint yielded strong fit ($SSR = 7.9117e07$) to the epidemic curves of each age group (Fig. 5A). This hypothesis is tempting as it seems reasonable that the older population would wish to isolate from the younger, potentially riskier, population. However, the optimal parameters for this constraint $[B_1, B_2]_{opt} = [0.2225, 0.2304]$ suggest that the older age group has more contacts than the younger age group – which is contrary to observed behaviour (17). The model yields this result because Group 2 ($I_2 = 16$) has far fewer active infectives at the beginning of the simulation (August 1, 2020) than Group 1 ($I_1 = 175$) and must therefore have a higher contact rate to produce the growth observed in Group 2 (Fig. 5A). This suggests that in order to explain the growth in the older population, without an unreasonably high contact rate, we

must have some degree of intergroup mixing such that the infectives in the younger group can contribute to the epidemic in the older group.

Next, we explored the hypothesis of complete (homogenous) mixing between groups, such that neither group makes any attempt to isolate from each other. Thus, the fraction of its total contact that an infective in group i will make with individuals in group j is the proportion of the total population that group j constitutes: $N_j / \sum_k N_k$. This constraint is summarized as:

$$\beta = \begin{bmatrix} B_1 \left(\frac{N_1}{N_1 + N_2} \right) & B_1 \left(\frac{N_2}{N_1 + N_2} \right) \\ B_2 \left(\frac{N_1}{N_1 + N_2} \right) & B_2 \left(\frac{N_2}{N_1 + N_2} \right) \end{bmatrix}, \qquad B_i \geq 0$$

This optimization resulted in worse fit (Fig. 5B; $SSR = 3.3015e08$) with optimal transmission parameters $[B_1, B_2]_{opt} = [0.1645, 0.4210]$. The aforementioned issue of overestimating contact behaviour for the older population relative to the younger population is even further exacerbated under this constraint. Due to a low number of active infectives, the older age group must make $B_2/B_1 = 2.6$ times as many contacts as the younger group in order to contribute proportionally to the larger epidemic in the younger population (as well as to its own smaller epidemic). However, this higher contact rate for the older population now overestimates the case growth in Group 2 (Fig. 5B), resulting in the worsened fit.

The condition of no mixing between age groups necessitates that Group 2 has high contact behaviour so as to contribute to its own epidemic. The condition of homogenous mixing between age groups, on the other hand, necessitates the same so as to contribute to the epidemic observed in Group 1. In either case, this prediction is impossible and neither hypothesis is able to reasonably explain the observed epidemic. Therefore, we must study the hypothesis of heterogenous mixing between age groups. In order to do so, we will introduce the idea of contact preference, $\mathcal{P}$ – where

the preference of an individual of group i to make contact with an individual in group j (relative to its own group) is defined to be $\mathcal{P}_{i,j}$. By definition, we have that $\mathcal{P}_{i,i} = 1$. We add that $0 \leq \mathcal{P}_{i,j} \leq 1$ for $i \neq j$ since it is unlikely for intergroup contacts to be preferred over intragroup contacts. Thus, the proportion of total contacts that an infective of group i makes with individuals of group j is the proportion of the total population that group j constitutes when weighted by contact preference: $\mathcal{P}_{i,j}N_j / \sum_j \mathcal{P}_{i,j}N_j$. This constraint is summarized as:

$$\beta = \begin{bmatrix} B_1\left(\frac{\mathcal{P}_{1,1}N_1}{\mathcal{P}_{1,1}N_1 + \mathcal{P}_{1,2}N_2}\right) & B_1\left(\frac{\mathcal{P}_{1,2}N_2}{\mathcal{P}_{1,1}N_1 + \mathcal{P}_{1,2}N_2}\right) \\ B_2\left(\frac{\mathcal{P}_{2,1}N_1}{\mathcal{P}_{2,1}N_1 + \mathcal{P}_{2,2}N_2}\right) & B_2\left(\frac{\mathcal{P}_{2,2}N_2}{\mathcal{P}_{2,1}N_1 + \mathcal{P}_{2,2}N_2}\right) \end{bmatrix}$$

$$= \begin{bmatrix} B_1\left(\frac{N_1}{N_1 + \mathcal{P}_{1,2}N_2}\right) & B_1\left(\frac{\mathcal{P}_{1,2}N_2}{N_1 + \mathcal{P}_{1,2}N_2}\right) \\ B_2\left(\frac{\mathcal{P}_{2,1}N_1}{\mathcal{P}_{2,1}N_1 + N_2}\right) & B_2\left(\frac{N_2}{\mathcal{P}_{2,1}N_1 + N_2}\right) \end{bmatrix},$$

$$B_i \geq 0, \qquad 0 \leq \mathcal{P}_{i,j} \leq 1$$

This optimization, unlike the previous, was ultimately sensitive to our initial guess and thus could not produce a stable global optimum. Instead, the model identified several different local optima. For example, the initial guess $[B_1, B_2, \mathcal{P}_{1,2}, \mathcal{P}_{2,1}]_0 = [0.2, 0.2, 0.1, 0.1]$ produced the optimal parameter set $[B_1, B_2, \mathcal{P}_{1,2}, \mathcal{P}_{2,1}]_{opt} = [0.1996, 0.3852, 0.0000, 0.2262]$ ($SSR = 3.17473e07$), whereas $[B_1, B_2, \mathcal{P}_{1,2}, \mathcal{P}_{2,1}]_0 = [0.5, 0.5, 0.1, 0.1]$ produced $[B_1, B_2, \mathcal{P}_{1,2}, \mathcal{P}_{2,1}]_{opt} = [0.1616, 0.5711, 0.3091, 1.0000]$ ($SSR = 5.0893e07$). These two optima represent significantly different transmission dynamics. Despite this, they are both able to simulate the epidemic (Fig. 5C,5D) with comparable efficacy (as measured by their SSR). The fact that two entirely different parameter inputs can produce the same simulated epidemic suggests that there is not a 1:1 mapping

of elements in the four-dimensional parameter space to elements in the model's output space. Therefore, we must conclude that our model, under the present heterogenous mixing constraint, is insufficiently determined by the reported case count data. In other words, without further constraint, our two-age group $SI_RI_{UR}R$ model is not able to clearly identify group transmission dynamics from reported case count data.

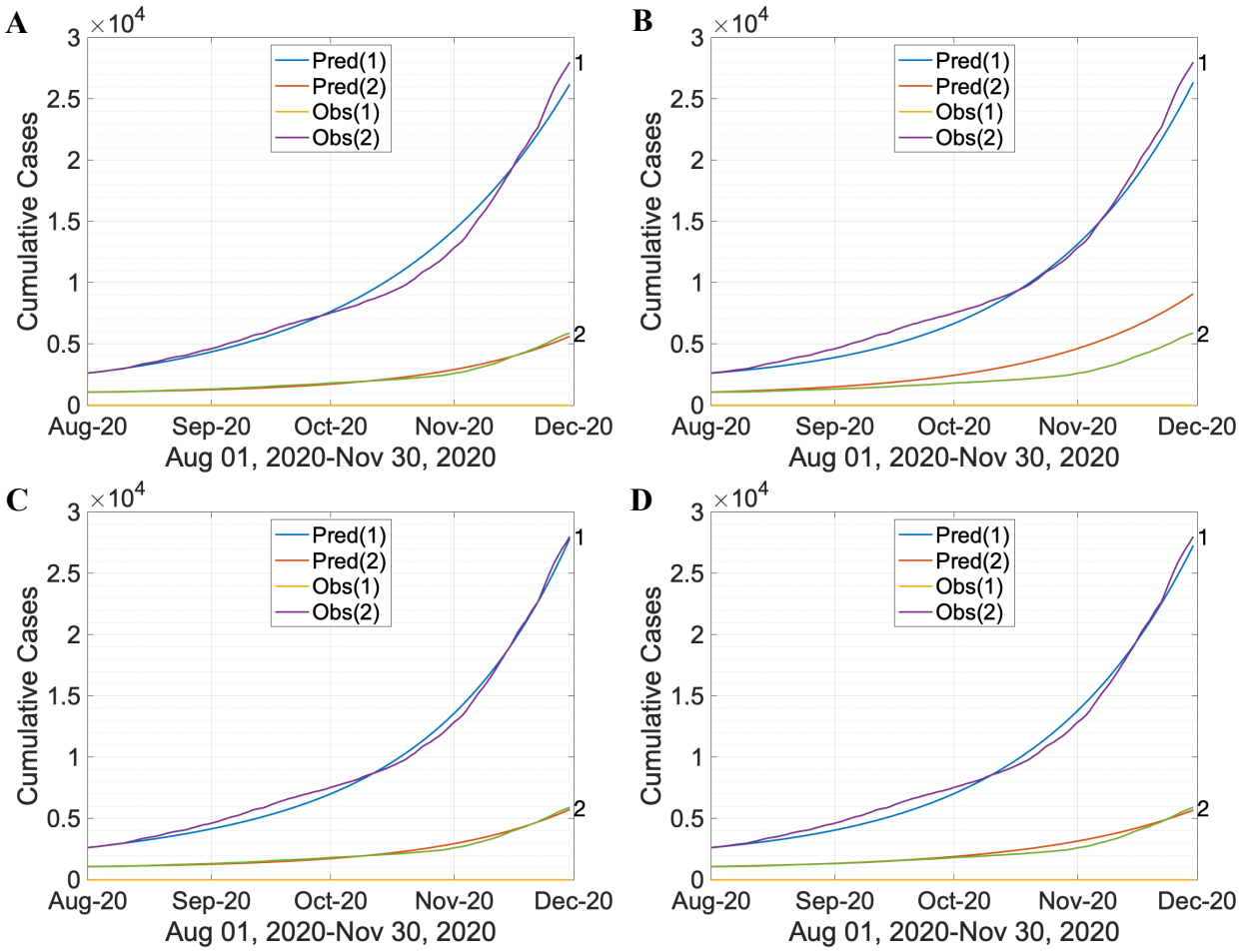

Figure 5: Fitting the two-age group $SI_RI_{UR}R$ model (with age-cutoff 60) to cumulative case data from August to November 2020 by optimizing the transmission parameter under various constraints. Panel **A** represents the constraint of no mixing between groups. Panel **B** represents the constraints of homogenous mixing between groups. Panels **C** & **D** represent the constraint of heterogenous mixing, optimized under different initial guesses. Optimization conducted by minimizing the SSR.

3.2.1.1 Parameter Constraint: Contact Behaviour

The primary objective of fitting an epidemiological model, or any statistical model, for that matter, to data is to make inferences on parameters which would otherwise be difficult to measure. However, an optimal fit of a model's output to data is not always sufficient to make inferences on its parameter input. Namely, if multiple combinations of the parameter set can produce the same (or comparable) output, we cannot be certain that the combination which produces the optimal fit to data is actually the most relevant. More generally, we say that our model is not identifiable if it fails to produce a 1:1 (injective) mapping from its parameter space to its output space (24). Models of even modest complexity may have too many degrees of freedom in their parameter space to achieve such a mapping. A common solution to this problem is to use external data, or reasonable assumptions, to further constrain the parameter space (and reduce its dimensionality).

The recent study by Brankston et al. (20) conducted electronic surveys on age-specific contact patterns of Canadians during the COVID-19 pandemic. The present report will use the contact data of individuals residing in B.C. at four different time-points (May, July, Sept, Dec 2020) to construct contact matrices, where $c_{i,j}$ describes the average number of daily contacts individuals in group i make with individuals in group j. The transmission parameter matrix is constructed from these contact matrices as $\beta_{i,j} = c_{i,j} \cdot p$. Therefore, in an effort to make our two-age group $SI_RI_{UR}R$ model identifiable under the constraint of heterogenous mixing, we will use these contact matrices to constrain our parameter space from four dimensions $[B_1, B_2, \mathcal{P}_{1,2}, \mathcal{P}_{2,1}]$ to just one $[p]$. The contact matrices for each time-point $t \in \{May, July, Sept, Dec\}$, $\beta(t) = p \begin{bmatrix} c_{1,1}(t) & c_{1,2}(t) \\ c_{2,1}(t) & c_{2,2}(t) \end{bmatrix}$, $0 \leq p \leq 1$, are summarized in Table B.1.1.

These contact matrices also allow us to directly calculate intergroup contact preferences, $\mathcal{P}_{i,j}$ for $j \neq i$, and make inferences on the extent to which the surveyed contact behavior is heterogenous. If $\mathcal{P}_{i,j} = 0$, we would infer that there is no mixing between groups; whereas, if $\mathcal{P}_{i,j} = 1$, we would infer that mixing between groups is entirely homogenous. If $0 \leq \mathcal{P}_{i,j} \leq 1$, we have that mixing between groups is heterogenous. The proportion of contacts c_i that group i makes with group j is $\Gamma_{\mathrm{i,j}} = c_{i,j} / \sum_k c_{i,k}$, which, under the heterogenous mixing constraint, is the same as $\mathcal{P}_{i,j} N_j / \sum_k \mathcal{P}_{i,k} N_k$. By computing $\Gamma_{\mathrm{i,i}}$, we have that

$$\Gamma_{\mathrm{i,i}} = \frac{c_{i,i}}{c_{i,i} + c_{i,j}} = \frac{N_i}{N_i + \mathcal{P}_{i,j} N_j}$$

$$\implies \mathcal{P}_{i,j} = \left(\frac{N_i}{N_j}\right) \left[\frac{1}{\Gamma_{\mathrm{i,i}}} - 1\right]$$

Table 1 summarizes the intergroup contact preferences calculated for each survey matrix and shows that the mixing behaviour was heterogenous throughout the course of the epidemic in 2020. During May and July of 2020, contact behaviour was low for both the younger and older populations (Fig. 6). At this point, both groups considerably restricted contact with each other ($\mathcal{P}_{1,2} \sim 50\%, \mathcal{P}_{2,1} \sim 40\%$). However, by September, contact behaviour in the younger population increased (while that in the older population stayed low). This increase in contact behaviour coincided with a period of case growth for the whole population in the fall (Fig. A.1). The younger population's preference for making contact with the older population at this time dropped ($\mathcal{P}_{1,2} \sim 30\%$), suggesting that most of these new contacts were with other young people. This is consistent with the nature of many workplaces and social settings. The older population, on the other hand, increased its preference for making contact with the younger population ($\mathcal{P}_{2,1} \sim 60\%$),

Survey Month	Intergroup Contact Preference	
	$\mathcal{P}_{1,2}(\%)$	$\mathcal{P}_{2,1}(\%)$
May	52.5	43.5
July	53.0	37.6
Sept	33.7	64.0
Dec	64.4	71.6

Table 1: Intergroup contact preferences for two-age group model (with age-cutoff 60) for each survey month. Calculated directly from contact matrices adapted from contact survey data surveyed from B.C. residents in 2020 (20).

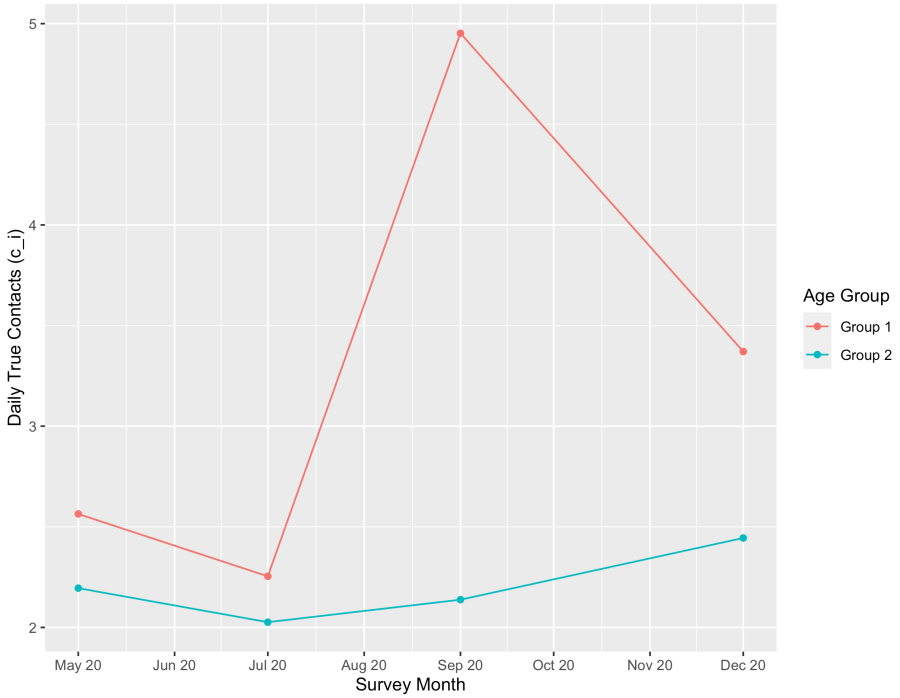

Figure 6: Average total daily contacts for each age group (Group 1, ages 0-59; Group 2, ages 60+). Adapted from contact survey data adapted from survey data in (20).

despite not increasing total contact behaviour (Fig. 6). This perhaps reflects fatigue from months of having to restrict contact with the younger (potentially riskier) population. By December, contact behaviour in the younger population dropped from its peak in September (Fig. 6). This is consistent with a province-wide prohibition on social gatherings enacted on November 7 (8). This restriction in contacts coincides with a decline in the epidemic observed since mid-November for the whole population (Fig. A.1). The decrease in social contact for the younger population increased its relative preference for contact with the older population ($\mathcal{P}_{1,2} \sim 60\%$).

We explored the extent to which the surveyed contact behaviour $\beta(t)$ (20) was able to explain the epidemic in B.C., from the month preceding to the month following each survey $t \in \{May, July, Sept, Dec\}$. In order to compare the fit between months, and between matrices, we optimized our model by minimizing the $NSSR$ (rather than SSR). With the exception of the $t = May$ matrix applied to the April data (during which time testing methods in B.C. were inconsistent (13)) each contact matrix provided excellent fit to the reported case data (Fig. 7). Therefore, these contact data provide insight into the heterogenous mixing patterns required to explain age group-transmission dynamics in the epidemic in B.C. from May 2020 to January 2021. Due to overlap, both the June and August data were fit with two survey matrices. For the June data, both the May and July matrices were able to explain the data, though the July matrix had slightly improved fit (Fig. 7A,7B). For the August data, the Sept matrix had superior fit to the July Matrix (Fig. 7B,7C). This suggests that the contact behaviour in August more closely matched the high-contact behaviour observed in September (than that in July).

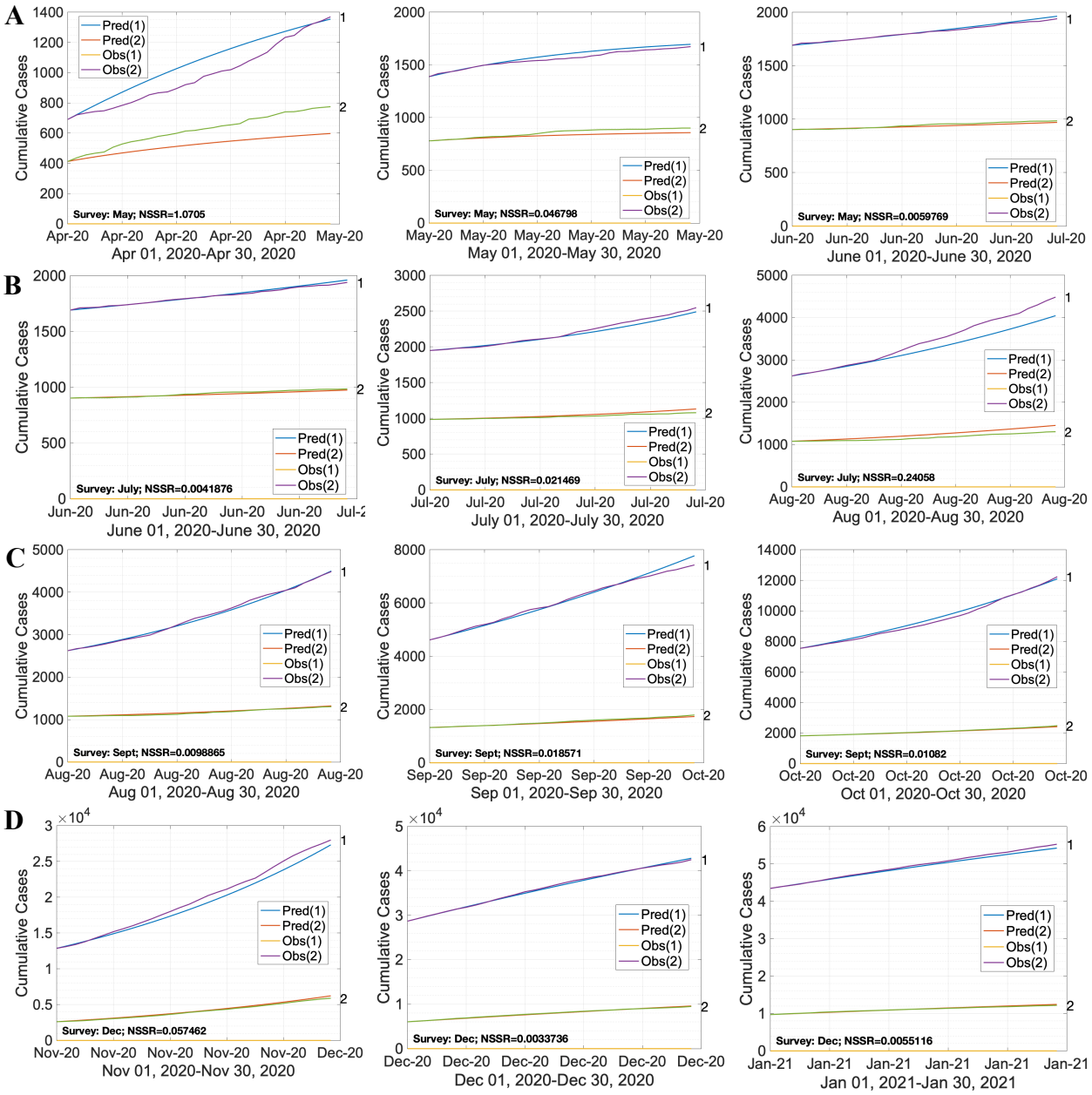

Figure 7: Fitting the two-age group $SI_RI_{UR}R$ model (with age-cutoff of 60) to the reported case data using the contact matrices adapted from survey data in (20). Each survey (**A**, May; **B**, July; **C**, Sept; **D**, Dec) was fit at a 30-day scale from the month preceding the survey to the month following. Optimization conducted by minimizing the $NSSR$.

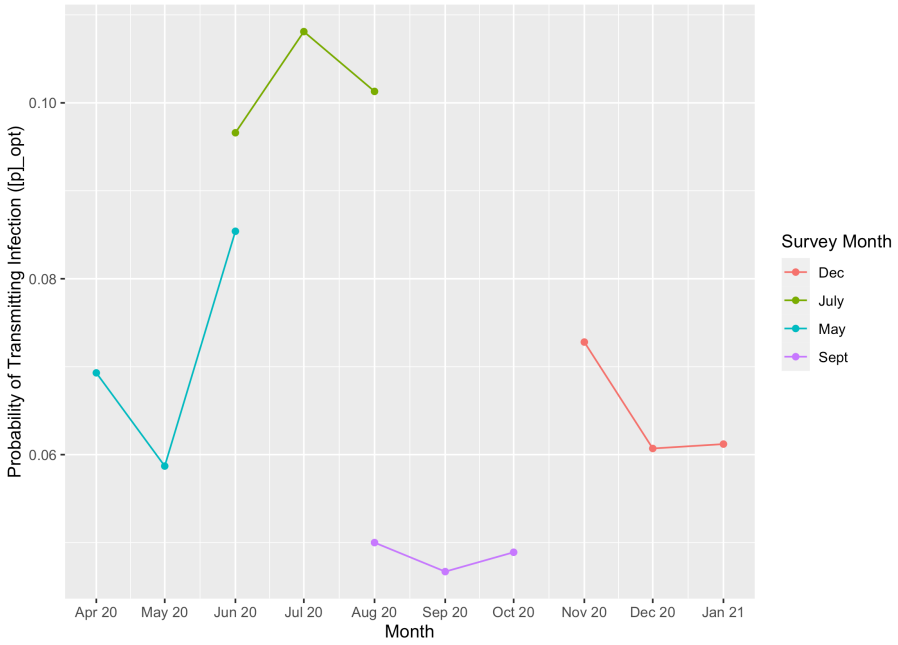

Figure 8: The optimal probability of infection yielded from optimizing the two-age group $SI_RI_{UR}R$ model (with age-cutoff of 60) to the reported case data for each contact matrix $t \in \{May, July, Sept, Dec\}$ adapted from survey data in (20). Optimization conducted by minimizing the $NSSR$.

Our optimization for each month should yield the probability of transmitting infection given contact between an infective and susceptible ($[p]_{opt}$). For consistent contact behaviour, this parameter should be a constant. In the current analysis, this parameter was approximately constant for each contact matrix, but varied between contact matrices (Fig. 8). Notably, the probability of transmitting infection was higher for the July contact matrix compared to the Sept matrix, despite having fewer absolute contacts. This may represent the fact that contacts during the summer months were risker (i.e., less masking) than in the fall.

3.2.1.2 Sensitivity Analysis: Age-structure

The model's sensitivity to the age-cutoff and the number of age-groups chosen was explored in depth in Appendix B. A consequential result is that the model's identifiability (or the $[p]_{opt}$ yielded from optimization) is stable with respect to age-structure (Fig. B.2.4).

3.2.1.3 Challenging Identifiability

The contact matrices adapted from (20) are sufficient to analyze intergroup contact preferences for the B.C. population (Table 1). However, in order to challenge the identifiability of our model, we will loosen the constraint applied on our parameter space by the contact matrices and attempt to identify $\mathcal{P}_{i,j}$. We can use the ratio of total contacts between the two age groups ($r = B_1/B_2 = c_1/c_2$), to constrain our parameter space such that:

$$\beta = \begin{bmatrix} B_1\left(\frac{N_1}{N_1 + \mathcal{P}_{1,2}N_2}\right) & B_1\left(\frac{\mathcal{P}_{1,2}N_2}{N_1 + \mathcal{P}_{1,2}N_2}\right) \\ B_2\left(\frac{\mathcal{P}_{2,1}N_1}{\mathcal{P}_{2,1}N_1 + N_2}\right) & B_2\left(\frac{N_2}{\mathcal{P}_{2,1}N_1 + N_2}\right) \end{bmatrix}$$

$$= \begin{bmatrix} B_2 r\left(\frac{N_1}{N_1 + \mathcal{P}_{1,2}N_2}\right) & B_2 r\left(\frac{\mathcal{P}_{1,2}N_2}{N_1 + \mathcal{P}_{1,2}N_2}\right) \\ B_2\left(\frac{\mathcal{P}_{2,1}N_1}{\mathcal{P}_{2,1}N_1 + N_2}\right) & B_2\left(\frac{N_2}{\mathcal{P}_{2,1}N_1 + N_2}\right) \end{bmatrix}$$

This results in a three-dimensional parameter space $[B_2, \mathcal{P}_{1,2}, \mathcal{P}_{2,1}]$ since r is calculated directly from the contact matrices. As with before, this constraint was insufficient to allow for identifiability (data not shown). Although, each of the survey months showed good fit to the reported data, the optimal intergroup contact preferences $\left[\mathcal{P}_{1,2}, \mathcal{P}_{2,1}\right]_{opt}$ were sensitive to our initial guess and not comparable to those calculated directly from the matrices (Table 1). In a final attempt to constrain intergroup dynamics from our two-age group $SI_RI_{UR}R$ model, we

constrained our parameter space further by fixing either of the intergroup contact preferences and attempting to identify the other. Yet again, this constraint was unable to reproduce the intergroup contact preferences in Table 1 (data not shown).

Since even partial knowledge on intergroup dynamics is insufficient to completely identify intergroup transmission dynamics from reported case data, we must conclude that our two-age group $SI_RI_{UR}R$ model is not identifiable under the heterogenous mixing constraint. Therefore, we will need to rely on age-specific contact patterns published in (20) to constrain our model for any further analysis.

3.2.1.4 R_0: Simulating the Epidemic

In subsection 3.1.2.1, we derived a formula to estimate R_0 for our two-age group $SI_RI_{UR}R$ model under the assumption of approximately equal infective counts in each age group:

$$R_0 = \frac{\beta_{1,1} + \beta_{1,2} + \beta_{2,1} + \beta_{2,2}}{\gamma_1 + \gamma_2}$$

In order to validate this estimate, we will test that the epidemic grows when and only when $R_0 > 1$. We fix $I_1 = I_2 = 10$ at the beginning of a hypothetical epidemic starting on March 1, 2020 and simulate the epidemic while varying our transmission parameters. Since $\gamma_i = 1/5$ (19), $R_0 > 1$ when and only when $\beta_{1,1} + \beta_{1,2} + \beta_{2,1} + \beta_{2,2} > 0.40$. When we fix $R_0 = 1.01$, the simulated epidemic grows (Fig. 9A). Conversely, when we fix $R_0 = 0.99$, the simulated epidemic goes to extinction (Fig. 9B). Indeed, this verifies that our estimate for R_0 is valid (under the assumption of equal numbers of active infectives in each age group).

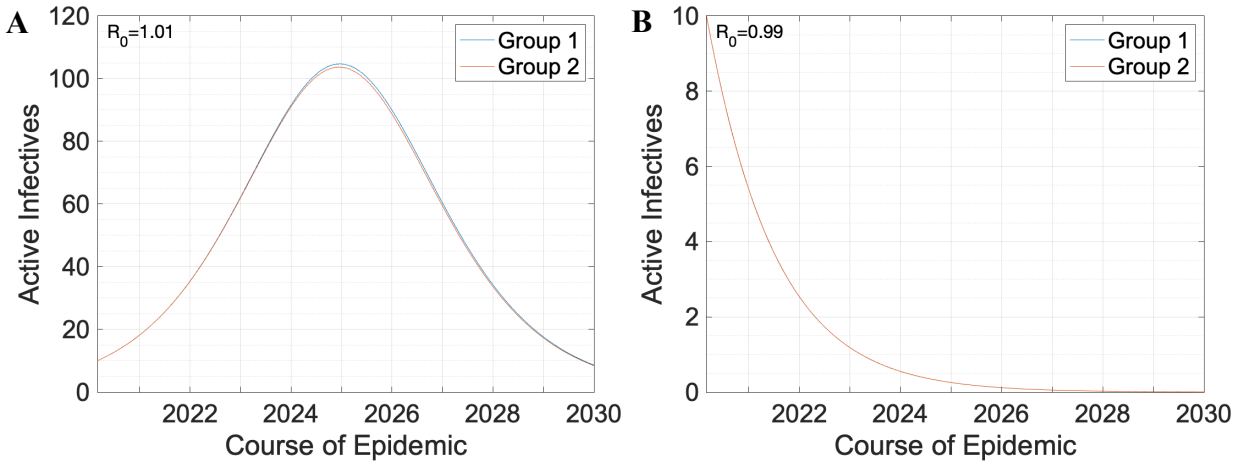

Figure 9: Simulating epidemics when fixing (A) $R_0 = 1.01$ and (B) $R_0 = 0.99$ using the estimate derived in subsection 3.1.2.1. Each group began with 10 active infectives.

3.2.2 Infection Ascertainment (θ)

Uncertainty around infection ascertainment (θ) remains a significant hurdle in understanding the epidemiology of COVID-19. In section 3.1, we parametrized our age-structured model with infection ascertainment (θ_i). In the current section, we will attempt to identify this parameter by fitting the two-age group $SI_RI_{UR}R$ model (with an age-cutoff of 60) to the reported case count data. To reduce the dimensionality of our parameter space, we will use the contact data in (20) as a constraint and fit to the August-October 2020 data using the Sept contact matrix.

We assumed an equal infection ascertainment θ_i for both age groups. Optimization under this constraint, however, yielded an exceptionally low ascertainment fraction $[p, \theta]_{opt} = [0.0521, 0.0147]$. This optimization yielded excellent fit ($NSSR = 0.42$) and was stable to different initial guesses. However, an ascertainment fraction of 1.5% is clearly unreasonable. Therefore, we sought to explain this result by exploring our model's sensitivity to the ascertainment fraction.

For values above 20%, the model shows little discernible sensitivity to the ascertainment fraction (Fig. 10A). In particular, the optimal probability of infection (p) given contact does not vary with the ascertainment fraction in this range. This result is consequential. It implies that if even 20% of infections were ascertained during this period, we would infer the same transmission dynamics from the reported cases as if 100% of infections were ascertained. On the other hand, for values below 20%, the model shows acute sensitivity to the ascertainment fraction. We observe that in order to fit the reported case data with an excessively low infection ascertainment, our virus must be considerably more transmissible. The model's fit to the observed epidemic is greatest in the lower range of ascertainment (below 20%), with an optimal ascertainment fraction of 1.4%

(Fig. 10B). This suggests that these lower values add some computational utility to the model while fitting. One possible explanation is that a lower ascertainment fraction implies a larger number of active infectives at the beginning of the simulation, which may allow the model more flexibility to fit to the observed case data by fine-tuning the transmission parameter.

This analysis was repeated by allowing each group to have a unique ascertainment fraction and still yielded excessively low optima (data not shown). Since we cannot trust the optimal ascertainment fraction yielded from optimization, we must conclude that our model is unable to identify the ascertainment fraction from reported case data alone.

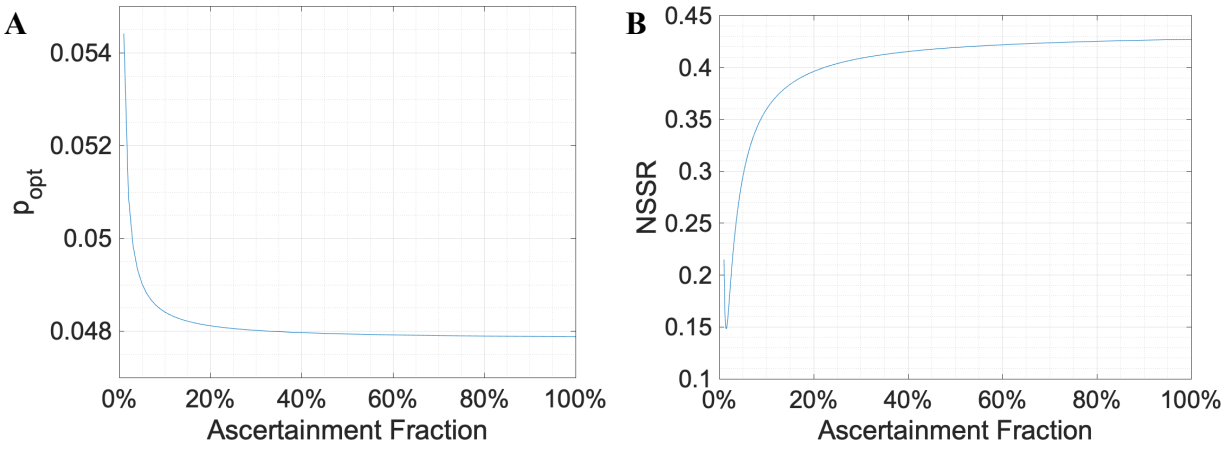

Figure 10: Sensitivity of the (**A**) optimal probability of infection and (**B**) its associated fit (as measured by the $NSSR$) to the ascertainment fraction $\theta = \theta_1 = \theta_2$ for both groups in the two-age group $SI_RI_{UR}R$ model (with age-cutoff 60). This analysis fit the Sept contact matrix to the August-October 2020 reported case data by minimizing the $NSSR$.

3.3 Estimating Infection Ascertainment

Estimating infection ascertainment remains an operationally difficult task. The most consistent methods rely on cross-sectional surveillance data randomly sampled from the population of interest (25). Serological surveys classically have been used to estimate the lifetime prevalence of a disease (i.e., the total number of individuals who have been infected). The ratio of total reported cases to this value yields the cumulative ascertain fraction. However, the cumulative fraction is only a measure of the average infection ascertainment (weighted by cases). Calculating the instantaneous ascertainment fraction, on the other hand, requires infection surveys which estimate the extent of active infectives at a given time (15). Nonetheless, these surveillance methods are expensive and underutilized in the ongoing COVID-19 pandemic (26). In the present section, we propose a novel method to calculate the instantaneous ascertainment fraction using only hospitalization and case data.

Since age is the greatest risk factor for severe disease from COVID-19 (17), individuals closer in age will experience this risk to a similar extent. As a result, the infection-hospitalization fraction (ϕ_I) for any age group with a sufficiently small age range should theoretically be a constant. The case-hospitalization fraction (ϕ_C), on the other hand, will vary over time as the proportion of infections ascertained by the health system (θ) varies. We see that

$$\phi_C = \frac{H}{C} = \frac{H}{I \cdot \theta} = \phi_I \cdot \frac{1}{\theta}$$

where H, I, and C represent the total number of hospitalizations, infections, and reported cases in a given time period. Thus, the ascertainment fraction for an age group (in a given time period) is calculated as $\theta = \phi_I / \phi_C$.

In the current section, we will apply this identity to estimate the ascertainment fraction of SARS-CoV-2 infections in B.C. throughout the course of the COVID-19 epidemic. Since the infection-hospitalization fraction is so sensitive to age (17), it will be advantageous to divide the B.C. population into as many age groups as feasible so as to allow for a more granular analysis. However, dividing the population into more age groups will also lead to more data scarcity – particularly in the younger age groups where we expect fewer hospitalizations. To bridge this balance, we will conduct our analysis with three age groups (ages 0-49, 50-69, 70+) whose cutoffs were decided based on k-means clustering of case-hospitalization fractions (see Appendix A).

In order to calculate the infection-hospitalization fraction for each age group, we needed to estimate the true extent of infection. The study in (21) reported that 0.56% (95% CI, 0.42-0.69%) of the B.C. population was seropositive for COVID-19 antibodies (from samples collected between May 9 and July 21, 2020). Although the data by age cohort was not available for B.C., the data from the whole of Canada suggested that there was no statistically significant difference between age groups. The current analysis will assume the same holds for the B.C. population and use this seropositivity to estimate the true extent of infection in each group – which in conjunction with the cumulative hospitalization curve yields our estimate for the infection-hospitalization fraction. Since the period of sample collection was so broad, we weighted our estimate by the number of blood samples collected each week (see Methods). We also applied an offset of 14 days to account for the delay from symptom onset and seroconversion (27). As expected, the infection-hospitalization fraction is highest in the older age groups and lowest in the younger age groups (Table 2).

Age Group	Infection-Hospitalization Fraction (95% CI)
0-49	0.5% (0.4-0.7%)
50-69	2.0% (1.7-2.7%)
70+	6.5% (5.3-8.7%)

Table 2: Infection-hospitalization fraction for each age group. Derived as the ratio of cumulative hospitalizations and cumulative infections as estimated from the seroprevalence study in (21). Seroprevalence estimates temporally weighted by the number of blood samples assayed (see Methods).

Due to scarcity of hospitalization events in the youngest age group (ages 0-49), we estimated the case-hospitalization fraction at the monthly level (Fig. 11). This may mask some rapid changes in testing behaviour (at the weekly or even daily scale), however, such changes are unlikely but for at the very outset of the epidemic. Then, we applied the infection-hospitalization fraction to this monthly case-hospitalization fraction to estimate the monthly ascertainment fraction $\theta = \phi_I/\phi_C$ for each age group in the B.C. healthcare system (Fig. 12).

As expected, infection ascertainment was lowest for all three age groups at the outset of the epidemic, before testing capacity was fully developed (March-May 2020; Fig. 12). In the summer months, infection ascertainment increased starkly in all age groups. This coincided with a period of relatively slow case growth (Fig. A.1). However, to note, the 50-69 age group yielded an exceptionally high ascertainment fraction for July which appears to be an outlier. This is likely due to the inherent stochasticity in hospitalization events – which is further exacerbated by low case numbers. As case growth began again (Sept 2020-Feb 2021) the ascertainment fraction stabilized at approximately 25-30% for each age group – which is to be expected of consistent testing habits within the B.C. health system.

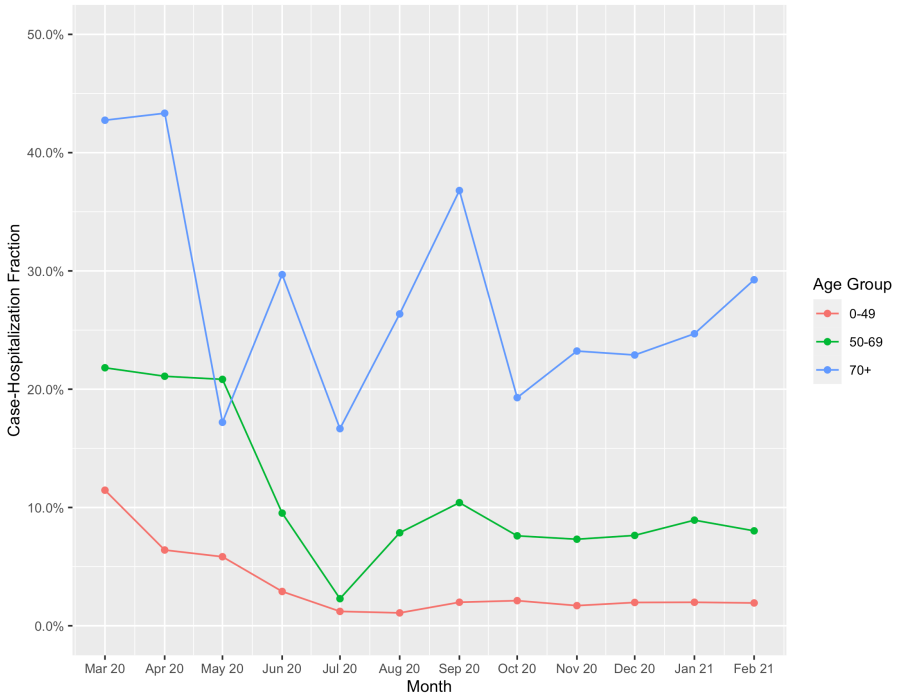

Figure 12: Case-hospitalization fraction (ϕ_C) computed for each age group (Group 1, ages 0-49; Group 2, ages 50-69; Group 3, 70+) at monthly scale.

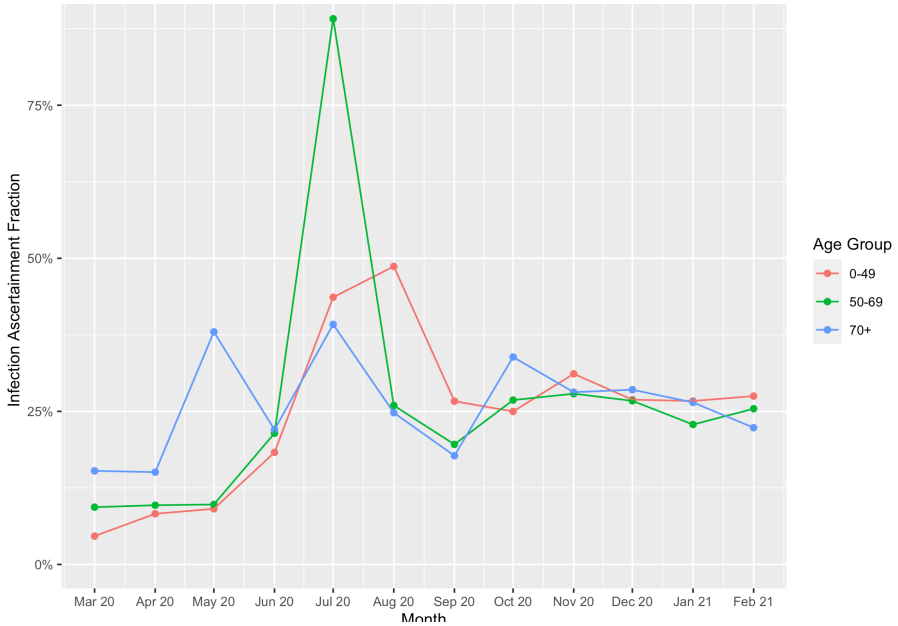

Figure 11: Infection ascertainment ($\theta = \phi_I/\phi_C$) computed for each age group (Group 1, ages 0-49; Group 2, ages 50-69; Group 3, 70+) at monthly scale. Error bars representing 95% CI excluded for clarity.

No efforts into quantifying the ascertainment fraction in B.C. have been reported in the literature thus far. However, such studies have been conducted for other health systems. Namely, the study in (15) used infection surveys to report that approximately 25% of infection were been ascertained in England before November 2020 (when testing was increased due to the risk of Variants of Concern). Although the epidemics of COVID-19 are in distinct in Canada and England, both high-income nations with comparable single-payer healthcare system (28). Therefore, it is reasonable to expect comparable testing capacities. Importantly, the study in (15) did not report a spike in infection ascertainment during the summer months (with the exception of a few regions). Therefore, it is possible that the stochasticity associated with low case numbers played a larger role in our estimation of infection ascertainment in the summer months.

3.3.1 The True Epidemic

By characterizing the ascertainment fraction for each month of the epidemic, we can back-calculate the true epidemic from the observed epidemic (Fig. 13). We are particularly interested in the size of the epidemic at its outset in March 2020. Our methods calculate that the epidemic in March was 14.4 (95% CI, 9.6-17.1) times larger than observed. This underscores the threat of the virus at this point and provides a retrospective justification of the initial lockdown (8).

In contrast, the epidemic at its peak in November 2020 was only 3.3 (2.2-3.9) times larger than observed (Fig. 13). Notably, since the ascertainment fraction for each age group stabilized from September onwards, the observed epidemic is sufficient to identify epidemiological trends. At the outset of the epidemic, on the other hand, testing capacity was building, and the ascertainment fraction varied, which muddles our ability to acutely identify trends.

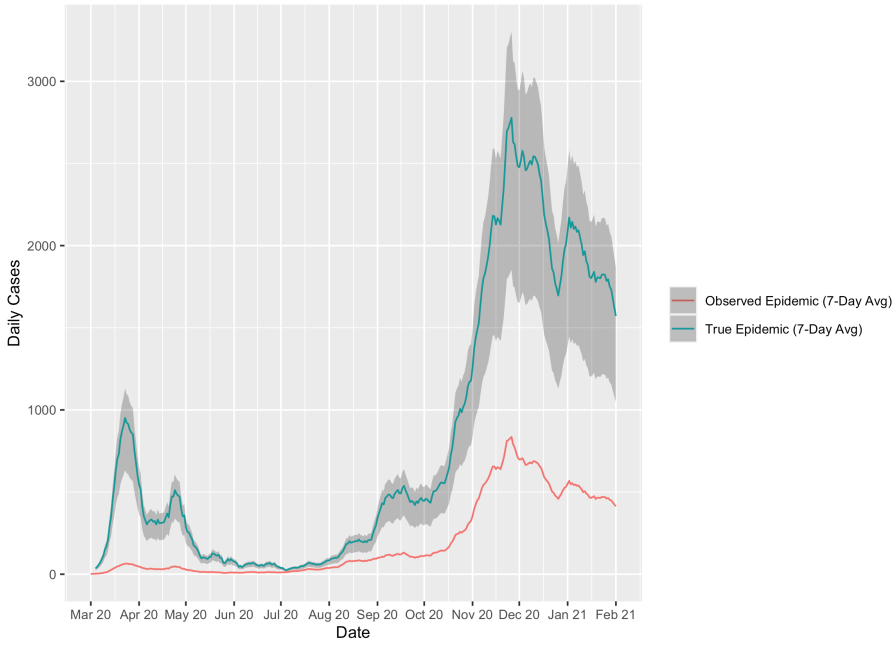

Figure 13: The true epidemic derived from the observed epidemic using monthly estimates for infection ascertainment in each age group. Shading represents 95% CI.

3.3.2 Validation

In order to verify that our estimates for the monthly ascertainment fraction for each age group are accurate, we compared our estimate for the cumulative burden of COVID-19 in the B.C. population (Fig. 14) to the literature. To date, the study in (21), which was used to estimate the infection-hospitalization fraction for each age group (Table 2), provides the only publicly available seroprevalence data for B.C. Since the length of sample collection in this study was broad (May 9-July 21, 2020), we must assume its estimate for seroprevalence (0.56%; 95% CI, 0.42-0.69%) holds for this entire period (minus the 14-day offset applied to account for delay to seroconversion; see Methods). Since our estimate for the lifetime prevalence of COVID-19 infection matches this seroprevalence estimate (Fig. 14), we can confirm that our methods are able to describe the true epidemic in B.C. at least during the first wave (Mar-May 2020).

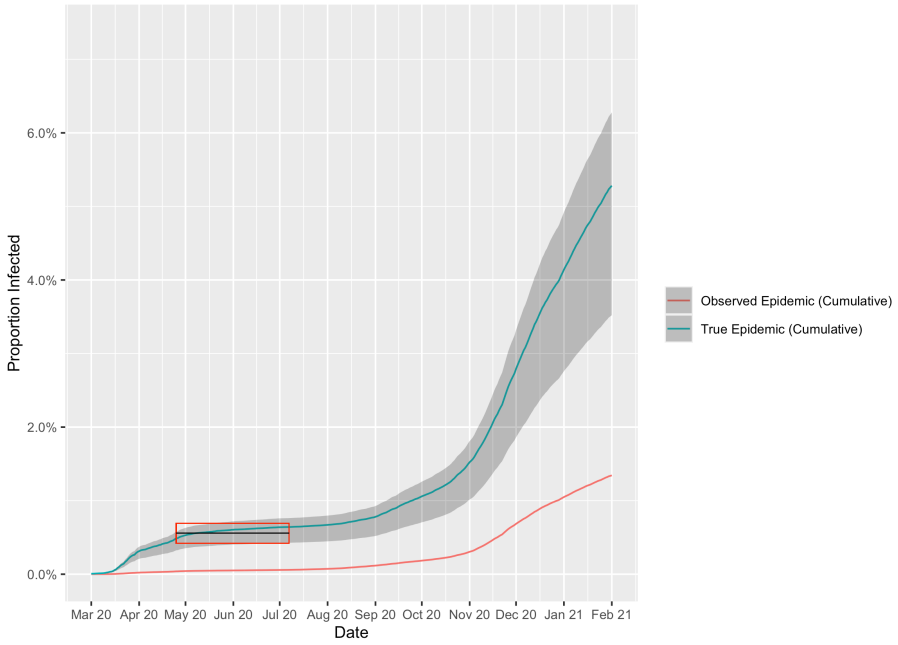

Figure 14: The lifetime prevalence of COVID-19 in the B.C. population as estimated using monthly estimates for infection ascertainment in each age group. Shading represents 95% CI. The seroprevalence estimate from (21) is marked by the black line (mean) and red rectangle (95% CI). Width of line and rectangle represents the duration of blood sample collection (shifted to account for 14-day delay from symptom onset to seroconversion).

As a secondary test, we artificially perturbed the ascertainment fraction values yielded for each age group by a constant amount in order to explore its effect on the lifetime prevalence estimate achieved by these methods. Since infection ascertainment was low in the beginning stages of the epidemic (Fig. 12), the system was very sensitive to even small changes in the ascertainment fraction (Fig. 15). In fact, increasing or decreasing the ascertainment fraction by as little as 2% makes our simulated curves miss the seroprevalence estimate in (21) – though they still fell within the 95% CI. This confirms the internal validity of our method for estimating ascertainment fraction.

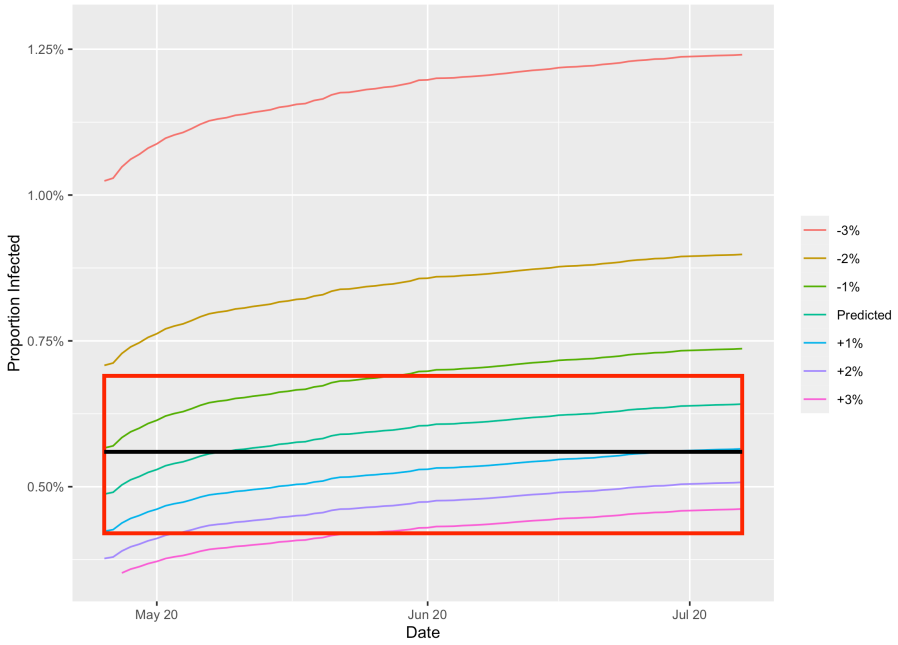

Figure 15: Sensitivity of lifetime prevalence of COVID-19 in the B.C. population. Predicted curve is as estimated using monthly estimates for infection ascertainment in each age group. The perturbations are added in the amounts labelled to the ascertainment fractions for each group for each month. The seroprevalence estimate from (21) is marked by the black line (mean) and red rectangle (95% CI). Width of line and rectangle represents the duration of blood sample collection (shifted to account for 14-day delay from symptom onset to seroconversion).

3.4 Vaccination

The primary objective of vaccination against a disease is elimination from a population (if not complete eradication worldwide). Whether or not eradication is possible with COVID-19 remains unknown (29). Regardless, vaccination can serve as a tool to prevent severe disease outcomes (such as hospitalization) in the short-term, which is of immediate interest to public health. In the current section, we will explore the effectiveness of various vaccination programs in preventing hospitalization events.

To this end, we will utilize our three-age group $V\sim SI_RI_{UR}HR$ model with age-cutoffs of 50 and 70 (motivated by k-means clustering of case-hospitalization fractions; see Appendix A). To date, four COVID-19 vaccines have been authorized for use by Health Canada (30). We will take the Pfizer-BioNTech (BNT162b2) vaccine as our model vaccine since the literature evaluating its real-world effectiveness is most robust. In particular, ongoing surveillance data from the Israel Ministry of Health suggests that the BNT162b2 vaccine is 94% effective at preventing asymptomatic infection and 97% effective at preventing severe disease (31). These values will be used to constrain the parameters in our model describing the extent to which vaccination blocks infection (δ_j) and prevents hospitalization (τ_j), respectively. Since the data for age is unavailable, we will assume that these values hold for all age groups. Quantifying the extent to which transmission is attenuated in vaccinated infectives (σ_j) is considerably more challenging. Studies have shown that vaccinated infectives carry up to four-fold lower viral loads compared to unvaccinated infectives (32). However, although viral loads have been shown to correlate with infectiousness, this correlation has not been sufficiently resolved so as to allow precise quantification (33). Therefore, the present analysis will conservatively assume that vaccination

attenuates 50% of transmission. The case-hospitalization fraction for each age group in our model is fixed from the final four months of data in 2020 (ages 0-49, 1.9%; 50-69, 7.7%; 70+, 23.5%).

3.4.1 Optimal Schedule

We speculate two (contrasting) vaccination strategies to prevent hospitalization events. The first strategy proposes to first vaccinate those at the highest risk of severe disease – namely the oldest population in descending order of age (17). The second strategy proposes to first vaccinate those most likely to transmit the virus – namely the youngest population, with the highest contact behaviour, in ascending order of age (20). In order to study these schedules, we will simulate the epidemic at the growth rate observed in November 2020 using the Sept contact matrix to describe age-specific contact patterns. November displayed the highest growth rate observed in B.C. to date; and the Sept contact survey displayed the highest number of contacts for the youngest population (ages 0-49) relative to the oldest (ages 70+). Therefore, these constraints represent a relatively high-contact scenario – reflecting the desire to uplift contact restrictions imposed by public health. Our simulation will arbitrarily run from November 1, 2020 to October 31, 2030, administrating 5 million vaccines (just under the total population of B.C., 5.1 million) at a uniform rate of 25,000 vaccines per day for 200 days.

As a benchmark, if the epidemic was allowed to grow without vaccination or additional public health restrictions) herd immunity would be achieved once 28.0% of the population became infected (ages 0-49, 32.6%; 50-69, 24.6%; 70+, 13.2%). Note that this value is computed for the level of contacts observed in November and thus not reflective of true herd immunity (which is computed given non-restrictive, pre-pandemic contact behaviour). This epidemic would result in a cumulative total of 61,759 hospitalizations (Table 3).

We observed that both vaccination strategies are highly effective at preventing hospitalization. If the oldest population is vaccinated first, the epidemic would produce 3,640 hospitalizations, successfully preventing 94.7% of hospitalizations. Conversely, if the youngest population is vaccinated first, the epidemic would only produce 1,658 hospitalizations, successfully preventing 97.8% of hospitalizations. These results suggest that the utility of blocking transmission in the younger population outweighs the risk associated with leaving the oldest population unprotected – albeit only marginally. It is important to note that this result appears to contradict the public health stance in B.C., which is actively vaccinating its population in descending order of age (9).

Vaccination Program	**Cumulative Hospitalizations**			
	Ages 0-49	**Ages 50-69**	**Ages 70+**	**All Ages**
No Vaccine	17,729	26,023	18,007	61,759
Oldest First	2,345	1,126	169	3,640
Youngest First	364	761	533	1,658

Table 3: Hospitalizations in each age group under different vaccination strategies.

3.4.1.1 Sensitivity Analysis: Attenuating Transmission (σ_j)

In order to better understand our model, we explored its sensitivity to the vaccine's ability to attenuate transmission (σ_j). We observe that both strategies are benefitted by even a small amount of attenuation (Fig. 16A). However, the marginal utility diminishes rapidly past just $\sigma_j = 15\%$. Notably, the strategy of vaccinating the youngest population first is favoured for all $\sigma_j > 13\%$, confirming that its relative utility is rooted in preventing transmission. Further, when we assume that the vaccine is unable to block infection (by fixing $\delta_j = 0$), vaccinating the younger population is only favoured for $\sigma_j > 23\%$ (Fig. 16B). Therefore, if the vaccine does not block infection, it

must attenuate transmission to a greater extent in order to justify vaccinating the youngest population first. This underscores the role of attenuating transmission in this strategy.

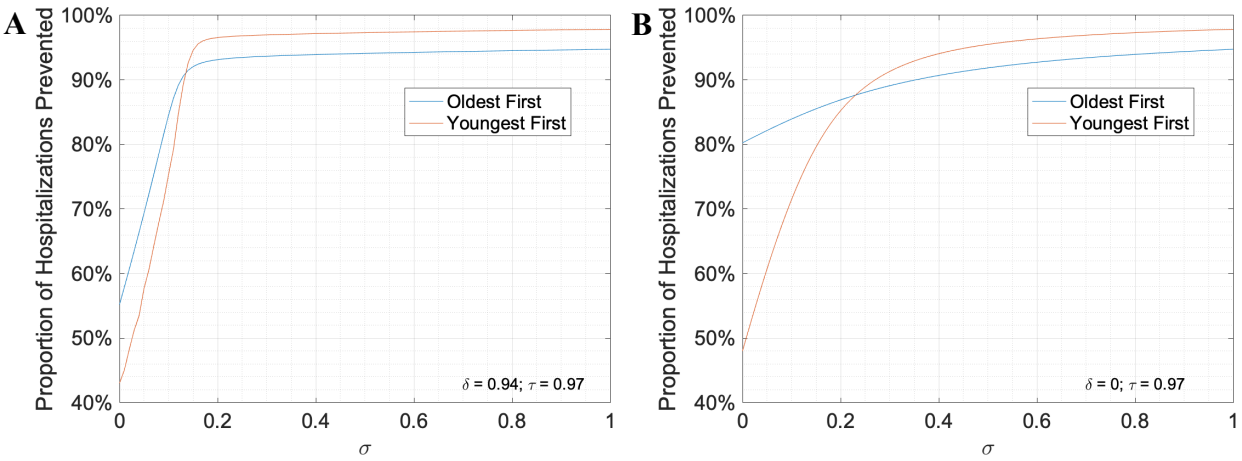

Figure 16: Sensitivity of vaccination programs to the vaccine's ability to attenuate transmission (σ). Panel **A** describes a scenario where the vaccine is able to block infection ($\delta = 0.94$); while panel **B** describes a scenario where the vaccine is wholly unable to do so ($\delta = 0$).

3.4.1.2 Sensitivity Analysis: Blocking Infection (δ_j)

Next, we tested the sensitivity of our model to the vaccine's ability to block infection (δ_j). We observe that neither strategy is particularly sensitive to this parameter, and that vaccinating the younger population is favoured throughout (Fig. 17A). When we remove the vaccine's ability to prevent hospitalization (fixing $\tau_j = 0$), both vaccination schedules are hindered. The strategy of vaccinating the older population first is especially afflicted and develops increased sensitivity to δ_j (Fig. 17B). Therefore, the vaccine's ability to ability to block infection is not a consequential parameter, under the condition that it still prevents hospitalization. Thus far, only the BNT162b2 vaccine has been definitively shown to block infection (31). These results confirm that other vaccines may yield equivalent utility to public health regardless of their ability to block infection (so long as they still prevent severe disease).

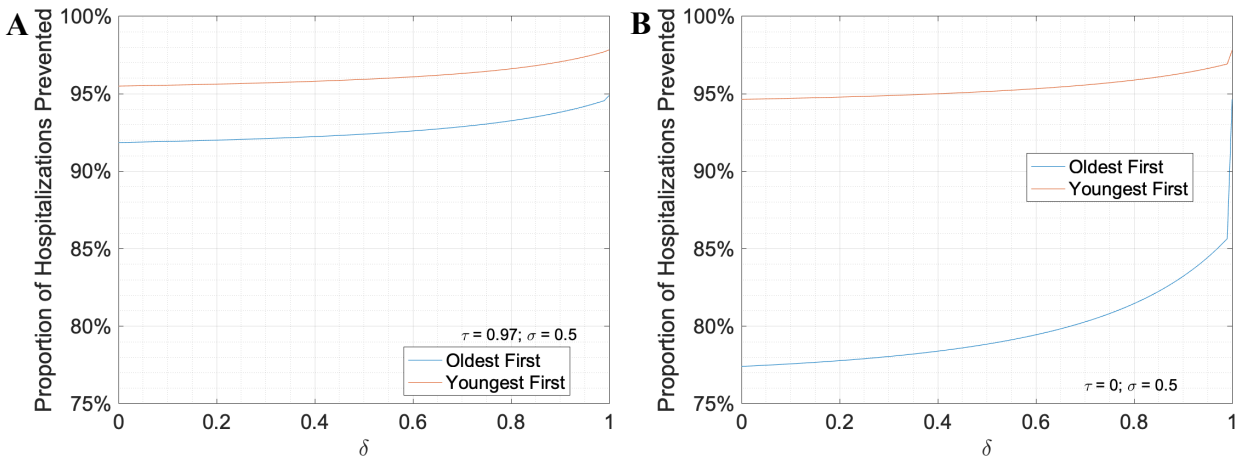

Figure 17: Sensitivity of vaccination programs to the vaccine's ability to block infection (δ). Panel **A** describes a scenario where the vaccine is able to prevent hospitalization ($\tau = 0.97$); while panel **B** describes a scenario where the vaccine is wholly unable to do so ($\tau = 0$).

3.4.1.3 Sensitivity Analysis: Preventing Hospitalization (τ_j)

Finally, we tested our model's sensitivity to the vaccine's ability to prevent hospitalization. The strategy of vaccinating the older population before the younger population shows considerable sensitivity to this metric, while the strategy of vaccinating the younger population first shows little thereof (Fig. 18A). When we eliminate the vaccine's ability to block infection (by fixing $\delta_j = 0$) the former strategy only becomes more sensitive to this metric (Fig. 18B). This underpins that the strategy of vaccinating the older population first provides direct protection to the most vulnerable population, while vaccinating the youngest population first provides indirect protection, by reducing transmission.

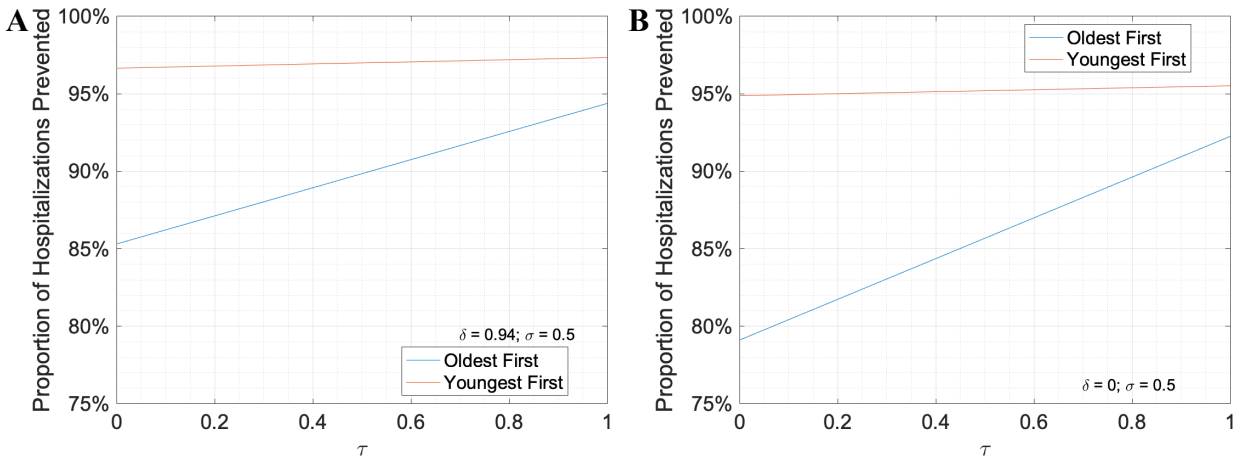

Figure 18: Sensitivity of vaccination programs to the vaccine's ability to prevent hospitalization (τ). Panel **A** describes a scenario where the vaccine is able to block infection ($\delta = 0.94$); while panel **B** describes a scenario where the vaccine is wholly unable to do so ($\delta = 0$).

3.4.1.4 Sensitivity Analysis: Contact Behaviour

The current model was simulated under the constraint that contact behaviour modelled the Sept contact survey conducted in (20). In this survey the youngest age group (ages 0-50) had 2.6 times as many daily contacts as the oldest age group (ages 70+). In the Dec contact survey this ratio was reduced to 1.9, representing that contact behaviour in the youngest age group was tamed between these months. The utility in vaccinating the youngest population relies on its high (observed) contact behaviour relative to the older population. Further, its utility may be sensitive to the level of contacts held. In order to test this, we explored the sensitivity of our model to the level of contacts for both the Sept and Dec contact matrices (Fig. 19). For both contact matrices, as the level of contacts decreases the proportion of hospitalizations prevented by either strategy also decreases. This is because when contact levels are sufficiently low the epidemic decays to extinction, in which scenario there are fewer simulated hospitalizations that can be prevented via vaccination. The level of contacts observed in December lies in this region, which is consistent with the observation that the observed epidemic was declining in December (Fig. A.1).

Using the Sept contact matrix to constrain contact behaviour, vaccinating the younger population first prevents more hospitalizations than vaccinating the older population at all contact levels (Fig. 19A). Furthermore, this strategy's utility relative to the alternative increases as contacts increase (when blocking transmission becomes more advantageous). Interestingly, for the Dec contact matrix constraint – which represents a scenario in which the younger population has tamed its contacts – the aforementioned strategy loses its utility and vaccinating the older population becomes the favoured strategy. This is especially true when the level of contacts is reduced, as observed in December (Fig. 19B). At the level of contacts observed in November, the difference between the two strategies diminishes.

These results suggest that if the younger population (relative to the older population) is exemplifying high-contact behaviour, then blocking chains of transmission in this population will lead to an overall reduced number of hospitalizations. However, if this population tames its contact behaviour, then the utility of this method is reduced and (especially for low levels of contacts) it will be preferred to vaccinate the older, more vulnerable, population. Notably, even under these stringent conditions the benefit in vaccinating the oldest population first is only marginally superior to the alternative.

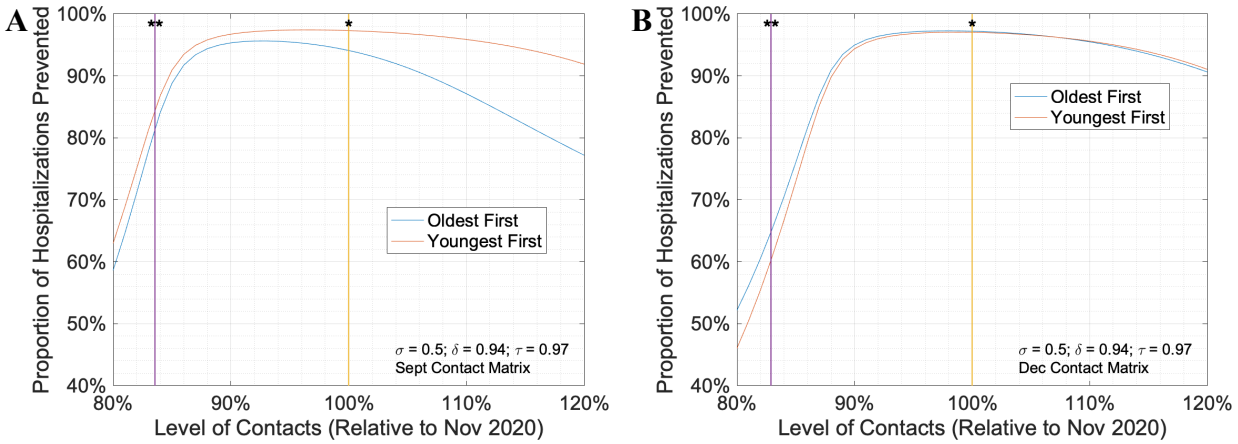

Figure 19: Sensitivity of vaccination programs to the level of contacts observed in the population, normalized to the level observed in November 2020. The lines annotated with (*) and (**) denote the levels of contact observed in November and December, respectively (20). Panel **A** explores the scenario where contact behaviour models the Sept contact matrix, in which the youngest population (ages 0-49) had 2.6 times as many contacts as the oldest population (ages 70+). Panel **B** explores the scenario where contact behaviour models the Dec contact matrix, in which the youngest population (ages 0-49) had only 1.9 times as many contacts as the oldest population (ages 70+).

Chapter 4: Discussion

The ongoing COVID-19 pandemic has posed an urgent demand to Mathematical Epidemiology as health systems around the world rely on models of SARS-CoV-2 transmission to make decisions in the interest of public health. Due to the necessarily radical nature of these decisions, they carry the weight of substantial social and economic ramifications. However, in the words of Statistician George E. P. Box, "All models are wrong, but some are useful" (34). Therefore, we must carefully construct our models to balance the extent to which they reflect reality and the extent to which they allow for useful analysis. Due to uncertainties around the role of age and infection ascertainment in the COVID-19 epidemic in B.C., we developed the classic Kermack-McKendrick SIR model (11) into the generalized n-age group $SI_RI_{UR}R$ model. This model, as with all other models, is necessarily wrong. In the current analysis, we explore the extent to which this model is useful.

4.1 Model Fitting

We fit our $SI_RI_{UR}R$ model to the reported case data under various constraints in order to determine the extent to which we could identify the transmission and ascertainment fraction parameters for each age group.

Our model was parametrized by n^2 transmission ($\beta_{i,j}$) parameters, each of which delineate intra- and intergroup transmission dynamics. We attempted to identify these parameters in the simplest case of $n = 2$ age groups (with age-cutoff 60) using various relevant constraints. The constraints of no mixing and homogenous mixing between age groups were sufficiently strict to allow us to identify a stable optimum. However, these optimizations yielded impossible results. Namely, both of these constraints necessitated that the older population has more contacts than the

younger population – which is inconsistent with observed behaviour (17). The heterogenous mixing constraint is the least restrictive and covers all theoretically possible contact patterns. However, the parameter space under this constraint contained far too many degrees of freedom to allow for an injective mapping to the model's output space (24). To understand this result, consider the two age groups as an example. The epidemic in the older age group could theoretically be simulated as being constructed entirely by its own infectives, entirely by infectives in the younger age group, or by some mix of the two – all with equivalent fit. Since multiple combinations of these parameters could construct the same epidemic, fitting to the observed epidemic cannot identify these parameters under the heterogenous constraint. This implies that the reported case data for each age groups – without external data to apply additional constraint – is insufficient to characterize the dynamics of COVID-19 infection between different age groups with any measurable degree of precision.

The current analysis employed the contact survey data in (20) to further constrain the parameter space. This constraint allowed our model to fit the epidemic from May 2020 to Jan 2021 with a high degree of accuracy. The fact that this fit was so spectacular is consequential as it verifies that our model (with this constraints) faithfully models the effect of contact behaviour on the observed COVID-19 epidemic in each age group.

Our model was also parametrized by n ascertainment fraction (θ_i) parameters, which give insight into testing behaviour. However, the optimal ascertainment fraction value were exceedingly low and thus do not hold real-world significance. This is not surprising as the literature does not show any analysis that has been able to identify these parameters from reported case data alone. However, some success has been reported by using hospitalization data in conjunction with case data (35).

4.2 Estimating Infection Ascertainment

To our knowledge, the method to estimate infection ascertainment as the ratio of the infection- and case-hospitalization fraction have yet to be employed in the literature. A similar analysis with infection- and case-fatalities was reported in (36), however an important distinction is that the current analysis is structured by age. If we were to ignore the effect of age, the case-hospitalization fraction for the whole population may change without changes in infection ascertainment. Namely, if infections shift towards the younger population, the case-hospitalization fraction would decrease. This would be computed as an artificial increase in infection ascertainment – confounding our estimation. Therefore, the current analysis employed three age groups – whose constituent age cohorts showed comparable case-hospitalization fractions – to circumvent this bias. In addition, the analysis in (36) did not use estimates for the infection-hospitalization fraction specific to the populations analyzed. Since the demographics and behaviour of each population is unique, it is best practice to estimate the infection-hospitalization fraction using data specific to the population of its interest.

Our estimates for infection ascertainment showed severe underreporting of infections during the first wave (before June 2020). This is consistent with the expected testing behaviour as testing is B.C. was limited to high-risk groups and individuals connected to travel (13). During the summer months, the epidemic slowed significantly, and infection ascertainment in each age increased considerably. One possible explanation for this may be that with fewer infection, the health system was less burdened and that contact tracing efforts were more efficient. Another possible explanation may be that with relatively few cases, stochasticity in the incidence of hospitalization given infection may play a larger role. In simpler terms, it could be sheer luck that the few infected in the summer months did not require hospitalization. This appears to be the case

for the 50-69 age group which reports an 80%+ ascertainment fraction for July. Nonetheless, from September onwards, the ascertainment fraction was consistently resolved between 25-30% for each age group. Since an estimated 17% of infections are asymptomatic (22), these results seem to suggest that at least 63% of symptomatic infections are not ascertained. This result suggests that either the extent of asymptomatic infection is underestimated or that there is a significant hesitancy to get tested. This may reflect the stigma attached to COVID-19, or people's commitments to work or family. Regardless, this result underlines a significant threat to B.C.'s contact tracing efforts.

We predicted that the older age group would have a higher ascertainment fraction since they are more likely to show severe disease and thus get tested. However, this was only observed at the very outset of the epidemic. One possible explanation may be that the older population is more likely to be exposed to occasional hospital outbreak since older individuals are more likely to seek hospital care due to reasons unrelated to COVID-19. However, it is uncertain to what extent hospital outbreaks played a role in the observed hospitalization events reported.

Using these estimates for infection ascertainment, we were able to back-calculate the true epidemic in B.C. We validated our resultant epidemic curve by comparing the lifetime prevalence of COVID-19 to a single seroprevalence estimates in the literature. If more seroprevalence data becomes available, we could further validate these methods. Since these methods only require case hospitalization data (along with a single time-point of seroprevalence), they are broadly applicable to other health system and recommended for use (once sufficiently validated).

It is worthwhile to mention that this analysis is confounded once COVID-19 variants of concern (VOC) play a role in the observed epidemic. This is because the true risk of hospitalization given infection with emerging VOC is largely unknown.

4.3 Vaccination

Our analysis shows that the Pfizer-BioNTech (BNT162b2) vaccine is unreasonably effective at protecting our population from adverse disease outcomes. We demonstrated that even after artificially decreasing the efficacy of the vaccine in one of the three metrics explored (its ability to prevent hospitalization, prevent infection, and attenuate transmission) the other two were sufficient to prevent hospitalization events in the population. Since hospitalization (and death) events are of the utmost importance to public health, this provides concrete evidence that even administering vaccines with lower efficacies would be greatly beneficial to the public health effort – discouraging vaccine hesitancy.

B.C.'s current vaccination strategy is to vaccinate the oldest population first, so as to directly protect those most vulnerable. The current analysis suggests that this strategy is only preferred over vaccinating the youngest population first when contacts are tamed in the younger population relative to the older population. Even with the level restraint observed in the Dec contact survey (20), vaccinating the older population is only marginally preferred over the vaccinating the younger population and its relative utility decreases as the level of contacts increase. Of course, the current analysis is an oversimplification of contact behaviour as contact behaviour is heterogenous even within age groups (20). Therefore, in order to reap the utility of both of these contrasting strategies, the present analysis suggests vaccinating high-contact individuals (regardless of age) simultaneously with the older, more vulnerable, population. This is consistent with a recently published preprint (37) which advocates to vaccinate frontline workers earlier than currently scheduled.

4.4 Conclusion

The current report explored the role of age and infection ascertainment in simulating the epidemic observed in B.C. Although the transmission dynamics between age groups and infection ascertainment could not be identified from reported case data alone, we were able to use external data to identify these parameters. Namely, surveyed contact data was sufficient to characterize transmission dynamics in the epidemic with a high degree of accuracy from May 2020 to January 2021. Further, we developed and validated a novel method to estimate the monthly ascertainment fraction for each age group using only case and hospitalization data (as well as a single time-point for seroprevalence). Due to its simplicity and relatively low cost, the present report recommends using these methods to characterize the true epidemic in other health systems. Finally, we simulated the effectiveness of different vaccination strategies under varying contact patterns. The current analysis advises that in order to minimize hospitalizations in B.C., we should prioritize vaccinating high-contact frontline workers alongside the oldest, most vulnerable population.

References

1. 武汉市卫健委通报肺炎疫情 湖北日报数字报 [Internet]. [cited 2021 Mar 5]. Available from: https://epaper.hubeidaily.net/pc/content/202001/01/content_15040.html

2. WHO Statement Regarding Cluster of Pneumonia Cases in Wuhan, China [Internet]. [cited 2021 Mar 5]. Available from: https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china

3. Chen X, Yu B. First two months of the 2019 Coronavirus Disease (COVID-19) epidemic in China: real-time surveillance and evaluation with a second derivative model. Glob Health Res Policy. 2020 Mar 2;5(1):7.

4. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. The species Severe acute respiratory syndrome-related coronavirus : classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020 Apr;5(4):536–44.

5. IHR Emergency Committee on Novel Coronavirus (2019-nCoV) [Internet]. [cited 2021 Mar 5]. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov)

6. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 [Internet]. [cited 2021 Mar 5]. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

7. BC COVID-19 Data [Internet]. [cited 2021 Mar 5]. Available from: http://www.bccdc.ca/health-info/diseases-conditions/covid-19/data

8. Engagement GC and P. COVID-19 province-wide restrictions [Internet]. Province of British Columbia; [cited 2021 Mar 5]. Available from: https://www2.gov.bc.ca/gov/content/safety/emergency-preparedness-response-recovery/covid-19-provincial-support/restrictions

9. Engagement GC and P. B.C.'s COVID-19 Immunization Plan [Internet]. Province of British Columbia; [cited 2021 Mar 5]. Available from: https://www2.gov.bc.ca/gov/content/covid-19/vaccine/plan

10. Anderson SC, Mulberry N, Edwards AM, Stockdale JE, Iyaniwura SA, Falcao RC, et al. How much leeway is there to relax COVID-19 control measures? [Internet]. Epidemiology; 2020 Jun [cited 2021 Mar 5]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.06.12.20129833

11. Kermack WO, McKendrick AG, Walker GT. A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser Contain Pap Math Phys Character. 1927 Aug 1;115(772):700–21.

12. Majumder MS, Mandl KD. Early Transmissibility Assessment of a Novel Coronavirus in Wuhan, China [Internet]. Rochester, NY: Social Science Research Network; 2020 Jan [cited 2021 Mar 5]. Report No.: ID 3524675. Available from: https://papers.ssrn.com/abstract=3524675

13. Phases of COVID-19 testing in BC [Internet]. [cited 2021 Mar 5]. Available from: http://www.bccdc.ca/health-info/diseases-conditions/covid-19/testing/phases-of-covid-19-testing-in-bc

14. Oran DP, Topol EJ. The Proportion of SARS-CoV-2 Infections That Are Asymptomatic. Ann Intern Med [Internet]. 2021 Jan 22 [cited 2021 Mar 5]; Available from: https://www.acpjournals.org/doi/full/10.7326/M20-6976

15. Colman E, Enright J, Puspitarani GA, Kao RR. Estimating the proportion of SARS-CoV-2 infections reported through diagnostic testing. medRxiv. 2021 Jan 1;2021.02.09.21251411.

16. Angulo FJ, Finelli L, Swerdlow DL. Estimation of US SARS-CoV-2 Infections, Symptomatic Infections, Hospitalizations, and Deaths Using Seroprevalence Surveys. JAMA Netw Open. 2021 Jan 5;4(1):e2033706–e2033706.

17. Kang S-J, Jung SI. Age-Related Morbidity and Mortality among Patients with COVID-19. Infect Chemother. 2020 Jun;52(2):154–64.

18. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. Appl Stat. 1979;28(1):100.

19. Davies NG, Klepac P, Liu Y, Prem K, Jit M, Eggo RM. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat Med. 2020 Aug;26(8):1205–11.

20. Brankston G, Merkley E, Fisman DN, Tuite AR, Poljak Z, Loewen PJ, et al. Quantifying Contact Patterns in Response to COVID-19 Public Health Measures in Canada. medRxiv. 2021 Mar 12;2021.03.11.21253301.

21. Saeed S, Drews SJ, Pambrun C, Yi Q-L, Osmond L, O'Brien SF. SARS-CoV-2 seroprevalence among blood donors after the first COVID-19 wave in Canada. Transfusion (Paris). 2021;61(3):862–72.

22. Byambasuren O, Cardona M, Bell K, Clark J, McLaws M-L, Glasziou P. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis. Off J Assoc Med Microbiol Infect Dis Can. 2020 Dec 1;5(4):223–34.

23. Rees EM, Nightingale ES, Jafari Y, Waterlow NR, Clifford S, B. Pearson CA, et al. COVID-19 length of hospital stay: a systematic review and data synthesis. BMC Med. 2020 Sep 3;18(1):270.

24. Hines KE, Middendorf TR, Aldrich RW. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. J Gen Physiol. 2014 Mar;143(3):401–16.

25. Gibbons CL, Mangen M-JJ, Plass D, Havelaar AH, Brooke RJ, Kramarz P, et al. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health. 2014 Feb 11;14(1):147.

26. Lytras T, Panagiotakopoulos G, Tsiodras S. Estimating the ascertainment rate of SARS-CoV-2 infection in Wuhan, China: implications for management of the global outbreak. medRxiv. 2020 Mar 26;2020.03.24.20042218.

27. Zhao J, Yuan Q, Wang H, Liu W, Liao X, Su Y, et al. Antibody Responses to SARS-CoV-2 in Patients With Novel Coronavirus Disease 2019. Clin Infect Dis Off Publ Infect Dis Soc Am. 2020 Nov 19;71(16):2027–34.

28. Wagstaff A, Neelsen S. A comprehensive assessment of universal health coverage in 111 countries: a retrospective observational study. Lancet Glob Health. 2020 Jan 1;8(1):e39–49.

29. Lee A, Thornley S, Morris AJ, Sundborn G. Should countries aim for elimination in the covid-19 pandemic? BMJ. 2020 Sep 9;370:m3410.

30. Canada H. COVID-19 Vaccines: Authorized vaccines [Internet]. aem. 2020 [cited 2021 Mar 24]. Available from: https://www.canada.ca/en/health-canada/services/drugs-health-products/covid19-industry/drugs-vaccines-treatments/vaccines.html

31. Real-World Evidence Confirms High Effectiveness of Pfizer-BioNTech COVID-19 Vaccine and Profound Public Health Impact of Vaccination One Year After Pandemic Declared | Pfizer [Internet]. [cited 2021 Mar 24]. Available from: https://www.pfizer.com/news/press-release/press-release-detail/real-world-evidence-confirms-high-effectiveness-pfizer

32. Levine-Tiefenbrun M, Yelin I, Katz R, Herzel E, Golan Z, Schreiber L, et al. Decreased SARS-CoV-2 viral load following vaccination. medRxiv. 2021 Feb 8;2021.02.06.21251283.

33. Marks M, Millat-Martinez P, Ouchi D, Roberts C h, Alemany A, Corbacho-Monné M, et al. Transmission of COVID-19 in 282 clusters in Catalonia, Spain: a cohort study. Lancet Infect Dis [Internet]. 2021 Feb 2 [cited 2021 Mar 24];0(0). Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30985-3/abstract

34. Box GEP. Robustness in the Strategy of Scientific Model Building. In: LAUNER RL, WILKINSON GN, editors. Robustness in Statistics [Internet]. Academic Press; 1979. p. 201–36. Available from: https://www.sciencedirect.com/science/article/pii/B9780124381506500182

35. Bhaduri R, Kundu R, Purkayastha S, Kleinsasser M, Beesley LJ, Mukherjee B. Extending the Susceptible-Exposed-Infected-Removed (SEIR) model to handle the high false negative rate and symptom-based administration of COVID-19 diagnostic tests: SEIR-fansy. medRxiv

[Internet]. 2020 Sep 25 [cited 2021 Apr 5]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7523173/

36. Russell TW, Golding N, Hellewell J, Abbott S, Wright L, Pearson CAB, et al. Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. BMC Med. 2020 Oct 22;18(1):332.

37. Mulberry N, Tupper P, Kirwin E, McCabe C, Colijn C. Vaccine Rollout Strategies: The Case for Vaccinating Essential Workers Early. medRxiv. 2021 Jan 1;2021.02.23.21252309.

Appendices

Appendix A Deciding Age-cutoff for Two-age Group Model

In order to explore the two-age group $SI_RI_{UR}R$ model, we must decide on an appropriate age-cutoff to define the two groups. This is a consequential decision. Ideally, we should include age cohorts with similar contact behaviours within the same age group. This will minimize the intragroup variability in transmission and ensure we don't miss potentially significant age-specific dynamics. However, we should also attempt to group age cohorts according to disease outcomes (i.e., risk of hospitalization). This will allow us to make more useful inferences on public health outcomes from our model. For most populations around the world, higher contact behaviour is skewed towards younger age cohorts and risk for severe disease outcomes is skewed towards older cohorts (17). However, the specific demographics of a population, including the relative population size and health status of each age cohort, can considerably affect these two variables. Therefore, in deciding an age-cutoff for our model, we will analyze age-specific infection and hospitalization dynamics for B.C. anew.

The epidemic in B.C. reached its first peak in March 2020 (Fig. A.1). Through a province-wide lockdown (8), B.C. was able to curtail growth and the epidemic diminished during the summer months. From August onwards, the epidemic began to grow once again, peaking in late November.

Assuming a constant ascertainment fraction across age cohorts, per capita case counts should serve as a proxy for studying age-specific contact behaviour. This is because age cohorts with more contacts will naturally incur more infections. However, we suspect that infection ascertainment does depends on age since older infectives are more likely to suffer from severe disease (and thus more likely to be tested and ascertained by the health system) (17). This is evident

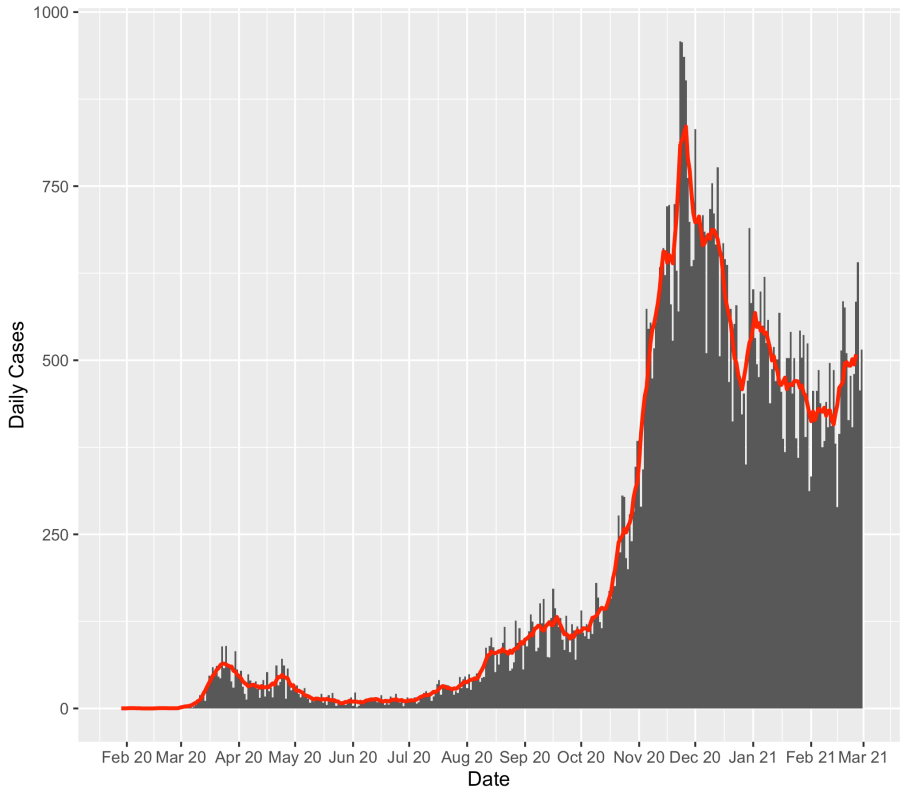

Figure A.1: Daily reported cases of COVID-19 in B.C. from outset of epidemic (January 26, 2020) to February 2021. Overlain red line represents 7-day moving average.

for B.C. as during the final four months of 2020, when testing practices were known to be consistent (13), the 90+ age cohort reported the highest per capita case count, followed by the 20-29 cohort (Fig. A.2). Clearly, the 20-29 age cohort reports high per capita case counts due to high contact behaviour while the 90+ age cohort does so due to higher infection ascertainment. Thus, per capita case counts are insufficient to delineate contact behaviour and we must turn to grouping age cohorts by severe disease outcomes.

Severe disease outcomes include hospitalization and death events. Our analysis will study hospitalization – through per capita hospitalization counts (Fig. A.3) and case-hospitalization fractions (Fig. A.4) – since the data for each age cohort in B.C. is more robust than for death events.

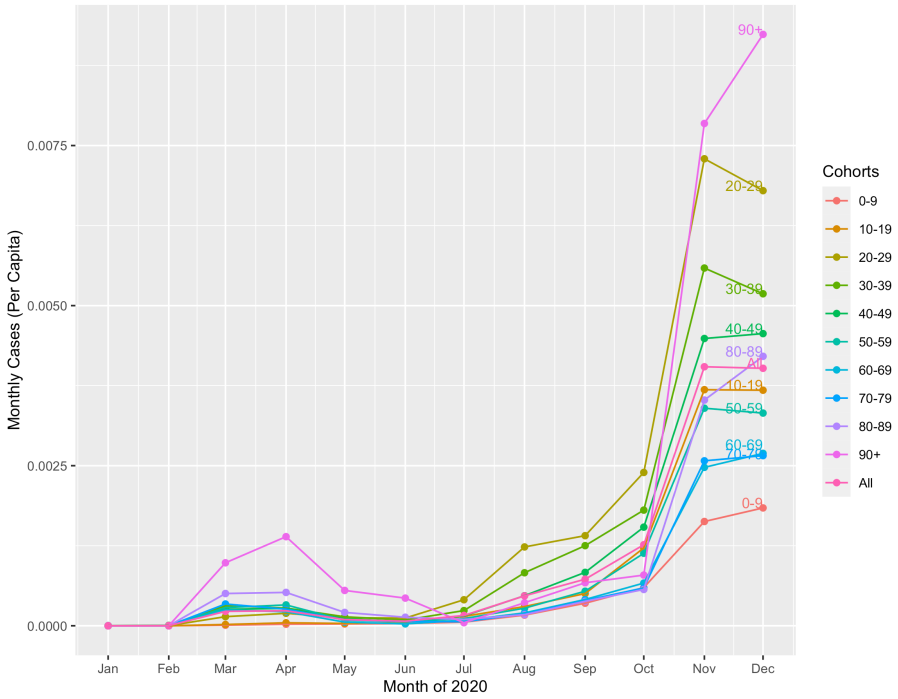

Figure A.2: Monthly reported cases (per capita) of COVID-19 for each age cohort in B.C. population, computed for each month of 2020.

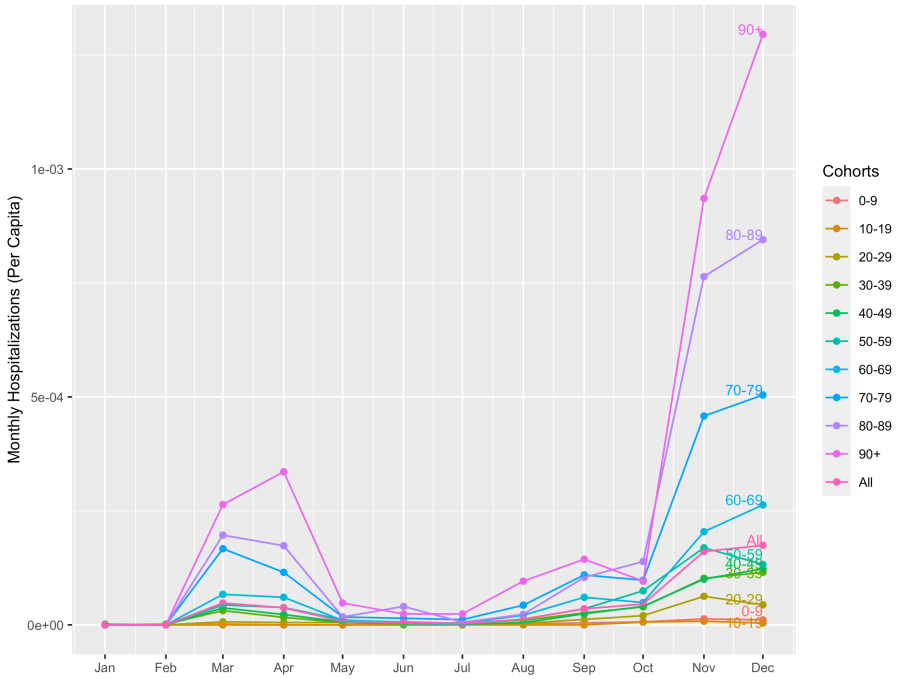

Figure A.3: Monthly (per capita) COVID-19 hospitalizations for each age cohort in B.C. population, computed for each month of 2020.

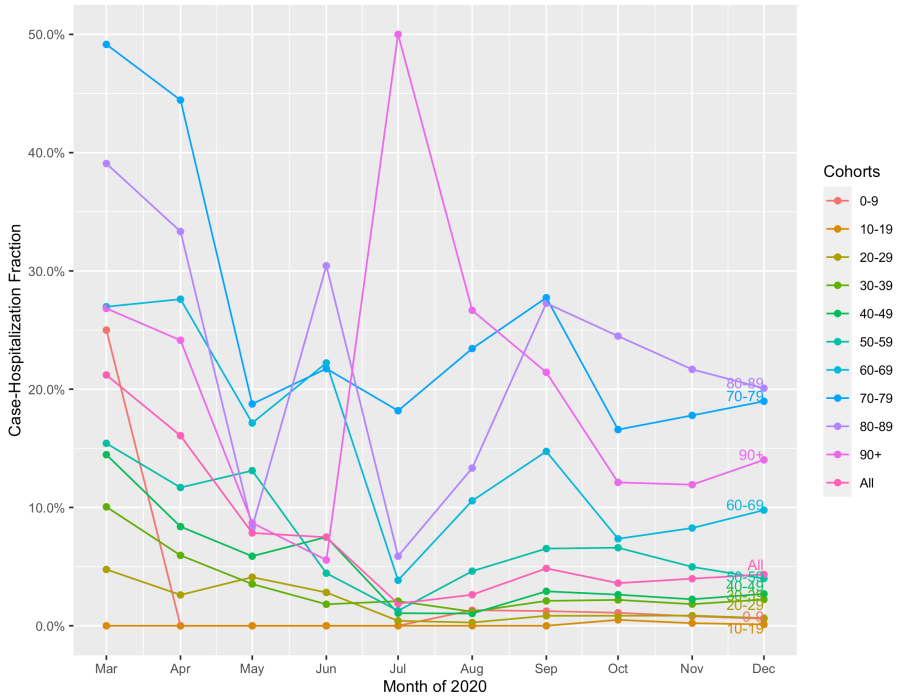

Figure A.4: COVID-19 case-hospitalization fraction for each age cohort in B.C. population, computed for each month of 2020.

The case-hospitalization fraction provides the purest measure of risk of severe disease. However, a high case-hospitalization fraction does not necessarily imply that an age cohort is severely afflicted by the disease as the cohort could be well isolated and incur few cases. Per capita hospitalization, on the other hand, measures the extent to which an age cohort is truly affected by the disease, given their true infection count and true risk of hospitalization given infection. The current study analyzes both metrics. Although we could visually group the age cohorts for each of these metrics, we employed a clustering algorithm to do so in an unbiased manner. k-means clustering groups observations so as to minimize the Euclidian distance between observations of a cluster to its centroid (18). We conducted k-means analysis on the per capita hospitalization (Fig. A.3) and case-hospitalization fraction (Fig. A.4) of each age cohort for the final four months of

2020. The per capita hospitalization data suggested an age-cutoff of 80 while the case-hospitalization fraction data suggested a cutoff of 70.

Both age-cutoffs suggested by the clustering analysis would leave far too small a proportion of the population in the older age groups (70, 12.1%; 80, 4.6%). Limited data in these groups would result in increased noise when attempting to fit our model to reported case counts. Therefore, we will begin our analysis with an age-cutoff of 60, such that Group 1 (74.8%) represents ages 0-59 and Group 2 (25.2%) represents ages 60+.

Appendix B Sensitivity to Age-structure

In subsection 3.2.1.1, the two-age group $SI_RI_{UR}R$ model, with an age-cutoff of 60, was fit to the reported data case using the contact survey data in (20) as a constraint. In the present section, we will test the sensitivity of this model to our choice for the age-cutoff and to the number of age groups.

B.1 Age-cutoff

Thus far, we have analyzed the two-age group $SI_RI_{UR}R$ model using an age-cutoff of 60. This cutoff was justified by a discourse (Appendix A) into the role of age in both the transmission and the public health impact (e.g., hospitalizations) of COVID-19, as well as with practical considerations to data limitations. In order to assess the sensitivity of our model to this age-cutoff, we will explore the effect of alternate cutoffs (50 & 70) on fitting to reported case data.

Setting the age-cutoff at 50 and 70 partitions the population such that the older population (Group 2) constitutes 40.5% and 12.1%, respectively. The contact matrices adapted from (20) for each of these age-cutoffs is summarized in Table B.1.1. Both the age-cutoff of 50 (Fig. B.1.1) and 70 (Fig. B.1.2) showed excellent fit to the reported case data. In fact, for some months, the fit for these age-cutoffs was even greater than for the age-cutoff of 60 previously analyzed (Fig. B.1.3); however, notably, the age-cutoff of 60 produced the most consistent results.

Survey Month	**Contact Matrix ($\beta = p \cdot c$)**		
	Age-cutoff = 50	**Age-cutoff = 60**	**Age-cutoff = 70**
May	$\beta = p \begin{bmatrix} 1.9385 & 0.7115 \\ 1.0125 & 1.1951 \end{bmatrix}$	$\beta = p \begin{bmatrix} 2.1780 & 0.3860 \\ 1.2355 & 0.9599 \end{bmatrix}$	$\beta = p \begin{bmatrix} 2.2636 & 0.2040 \\ 1.5600 & 0.6900 \end{bmatrix}$
July	$\beta = p \begin{bmatrix} 1.8669 & 0.6078 \\ 0.8482 & 0.9773 \end{bmatrix}$	$\beta = p \begin{bmatrix} 1.9121 & 0.3420 \\ 1.0678 & 0.9584 \end{bmatrix}$	$\beta = p \begin{bmatrix} 2.1405 & 0.1241 \\ 0.8600 & 0.7300 \end{bmatrix}$
Sept	$\beta = p \begin{bmatrix} 3.7869 & 1.2867 \\ 1.5427 & 1.3283 \end{bmatrix}$	$\beta = p \begin{bmatrix} 4.4457 & 0.5066 \\ 1.3993 & 0.7387 \end{bmatrix}$	$\beta = p \begin{bmatrix} 4.0661 & 0.2232 \\ 1.4000 & 0.5500 \end{bmatrix}$
Dec	$\beta = p \begin{bmatrix} 2.2021 & 1.1966 \\ 1.5321 & 1.1811 \end{bmatrix}$	$\beta = p \begin{bmatrix} 2.7687 & 0.6022 \\ 1.6612 & 0.7833 \end{bmatrix}$	$\beta = p \begin{bmatrix} 3.0345 & 0.2396 \\ 1.4200 & 0.3600 \end{bmatrix}$

Table B.1.1: Contact matrices for two-age group model with age-cutoffs 50, 60, and 70. Adapted from contact data surveyed from B.C. residents in 2020 (20).

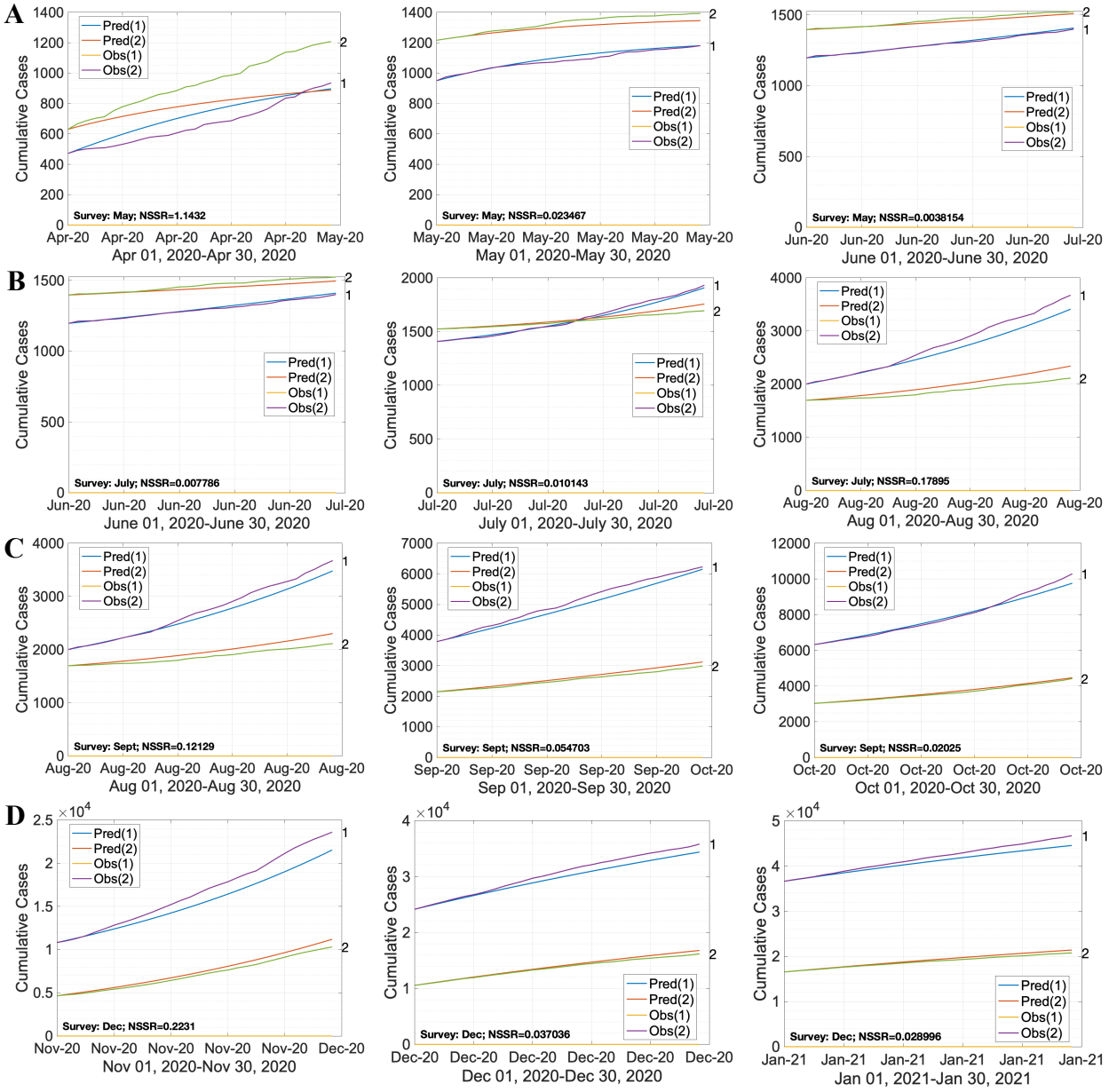

Figure B.1.1: Fitting the two-age group $SI_RI_{UR}R$ model (with age-cutoff of 50) to the reported case data using the contact matrices adapted from survey data in (20). Each survey (**A**, May; **B**, July; **C**, Sept; **D**, Dec) was fit at a 30-day scale from the month preceding the survey to the month following. Optimization conducted by minimizing the $NSSR$.

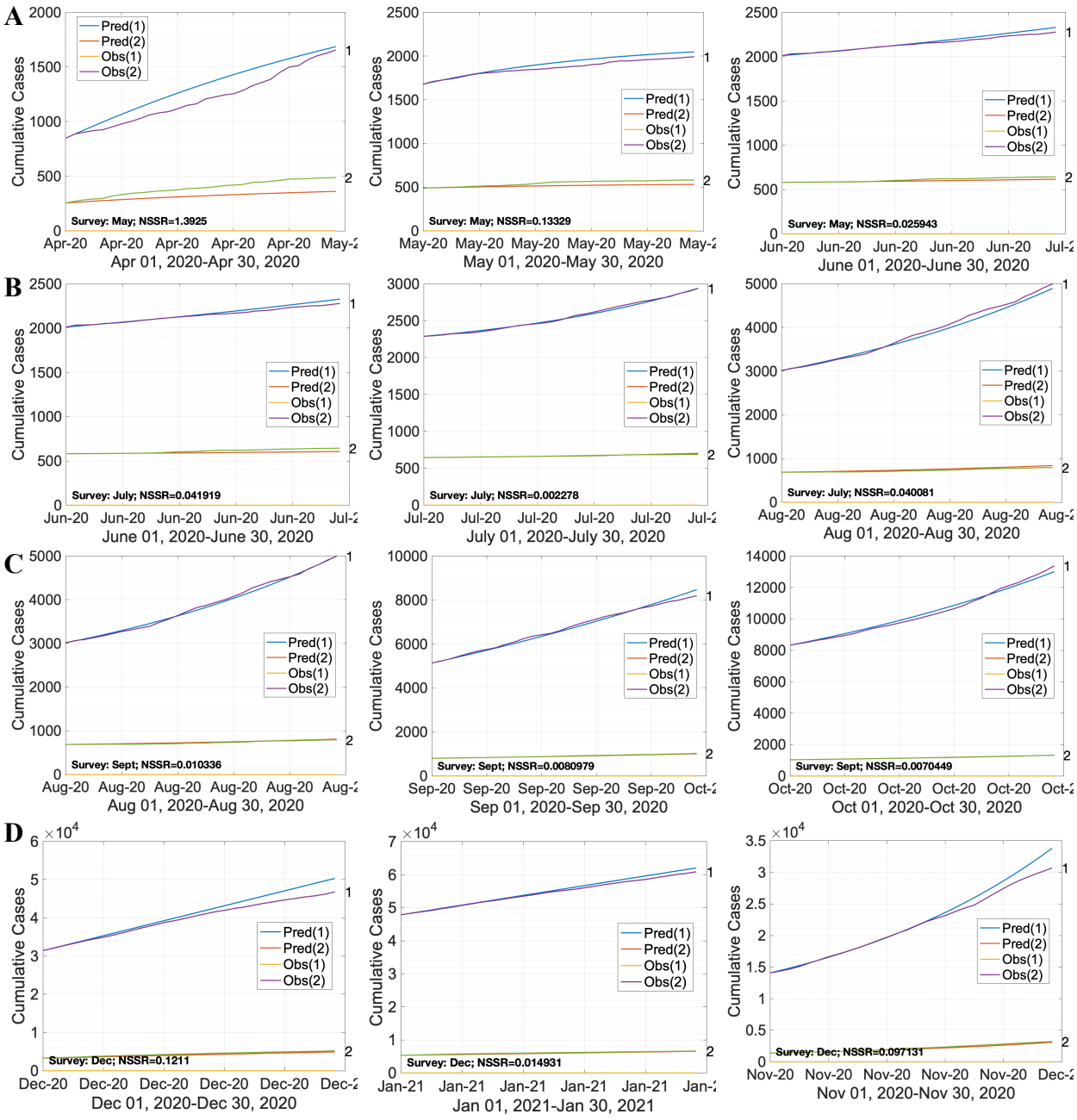

Figure B.1.2: Fitting the two-age group $SI_RI_{UR}R$ model (with age-cutoff of 70) to the reported case data using the contact matrices adapted from survey data in (20). Each survey (**A**, May; **B**, July; **C**, Sept; **D**, Dec) was fit at a 30-day scale from the month preceding the survey to the month following. Optimization conducted by minimizing the $NSSR$.

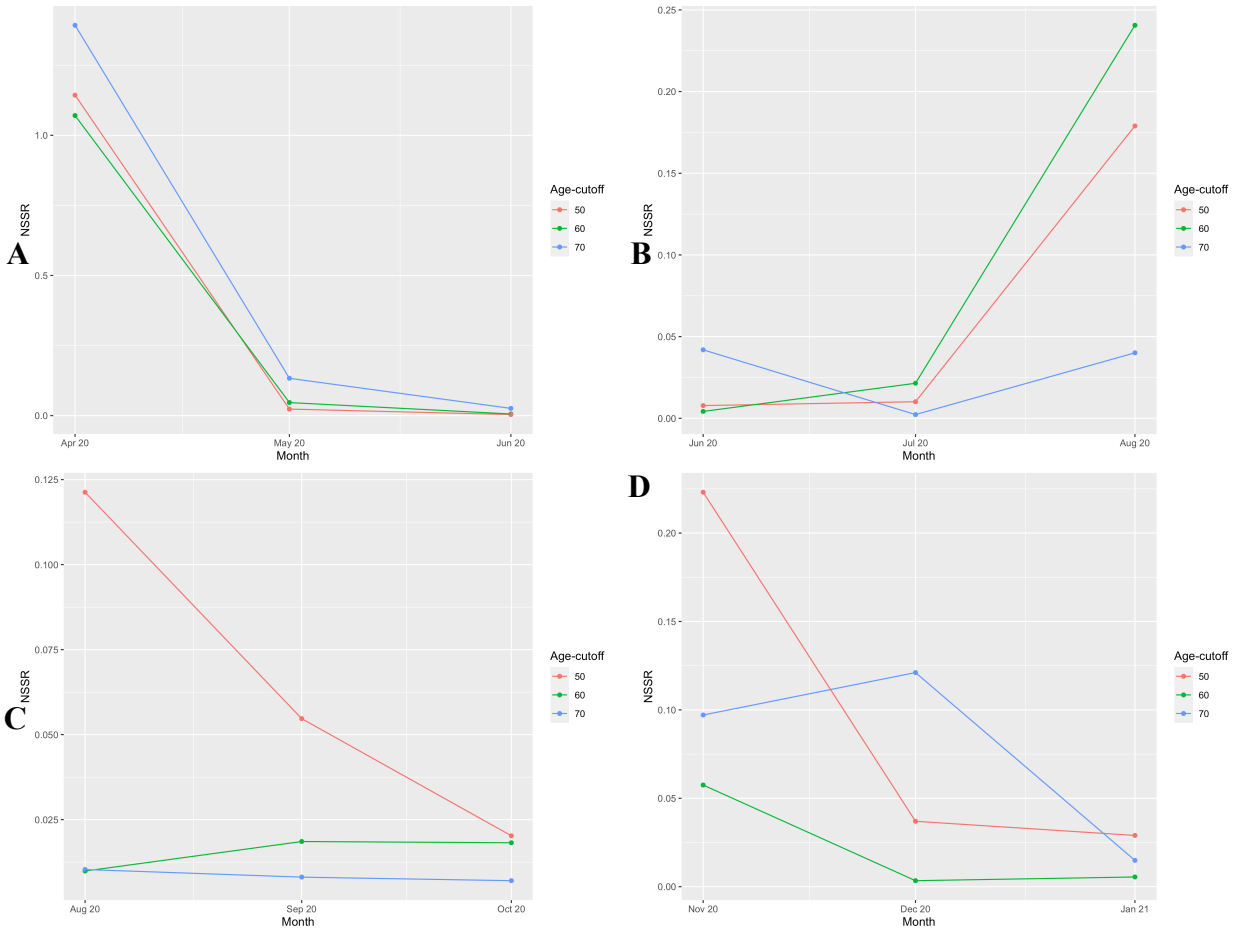

Figure B.1.3: Sensitivity of fit (as measured by the $NSSR$) to the age-cutoff selected the two-age group $SI_RI_{UR}R$ model, for each contact matrix (**A**, May; **B**, July; **C**, Sept; **D**, Dec) adapted from (20).

The different age-cutoffs were able to produce similar estimates for the probability of transmitting infection for each contact matrix (Fig. B.1.4). This is even more consequential than goodness of fit, as it confirms that the identifiability of our two-age group $SI_RI_{UR}R$ model is not influenced by our selection of age groups. This may be advantageous if we wish to fit our model to data from a region which formats age data differently than B.C.

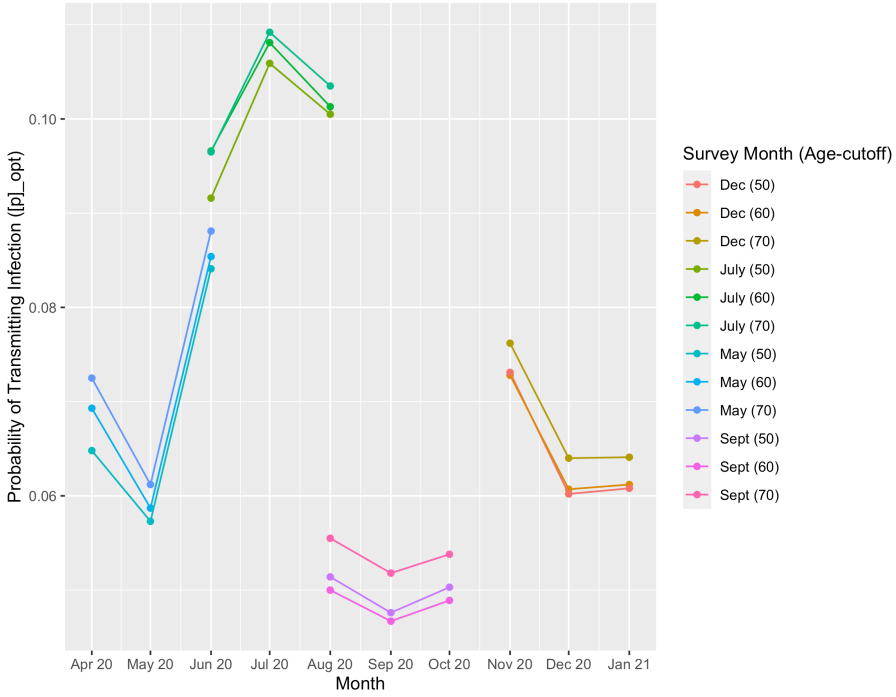

Figure B.1.4: Sensitivity of the optimal probability of infection to the age-cutoff selected for the two-age group $SI_RI_{UR}R$ model, for each contact matrix $t \in \{May, July, Sept, Dec\}$ adapted from (20). Optimization conducted by minimizing the $NSSR$.

B.2 Number of Age Groups

Structuring our model by more age groups allows us to acutely identify age-specific trends in transmission that might otherwise be lost. In addition, since the risk of severe disease with COVID-19 is so sensitive to age (17), a more granular analysis will allow us to more accurately model this risk for specific age groups, which is of great relevance to public health.

In the present section, we will explore the three-age group $SI_RI_{UR}R$ model. We will choose our age groups so as to most accurately model risk of hospitalization from COVID-19. This decision will be informed from unbiased k-means clustering analysis (similar to that conducted in Appendix A) on the per capita hospitalization (Fig. A.3) and case-hospitalization fraction (Fig. A.4) of each age cohort for the final four months of 2020. The per capita hospitalization data suggested age-cutoffs of 30 and 70, whereas the hospitalization rate data suggested age-cutoffs of 50 and 70. In the interest of balancing data sizes between the populations, we will conduct our analysis with the age-cutoffs 40 and 70, such that Group 1 (ages 0-39), Group 2 (ages 40-69), and Group 3 (ages 70+) comprise 46.2%, 41.7%, and 12.1% of the population, respectively.

We constrained our three-age group $SI_RI_{UR}R$ model using the contact data surveyed in (20), recorded in Table B.2.1, and fit it to the reported case data from the month prior to the month following each survey time-point $t \in \{May, July, Sept, Dec\}$. So as to compare the fit between different months, and different contact matrices, we optimized our model by minimizing $NSSR$.

Survey Month	Contact Matrix ($\beta = p \cdot c$)	
	Age-cutoff = 40, 70	**Age-cutoff = 50, 70**
May	$\beta = p \begin{bmatrix} 1.6634 & 0.8857 & 0.1300 \\ 0.9662 & 1.1009 & 0.2548 \\ 0.6400 & 0.9200 & 0.6900 \end{bmatrix}$	$\beta = p \begin{bmatrix} 1.9385 & 0.5487 & 0.1628 \\ 1.0579 & 0.8614 & 0.2674 \\ 0.9200 & 0.6400 & 0.6900 \end{bmatrix}$
July	$\beta = p \begin{bmatrix} 1.3834 & 0.9829 & 0.0700 \\ 0.9155 & 1.0371 & 0.1691 \\ 0.2200 & 0.6400 & 0.7300 \end{bmatrix}$	$\beta = p \begin{bmatrix} 1.8669 & 0.5235 & 0.0844 \\ 0.9904 & 0.7113 & 0.1939 \\ 0.3700 & 0.4900 & 0.7300 \end{bmatrix}$
Sept	$\beta = p \begin{bmatrix} 3.2528 & 1.7653 & 0.1066 \\ 1.8256 & 1.7781 & 0.2798 \\ 0.3600 & 1.0400 & 0.5500 \end{bmatrix}$	$\beta = p \begin{bmatrix} 3.7869 & 1.0967 & 0.1899 \\ 1.9302 & 1.1210 & 0.2644 \\ 0.7400 & 0.6600 & 0.5500 \end{bmatrix}$
Dec	$\beta = p \begin{bmatrix} 1.8494 & 1.3599 & 0.2847 \\ 1.2989 & 1.6466 & 0.2166 \\ 0.6800 & 0.7400 & 0.3600 \end{bmatrix}$	$\beta = p \begin{bmatrix} 2.2021 & 0.9505 & 0.2461 \\ 1.8047 & 1.0738 & 0.2309 \\ 0.8900 & 0.5300 & 0.3600 \end{bmatrix}$

Table B.2.1: Contact matrices for three-age group model with age-cutoffs 40 & 70 and 50 & 70. Adapted from contact data surveyed from B.C. residents in 2020 (20).

The three-age group model showed good fit to the reported case data for every month (excluding April) (Fig. B.2.1). Due to the quadratic nature of the calculation of $NSSR$ and SSR, the three-age group model will always return a higher error (for equivalent fit) than the two-age group model. However, visually speaking the fit for the three-age group model (with age-cutoffs 40 and 70) was inferior to the fit for the two-age group model (for all age-cutoffs previously tested; Fig. 7, Fig. B.1.1, Fig. B.1.2). This is unsurprising as we are requiring the optimization to fit to more curves. Nonetheless, in order to confirm that this was not an artefact of our choice for the age-cutoff, we also tested the age-cutoff combination of 50 and 70 (Fig. B.2.2; contact matrices in appendix). Although this combination improved fit for some months and worsened it for other (Fig. B.2.3), there was ultimately no superior choice for the age-cutoffs and both combinations performed comparably.

Notably, however, the probability of transmitting infection identified from the three-age group model (for both age-cutoff combinations tested) was entirely comparable to the two-age

group model (for all age-cutoffs tested) (Fig. B.2.4). This is consequential as it confirms that the identifiability of our model is not influenced by the number of age groups.

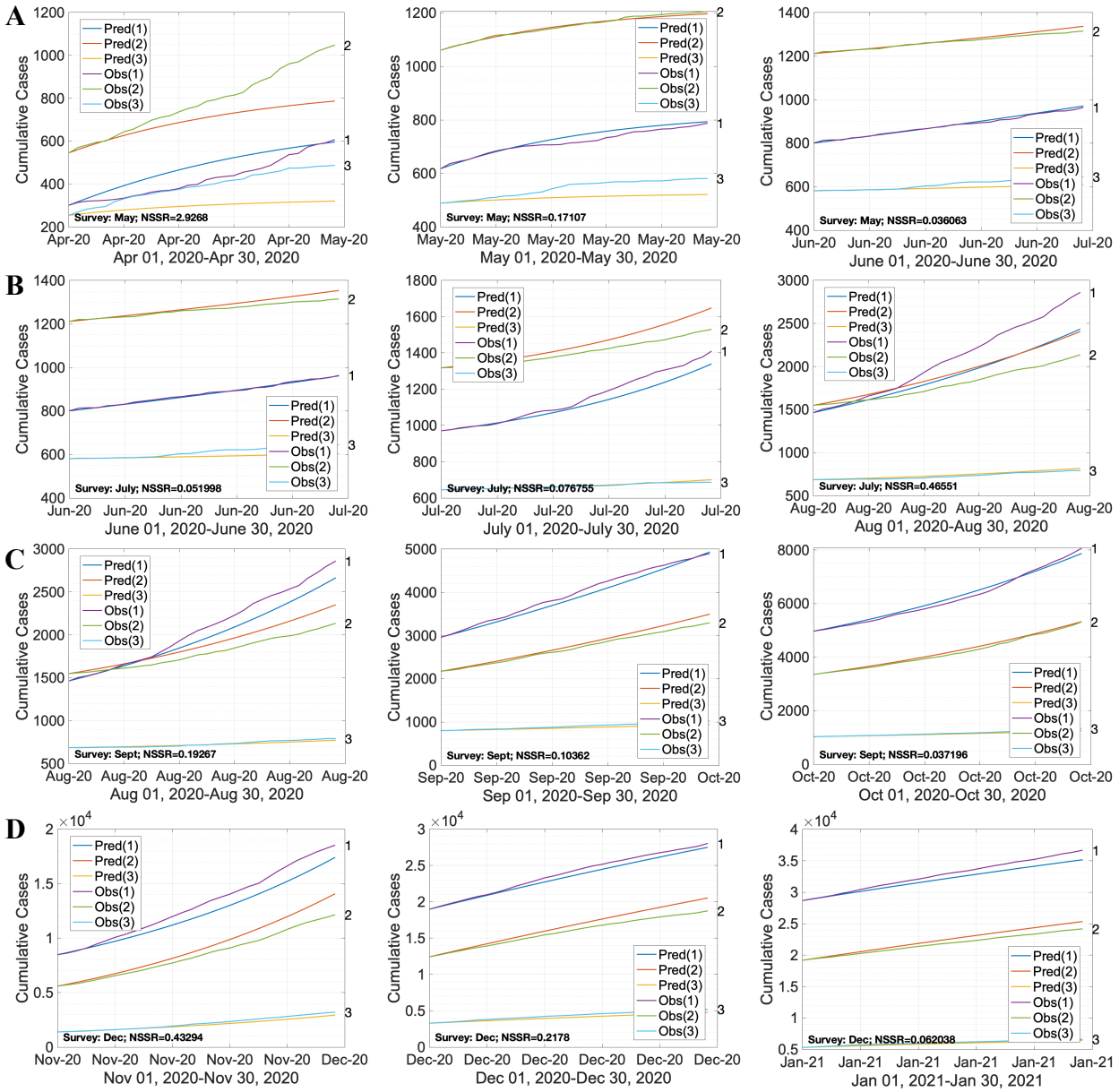

Figure B.2.1: Fitting the three-age group $SI_RI_{UR}R$ model (with age-cutoffs of 40 & 70) to the reported case data using the contact matrices adapted from survey data in (20). Each survey (**A**, May; **B**, July; **C**, Sept; **D**, Dec) was fit at a 30-day scale from the month preceding the survey to the month following. Optimization conducted by minimizing the $NSSR$.

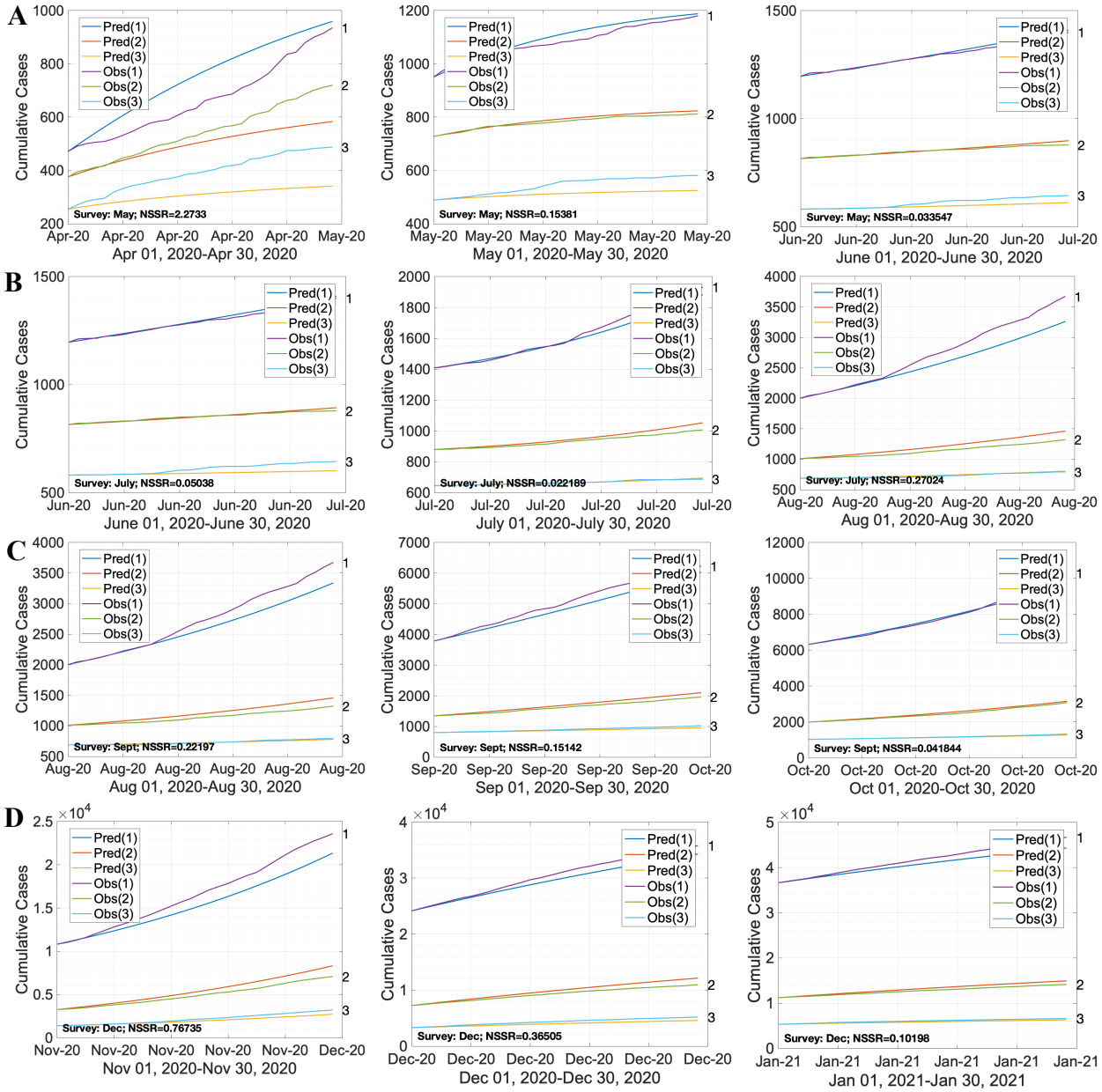

Figure B.2.2: Fitting the three-age group $SI_RI_{UR}R$ model (with age-cutoffs of 50 & 70) to the reported case data using the contact matrices adapted from survey data in (20). Each survey (**A**, May; **B**, July; **C**, Sept; **D**, Dec) was fit at a 30-day scale from the month preceding the survey to the month following. Optimization conducted by minimizing the $NSSR$.

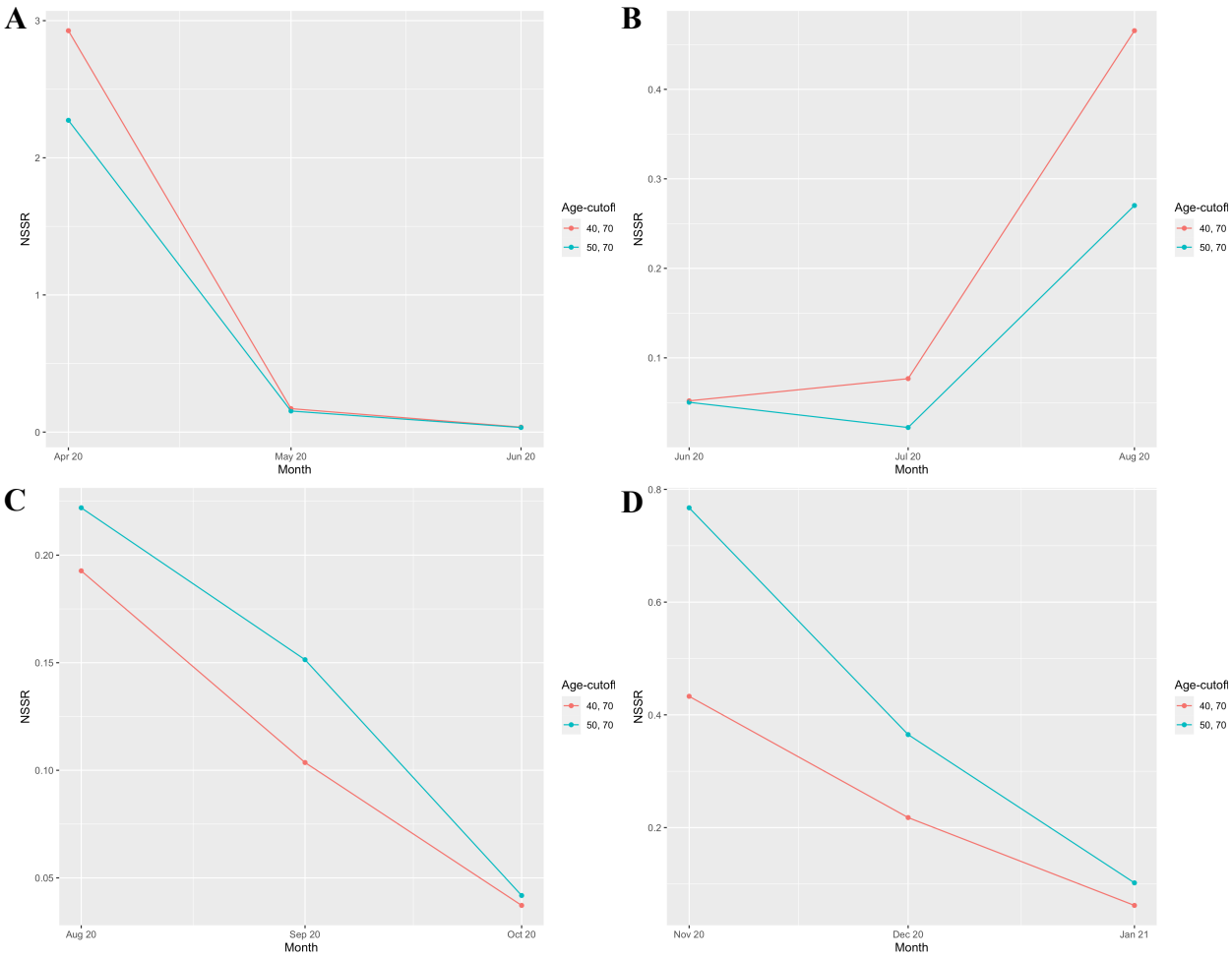

Figure B.2.3: Sensitivity of fit (as measured by the $NSSR$) to the age-cutoffs selected for the three-age group $SI_RI_{UR}R$ model, for each contact matrix (**A**, May; **B**, July; **C**, Sept; **D**, Dec) adapted from (20).

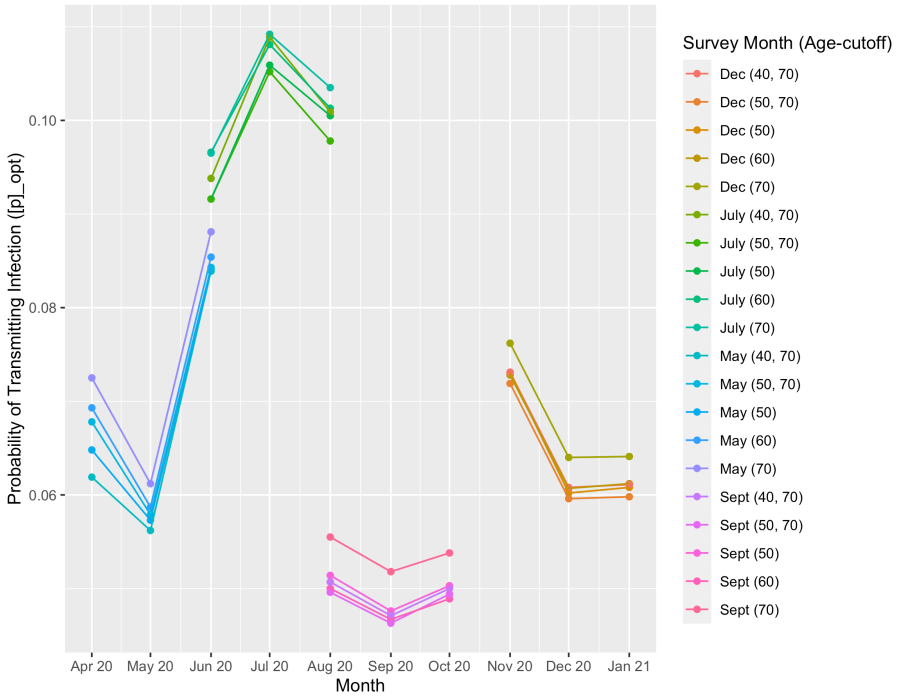

Figure B.2.4: Sensitivity of the optimal probability of infection to the age-cutoffs selected for the two- and three-age group $SI_RI_{UR}R$ model, for each contact matrix $t \in \{May, July, Sept, Dec\}$ adapted from (20). Optimization conducted by minimizing the $NSSR$.