

1 **Running Title:** Modeling to predict SARS-CoV-2 transmission dynamics

2 **Keywords:** SARS-CoV-2, COVID-19, cycle threshold, viral load, transmission, testing, public  
3 health, vaccine, variant of concern

4 **Prediction of SARS-CoV-2 transmission dynamics based on population-level cycle**  
5 **threshold values: A Machine Learning and mechanistic modeling study**

6  
7 **Authors:** Afraz A. Khan, MSc<sup>1</sup>, Hind Sbihi, PhD<sup>1,2</sup>, Michael A. Irvine, PhD<sup>1,2</sup>, Agatha N.  
8 Jassem, PhD<sup>1,3</sup>, Yayuk Joffres, MPH<sup>1,2</sup>, Braeden Klaver, MPH<sup>1,2</sup>, Naveed Janjua, MBBS,  
9 DrPH<sup>1,2</sup>, Aamir Bharmal, MD, MPH<sup>2,4</sup>, Carmen H. Ng, MSc<sup>4</sup>, Amanda Wilmer, MD<sup>5</sup>, John  
10 Galbraith, MD<sup>6</sup>, Marc G. Romney, MD<sup>3,7</sup>, Bonnie Henry, MD, MPH<sup>8</sup>, Linda M. N. Hoang, MD,  
11 MSc<sup>1,3</sup>, Mel Krajden, MD<sup>1,3</sup>, Catherine A. Hogan, MD, MSc<sup>1,3\*</sup>

12 **Affiliations:**

13 <sup>1</sup>British Columbia Centre for Disease Control, Vancouver, British Columbia, Canada

14 <sup>2</sup>School of Population and Public Health, University of British Columbia, Vancouver, British  
15 Columbia, Canada

16 <sup>3</sup>Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver,  
17 British Columbia, Canada

18 <sup>4</sup>Office of the Medical Health Officer, Fraser Health, Surrey, British Columbia, Canada

19 <sup>5</sup>Division of Medical Microbiology, Kelowna General Hospital, Kelowna, British Columbia,  
20 Canada

21 <sup>6</sup>Division of Microbiology and Molecular Diagnostics, Victoria General Hospital, Victoria,  
22 British Columbia, Canada

23 <sup>7</sup>Division of Medical Microbiology and Virology, St. Paul's Hospital, Vancouver, British

24 Columbia, Canada

25 <sup>8</sup>Ministry of Health, Victoria, British Columbia, Canada

26

27 \*Corresponding author

28 **Address for correspondence:**

29 Catherine A. Hogan

30 655 W 12th Avenue, Room 2054

31 Vancouver, BC, Canada, V6R 2M7

32 Phone (604) 802-5183

33 [catherine.hogan@bccdc.ca](mailto:catherine.hogan@bccdc.ca)

34 **Abstract** (169 words)

35 **Background:** Polymerase chain reaction (PCR) cycle threshold (Ct) values can be used to  
36 estimate the viral burden of Severe Acute Respiratory Syndrome Coronavirus type 2 (SARS-  
37 CoV-2) and predict population-level epidemic trends. We investigated the use of machine  
38 learning (ML) and epidemic transmission modeling based on Ct value distribution for SARS-  
39 CoV-2 incidence prediction during an Omicron-predominant period.

40 **Methods:** Using simulated data, we developed a ML model to predict the reproductive number  
41 based on Ct value distribution, and validated it on out-of-sample province-level data. We also  
42 developed an epidemiological model and fitted it to province-level data to accurately predict  
43 incidence.

44 **Results:** Based on simulated data, the ML model predicted the reproductive number with highest  
45 performance on out-of-sample province-level data. The epidemiological model was validated on  
46 outbreak data, and fitted to province-level data, and accurately predicted incidence.

47 **Conclusions:** These modeling approaches can complement traditional surveillance, especially  
48 when diagnostic testing practices change over time. The models can be tailored to different  
49 epidemiological settings and used in real time to guide public health interventions.

50

51 **Funding:** This work was supported by funding from Genome BC, Michael Smith Foundation for  
52 Health Research and British Columbia Centre for Disease Control Foundation to C.A.H. This  
53 work was also funded by the Public Health Agency of Canada COVID-19 Immunity Task Force  
54 COVID-19 Hot Spots Competition Grant (2021-HQ-000120) to M.G.R.

55 *Introduction*

56 SARS-CoV-2 viral burden can be quantitated by the use of polymerase chain reaction (PCR)  
57 cycle threshold (Ct) values, which are inversely proportional to the amount of target viral  
58 sequence present in the patient sample. Although this information is frequently available from  
59 routine molecular methods for the diagnosis of SARS-CoV-2 infection, clinical results are  
60 almost universally reported qualitatively as present or absent due to sources of sampling  
61 variability, lack of inter-test standardization, insufficient supporting clinical correlation data, and  
62 lack of regulatory approval for purposes other than qualitative reporting, all of which limit  
63 interpretation of Ct values for clinical care. Though the use of Ct values to guide individual-level  
64 management is not currently routinely recommended (1, 2), the assessment of aggregated Ct  
65 values at a population level may be useful to help assess early epidemiological transmission  
66 trends to improve epidemic forecasting (3-5), and parallels the concept of measuring community  
67 viral load used for other viruses (6-8). Accurate projection of epidemic trends is critical to  
68 effectively plan public health efforts including healthcare resource allocation. Indeed, an  
69 epidemic in the growth phase is more likely to be associated with high viral load burden at a  
70 population level; conversely, the decline phase of an epidemic is likely to demonstrate lower  
71 viral burden. A modeling approach was previously published to inform epidemic SARS-CoV-2  
72 trajectory based on aggregated Ct value data (3), and supported the usefulness of population-  
73 level Ct value analysis. However, SARS-CoV-2 testing practices globally have evolved  
74 substantially during the pandemic, most frequently by restricting testing to symptomatic  
75 individuals, which limits the usefulness of modeling approaches that rely on stable population  
76 sampling strategies. Starting in December 2021 in British Columbia (BC), use of PCR testing  
77 was partially restricted in the context of roll-out of rapid antigen tests, limiting understanding of

78 population trends. New tools are needed to estimate incidence in a manner that is independent of  
79 the biases associated with testing guidance. This includes modeling approaches robust to varying  
80 testing guidelines, sample selection strategies and epidemiologic settings, and that account for  
81 other variables that impact viral burden such as variant of concern (VoC) and vaccination status.  
82 The main variants of concern (VoC) described to date have been associated with varying impact  
83 on viral burden, with the Delta and certain Omicron subvariants associated with highest viral  
84 load (9-14). Furthermore, evidence suggests that SARS-CoV-2 vaccination is associated with a  
85 reduction in viral burden, and correspondingly higher Ct values and potentially lower  
86 transmission risk, in individuals who develop post-vaccination infection (14-19). In this study,  
87 we investigated two modeling approaches based on Ct value distribution of asymptomatic  
88 individuals, machine learning and epidemic transmission modeling, to predict SARS-CoV-2  
89 incidence based on province-wide data and an outbreak in a long-term care facility in British  
90 Columbia, Canada. We assessed the novel application of five machine learning models (Lasso,  
91 LGBM, XGBoost, CatBoost, RF), and validated two previously-described epidemic models  
92 (SEIR) (3), to determine the highest performing models across a range of epidemiological  
93 settings to predict SARS-CoV-2 incidence.

94

## 95 *Methods*

### 96 **Study design**

97 Individuals with polymerase chain reaction (PCR)-confirmed SARS-CoV-2 infection by  
98 nasopharyngeal swab or saline gargle between November 19<sup>th</sup> 2021 and January 8<sup>th</sup> 2022 were  
99 included, capturing emergence of Omicron wave in the province. Descriptive analyses of Ct  
100 value distribution included the two main specimen type categories: nasopharyngeal swabs and

101 saline gargles, while modeling analyses focused on nasopharyngeal swabs given the higher  
102 diagnostic yield and collection standardization. Analysis was based on province-wide data  
103 incorporating all SARS-CoV-2 diagnostic tests based on the *E* gene target performed in BC.  
104 Three pandemic phases in BC were considered based on vaccination roll-out and VoC  
105 distribution (**Supplemental Table 1**). To capture the largest representation of asymptomatic  
106 individuals in BC throughout the pandemic, the study focused on phase 3. These individuals  
107 were tested in the context of occupational screening or pre-travel. The study population sampled  
108 thus represented a heterogeneous mix of vaccinated and unvaccinated individuals, and  
109 predominantly Omicron (BA.1) variant (**Figures 1A and 1B**).

110  
111 Given the importance of population composition in informing model selection, the selected  
112 models were chosen based on several factors including sampling type and frequency, sample  
113 size, and computational complexity. Current models which make use of cross-sectional Ct values  
114 to infer epidemic trajectories (3) rely on random sampling of the population to accurately predict  
115 epidemic trends. However, in the context of symptom-based testing, the distribution of Ct values  
116 is not a complete representation of the infected population. Thus, this study was performed on a  
117 population of asymptomatic individuals as they served as the best proxy for frequent non-  
118 symptom-based sampling.

119  
120 **Testing practices and public health measures**  
121 COVID-19 diagnostic testing and laboratory test guidelines changed in BC over the course of the  
122 pandemic, and can be summarized as follows: 1) exposure-based testing (onset of the pandemic),  
123 2) targeted testing (starting March 16, 2020), 3) expanded testing (starting April 9, 2020), 4)

124 symptom-based testing (starting April 21, 2020), 5) revised symptom-based testing (starting  
125 December 17, 2020), and 6) High risk/targeted population testing (increased risk of severe  
126 disease or work in a high-risk setting) (starting January 18, 2021) (20, 21). Thus, asymptomatic  
127 and mildly symptomatic testing was initiated starting in December 2021 with the organized roll-  
128 out of rapid antigen tests, which corresponds with the current study. While COVID-19 testing  
129 was initially centralized at the BCCDC Public Health Laboratory (PHL), testing capacity and  
130 data capture reflects results from all provincial testing laboratories which generate *E* gene Ct  
131 values.

132

### 133 **Laboratory data - SARS-CoV-2 diagnostic testing**

134 SARS-CoV-2 diagnostic testing was performed in laboratories throughout all five health  
135 authorities in BC, and only assays based on the *E* gene target were included for this study. The  
136 testing strategy and test result interpretation criteria used for the participating laboratories are  
137 described separately (**Supplemental Table 2**), and included the BCCDC PHL laboratory-  
138 developed test (LDT) (22), LightMix SarbecoV E-gene plus EAV control assay (TIB Molbiol,  
139 Berlin, Germany), Alinity m SARS-CoV-2 (Abbott, Chicago, IL), BD SARS-CoV-2 (Becton,  
140 Dickinson and Co., Franklin Lakes, NJ), cobas 6800 and 8800 (Roche Diagnostics, Basel  
141 Switzerland), GeneXpert Xpress SARS-CoV-2 (Cepheid, Sunnyvale, CA), Panther Fusion  
142 (Hologic, San Diego, CA) and Allplex SARS-CoV-2 (Seegene, Seoul, South Korea). For  
143 individuals having undergone repeat SARS-CoV-2 testing within a one-week period, only the  
144 first positive test per person was included.

145

### 146 **Laboratory data - Variant of Concern identification**

147 The BCCDC Public Health Laboratory (PHL) continuously monitors for variants of concern  
148 (VOCs), variants of interest (VOIs), and variants under monitoring (VUMs). Various approaches  
149 were used over time including VoC screening and confirmation by whole genome sequencing  
150 (WGS) when applicable at the BCCDC PHL as previously described (22). Testing strategy is  
151 optimized based on available capacity and clinical and public health needs, and changed over the  
152 course of the SARS-CoV-2 pandemic. One such strategy included deployment of in brief, a  
153 subset of samples in the earlier phase of the epidemic (January 2021 to May 2021) was tested by  
154 targeted single nucleotide polymorphism (SNP) quantitative polymerase chain reaction (qPCR)  
155 for VoC screening, followed by confirmation by WGS. From June 2021 onward, sample VoC  
156 status was detected by WGS alone. From September 2021, owing to increased case burden and  
157 limited capacity, there was a transition from WGS of all samples to a subset positive SARS-  
158 CoV-2 samples. This subset comprised of targeted surveillance (cases from outbreaks, vaccine  
159 escape, reinfection and travel-related), and representative baseline surveillance. In addition,  
160 100% of positive samples underwent WGS in the first week of each month. Starting November  
161 15 2021 in the context of the Omicron variant emergence, WGS was resumed for all samples.  
162 Owing to the high transmissibility of Omicron and the surge in case load, starting December 21  
163 2021, there was transition from full sequencing to sequencing a subset of representative positive  
164 samples in addition to priority cases (including outbreaks, long-term care, vaccine escape, travel-  
165 related, hospitalization)). Full VoC characterization for the province of BC is described  
166 separately (**Supplemental Figure 2**).

167

168 **Vaccination status**



169 Vaccination status was defined based on the date of vaccine receipt relative to the date of the  
170 sample collection included for the study (**Supplemental Figure 3**) (23). For primary dose series  
171 all mRNA (Pfizer, Moderna) and viral vector vaccines (AstraZeneca, Janssen) were considered.  
172 For the Janssen vaccine only, fully vaccinated status was defined as having received one dose 14  
173 days or more prior to sample collection. For all other vaccines, **Unvaccinated status** was defined  
174 as having received no SARS-CoV-2 vaccine, or having received a SARS-CoV-2 vaccine less  
175 than 21 days prior to the sample collection date. **Partially vaccinated** status was defined as  
176 having received the SARS-CoV-2 vaccine dose 1 greater or equal to 21 days prior to sample  
177 collection, but having received dose 2 less than 14 days prior to the sample collection. **Fully**  
178 **vaccinated** status was defined as greater or equal to 14 days since the receipt of dose 2, but  
179 having received dose 3 less than 14 days prior to the sample collection. Cross-over vaccination  
180 was considered in the same category as homologous vaccine schedules.

181

## 182 **Outbreak case study**

183 To further validate the models, a separate analysis was performed using a well-characterized  
184 outbreak in a long-term care facility that occurred in BC. This outbreak was selected on the  
185 basis of time of occurrence of pre-vaccination roll-out to the general population, large size and  
186 generalizability of the affected population. This outbreak included large-scale asymptomatic  
187 testing. Testing was done weekly until no additional cases were identified within 14 days of the  
188 last exposure. There were 7 rounds of weekly testing at the outbreak facility, all negative  
189 residents and staff were tested for each round. Anyone who developed symptoms was also tested.  
190 The epidemiologic data and curve describing the outbreak are presented separately  
191 (**Supplemental Figure 4**). As for the main study, analysis was based on SARS-CoV-2

192 diagnostic tests based on the *E* gene target. However, due to missing data in the long-term care  
193 facility data, wherever the *E* gene target was unavailable the *ORF1* gene target was used instead.

194

## 195 **Data sources**

196 Two main data sources were employed for this study: 1) the Provincial Health Laboratory  
197 Viewer and Reporter (PLOVER) database which includes the laboratory diagnostic datasets, and  
198 2) the Provincial Immunization Registry (PIR) dataset which includes vaccination data. The  
199 laboratory datasets house data on SARS-CoV-2 testing (including date of collection, specimen  
200 type, diagnostic quantitative PCR gene target results, VoC screening, and SARS-CoV-2 lineage  
201 based on WGS), and individual-level epidemiological data (including age, sex, patient as well as  
202 ordering physician health authority). Gene target results include Ct values of the *E* and ORF1  
203 targets, and the internal control RNaseP. The PIR dataset includes individual-level vaccination  
204 data (including number, type, series, dose and date of each vaccine received). Both of these  
205 datasets form the basis of the covariates which inform the ML models in the study. For the  
206 outbreak case study, additional data were directly gathered from public health partners  
207 (**Supplemental Figure 4**) as these were not otherwise available through provincial datasets. Data  
208 linkages were performed between the laboratory and PIR datasets through a sequential  
209 deterministic linkage based on a minimum of three personal identifiers (personal health number,  
210 last name with first three digits of first name, and date of birth). These linkages were performed  
211 prospectively on a weekly basis, and specimens with unsuccessful linkages were excluded from  
212 the study.

213

## 214 **Data & Code Availability**

215 The genomic sequencing data are publicly available in GISAID under the submitter British  
216 Columbia Center for Disease Control Public Health Laboratory (BCCDC PHL). The individual  
217 level demographic and epidemiological data can be made accessible following the data  
218 governance and data access policy guidelines ([http://www.bccdc.ca/about/accountability/data-](http://www.bccdc.ca/about/accountability/data-access-requests)  
219 [access-requests](http://www.bccdc.ca/about/accountability/data-access-requests)). Code for this study is available (<https://github.com/Afraz496/Vital-E-paper>).

220

## 221 **Ethics**

222 This research was approved by University of British Columbia Research Ethics (H20-0297  
223 BCC19C-COVID-19 Research).

224

## 225 **Models**

### 226 **Machine learning and epidemic transmission models**

227 This study compared two different approaches for inference, machine learning (ML) and  
228 epidemic transmission modeling, both of which were used to predict the reproductive number  
229 ( $R_t$ ) to estimate incidence. The first modeling approach was based on a collection of ML  
230 approaches, including Lasso (24), Random Forest (RF), Light Gradient Boosting Modeling  
231 (LGBM), eXtreme Gradient Boosted Modeling (XGBM) and CatBoost. Due to the testing  
232 guidelines which were tailored for mainly symptomatic individuals, and given the absence of  
233 sufficiently-large random samples, analysis was pursued with simulated data instead. Ct data  
234 were generated to simulate a sufficiently large random sample of a population using the  
235 virosolver package (R software, version 4.1.2). This was applied on varying sample sizes (100,  
236 1000 and 10000) on a simulated population of 50,000 individuals. The simulation horizon was  
237 set at 140 calendar days to encompass a typical single COVID-19 wave. These data were used to  
238 create summary statistics of the Ct distribution including mean, median, variance, skewness and

239 kurtosis. The trained data was generated from a unique simulation file with a fixed random seed  
240 and three distinct sample sizes, so three models were investigated in this study. Hyperparameter  
241 tuning was performed via a grid search of hyperparameters on each model (**Supplemental Table**  
242 **5**). The best performing model was chosen by finding the optimal set of hyperparameters for  
243 which the Mean squared error (MSE) between the true simulated  $R_t$  and Predicted simulated  $R_t$   
244 was minimized. SHapley Additive exPlanation (SHAP) analysis was performed for feature  
245 ranking and importance (25). The second modeling approach was adapted from an existing  
246 methodology (3), and is based on a single epidemic model. The compartmental SEIR model  
247 captures different stages in individual infections (namely **Susceptible**, **Exposed**, **Infectious** and  
248 **Recovered**). The SEIR model was validated on a patient outbreak facility in BC where point  
249 prevalence testing was done in infrequent intervals. The SEIR model was then fitted to  
250 provincial data from asymptomatic individuals. Modifications to the viral kinetics for the SEIR  
251 model were applied to these provincial data to account for the specific nature of the Omicron  
252 (BA.1) variant.

253

## 254 *Results*

### 255 Cohort description

256 During the study period, a total of 331,785 SARS-CoV-2 tests were performed in BC, of which  
257 79,443 were positive. Restricting these to the first positive test per person, there were 71,642  
258 included in the study (**Figures 1A and 1B**). Of these, 35,369 were nasopharyngeal specimens  
259 and 36,108 were saline gargle specimens (**Table 1 and Figure 2**). The cohort was predominantly  
260 composed of adults aged 18-59 years (72.2%), followed by adults aged 60 years and above  
261 (12.1%), and children 0-17 years (15.7%). Over half the cases resided in two of the five health  
262 authorities accounting for 35.9% and 30.9%, respectively. The Omicron (BA.1) variant

263 predominated throughout the study (**Table 1 and Supplemental Figure 2**). By the end of the  
264 study period, a total of 18,459 (27.3%) were unvaccinated, 1,540 (2.3%) had received 1 dose of  
265 vaccine, and 47,580 (70%) were fully vaccinated.

266

### 267 **First modeling approach: machine learning**

268 The fitted ML models were applied to out-of-sample Ct data from the simulated Ct values  
269 (**Figure 3**). With increasing sample sizes, the MSE across each model reduced by 82% showing  
270 an increased ability in higher moments (mean, median, variance, skewness and kurtosis) of the  
271 Ct distribution to predict epidemic trends. Random Forest showed the largest improvement in  
272 MSE performance while demonstrating lowest performance in smaller sample size. Besides the  
273 smallest sample size, all models generally perform similarly across an increased sample size  
274 (**Figure 3**). For the largest sample size, apart from Lasso all other models have a much tighter  
275 IQR and smaller MSE median score (at around 0.03). Across all sample sizes, the variance of the  
276 Ct distribution was the top ranking feature (**Supplemental Figure 5**).

277

### 278 **Second modeling approach: epidemic transmission models**

#### 279 **SEIR**

280 The most precise results were observed with sampling from a total of five horizons. The model  
281 posteriors indicated an incidence peak from December 27 2021 to January 1 2022, which  
282 overlapped with the observed peak of reported cases in the province (**Figure 4**). Similarly, the  
283 exponential growth phase coincided with the increase in reported cases from our cohort from  
284 December 20 2021 to December 27 2021, and the decline of the incidence coincided with the  
285 decline in cohort cases from January 1 2022 to January 5 2022. The posterior predictive Ct

286 distribution also closely matched the observed Ct distribution on each of the time horizons,  
287 supporting accurate incidence projection independent of biases of testing guidance.

### 288 **Outbreak case study**

289 This outbreak occurred in a long-term care facility, and resulted in a total of 156 individuals (93  
290 residents and 63 staff) infected with SARS-CoV-2 (**Supplemental Figure 4**). Of these  
291 individuals, 58.1% of infections were asymptomatic in the residents, whereas 9.5% were  
292 asymptomatic within the staff. There were 26 (28.0%) deaths in the residents group, and no  
293 deaths among the staff. A multiple cross-section SEIR model was fitted to the outbreak data, and  
294 showed a peak in incidence on the 12th day of the outbreak which preceded by two days the  
295 observed peak at the outbreak facility (**Figure 5**). The real incidence fell within the 95% credible  
296 interval of the predicted MCMC chains of the SEIR model. The model also accurately predicted  
297 the decline in cases by the 20th day of the outbreak (**Figure 5**).

298

### 299 *Discussion*

300 In this study, we demonstrated the utility of two distinct modeling approaches based on  
301 aggregated cycle threshold values, machine learning and epidemic transmission modeling, to  
302 predict epidemic trends across varying sampled patient populations, random, and targeted and  
303 non-random testing. Based on out-of-sample mean squared error (MSE) change in the  
304 reproduction number between the true and predicted values, the ML model performed best on  
305 randomly-sampled province-level data. Within epidemic transmission models, the SEIR model  
306 performed highest with randomly-sampled outbreak data. Taken together, these approaches  
307 accurately predicted epidemic transmission dynamics at the outbreak case study level, and at a  
308 provincial level for the province of BC, Canada. Early in the pandemic, SARS-CoV-2 diagnostic

309 molecular testing was more largely based on random sampling which, despite possible  
310 underascertainment due to lack in testing access, could be used to estimate full case counts to  
311 monitor and predict transmission dynamics. As testing needs overwhelmed laboratory capacity  
312 with increasing case burden and the emergence of variants of concern, molecular testing practice  
313 recommendations shifted to testing individuals who were symptomatic and/or with a minimal  
314 illness severity, resulting in sampling of a selected population. These changes in testing  
315 indications, foremost predicated on symptom-based testing, led to substantially more limited  
316 capacity to assess case counts for epidemic monitoring, generating a critical unmet need for other  
317 approaches to infer epidemic trends to support clinical and public health planning.

318  
319 This study comprehensively investigated varying sampling types and modeling approaches,  
320 drawing on both previously-published work and description of the novel application of machine  
321 learning modeling for SARS-CoV-2 transmission dynamics prediction. Our work identified that  
322 diagnostic testing indication, sampling type, and the individual population tested are critical  
323 factors, and that model selection must be tailored to the epidemiological circumstances of  
324 testing. More specifically, random vs targeted or non-random sampling must be accounted for to  
325 ensure appropriate model selection, as SEIR modeling was only suitable for random sampling.  
326 For example, performance of the SEIR model in this study was robust across sample sizes and  
327 the long-term care facility dataset. Results from this modeling approach demonstrated a slight  
328 difference in incidence peak timing and amplitude. This is likely explained due to the lag time  
329 between onset of the infectious period and reporting given that site-wide facility testing was  
330 performed at set time periods rather than on a daily basis, and represents a pragmatic approach to  
331 real-world settings. Thus, this approach is well suited for long-term care or assisted living or

332 community-living facility outbreak investigations such as shelters, or within small hospital  
333 systems. In contrast, the novel application of machine learning approaches described in this  
334 study performed the best with large datasets (>1,000 COVID-19 positive cases), making this the  
335 approach of choice for large population settings such as at the province, state or large hospital  
336 network system level. Indeed, machine learning models can offer greater flexibility by  
337 incorporating different summary statistics and other data as features, fully harnessing the  
338 potential of larger datasets.

339  
340 Importantly, all approaches described in this study could predict future trends within a one to  
341 four-week timeframe, demonstrating utility for timely prediction of SARS-CoV-2 transmission  
342 dynamics that could be harnessed to help inform future outbreak resource allocation and  
343 decision-making. Thus, use of these models can be used to support critical decision-making  
344 across several settings, including hospitals, long-term care facilities, public health departments  
345 and others, to help inform planning of resource allocation, vaccination efforts, and isolation  
346 practices. More specifically, this approach lays the groundwork for a sentinel surveillance  
347 monitoring strategy that could be automated and alert appropriate authorities at pre-determined  
348 signals of predicted incidence changes, and may be expanded to other infections for which  
349 testing is widespread and predictive tools are needed.

350  
351 This study focused on a time period of Omicron (BA.1) predominance, and revealed that despite  
352 its shorter incubation period compared to other variants of concern, the Ct distribution of this  
353 variant could successfully be described through an SEIR compartmental model and machine  
354 learning approaches. Furthermore, in the context of a sampled population with heterogeneous



355 vaccination status, the current study demonstrated accurate prediction of incidence based on  
356 overall Ct distribution and viral kinetics without incorporating individual-level vaccination  
357 status. Further work is necessary to study the impact of vaccination status on accuracy of  
358 incidence prediction.

359  
360 The main strength of this study is that it provides a comprehensive modeling toolkit that can be  
361 leveraged across population and sampling settings, and that may incorporate covariates such as  
362 variant of concern and vaccination status. This approach could predict transmission dynamics in  
363 a way that could not be performed through case count analysis from biased sampling as was  
364 occurring in the province of BC. This modeling is also advantageous as it can be performed in  
365 real-time, rather than rely on monitoring of clinical indicators of severity such as hospitalization  
366 and intensive care unit admission which considerably lag behind true incidence rise. A limitation  
367 of previous studies is the use of a single or limited methodology for analysis that may perform  
368 well in a specific setting such as long-term care facilities, but lacked flexibility and predictive  
369 performance for generalizability to larger settings and in the context of changing testing practices  
370 (3). Our body of work filled this gap and further presented a methodology to incorporate  
371 assessment of variant of concern and vaccination status, two important potential confounders on  
372 Ct value distribution, although these characteristics were noted to be less important than the  
373 moments of infection (mean, median, variance, skewness and kurtosis of the aggregated Ct  
374 distribution). Additional strengths of this study also include the independent assessment of the  
375 models in a long-term care facility outbreak to validate the previously-published models (3).  
376 Furthermore, the main analysis drew on a provincial dataset linking laboratory data and  
377 vaccination status in real-time, thus leveraging the design for the highest possible public health

378 uptake and impact. Taken together, this approach lays the framework for expansion to use for  
379 other pathogens for which surveillance needs are critical including other respiratory pathogens  
380 and monkeypox. Indeed, further work may also build on this approach and further integrate  
381 complementary datasets including wastewater Ct distribution to further enhance prediction  
382 ability.

383  
384 However, there are several limitations. Firstly, the methodology is based on the assumption of  
385 random or random sampling which is challenging to confirm. Indeed, testing practices were  
386 modified following clinical and public health guidance of the province, and may have led to bias  
387 in sampling. Restriction of the study population to the asymptomatic subgroup consisting of  
388 travelers and occupational health testing led to greater confidence in the employed sampling  
389 strategy tested and the validity of this assumption. The need for random sampling remains a  
390 limitation for broader uptake of this approach, though it may be more attainable in the context of  
391 outbreak investigation where full populations are sampled at once. Nonetheless, even when a full  
392 population is sampled there may be specific population-level characteristics that need to be  
393 accounted for. One such limitation in the current work is that although the long-term care  
394 environment provides more a consistent testing environment, it tends to be a highly vaccinated  
395 population which may introduce bias. Finally, this study aggregated Ct-level data across multiple  
396 laboratories and assays, which may not adequately capture intra- and inter-assay variation.

397  
398 In summary, this study proposes a comprehensive suite of modeling strategies based on  
399 population-level Ct values to accurately predict SARS-CoV-2 transmission dynamics across  
400 epidemiological settings. These modeling approaches can be used in real time to guide clinical

401 and public health interventions. Such tools are needed to estimate incidence in a manner that is  
402 independent of the biases associated with testing guidance, and to complement traditional  
403 surveillance based on case numbers or clinical indicators. Further work will be needed to expand  
404 validation of the machine learning models based on larger datasets and different settings with  
405 newly-emerging variants, to assess real-time predictive power for direct clinical and public  
406 health impact.

407

408

409 **Funding:** This work was supported by funding by Genome BC, Michael Smith Foundation for  
410 Health Research and British Columbia Centre for Disease Control Foundation to C.A.H. This  
411 work was also funded by the Public Health Agency of Canada *COVID-19 Immunity Task Force*  
412 *COVID-19 Hot Spots Competition Grant (2021-HQ-000120)* to M.G.R.

413 **Acknowledgments:** We thank the laboratory teams (virology, bacteriology and molecular) at the  
414 BCCDC Public Health Laboratory for their contribution toward testing, on which this research is  
415 based. We also thank the data analytics team for supporting the data infrastructure and review  
416 that enabled this work. Finally, we thank the British Columbia Association of Medical  
417 Microbiologists for testing and sharing samples and data that enabled province-wide data  
418 collection, and public health partners throughout the province for their dedicated effort to  
419 outbreak management and infection control, and for sharing outbreak-level data that supported  
420 this research.

421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466

## References

1. Infectious Diseases Society of America and Association of Molecular Pathology. 2021. IDSA and AMP joint statement on the use of SARS-CoV-2 PCR cycle threshold (Ct) values for clinical decision-making. <https://www.idsociety.org/globalassets/idsa/public-health/covid-19/idsa-amp-statement.pdf>. Accessed August 19 2022.
2. American Association for Clinical Chemistry. 2021. AACC Recommendation for Reporting SARS-CoV-2 Cycle Threshold (CT) Values. <https://www.aacc.org/science-and-research/covid-19-resources/statements-on-covid-19-testing/aacc-recommendation-for-reporting-sars-cov-2-cycle-threshold-ct-values>. Accessed August 19 2022.
3. Hay JA, Kennedy-Shaffer L, Kanjilal S, Lennon NJ, Gabriel SB, Lipsitch M, Mina MJ. 2021. Estimating epidemiologic dynamics from cross-sectional viral load distributions. medRxiv doi:10.1101/2020.10.08.20204222:2020.10.08.20204222.
4. Walker AS, Pritchard E, House T, Robotham JV, Birrell PJ, Bell I, Bell JI, Newton JN, Farrar J, Diamond I, Studley R, Hay J, Vihta KD, Peto TE, Stoesser N, Matthews PC, Eyre DW, Pouwels KB, team C-IS. 2021. Ct threshold values, a proxy for viral load in community SARS-CoV-2 cases, demonstrate wide variation across populations and over time. *Elife* 10.
5. Phillips MC, Quintero D, Wald-Dickler N, Holtom P, Butler-Wu SM. 2022. SARS-CoV-2 cycle threshold (Ct) values predict future COVID-19 cases. *J Clin Virol* 150-151:105153.
6. Herbeck J, Tanser F. 2016. Community viral load as an index of HIV transmission potential. *Lancet HIV* 3:e152-4.
7. Das M, Chu PL, Santos GM, Scheer S, Vittinghoff E, McFarland W, Colfax GN. 2010. Decreases in community viral load are accompanied by reductions in new HIV infections in San Francisco. *PLoS One* 5:e11068.
8. Jordan AE, Perlman DC, Cleland CM, Wyka K, Schackman BR, Nash D. 2020. Community viral load and hepatitis C virus infection: Community viral load measures to aid public health treatment efforts and program evaluation. *J Clin Virol* 124:104285.
9. Wang Y, Chen R, Hu F, Lan Y, Yang Z, Zhan C, Shi J, Deng X, Jiang M, Zhong S, Liao B, Deng K, Tang J, Guo L, Jiang M, Fan Q, Li M, Liu J, Shi Y, Deng X, Xiao X, Kang M, Li Y, Guan W, Li Y, Li S, Li F, Zhong N, Tang X. 2021. Transmission, viral kinetics and clinical characteristics of the emergent SARS-CoV-2 Delta VOC in Guangzhou, China. *EClinicalMedicine* 40:101129.
10. Teyssou E, Delagrèverie H, Visseaux B, Lambert-Niclot S, Briclher S, Ferre V, Marot S, Jary A, Todesco E, Schnuriger A, Ghidaoui E, Abdi B, Akhavan S, Houhou-Fidouh N, Charpentier C, Morand-Joubert L, Boutolleau D, Descamps D, Calvez V, Marcelin AG, Soulie C. 2021. The Delta SARS-CoV-2 variant has a higher viral load than the Beta and the historical variants in nasopharyngeal samples from newly diagnosed COVID-19 patients. *J Infect* 83:e1-e3.
11. Ong SWX, Chiew CJ, Ang LW, Mak TM, Cui L, Toh M, Lim YD, Lee PH, Lee TH, Chia PY, Maurer-Stroh S, Lin RTP, Leo YS, Lee VJ, Lye DC, Young BE. 2021. Clinical and virological features of SARS-CoV-2 variants of concern: a retrospective cohort study comparing B.1.1.7 (Alpha), B.1.315 (Beta), and B.1.617.2 (Delta). *Clin Infect Dis* doi:10.1093/cid/ciab721.

- 467 12. King KL, Wilson S, Napolitano JM, Sell KJ, Rennert L, Parkinson CL, Dean D. 2022.  
468 SARS-CoV-2 variants of concern Alpha and Delta show increased viral load in saliva.  
469 PLoS One 17:e0267750.
- 470 13. Ito K, Piantham C, Nishiura H. 2022. Relative instantaneous reproduction number of  
471 Omicron SARS-CoV-2 variant with respect to the Delta variant in Denmark. J Med Virol  
472 94:2265-2268.
- 473 14. Puhach O, Adea K, Hulo N, Sattonnet P, Genecand C, Iten A, Jacqueroiz F, Kaiser L,  
474 Vetter P, Eckerle I, Meyer B. 2022. Infectious viral load in unvaccinated and vaccinated  
475 individuals infected with ancestral, Delta or Omicron SARS-CoV-2. Nat Med 28:1491-  
476 1500.
- 477 15. Adamson PC, Pfeiffer MA, Arboleda VA, Garner OB, de St Maurice A, von Bredow B,  
478 Flint J, Kruglyak L, Currier JS. 2021. Lower Severe Acute Respiratory Syndrome  
479 Coronavirus 2 Viral Shedding Following Coronavirus Disease 2019 Vaccination Among  
480 Healthcare Workers in Los Angeles, California. Open Forum Infect Dis 8:ofab526.
- 481 16. Brown CM, Vostok J, Johnson H, Burns M, Gharpure R, Sami S, Sabo RT, Hall N,  
482 Foreman A, Schubert PL, Gallagher GR, Fink T, Madoff LC, Gabriel SB, MacInnis B,  
483 Park DJ, Siddle KJ, Harik V, Arvidson D, Brock-Fisher T, Dunn M, Kearns A, Laney  
484 AS. 2021. Outbreak of SARS-CoV-2 Infections, Including COVID-19 Vaccine  
485 Breakthrough Infections, Associated with Large Public Gatherings - Barnstable County,  
486 Massachusetts, July 2021. MMWR Morb Mortal Wkly Rep 70:1059-1062.
- 487 17. Regev-Yochay G, Amit S, Bergwerk M, Lipsitch M, Leshem E, Kahn R, Lustig Y,  
488 Cohen C, Doolman R, Ziv A, Novikov I, Rubin C, Gimpelevich I, Huppert A, Rahav G,  
489 Afek A, Kreiss Y. 2021. Decreased infectivity following BNT162b2 vaccination: A  
490 prospective cohort study in Israel. Lancet Reg Health Eur 7:100150.
- 491 18. Levine-Tiefenbrun M, Yelin I, Alapi H, Katz R, Herzog E, Kuint J, Chodick G, Gazit S,  
492 Patalon T, Kishony R. 2021. Viral loads of Delta-variant SARS-CoV-2 breakthrough  
493 infections after vaccination and booster with BNT162b2. Nat Med 27:2108-2110.
- 494 19. Jung J, Kim JY, Park H, Park S, Lim JS, Lim SY, Bae S, Lim YJ, Kim EO, Kim J, Park  
495 MS, Kim SH. 2022. Transmission and Infectious SARS-CoV-2 Shedding Kinetics in  
496 Vaccinated and Unvaccinated Individuals. JAMA Netw Open 5:e2213606.
- 497 20. British Columbia Ministry of Health. 2022. B.C. COVID-19.  
498 <https://experience.arcgis.com/experience/a6f23959a8b14bfa989e3cda29297ded>.  
499 Accessed September 23 2022.
- 500 21. British Columbia Centre for Disease Control BCMoH. 2020. Summary of evidence for  
501 updated COVID-19 testing guidance in British Columbia. [http://www.bccdc.ca/Health-  
502 Professionals-Site/Documents/Summary\\_evidence\\_COVID-19\\_testing\\_guidance.pdf](http://www.bccdc.ca/Health-Professionals-Site/Documents/Summary_evidence_COVID-19_testing_guidance.pdf).
- 503 22. Hogan CA, Jassem AN, Sbihi H, Joffres Y, Tyson JR, Noftall K, Taylor M, Lee T, Fjell  
504 C, Wilmer A, Galbraith J, Romney MG, Henry B, Kraiden M, Galanis E, Prystajek N,  
505 Hoang LMN. 2021. Rapid Increase in SARS-CoV-2 P.1 Lineage Leading to  
506 Codominance with B.1.1.7 Lineage, British Columbia, Canada, January-April 2021.  
507 Emerg Infect Dis 27:2802-2809.
- 508 23. Government of Canada. 2022. COVID-19 vaccine: Canadian Immunization Guide.  
509 [https://www.canada.ca/en/public-health/services/publications/healthy-living/canadian-  
510 immunization-guide-part-4-active-vaccines/page-26-covid-19-vaccine.html](https://www.canada.ca/en/public-health/services/publications/healthy-living/canadian-immunization-guide-part-4-active-vaccines/page-26-covid-19-vaccine.html). Accessed  
511 August 26 2022.

- 512 24. Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. Journal of the  
513 Royal Statistical Society Series B (Methodological) 58:267-88.
- 514 25. Lundberg S. LS. 2017. A unified approach to interpreting model predictions, abstr 31st  
515 Conference on Neural Information Processing Systems (NIPS 2017),

516

517 **Table 1.** Epidemiological, clinical and laboratory data of the cohort of asymptomatic individuals  
 518 tested during the test period of the study.

519

<b>Group</b>	<b>Subgroup</b>	<b>Phase 3</b>
Testing	First positives	71642
	Negatives	252342
	Repeats	7801
Specimen type	NP	35369
	SG	36108
	Other	165
No <i>E</i> gene result		21068
Age (years)	0-4	1963
	5-18	9161
	19-39	31914
	40-59	19864
	60-79	7453
	80+	1279
Sex	Male	33415
	Female	37653
	Unknown	574
Patient health authority	1	31841
	2	8951
	3	3635

	4	17202
	5	9752
	Unknown	261
Vaccination status**	Unvaccinated	18459
	1 dose	1540
	Fully vaccinated	47580
	Unknown	4063
VoC lineage	Alpha	0
	Beta	0
	Delta	9049
	Gamma	0
	Omicron (BA.1)	12945
	Unknown	28580

520 \*For all group variables except testing, data presented as first positive result per person

521 \*\*Does not include individuals who received  $\geq 3$  doses of vaccine

522

523 NP: nasopharyngeal; SG: saline gargle; VoC: variant of concern



524 **Table 2.** SEIR and ML model comparisons for SARS-CoV-2 incidence prediction  
 525

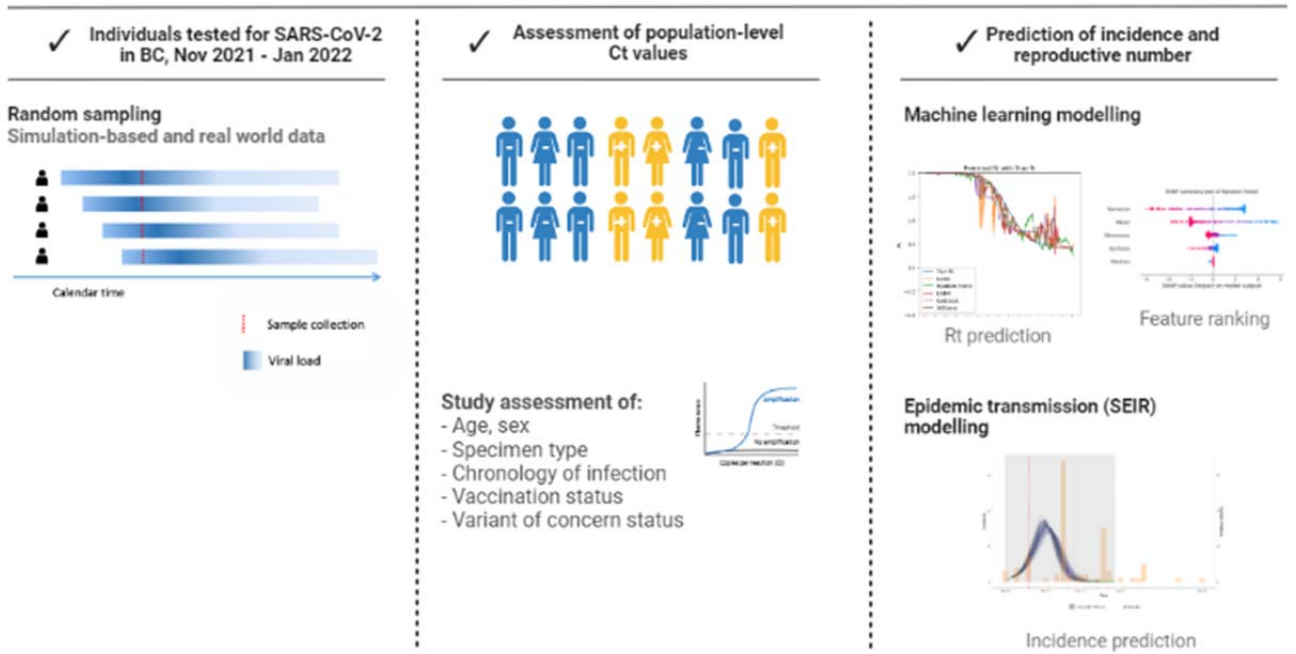
Model		SEIR	ML
<b>Sampling type</b>		Random sampling	
<b>Number of COVID-positive samples</b>		Small (~30)	Large (>1000)
<b>Sampling frequency</b>		Single/multiple snapshots	Daily snapshots
<b>Flexibility:</b>	<b>i) Modeling of Transmission</b>	i) Fixed in time	i) Time-independent
	<b>ii) Ability to add in multiple predictors</b>	ii) Unable to incorporate multiple predictors	ii) Allows incorporation of various predictors and is flexible in their representation
<b>Scalability</b>		Single outbreak setting	Population level
<b>Computational Complexity*</b>		Low	Low-moderate
<b>Predictive power requirements</b>		Good in single setting with well-mixed population and stable contact behavior/infection control	No requirements other than sufficient sample size for Ct summary statistics by snapshot
<b>Additional Sampling Requirements</b>		None	Ordered in time, restricted to fixed interval sampling

526 \* Relative computational complexity based on assumed sample size provided in scalability row

527 COVID-19: coronavirus disease 2019; SEIR: susceptible-exposed-infected-recovered; ML:  
 528 machine learning; Ct: cycle threshold  
 529

530

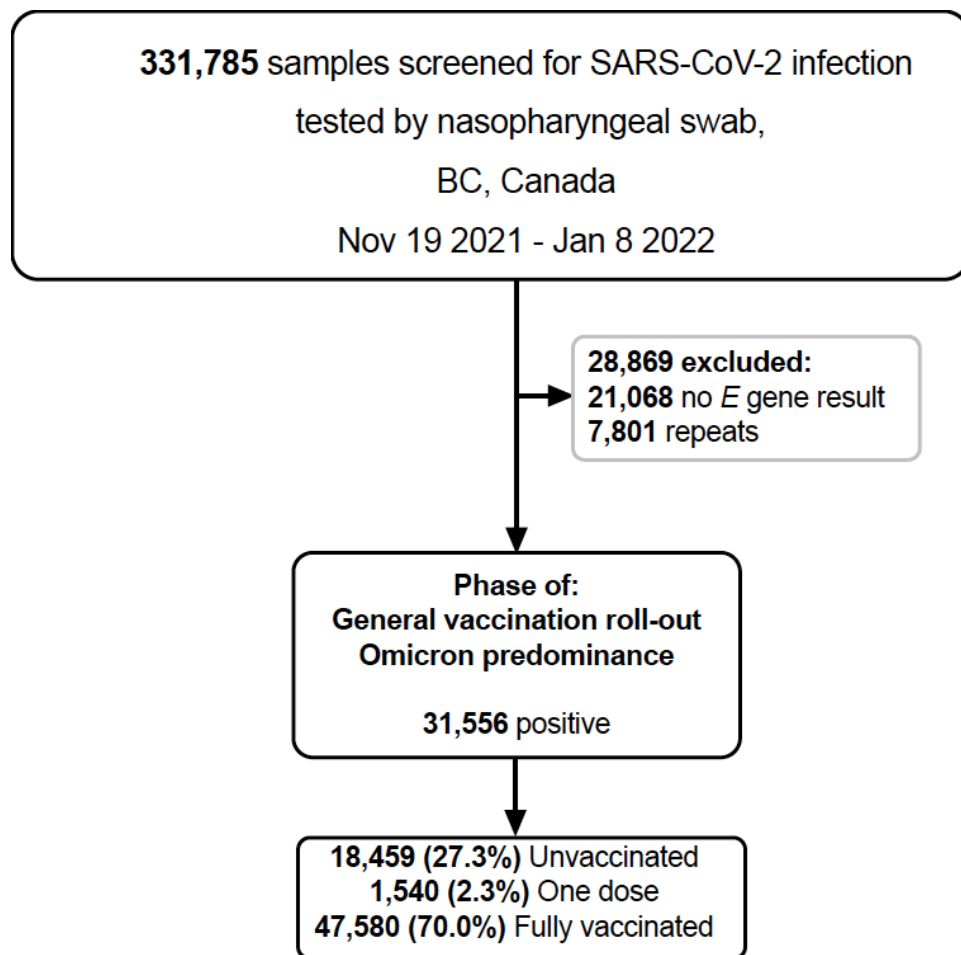
A.



531

532

533 B.



534

535

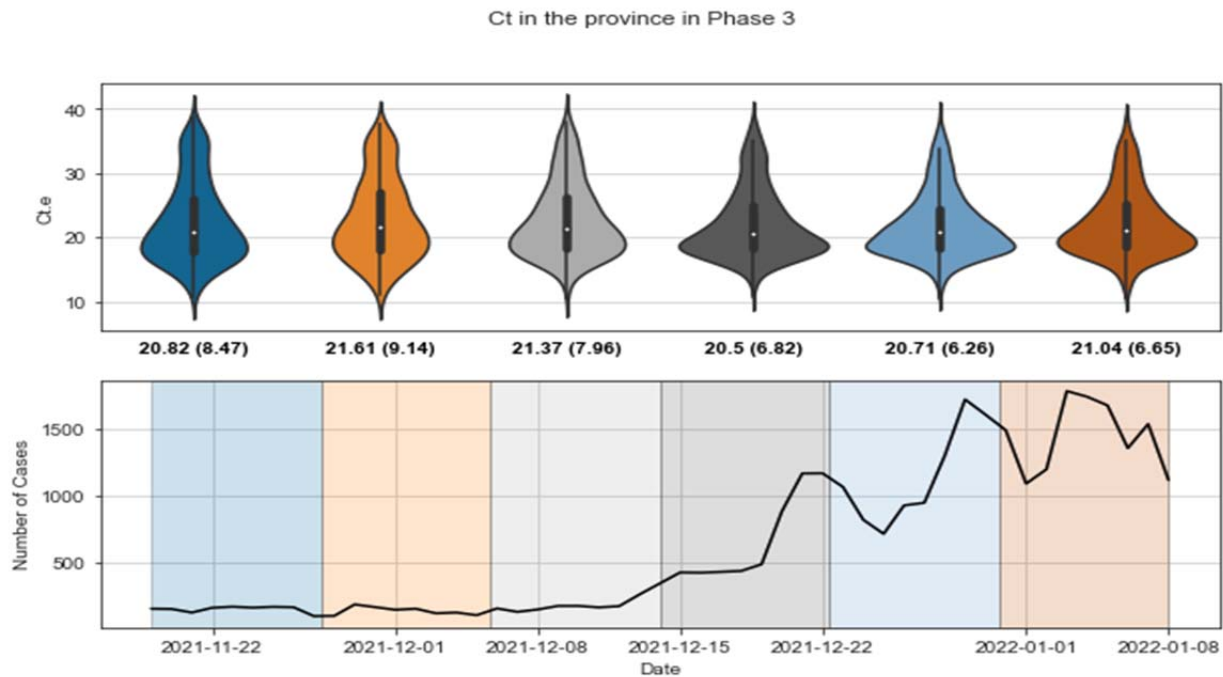
536 **Figure 1.** Overall design (A) and flowchart (B) of the study.

537 BC: British Columbia; E gene: Envelope gene; SARS-CoV-2: Nov: November; Rt: reproductive

538 number; SARS-CoV-2: severe acute respiratory syndrome coronavirus type 2

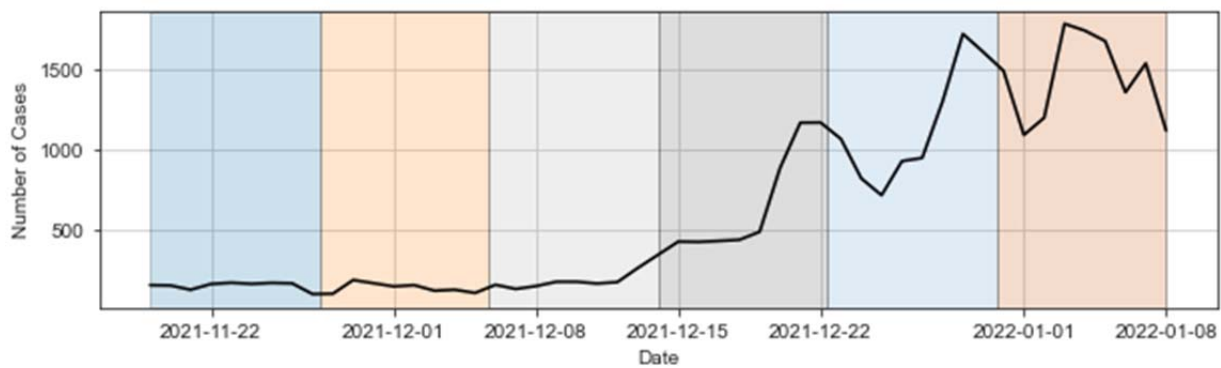
539

540 A.



541

542 B.



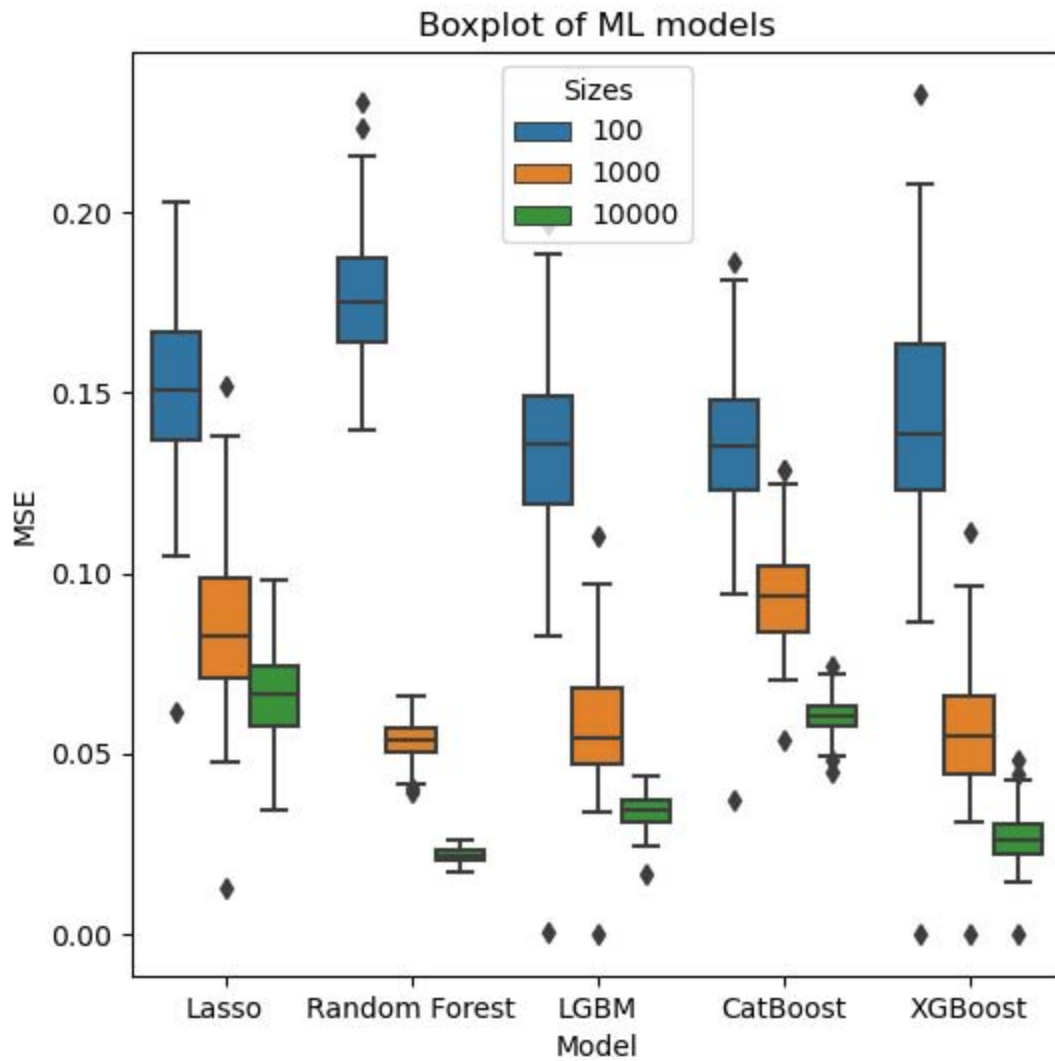
543

544 **Figure 2.** Violin plots demonstrating the cycle threshold value distribution (A) and absolute  
545 number of cases of confirmed SARS-CoV-2 infection (B) in British Columbia across different  
546 time points of the study period.

547 Ct. e: Envelope (*E*) gene cycle threshold value; SARS-CoV-2: SARS-CoV-2: severe acute  
548 respiratory syndrome coronavirus type 2

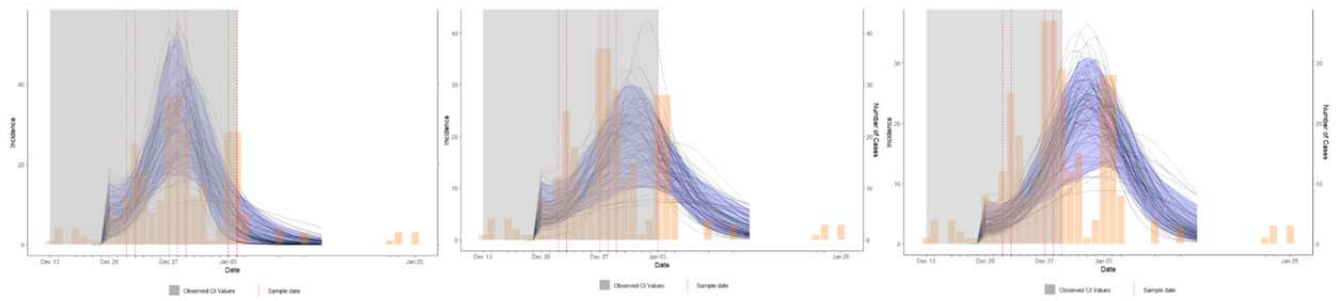
549



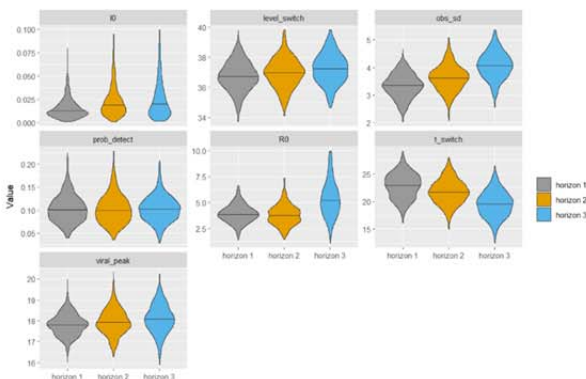


551  
552 **Figure 3.** Boxplot representation of MSE scores across models on out-of-sample simulated cycle  
553 threshold data.  
554 LGBM: Light Gradient Boosting Model; MSE: Mean squared error; XGBM: eXtreme Gradient  
555 Boosting Model, ML: Machine Learning  
556

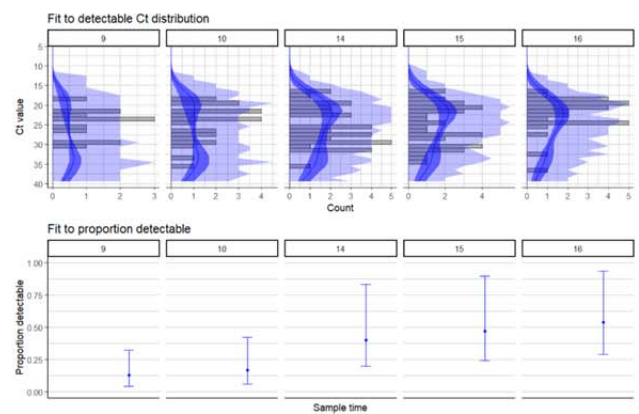
I.



II.



III.

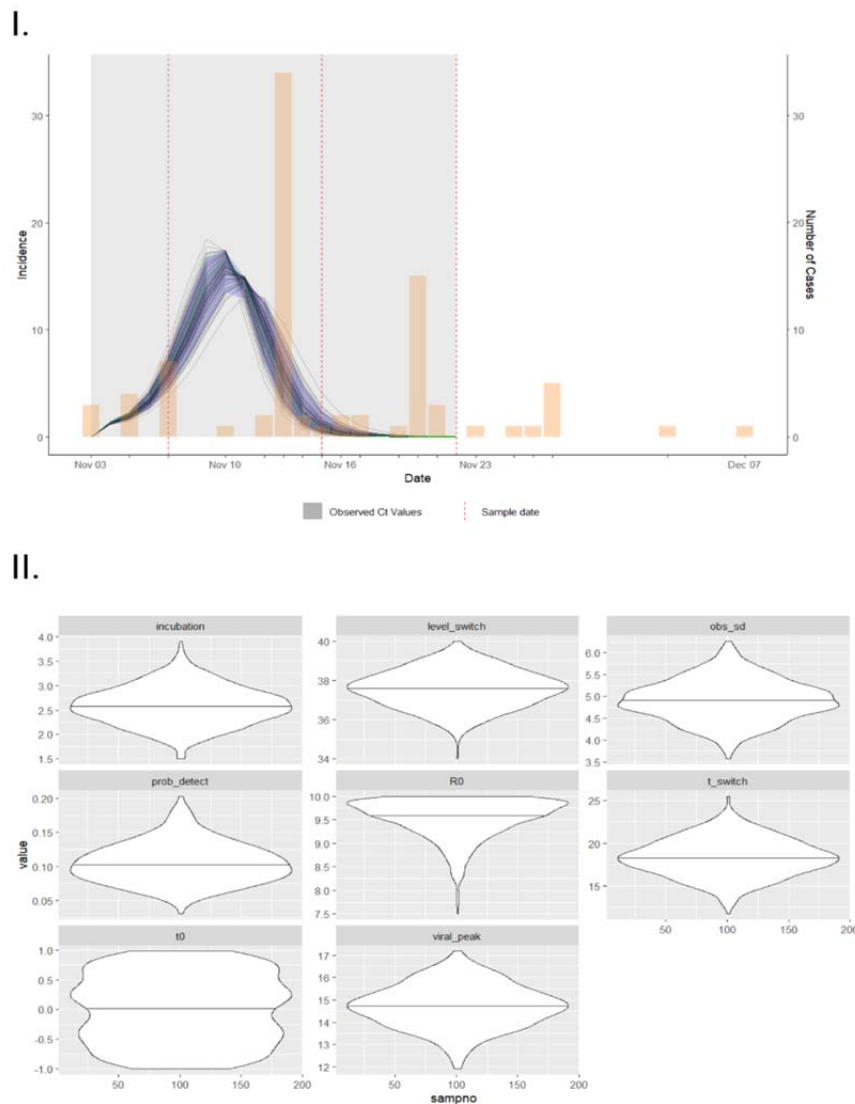


557

558

559 **Figure 4.** Overall population modeling findings. A multiple-cross section SEIR model was fitted  
560 to the overall population-level data (I), and showed an incidence peak from December 27 2021 to  
561 January 1 2022, which overlapped with the observed peak of reported cases in the province. The  
562 Monte Carlo chain model-predicted incidence curve is represented (black lines), and was  
563 overlaid with the reported number of confirmed SARS-CoV-2-positive (yellow bars) cases.  
564 Violin plots of the viral kinetic parameters for the SEIR model are presented (II). Three unique  
565 time horizons were chosen, each of which is depicted by a different color. The MCMC approach  
566 searches over the viral kinetic parameters presented above, and is based on prior values  
567 described separately (**Supplemental Table 6**). To align with the described Omicron viral  
568 kinetics, the incubation period was fixed and set at three days, and the infectious viral kinetic

569 parameter was fixed. An upper bound of  $I_0$  was set at 0.100. The fit to detectable cycle threshold  
570 distribution and the fit to proportion variable are presented over different time points (A).  
571  
572 Ct: cycle threshold; SARS-CoV-2: severe acute respiratory syndrome coronavirus type 2; SEIR:  
573 susceptible-exposed-infected-recovered



574  
575 Figure 5. Long-term care facility outbreak investigation modeling findings. a multiple-cross  
576 section SEIR model was fitted to the outbreak data (I), and showed a peak in incidence on the



577 12th day of the outbreak which preceded by two days the observed peak at the outbreak facility.

578 The population included in this outbreak investigation was sampled at three pre-determined time

579 points (dashed red lines). The Monte Carlo chain model-predicted incidence curve is represented

580 by black lines, and was overlaid with the reported number of confirmed SARS-CoV-2-positive

581 cases in this outbreak setting (yellow bars). Violin plots of the viral kinetic parameters for the

582 SEIR model are also presented in the outbreak case study (II).

583

584 Ct: cycle threshold; SARS-CoV-2: severe acute respiratory syndrome coronavirus type 2; SEIR:

585 susceptible-exposed-infected-recovered

586

587 **Supplemental Table 1.** Vaccination phase definitions used for the study

<b>SARS-CoV-2 phase</b>	<b>Vaccination Phase*</b>	<b>Wave**</b>	<b>Target population</b>
Phase 1	Phase 1	Waves 1, 2, 3 Dec 14 2020 - Mar 7 2021	Residents, staff and essential visitors to long-term care settings; individuals assessed and awaiting a long-term care placement; health care workers providing care for COVID-19 patients; and remote and isolated Indigenous communities.
	Phase 2	3 Mar 8 2021 - Apr 2021	Individuals age $\geq 80$ ; Indigenous peoples age $\geq 65$ and Indigenous Elders; Indigenous communities; hospital staff, community general practitioners and medical specialists; vulnerable populations in select congregate settings; and staff in community home support and nursing services for seniors.
	Phase 3	3 Apr 15 2021 - May 10 2021	Individuals aged 60-79 years, Indigenous peoples aged 18-64 and people aged 16-74 who are clinically extremely vulnerable.
Phase 2.1	Phase 4	Waves 3, 4 May 11 2021- Jul 17 2021	Everyone aged $\geq 12$ years-old. From September 2021, third vaccine dose available for people who are clinically extremely vulnerable
Phase 2.2	Phase 4	4 Jul 18 2021 - Nov 18 2021	
Phase 3*	Phase 5	4 Nov 19 2021 - Jan 8 2022	Everyone aged $\geq 5$ years-old. From the end of November 2021, children aged 5-11 are eligible for vaccination. Everyone aged $\geq 18$ and invited to get a 'booster' (third vaccine dose) within 6-8 months after receipt of their second dose.

588 \*The current study included Phase 3 only

589 \*\*Vaccination phases were defined by vaccine eligibility of the target populations in BC, and are  
590 detailed separately (21)

591

592 SARS-CoV-2: SARS-CoV-2: severe acute respiratory syndrome coronavirus type 2

593 **Supplemental Table 2.** SARS-CoV-2 diagnostic testing strategy based on the envelope (*E*) gene target and test result interpretation  
 594 criteria used for the participating laboratories

Type of assay	Assay used (manufacturer)	Extraction	PCR	Ct Interpretation criteria
Laboratory-developed	BCCDC PHL LDT	MagMax	ABI 7500	Ct threshold for positivity: 38
		MagNa Pure 24	ABI 7500	
Commercial	LightMix SarbecoV E-gene plus EAV control assay (TIB Molbiol)	MagNA Pure Compact or MagNA Pure 96	LightCycler 480	Manufacturer recommended threshold
	Allplex SARS-CoV-2 assay (Seegene)	STARlet	CFX 96	
	Cobas 6800 and 8800 SARS-CoV-2 Test (Roche Molecular Diagnostics)			
	Xpert Xpress SARS-CoV-2 (Cepheid)			
	BD MAX SARS-CoV-2 (BD)			
	Respiratory Panel 2.1 (BioFire)			
Panther Fusion (Hologic)				

595

596 BCCDC PHL: British Columbia Centre for Disease Control Public Health Laboratory; Ct: cycle threshold; FHA: Fraser Health  
 597 Authority; IHA: Interior Health Authority; LDT: laboratory-developed test; NHA: Northern Health Authority; PCR: polymerase chain  
 598 reaction; SARS-CoV-2: severe acute respiratory syndrome coronavirus type 2; SPH: St. Paul's Hospital; VCH: Vancouver Coastal  
 599 Health.

600 **Supplemental Table 3.** Epidemiological, clinical and laboratory data of the earlier British Columbia SARS-CoV-2 pandemic phases

<b>Group</b>	<b>Subgroup</b>	<b>Phase 1 (Dec 14 2020 - May 10 2021)</b>	<b>Phase 2.1 (May 11 2021 - July 17 2021)</b>	<b>Phase 2.2 (July 18 2021 - Nov 18 2021)</b>
Testing	Positives	55291	6529	31613
	Negatives	815993	190679	599953
	Repeats	53740	5263	18758
No <i>E</i> gene Result		1623	257	1611
Age	0-4	4203	712	3200
	5-18	23586	2849	12866
	19-39	15437	1709	8173
	40-59	7950	821	4366
	60-79	2150	181	1391
	80+	6183	2312	47
Sex	Male	29102	3492	16357
	Female	25729	2969	14573
	Unknown	460	68	683

Patient health authority	1	31045	3387	7851
	2	4233	987	9761
	3	1922	275	4317
	4	337	45	230
	5	15370	1680	6065
Vaccination status	Unknown	2384	155	2389
	1 dose	100	99	9665
	Fully vaccinated	16293	2462	16174
VoC lineage	Alpha	74	10	0
	Beta	253	483	22220
	Delta	4595	2134	99
	Gamma	0	0	0
	Omicron	293	39	307

601

602 VoC: variant of concern

603 **Supplemental Table 5.** Hyperparameter selection

Hyperparameter	Lasso	RF	LGBM	XGBoost	Catboost
Alpha	[0.001,0.01,0.1]	-	-	-	-
Max Depth	-	[2,4,8,16,32]	[2,4,8]	[6,10]	-
Number of Estimators	-	[4,16,64,256]	-	-	-
Minimum Sample Split	-	[2,4,8,16,32]	-	-	-
Number of Leaves	-	-	[4,8,16,32,64,128]	-	-
Minimum Data in Leaf	-	-	[2,4,8,16,32]	-	-
Depth	-	-	-	-	[1,5,10]
Iterations	[500,1000,2000]	-	-	-	[250,500,1000]
Learning Rate	-	-	-	-	[0,001, 0.01, 0.1]
L2 Regularization	-	-	-	-	[1,5,10]
Min child weight	-	-	-	[1,3]	-

604

605 RF: Random Forest; LGBM: Light Gradient Boosting Model; XGBM: eXtreme Gradient

606 Boosting Model

607

608 **Supplemental Table 6.** Control table of priors for SEIR model

Values	Names*	Fixed	Lower bound	Upper bound	Steps	Lower start	Upper start
0	tshift	1	0	3	0.1	0	10
5	desired_mode	1	0	7	0.1	0	10
19.73	viral_peak	0	0	40	0.1	15	25
5	obs_sd	0	0	25	0.1	1	10
0.79	sd_mod	1	0	1	0.1	0.4	0.6
14	sd_mod_wane	1	0	14	0.1	0	14
40	true_0	1	40	100	0.1	40	100
40	intercept	1	35	100	0.1	35	100
3	LOD	1	0	10	0.1	0	10
5	incu	1	0	10	0.1	0	10
13.29	t_switch	0	0	30	0.1	10	30
38	level_switch	0	0	40	0.1	33	40
1000	wane_rate2	1	0	10000	0.1	10	50
0.103	prob_detect	0	0	1	0.1	0.01	0.1
1	t_unit	1	0	1	0.1	0	1
2	R0	0	1	10	0.1	1.5	3
4	infectious	1	0	25	0.1	5	10
3	incubation	1	0	25	0.1	5	10
1	t0	1	0	100	0.1	0	50
0.0001	I0	0	0	0.1	0.1	0	0.001

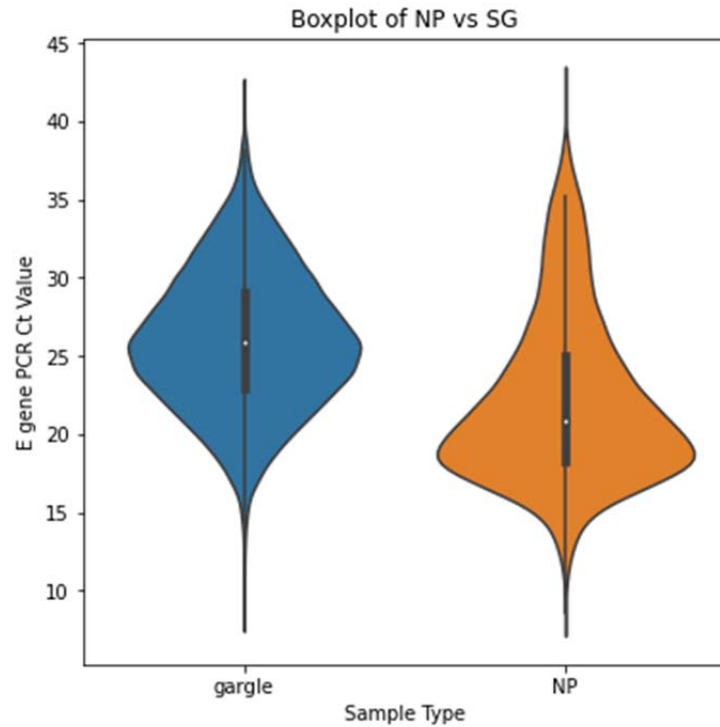
609 \*Adapted parameters (3)

610

611 SEIR: susceptible-exposed-infected-recovered

612





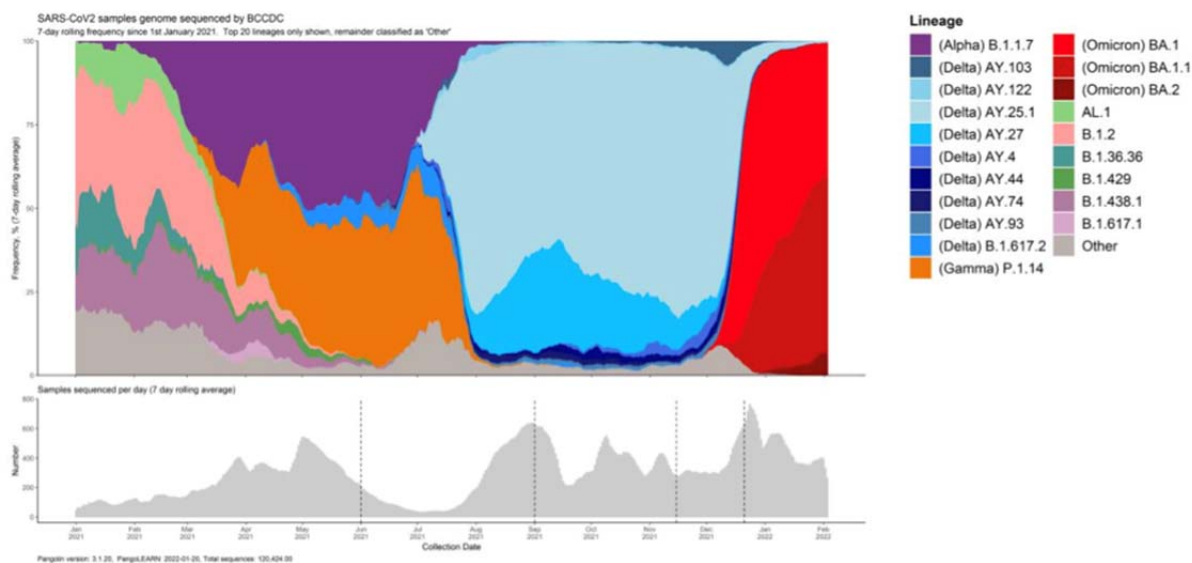
613  
614

615 Supplemental Figure 1. Violin plot demonstrating the overall cycle threshold value distribution  
616 for saline gargle compared to nasopharyngeal specimens for the entire study period.

617 Ct. e: Envelope (*E*) gene cycle threshold value; NP: nasopharyngeal

618  
619  
620

621



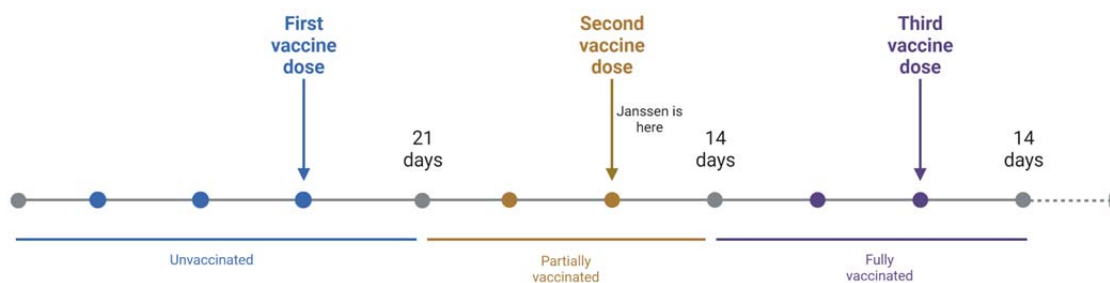
622

623 **Supplemental Figure 2.** Twenty most prevalent SARS-CoV-2 variant of concern lineages in  
624 British Columbia from January 2021 to January 2022. The current study was performed during a  
625 time of Omicron variant predominance, from November 19 2021 to January 8 2022.

626 BCCDC: British Columbia Centre for Disease Control; SARS-CoV-2: severe acute respiratory  
627 syndrome coronavirus type 2

628

629



630  
631 **Supplemental Figure 3.** Vaccination status definitions. Vaccination status was defined based on  
632 the date of vaccine receipt relative to the date of the sample collection included for the study. For  
633 the Janssen vaccine only, fully vaccinated status was defined as having received one dose 14  
634 days or more prior to sample collection. For all other vaccines, **Unvaccinated status** was defined  
635 as having received no SARS-CoV-2 vaccine, or having received a SARS-CoV-2 vaccine less  
636 than 21 days prior to the sample collection date. **Partially vaccinated** status was defined as  
637 having received the SARS-CoV-2 vaccine dose 1 greater or equal to 21 days prior to sample  
638 collection, but having received dose 2 less than 14 days prior to the sample collection. **Fully**  
639 **vaccinated** status was defined as greater or equal to 14 days since the receipt of dose 2, but  
640 having received dose 3 less than 14 days prior to the sample collection.

641  
642 SARS-CoV-2: severe acute respiratory syndrome coronavirus type 2

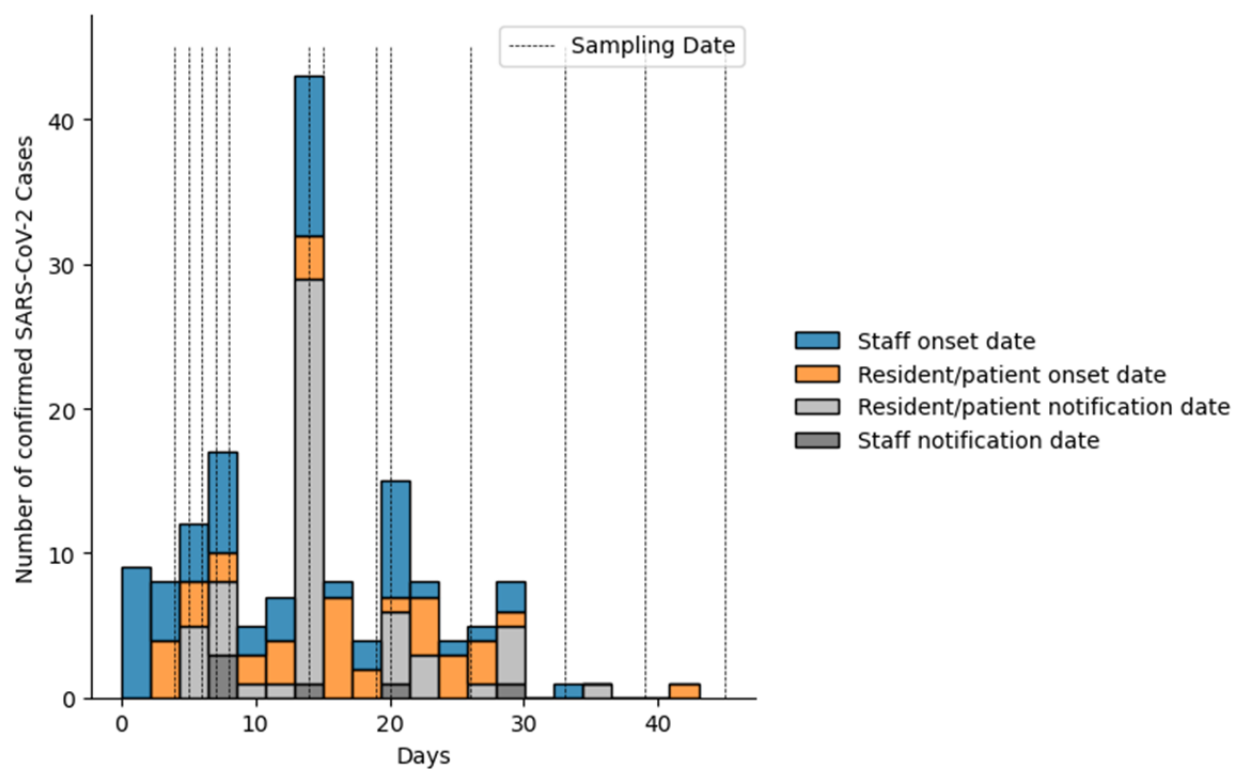
643

644 A.

	<b>Residents</b> (n=93)	<b>Staff</b> (n=63)
Age (Median, range)	89 (46-100)	42 (19-71)
Female (%)	68 (73.1%)	54 (85.7%)
Asymptomatic at the time of assessment (%)	54 (58.1%)	6 (9.5%)
Ever hospitalized	3 (3.2%)	0
Deceased	26 (28.0%)	0

645

646 B.

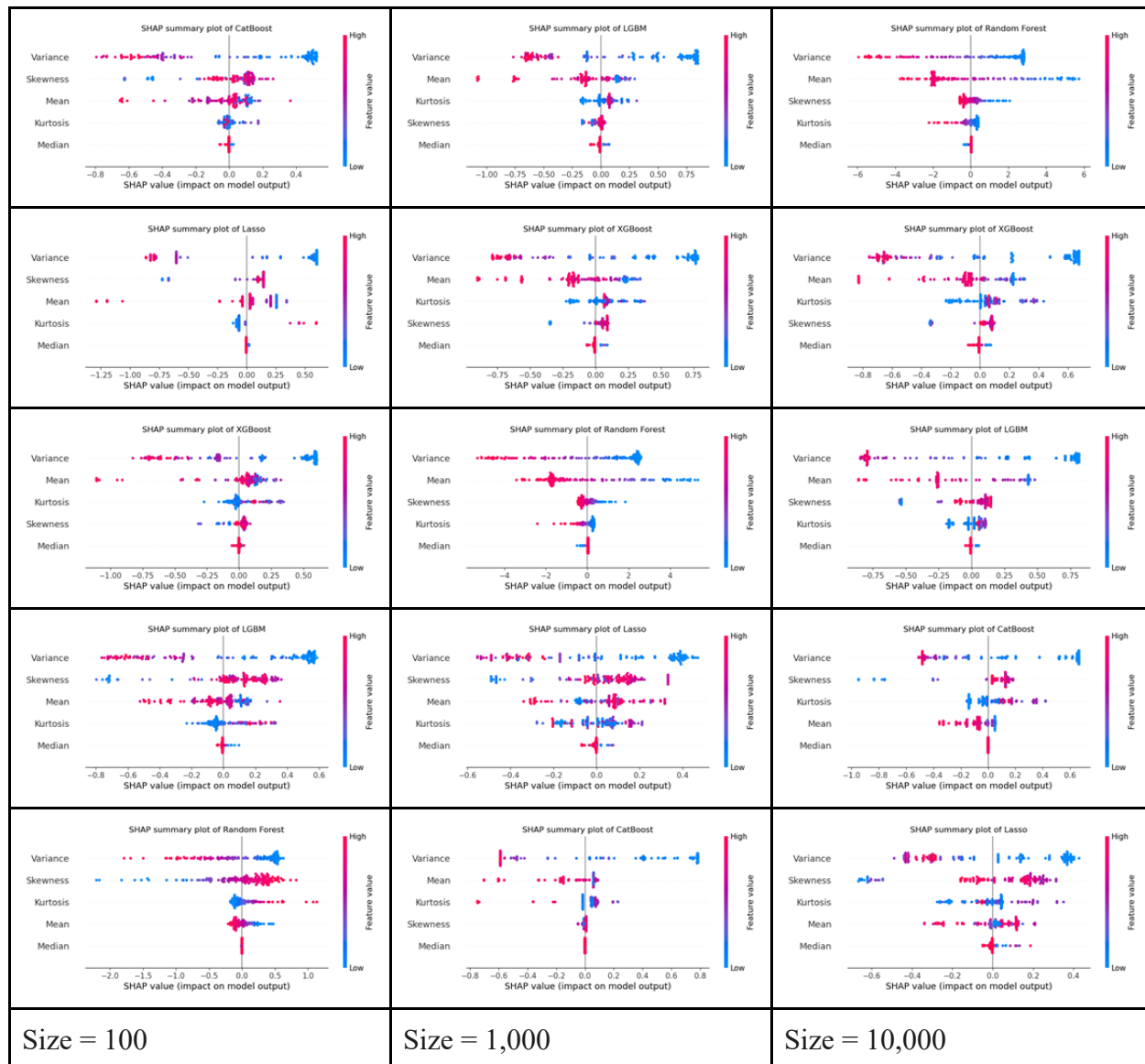


647

648 **Supplemental Figure 4.** Case study epidemiological data (A) and epidemic curve (B) for the

649 156 infected individuals in the long term care facility outbreak.

650



651

652 **Supplemental Figure 5.** SHAP summary outputs explaining the machine learning output based  
653 on simulated cycle threshold (Ct) data. Results are presented stratified by three different  
654 population sizes: 100, 1,000 and 10,000 with each column in descending order of performance.

655 Of the five features explored, the top ranking feature across all models was the variance of the Ct  
656 data.

657 SHAP: SHapley Additive exPlanations; LGBM: Light Gradient Boosting Model; XGBM:

658 eXtreme Gradient Boosting Model

659

660