Network Modeling Analysis in Health Informatics and Bioinformatics (2022) 11:32
https://doi.org/10.1007/s13721-022-00375-1

ORIGINAL ARTICLE

SARS-CoV-2 transmission in university classes

William Ruth[1] · **Richard Lockhart**[1]

Received: 11 March 2022 / Revised: 29 July 2022 / Accepted: 10 August 2022 / Published online: 27 August 2022

Abstract
We investigate transmission dynamics for SARS-CoV-2 on a real network of classes at Simon Fraser University. Outbreaks are simulated over the course of one semester across numerous parameter settings, including moving classes above certain size thresholds online. Regression trees are used to analyze the effect of disease parameters on simulation outputs. We find that an aggressive class size thresholding strategy is required to mitigate the risk of a large outbreak, and that transmission by symptomatic individuals is a key driver of outbreak size. These findings provide guidance for designing control strategies at other institutions, as well as setting priorities and allocating resources for disease monitoring.

Keywords Disease modelling · Individual-level models · Network analysis · Stochastic simulation

1 Introduction

Key findings include that moving classes online has a major impact on the severity of potential outbreaks. Additionally, most of our ability to anticipate this severity for a particular threshold is captured by a small number of disease parameters. This suggests that a small number of parameters give us most of the available information about outbreak severity. Under both measures of severity, we find that the most important parameters for prediction are the infectiousness and duration for symptomatic cases. This finding is fortunate from a public health management perspective, since symptomatic cases are easier to detect and quarantine.

1.1 Past work

Statistical and mathematical models are powerful tools for studying the SARS-CoV-2 pandemic. Many authors have developed sophisticated models to predict the spread of the disease, which have influenced policy and, ultimately, saved lives (Vespignani et al. 2020). The problem of disease modelling is large and multifaceted. Here, we focus exclusively on transmission within a the context of a university. Specifically, we investigate the effect of moving certain classes online, to see whether limited in-person instruction can be maintained while preventing a major outbreak. Our data come from Simon Fraser University in Burnaby, Canada, but the framework can be applied to other institutions.

Giving a complete overview of the SARS-CoV-2 modelling literature here would be impossible. We provide only a brief summary of some closely related work. Models can be broadly classified into two categories: individual level, and differential equation based (Brauer et al. 2019). Individual-level models investigate the effects of individual agents' actions on population level outcomes, whereas differential equation-based techniques involve directly modelling population level phenomena. We work entirely within the individual-level model framework. See Estrada (2020) for an overview of differential equation models for SARS-CoV-2 spread. See Kiss et al. (2017) for a thorough overview of modelling and analysis of disease spread on networks. Fractional derivatives have been applied in a variety of contexts to extend the differential equation framework; for example, to model the early stages of the pandemic in Pakistan (Naik et al. 2020). More complicated models have been used to investigate the effect of various disease reservoirs (Naik et al. 2021a), environmental infection (Naik et al. 2021b), and interaction of quarantine measures with diabetes complications (Özköse and Yavuz 2022).

Many individual-level models have a compartmental structure, such as SIR (susceptible, infectious, removed) or SEIR (susceptible, exposed, infectious, removed) (Brauer 2008; Deardon et al. 2010), where individuals are assigned

✉ William Ruth
wruth@sfu.ca

[1] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

to a category based on their disease status, and the researcher models how individuals move between categories. While much work has been done on modelling community transmission (see, e.g., BCCDC 2021; Chang et al. 2020; Rădulescu et al. 2020; Tuite et al. 2020), some authors have instead directed their efforts toward understanding outbreaks on university campuses (Gressman and Peck 2020; Zhou et al. 2021; Kharkwal et al. 2020; Borowiak et al. 2020; Bahl et al. 2020; Frazier 2020; Ambatipudi et al. 2021; Christensen et al. 2020; Weeden and Cornwell 2020). Gressman and Peck (2020) simulated social dynamics within a university and the corresponding infection rates. They examined the effects of various interventions, including mask wearing, remote instruction, and random testing; with particular attention paid to the test's false-positive rate (i.e., specificity). Zhou et al. (2021) use simulation to investigate the effects of several control strategies with a simplified model of SARS-CoV-2 dynamics. Kharkwal et al. (2020) developed a detailed framework for simulating infections, which integrates models of various phenomena related to the disease. Borowiak et al. (2020) studied the effects of different strategies for grouping students in dorms and classes. Bahl et al. (2020) developed a detailed model of how students and faculty interact on a small university campus. In a seminar presentation, Frazier (2020) discussed both individual-level and differential equation models for disease spread at a large campus; paying particular attention to universal testing schemes and strategies for contact tracing. Ambatipudi et al. (2021) developed a framework for assessing risk of infection over the course of a semester based on room crowding and air circulation. Christensen et al. (2020) give a rapid review of studies modelling COVID transmission in universities.

Weeden and Cornwell (2020) took a different approach to investigating enrollment at Cornell University. Instead of studying disease transmission directly, they measured numerous graph-theoretic properties of the enrollment network. Their focus was on measuring the connectedness of the network.

1.2 Our contribution

We received data on enrollments at Simon Fraser University (SFU), a medium-sized school located just outside of Vancouver, Canada. Our dataset contains enrollment records from the fall term in 2019 consisting of 110,000–120,000 entries, where each entry corresponds to a specific course taken by a specific student.[1] We also have records of the days on which each class meets, but not at what time. Our dataset does not include any distance learning courses, co-op courses (a.k.a. work experience), or courses that do not meet at one of SFU's main campuses. A number of classes in the dataset do not have any meeting days. Section 1 of the Supplemental Material discusses network properties of our dataset at length. Preliminary analysis (not shown) did not identify strong association between these network summaries and our other outcomes of interest. The data used in our study are available on the associated Github repository (Ruth and Lockhart 2022).

Along with many other universities, SFU adopted a near-total lockdown policy in response to the SARS-CoV-2 virus and has only recently returned to in-person instruction. Although this lockdown dramatically reduced the possibility of on-campus transmission, it has also adversely impacted students' learning. There was also an interim period where some classes were held on campus while others remained virtual. The partial return to campus model has the clear advantage of allowing many classes to meet in-person, but also carries an increased risk of infection. Particularly catastrophic would be an outbreak on campus, where a large proportion of the student body becomes infected.

The goal of our study is to investigate potential outbreaks on SFU's main campus when a limited number of smaller classes are allowed to meet in-person. We focus particularly on how properties of these outbreaks vary as the size of in-person classes varies. Although there are countless ways in which students can infect each other on- and off-campus, we focus on disease transmission through classes. As such, we omit all classes which do not have a scheduled meeting day. Ideally, we would have investigated these courses further. However, for privacy reasons we do not have identifying information for any of these courses, and are thus unable to learn any more about them. Removing these courses will undoubtedly have changed the structure of the enrollment network at SFU, but not in a way that impacts person-to-person contact and thus disease transmission (ostensibly, there is no in-person interaction in a course with no meeting days). We treat labs and tutorials as distinct classes with no inherent connection to the main course with which they are affiliated (other than overlapping enrollment), since each meeting, be it lecture or tutorial, is a separate opportunity for disease spread.

The enrollment network at SFU contains a number of isolated groups of students. That is, groups of students who share classes with each other, but not with anyone outside the group. In graph theory, these groups are called connected components of the network (Clark and Holton 1991). Since the only avenue of disease transmission that we study is via shared classes, and there are no shared classes between components, we focus on one connected component at a time. It turns out that each term's network is dominated by a single large component, so we keep only this main component and omit all the small ones. Note that the inclusion or exclusion

[1] We also have data for spring 2019 and 2020 terms, but limit our investigation to fall 2019 for the sake of brevity.

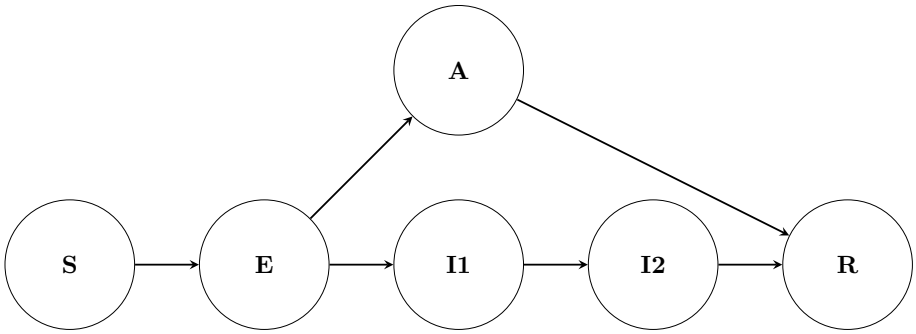

Fig. 1 Modelled disease trajectory. Arrows represent possible transitions

of tutorials has no effect on which students belong to the same connected component (Alice and Bob share a tutorial if and only if they also share the class with which that tutorial is affiliated).

To model disease spread, we simulate transmission over the course of one term. We start by infecting ten randomly chosen individuals. We then track how the disease spreads through classes over 90 days (roughly the duration of the pre-exam portion of a term at SFU), with numerous different regimes for the epidemiological properties of the disease. We also consider several schemes for moving large classes online to slow the spread of the infection. Parameter values are given in Table 2, and were chosen to reflect a reasonable range of values from the literature. Multiple simulations are run under each regime, then various numerical and graphical summaries are reported.

It is important to note that the results of our simulations should not be taken literally. There are many factors that influence how a disease might spread across a university campus, and we can't hope to model all of them. As such, our findings are meant to be interpreted qualitatively; suggesting trends across variables, rather than as a tool to set specific policy strategies.

Computation is done using the `R` (R Core Team 2021) and `Julia` (Bezanson et al. 2017) programming languages.

The rest of this paper proceeds as follows. Section 2 discusses our disease model and computational framework. Section 3 describes the analysis we perform, and Sect. 4 presents the results. Section 5 contains interpretation of our results and some discussion of the limitations of our study. Finally, 6 gives some broader implications of our work.

2 Simulation study

To investigate the relationship between network structure and disease transmission, we carry out a simulation study. We use an **SEAIR** compartment model for the behaviour of SARS-CoV-2. Respectively, these compartments correspond to individuals who are **S**usceptible, **E**xposed but not infectious, **A**symptomatic and infectious, **I**nfectious and symptomatic, or **R**ecovered. The **I** compartment is further subdivided into individuals who are not yet symptomatic (i.e., presymptomatic) and those who are fully symptomatic, denoted **I1** and **I2**, respectively. See Martcheva (2015) for a detailed overview of compartment models for disease. An earlier draft of this manuscript gives an alternative disease model with fewer compartments; see (Ruth and Lockhart 2021).

2.1 Infection dynamics

Figure 1 shows which transitions are allowed in our model. Individuals can progress forward but not backward along each arrow. In short, susceptible individuals can only transition to exposed. Exposed individuals transition to either asymptomatic or presymptomatic. The asymptomatic individuals transition directly to recovered, whereas those who are presymptomatic will transition through symptomatic before finally becoming recovered.

We model holding times in the **E**, **A**, **I1** and **I2** compartments using geometric random variables (supported on the positive integers, excluding zero), with a different success probability for each compartment. Call these probabilities q_E, q_A, q_{I1} and q_{I2}, respectively. Thus, the number of individuals transitioning out of compartment X on any particular day follows the Binomial(N_X, q_X) distribution, where X is a compartment other than **S** or **R** (see below for details on transitioning out of **S**; no transitions out of **R** occur), and N_X is the number of individuals in compartment X on that day. The specific individuals who transition out of a compartment are chosen uniformly at random from the members of that compartment.

We also use Bernoulli random variables to choose a destination when individuals transition out of **E**. Call q_{EA} the probability that a transition from **E** is to **A**. Thus, among those individuals transitioning out of **E**, the number that transition to **A** follows a binomial distribution. The remainder of those leaving **E** enter compartment **I1**.

Holding times in the **S** compartment are more complicated. Specifically, the probability of a susceptible individual transitioning to exposed on a particular day depends on both

the sizes of classes in which the susceptible is enrolled and the number of contagious individuals who are also enrolled in these classes.

Consider a class with one susceptible student, and some number of students in the contagious compartments, **A**, **I1** and **I2**. We assume that all possible transmission events are independent. On a particular day, each contagious individual has some probability of transmitting the disease to our susceptible student. This probability depends on which compartment the contagious individual is in. We model the transmission probability in a class between a single susceptible-contagious pair as inversely proportional to the square root of the class size, with a different proportionality constant for each contagious compartment: θ_A, θ_{I1} and θ_{I2}. Letting τ_X be the pairwise transmission probability for a contagious individual in compartment X, the infection probability for a single susceptible on a particular day is then $\tau_* = 1 - (1-\tau_A)^{M_A}(1-\tau_{I1})^{M_{I1}}(1-\tau_{I2})^{M_{I2}}$, where M_X is the number of individuals in the class who are in compartment X. Finally, the number of new cases in this class follows the Binomial(M_S, τ_*) distribution. After determining how many individuals will transition out of **S** on a particular day, we select the particular individuals uniformly at random from this compartment.

The process just described covers how to generate transitions in a single class. To simulate a full day, we run this process independently in every class that meets on that day. Note that it is possible under our framework for an individual to become infected in more than one of their classes. When this happens, we simply move this individual to compartment **E**, and ignore any multiplicity effects.

See Section 2 of the Supplemental Material for a pseudocode description of our algorithm. The probabilistic model induced by our simulation algorithm has eight parameters, four probabilities for geometric holding time distributions, one probability for Bernoulli trials to choose transition destinations, and three proportionality constants for transmission probabilities. To simplify identification of parameter values from the literature, we re-parameterize the infectiousness parameters for compartments **A** and **I1** to be proportional to the infectiousness of individuals in compartment **I2**. That is, we write $\theta_A = \rho_A \theta_{I2}$ or, equivalently, $\tau_A = \rho_A \tau_{I2}$. We define ρ_{I1} similarly.

We also consider a control strategy where classes above a certain size, ϕ, are moved online, thereby preventing transmission between students in these classes.[2] We arbitrarily choose the threshold values 20, 50, 100 and ∞, as these are qualitatively different class sizes (the maximum class size is 481). After removing classes above the specified threshold from the network, we find the largest component of the new network and remove any students who are not connected to this main component. Table 1 gives the number of students remaining in both the full network and the largest connected component after removing classes above the specified threshold. We focus on only the largest component of each network; the size of the full network is included only for completeness. The threshold size is another parameter for our model, giving a total of nine.

Table 1 Sizes of networks for various class size thresholds, both before and after removing isolated components

Threshold	Size of network	Size of largest component
20	17,851	16,866
50	25,470	23,660
100	26,540	24,752
∞	27,307	25,627

Table 2 Candidate values and their sources for each parameter in our model

Parameter	Values	Source
θ_{I2}	0.141, 0.198, 0.240	Thompson et al. (2021)
ρ_A	0.4, 0.75, 1	Johansson et al. (2021)
ρ_{I1}	0.18, 0.63, 2.26	Buitrago-Garcia et al. (2020)
q_E	0.168, 0.182, 0.196	Xin et al. (2021)
q_A	0.115, 0.138, 0.169	Byrne et al. (2020)
q_{I1}	0.333, 0.435, 0.833	Byambasuren et al. (2020); Xin et al. (2021)
q_{I2}	0.063, 0.075, 0.092	Byambasuren et al. (2020)
q_{EA}	0.09, 0.18, 0.26	Byambasuren et al. (2020)
ϕ	20, 50, 100, ∞	

Table 2 lists the values used for our parameters, as well as their sources. For each parameter related to the disease, we use three plausible values based on a literature review. We also include four different class size thresholds. See Section 3 of the Supplemental Material for more details.

2.2 Simulation details

Our simulation is run in discrete time over a period of 90 days, which corresponds roughly to a semester at SFU. On each day, we simulate the dynamics described in Sect. 2.1. On a particular day, new cases can only arise in classes which meet on that day, but we do allow all individuals who are infected to (possibly) progress to a more advanced stage

[2] Recall that we treat tutorials and labs independently of the courses with which they are associated. We also apply this strategy to the removal of classes from the network. That is, it is possible for the lecture portion of a course to be moved online while labs continue to meet in person. This is consistent with SFU's early strategy of prioritizing in-person meeting of classes with experiential components.

of the disease. We initialize our simulation by randomly moving ten individuals to the **I2** compartment[3] (i.e., symptomatic infected). At each time step, we track the number of individuals in each compartment.

We repeat our simulation 10 times at each parameter combination. This gives us a total of 26,244 sets of 10 disease trajectories. Computation took approximately 18 hours using the Digital Research Alliance of Canada's Cedar cluster (https://alliancecan.ca/en).

3 Analysis

In this section, we describe the analysis we perform on the output of our simulation. This includes the selection of a small number of summary statistics of the disease trajectories, as well as the associated descriptive analysis and modelling of these summaries. The results of our analysis are presented in Sect. 4, and interpretation is presented in Sect. 5.

3.1 Summarizing the trajectories

To avoid characterizing the entire trajectories simultaneously, we summarize each curve with a pair of statistics: the proportion of the population who ever becomes infected, and the peak infection size. The former, defined as the proportion of individuals who leave compartment **S** by the end of term, is referred to in the epidemiology literature as the cumulative incidence of infection, or CII (Cowling and Wong 2020). The latter measure, peak infection size, is defined as the largest proportion of individuals simultaneously outside compartments **S** and **R** (i.e., the proportion of individuals among **E**, **A**, **I1** and **I2**). While the peak infection size is closely connected to CII, the CII measures impact of the disease across the entire term, while peak infection size measures the largest instantaneous number of cases.

Both of our summaries are defined as proportions of the population size. However, there is some ambiguity in the definition of these proportions, since the number of students changes for different class size thresholds. We also restrict attention to the largest connected component in each network, so the population size is not even the number of students remaining after thresholding. Unless stated otherwise, when we discuss a proportion or a population size, it is taken with respect to the number of students in the largest connected component after thresholding.

[3] We use 10 instead of 1 initial case because we want to investigate properties of outbreaks. A similar study with 1 initial case would be better able to investigate the probability of an outbreak occurring, at the expense of having less data for studying the outbreaks themselves. We discuss this in more detail in Sect. 5.2.

3.2 Statistical analysis

Recall that the purpose of our analysis is to provide interpretable results to help inform policy decisions. As such, our modelling choices favor ease of interpretation over statistical optimality.

The analysis of our two response variables is similar, so we describe the common methodology here. We begin by constructing side-by-side boxplots for each simulation parameter, summarizing the distribution of the response within each parameter level. This gives a preliminary qualitative understanding of the marginal relationships between simulation parameters and the response.

For both summaries, the difference across levels of ϕ, the class size threshold, is much greater than across levels of the other parameters. As such, we emphasize the effect of class size threshold as a predictor throughout our analysis. Since the differences are so large across thresholds, we produce a histrogram of the outcome at each threshold level.

One extreme outlier was detected for both outcomes in the threshold = 100 group. This outbreak has an order of magnitude fewer infections that the other simulation runs with the same parameter settings. This behaviour is due to the infection being slow to get going (although the outbreak never becomes extinct). Because of its wildly different behaviour from similar simulation runs, we opt to remove the outlier from analysis. We do, however, retain the other runs at this parameter combination.

Next, we fit a regression tree model to explain the mean response using our parameters as covariates (see, e.g., Breiman et al. 1984). Regression trees have the advantage of being easily interpretable, even in high-dimensional settings. This model also highlights the relative importance of each variable for predicting the response. Briefly, a regression tree recursively partitions the predictor space by choosing a predictor and a value of that predictor with which to divide the space into 'low' and 'high'. Within each new subregion, the response is predicted by its mean over the sample observations in that region. The choice of predictor and dividing value is made to minimize the global sum of squared errors over all such splits. This partitioning process is then applied repeatedly, with splits at each step being chosen from among those that could be made in any of the subregions defined up to that point (i.e., recursive partitioning). The result of the recursive partitioning algorithm can be visualized in a tree shape, starting with a 'root' node, and with each split replacing an existing node with two 'child nodes'. It is common to continue splitting until a very large tree is produced (which almost certainly overfits the data), then to choose a smaller

subtree using cross-validation.[4] The selection of a subtree is evocatively called 'pruning'.

Since both of our responses are proportions, before doing any splitting we apply a logit transformation (i.e., $x \mapsto \log[x/(1-x)]$) to get a range more compatible with the squared error loss used by regression trees. As mentioned above, class size threshold is much more strongly associated with CII than any epidemiological parameters, so we divide our data into groups based on the class size threshold and fit a separate model within each group.

At each threshold level, we begin by fitting a large tree, then prune this tree using the usual cross-validation (CV) strategy (For an introduction to cross-validation and more details, see Hastie et al. (2009). When choosing subtrees, we first find the tree with minimum average error across CV folds. Call this the CV-min tree. Then, we compute the standard deviation of the CV error across folds at the minimizer, and select the smallest tree with mean CV error no larger than minimum plus 1 standard error. Call this the CV-1se tree.). We also consider pruning to each of a small number of sizes which lend themselves well to interpretation. Specifically, we consider trees with 10, 25, 50, 100 and 200 splits. To illustrate the behaviour of these small trees, we plot the trajectory of root mean squared error (CV-RMSE) across tree sizes up to 200. The performance of these trees gets reasonably close to that of the minimum CV-RMSE tree.

Next, we report measures of variable importance and goodness-of-fit for the interpretable and CV-optimal trees discussed above. See Therneau and Atkinson (2019) and Breiman et al. (1984) for details of how variable importance is measured with regression trees. For goodness-of-fit, we report the CV-RMSE of each tree. Note that, since the tuned trees are chosen to optimize the CV-RMSE, this performance metric is biased for those trees and should be interpreted cautiously (see, e.g., Hastie et al. 2009).

Finally, for each class size threshold we choose a small tree which performs fairly well, and explore which splits are actually made. Since this conveys similar information to the variable importances discussed above but with more detail, we include plots of the splits made for the chosen trees in Section 4 of the Supplemental Material.

Computation for our analysis was done in `R` (R Core Team 2021) using the `rpart` (Therneau and Atkinson 2019) package and ran nearly instantaneously on an author's laptop.

[4] For an introduction to cross-validation and more details, see Hastie et al. (2009). When choosing subtrees, we first find the tree with minimum average error across CV folds. Call this the CV-min tree. Then, we compute the standard deviation of the CV error across folds at the minimizer, and select the smallest tree with mean CV error no larger than minimum plus 1 standard error. Call this the CV-1se tree.

4 Results

In this section, we present the output of our analysis without any discussion. See Sect. 5 for interpretation of these results. We first present all results for the cumulative incidence of infection (CII), then move on to peak outbreak size.

4.1 Cumulative incidence of infection

In Fig. 2, we give boxplots of the CII across levels for each of the simulation parameters. Each subplot gives the CII at each level of one of the parameters, averaged across levels of all other parameters.

Figures 3 and 4 give histograms of the CII for each class size threshold. Axis scales are held fixed in Fig. 3 and allowed to vary between plots in Fig. 4. These two histograms highlight, respectively, the differences between threshold levels and features of the distribution within each threshold.

Table 3 gives the number of splits and the root mean squared CV error (CV-RMSE) for both CV trees across all class size thresholds when predicting logit-CII. Recall that the logit transformation is applied to the response before any model fitting is performed, and that CV-RMSEs are reported on the logit-scale.

Figure 5 gives the logit-CII CV-RMSE as a function of tree size, focusing attention on trees with few splits (i.e., at most 200). The curve gives the CV-RMSE for each split. The horizontal line gives the minimum CV-RMSE over all numbers of splits. The vertical lines correspond to subtrees with 10, 25, 50, 100 and 200 splits. The horizontal ticks along the vertical axis give the CV-RMSE of each of the five subtrees of interest.

Table 4 gives variable importance measures for some trees at each class size threshold when predicting logit-CII. The values for each tree have been re-scaled to sum to one across variables. See Sect. 2.1 for parameter definitions.

Table 5 gives the CV-RMSEs of some selected trees at each class size threshold when predicting logit-CII. As discussed in Sect. 3.2, reported CV-RMSEs for the CV-1se and CV-min trees are biased due to the optimization involved in selecting these trees.

Taken together, the above results suggest that 25 splits provides a good balance between interpretability and capturing most of the possible improvement in CV-RMSE when predicting logit-CII. Ideally, we would use a larger tree, but adding more splits quickly makes the tree infeasible to visualize and interpret. Furthermore all trees CV-RMSE values are quite small. See Section 4 of the Supplemental Material for the pruned trees with 25 splits at each threshold level.

All trees split first on θ_{I2}, the infectiousness parameter for symptomatic cases. For threshold levels of 20 and 50, the

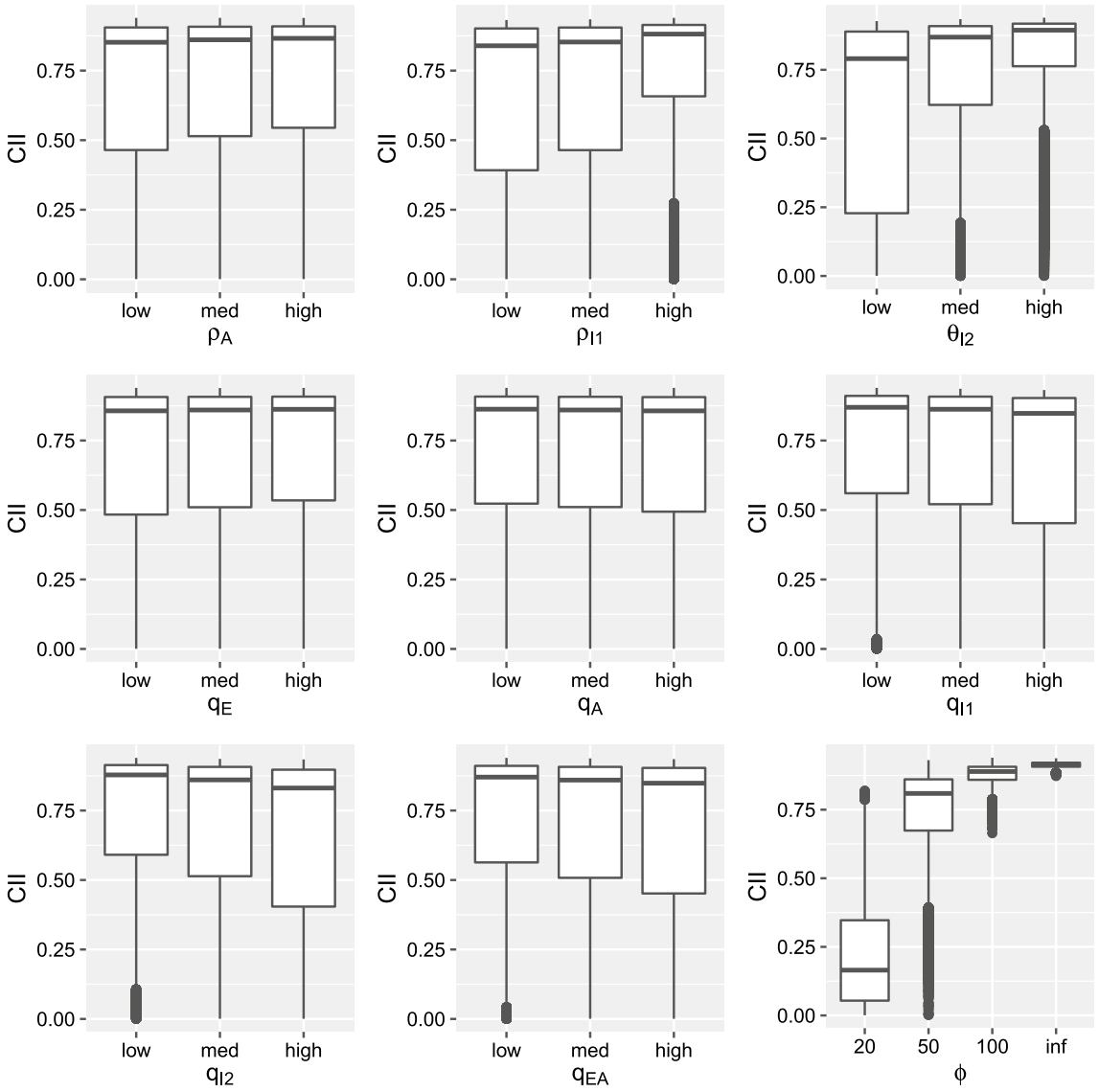

Fig. 2 Boxplots of CII across levels for each simulation parameter

next step includes splitting up whichever pair of θ_{I2} levels remained together after the first split. These trees then split on ρ_{I1}, the relative infectiousness of presymptomatic individuals, followed by the holding time parameters for presymptomatic and sympomatic individuals. For threshold levels of 100 and ∞, the second stage splits on q_{I2}, the holding time parameter for symptomatic individuals. These trees then split on θ_{I2} if possible (i.e., in the group with two remaining levels of this predictor), and on ρ_{I1}. There is remarkable similarity between the trees fit at threshold levels of 20 and 50, as well as between thresholds of 100 and ∞.

4.2 Peak outbreak size

In Fig. 6, we give boxplots of the peak outbreak size across levels for each of the simulation parameters. Each subplot gives the peak outbreak size at each level of one of the parameters, averaged across levels of all other parameters.

Figures 7 and 8 give histograms of the peak outbreak size for each class size threshold. Axis scales are held fixed in Fig. 7 and allowed to vary between plots in Fig. 8. These two histograms highlight, respectively, the differences between threshold levels and features of the distribution within each threshold.

Table 6 gives the number of splits and the CV-RMSE for both CV trees across all class size thresholds when predicting logit-peak outbreak size. The curve gives the CV-RMSE

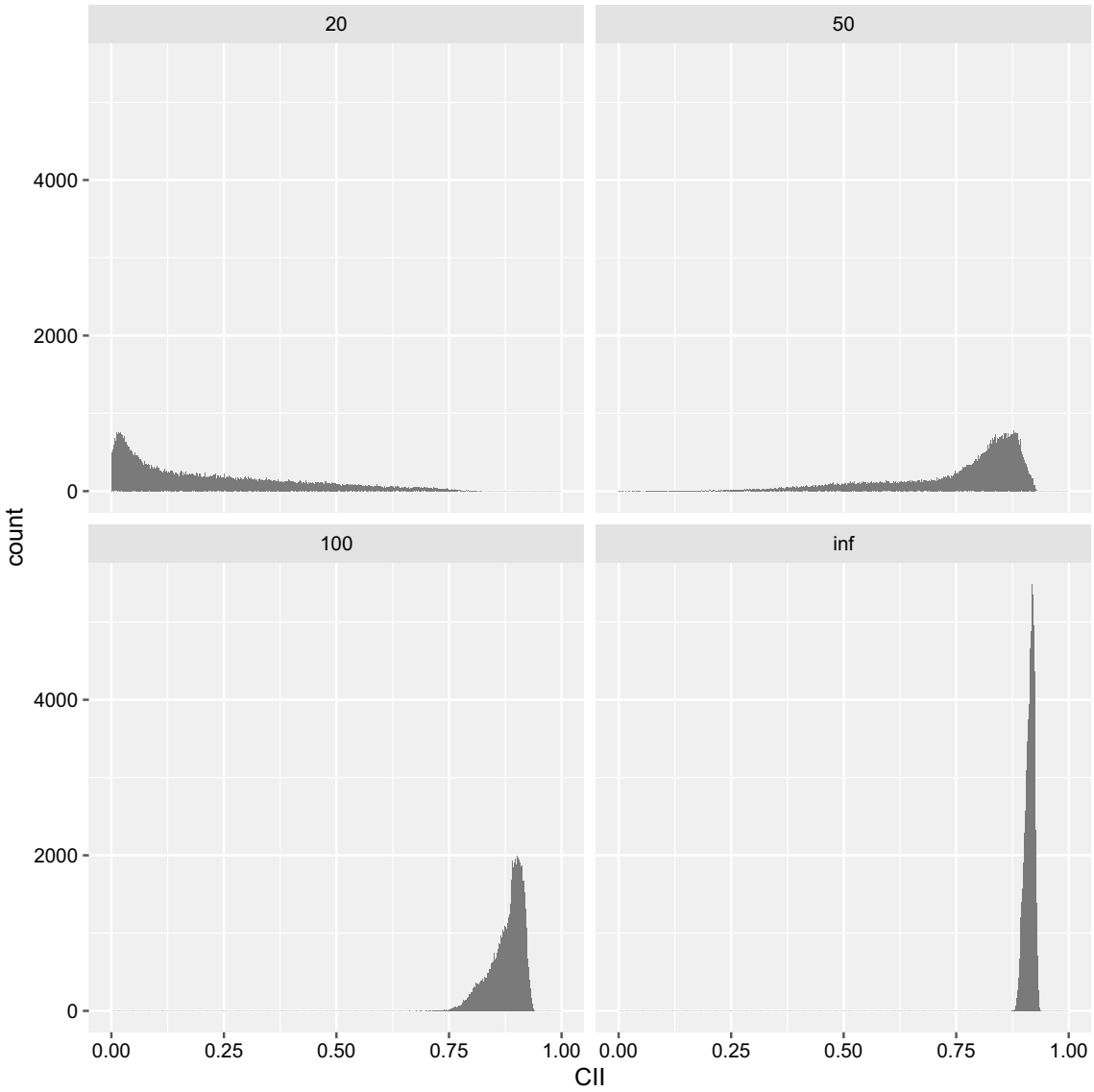

Fig. 3 Histograms of CII within each class size threshold. Axis scales held fixed across plots

for each split. The horizontal line gives the minimum CV-RMSE over all numbers of splits. The vertical lines correspond to subtrees with 10, 25, 50, 100 and 200 splits. The horizontal ticks along the vertical axis give the CV-RMSE of each of the five subtrees of interest.

Figure 9 gives the logit-peak outbreak size CV-RMSE as a function of tree size, focusing attention on trees with few splits (i.e. at most 200). Vertical lines are given at 10, 25, 50, 100 and 200 splits,[5] and ticks on the *Y*-axis show these trees' error rates. The global CV-RMSE is given by a horizontal line.

Table 7 gives variable importance measures for some trees at each class size threshold when predicting logit-peak outbreak size. The values for each tree have been re-scaled to sum to one across variables. See Sect. 2.1 for parameter definitions.

Table 8 gives the CV-RMSEs of some selected trees at each class size threshold when predicting logit-peak outbreak size. As discussed in Sect. 3.2, reported CV-RMSEs for the CV-1se and CV-min trees are biased due to the optimization involved in selecting these trees.

The above results suggest that 25 splits provides a good balance between interpretability and capturing most of the possible improvement in CV-RMSE when predicting

[5] All trees with 10 splits are worse than the best tree with 9 splits with respect to the criterion used for tuning. As such, the optimal 9-split tree is used in place of a 10-split tree. For consistency, we still refer to this as the 10-split tree.

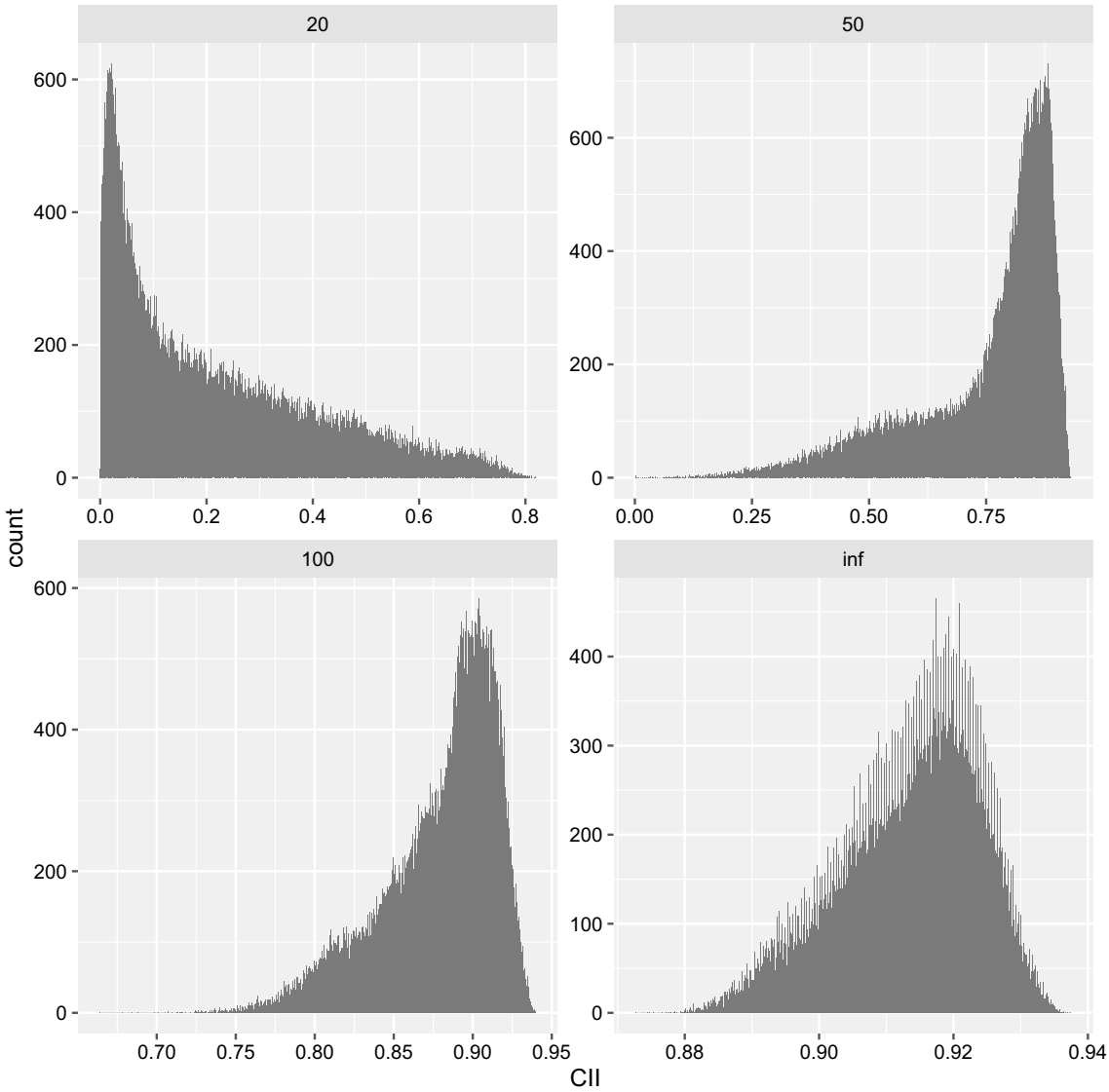

Fig. 4 Histograms of CII within each class size threshold. Axis scales differ across plots

Table 3 Summaries of CV-tuned trees for predicting logit-CII across class size thresholds

	CV-1se		CV-min	
Threshold	Splits	CV-RMSE	Splits	CV-RMSE
20	184	0.64	348	0.64
50	591	0.22	5660	0.21
100	1751	0.05	3629	0.05
∞	520	0.04	854	0.04

logit-peak outbreak size. Although the relative difference in CV-RMSE between the 25 split and CV-min trees appears large in Fig. 9, the absolute difference is quite small (see, e.g. Table 8 and the corresponding results for CII in Table 5). Ideally, we would use a larger tree, but adding more splits quickly makes the tree infeasible to visualize and interpret. Furthermore all trees CV-RMSE values are quite small. See Section 4 of the Supplemental Material for the pruned trees with 25 splits at each threshold level.

When predicting peak outbreak size, the three thresholded groups (i.e. threshold level of 20, 50 or 100) are quite similar, while the unthresholded group is different. In all threshold levels other than ∞, the first split is on θ_{I2}, the infectiousness parameter for symptomatic individuals. However, in the unthresholded group, the first split is on q_{I2}, the holding time parameter for symptomatic individuals. At the next step, splits are made either on q_{I2} for the thresholded trees, or on θ_{I2} and q_{I1}, the holding time parameter for presymptomatic individuals, when no thresholding is applied. The trees start to diverge

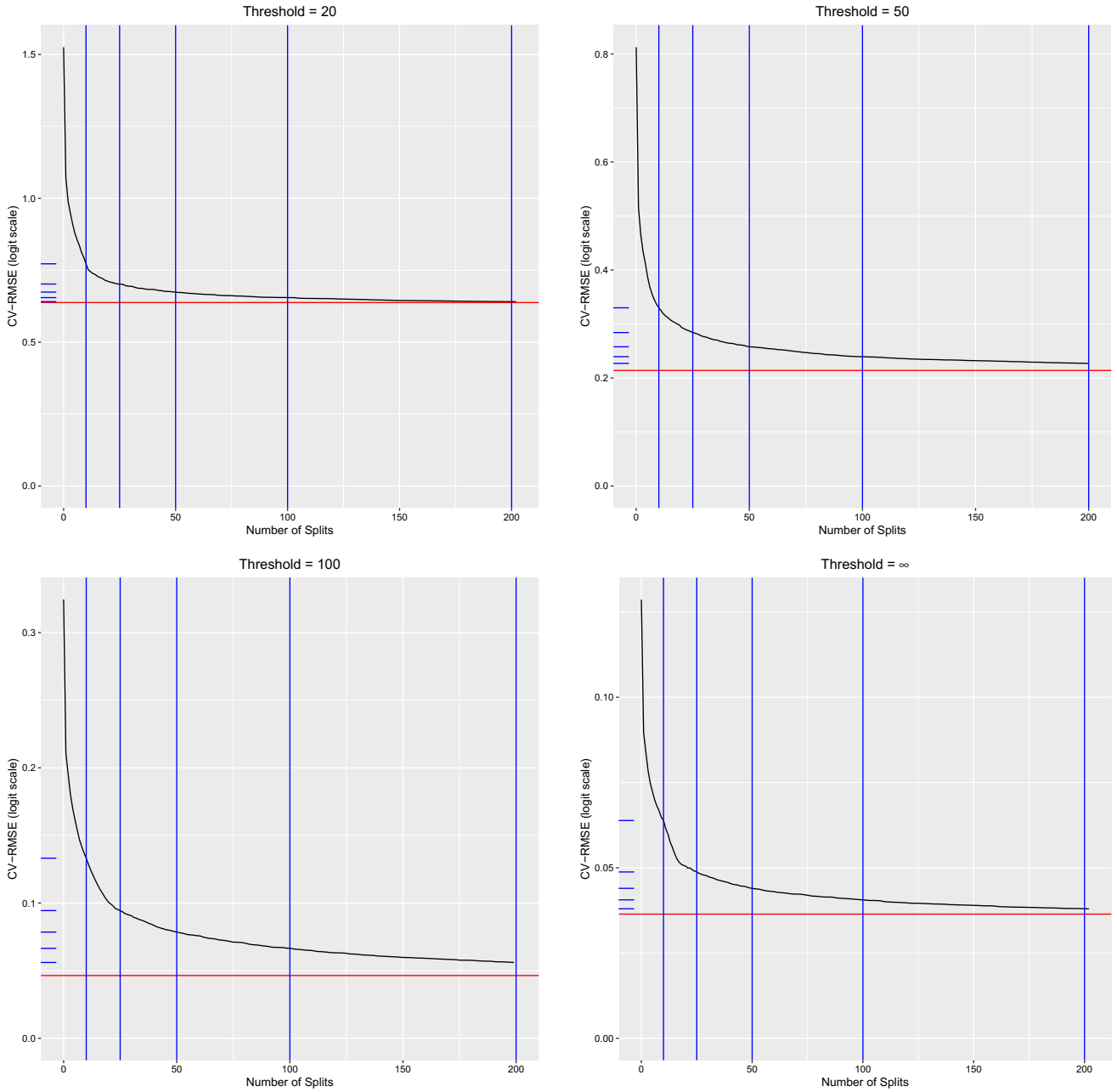

Fig. 5 CV-RMSE for predicting logit-CII across tree sizes for each class size threshold. Vertical lines correspond to trees with 10, 25, 50, 100 and 200 splits, with ticks on the *Y*-axis at these trees' CV-RMSE values. The horizontal line is the global minimum

at the next level, with splits being made on θ_{I2}, q_{I1}, q_{I2} and ρ_{I1}, the relative infectiousness of presymptomatic to symptomatic individuals.

5 Discussion

In this section we discuss the findings of our study and some ideas for future work. Section 6 gives interpretation and implications of these findings.

The results of our simulation show that moving classes online is strongly associated with lessening the severity of a disease outbreak. Conversely, the differences across levels of any single epidemiological parameter are small when averaged across the other parameters. This is true whether severity is measured by total number infected (a.k.a. cumulative incidence of infection, or CII) or by peak simultaneous case count. Differences across class size thresholds is most pronounced for the CII, see Figs. 2, 3, 4, where if all classes are allowed to meet in person,

Table 4 Variable importance measures for selected trees of interest in each class size threshold for predicting logit-CII. Values of ≈ 0 round to 0. Blank cells indicate that no splits were made on that variable by that tree

Threshold	Tree	ρ_A	ρ_{I1}	θ_{I2}	q_E	q_A	q_{I1}	q_{I2}	q_{EA}
20	10		0.12	0.78			0.06	0.04	
	25		0.11	0.73			0.06	0.09	≈ 0
	50	≈ 0	0.11	0.72			0.06	0.09	0.01
	100	0.01	0.11	0.71		≈ 0	0.06	0.09	0.02
	200	0.01	0.11	0.70	≈ 0	≈ 0	0.06	0.09	0.02
	CV-1se	0.01	0.11	0.70	≈ 0	≈ 0	0.06	0.09	0.02
	CV-min	0.01	0.11	0.69	≈ 0	≈ 0	0.06	0.09	0.02
50	10		0.12	0.77			0.05	0.07	
	25	≈ 0	0.13	0.73			0.05	0.09	0.01
	50	0.01	0.13	0.71			0.05	0.09	0.02
	100	0.01	0.13	0.70			0.05	0.09	0.02
	200	0.01	0.13	0.69	≈ 0	≈ 0	0.05	0.09	0.03
	CV-1se	0.02	0.13	0.68	≈ 0	≈ 0	0.05	0.08	0.03
	CV-min	0.02	0.13	0.68	0.01	0.01	0.06	0.08	0.03
100	10		0.07	0.76			0.03	0.14	
	25	≈ 0	0.10	0.69			0.05	0.14	0.01
	50	0.01	0.10	0.67			0.05	0.14	0.02
	100	0.01	0.10	0.66			0.06	0.14	0.03
	200	0.01	0.10	0.65		≈ 0	0.06	0.14	0.03
	CV-1se	0.02	0.10	0.65	≈ 0	≈ 0	0.06	0.14	0.03
	CV-min	0.02	0.10	0.65	≈ 0	≈ 0	0.06	0.14	0.03
∞	10		0.10	0.73			0.02	0.15	
	25		0.11	0.67			0.06	0.15	0.01
	50	0.01	0.10	0.65			0.06	0.16	0.02
	100	0.01	0.11	0.63			0.07	0.15	0.03
	200	0.01	0.11	0.62		≈ 0	0.07	0.15	0.03
	CV-1se	0.02	0.11	0.62	≈ 0	≈ 0	0.07	0.15	0.04
	CV-min	0.02	0.11	0.62	≈ 0	≈ 0	0.07	0.15	0.04

Table 5 CV-RMSE for predicting logit-CII using selected trees across class size thresholds. *CV-RMSEs for trees chosen based on this metric are optimistically biased

Threshold	10	25	50	100	200	CV-1se*	CV-min*
20	0.77	0.70	0.67	0.66	0.64	0.64	0.64
50	0.33	0.28	0.26	0.24	0.23	0.22	0.21
100	0.13	0.09	0.08	0.07	0.06	0.05	0.05
∞	0.06	0.05	0.04	0.04	0.04	0.04	0.04

most students become infected over the course of a term. However, if all classes with more than 20 students are moved online, most simulation runs have well below 50% infections. The effect is weaker for peak infection size, see Figs. 6, 7, 8, but it is still clearly preferable to keep large classes online from this perspective.

To elaborate, for both CII and peak outbreak size, we see a strong qualitative difference between a threshold of 20 versus the other levels. Specifically, thresholding at 20 gives a right-tailed distribution with most of the mass concentrated near 0. Increasing the threshold gives either a left-tailed or symmetric distribution for CII or peak outbreak size, respectively, with values concentrated away from 0. Unsurprisingly, as the threshold level increases, the distribution moves farther from 0. Said differently, with more classes allowed to meet in person, more students become infected, regardless of whether total or peak case numbers are being counted. A cursory analysis (not shown) indicates that the low proportions of cases among more severe thresholding levels seen in Fig. 3 are because the outbreak did not have time to finish, not because it stalled (i.e. there are still many contagious individuals).

In our fitted tree models, most of the predictive power is captured by a small number of splits, relative to the performance of a full-sized tree (with the possible exception of the peak outbreak size at larger threshold levels, although

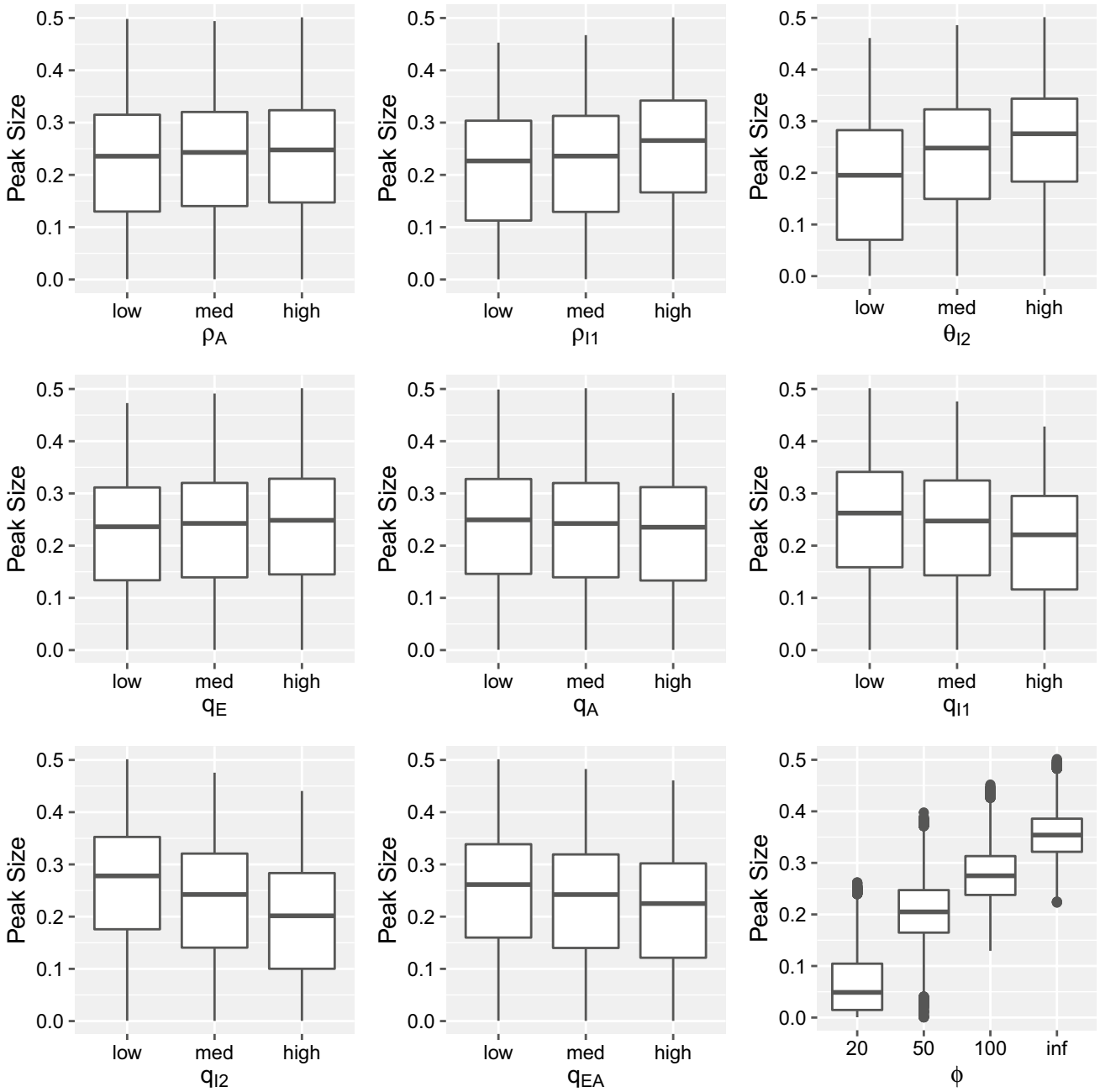

Fig. 6 Boxplots of peak outbreak size across levels for each simulation parameter

the absolute differences there are small). This is fortunate, since it means that any monitoring work done to anticipate severity of potential outbreaks can focus on a small number of parameters. Furthermore, being able to capture most of the information from a tree in only a small number of splits prevents us needing to interpret large trees.

Relative contributions of the various epidemiological parameters differ across thresholds and across responses (see Tables 4 and 7). For CII, we see most of the importance concentrated on one predictor; specifically, p_{I2}, the infectiousness parameter for symptomatic cases. For peak outbreak size, we start with most of the importance concentrated on p_{I2} when thresholding at 20 students, but as we allow larger classes, p_{I2} becomes less important and other variables become more important. Specifically, q_{I2}, the holding time parameter for the sympomatic compartment, matches the importance of p_{I2} when thresholding at 100 students, and exceeds the importance of p_{I2} when no thresholding is applied. We also see ρ_{I1} and q_{I1}, the relative infectiousness and holding time parameters, respectively,

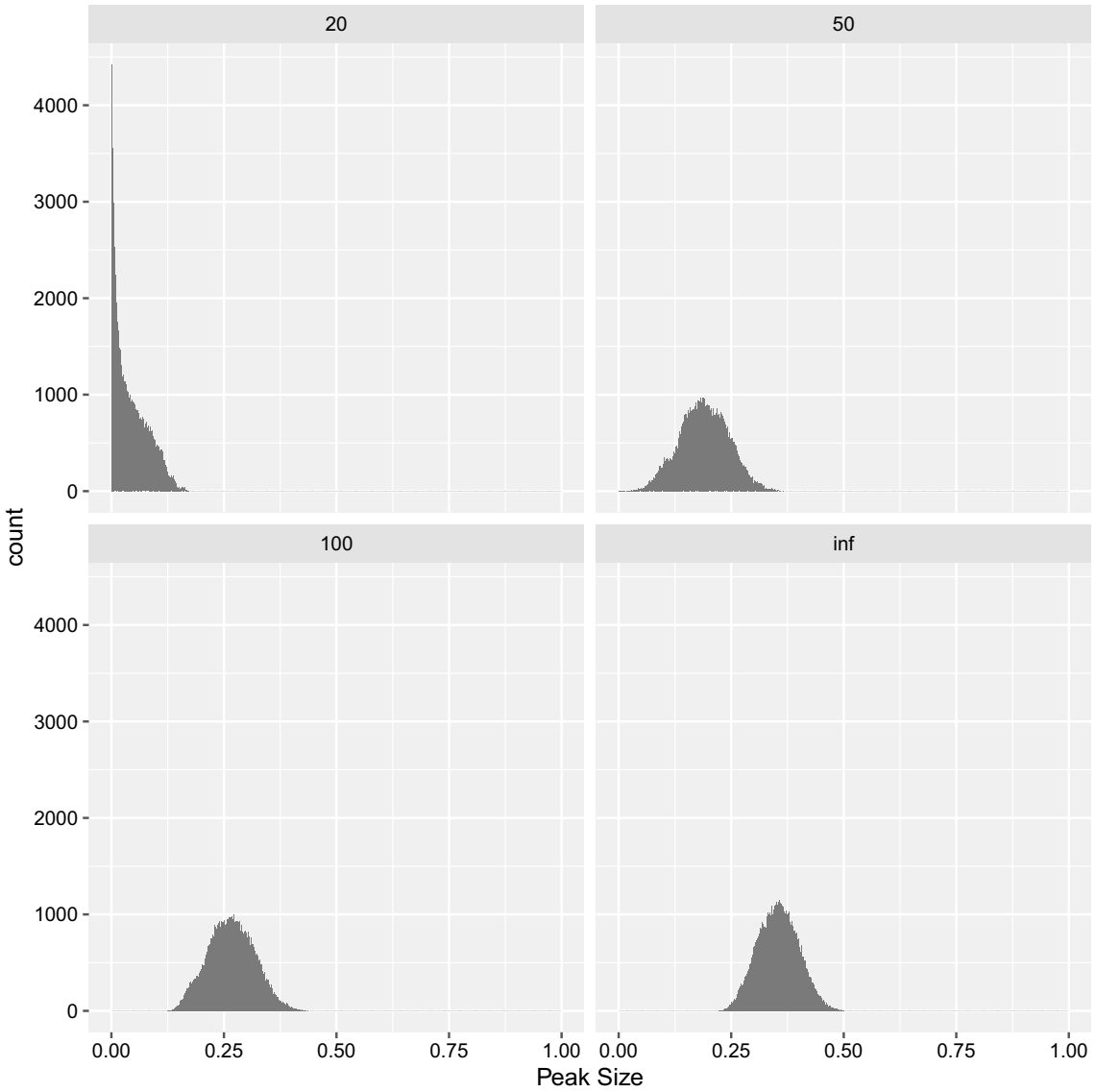

Fig. 7 Histograms of peak outbreak size within each class size threshold. Axis scales held fixed across plots

for the presymptomatic compartment, become much more important relative to θ_{I2} as thresholding is weakened. These results are consistent with the actual splits made early in tree fitting; see Section 4 of the Supplemental Material.

A high importance score for θ_{I2} does not necessarily tell us that symptomatic cases are the primary driver of infection in our model. In fact, we expect θ_{I2} to appear important because we have parameterized the other compartments' infectiousness values relative to that of the symptomatic compartment. That is, if we change the infectiousness of sympomatic cases, we also change the infectiousness of the other contagious compartments, while the converse is not true. However, the high importance score of q_{I2} for high threshold levels with peak outbreak size suggest that the symptomatic compartment is, in fact, an important determinant of the infection's peak severity over the course of a term.

Several variables either are not selected for splitting or have a very low importance score: specifically, ρ_A, the relative infectiousness of asymptomatic cases, and q_E and q_A, the holding time parameters for exposed status and asymptomatic cases, respectively. The low importance of duration spent in the exposed compartment is unsurprising, since this compartment neither transmits nor receives infection. If exposed durations were often of a similar order to the duration of the simulation (90 days), then we would expect to see a larger effect, where many individuals never progress to the contagious phase of the disease. However, under our chosen parameter values, the mean time spent in the exposed compartment never exceeds 6 days. The low

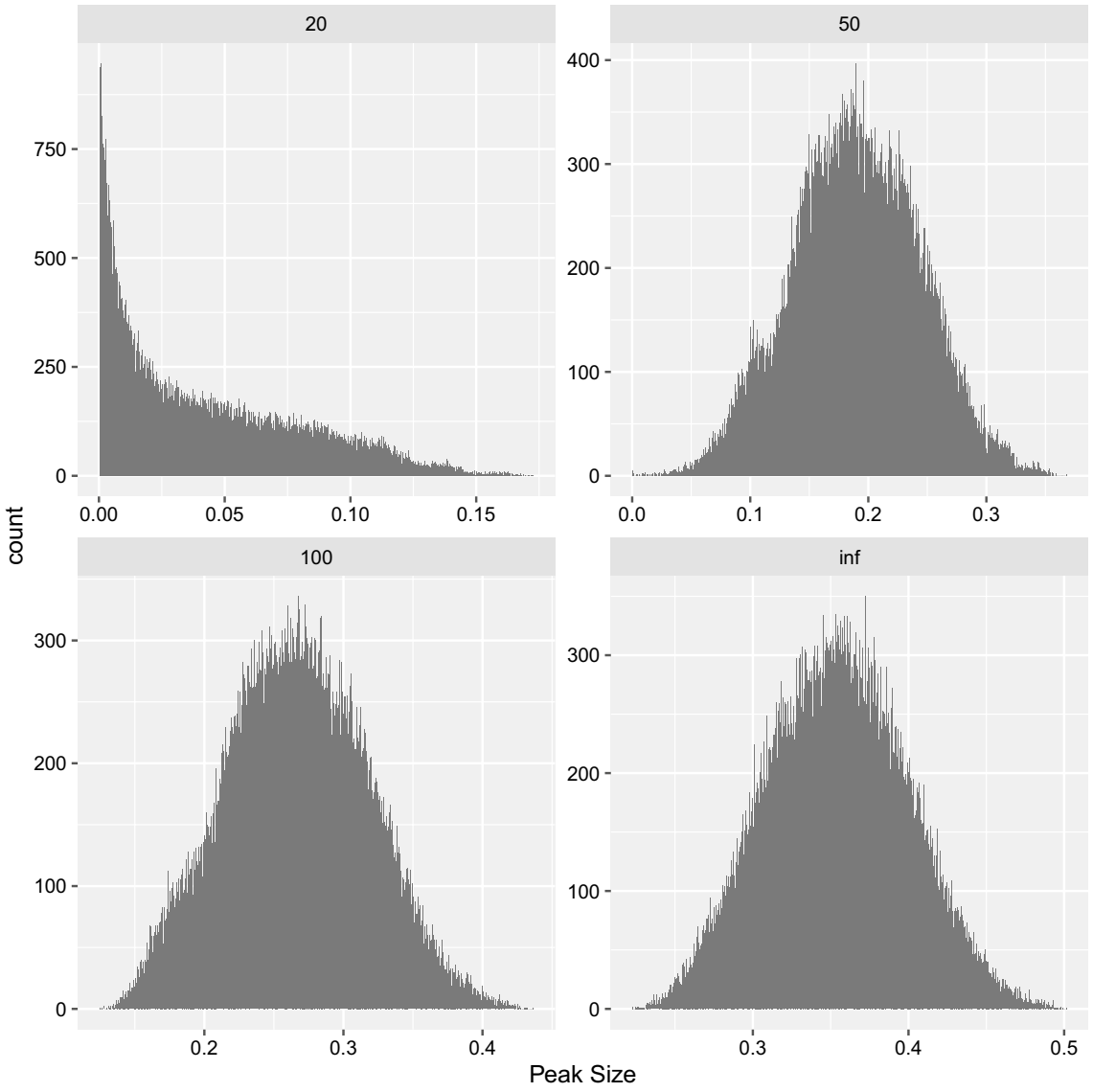

Fig. 8 Histograms of peak outbreak size within each class size threshold. Axis scales differ across plots

Table 6 Summaries of CV-tuned trees for predicting logit-peak outbreak size across class size thresholds

	CV-1se		CV-min	
Threshold	Splits	CV-RMSE	Splits	CV-RMSE
20	223	0.59	508	0.59
50	565	0.11	6228	0.10
100	4616	0.03	6057	0.03
∞	4485	0.02	6033	0.02

importance of parameters associated with asymptomatic cases is also unsurprising due to the relatively low infectiousness of individuals in this compartment (see Table 2). In fact, if we remove the single large value of ρ_{I1}, the relative infectiousness of presymptomatic cases, then the importance of this variable drops to be closer to that of ρ_A.

Our control strategy of removing classes above a specified threshold eliminates a substantial proportion of enrollments from the network. To ensure that any improvement seen is not only due to this 'thinning' of paths along which the infection can spread, we repeat our simulation with enrollments removed uniformly at random instead of according to the more systematic class size threshold strategy. More specifically, for each class size threshold other than ∞, we remove enrollments chosen uniformly at random until the number remaining matches the number of enrollments among classes below that threshold. We call this process 'thinning the network'. Note that thinning is applied after removing classes with only one

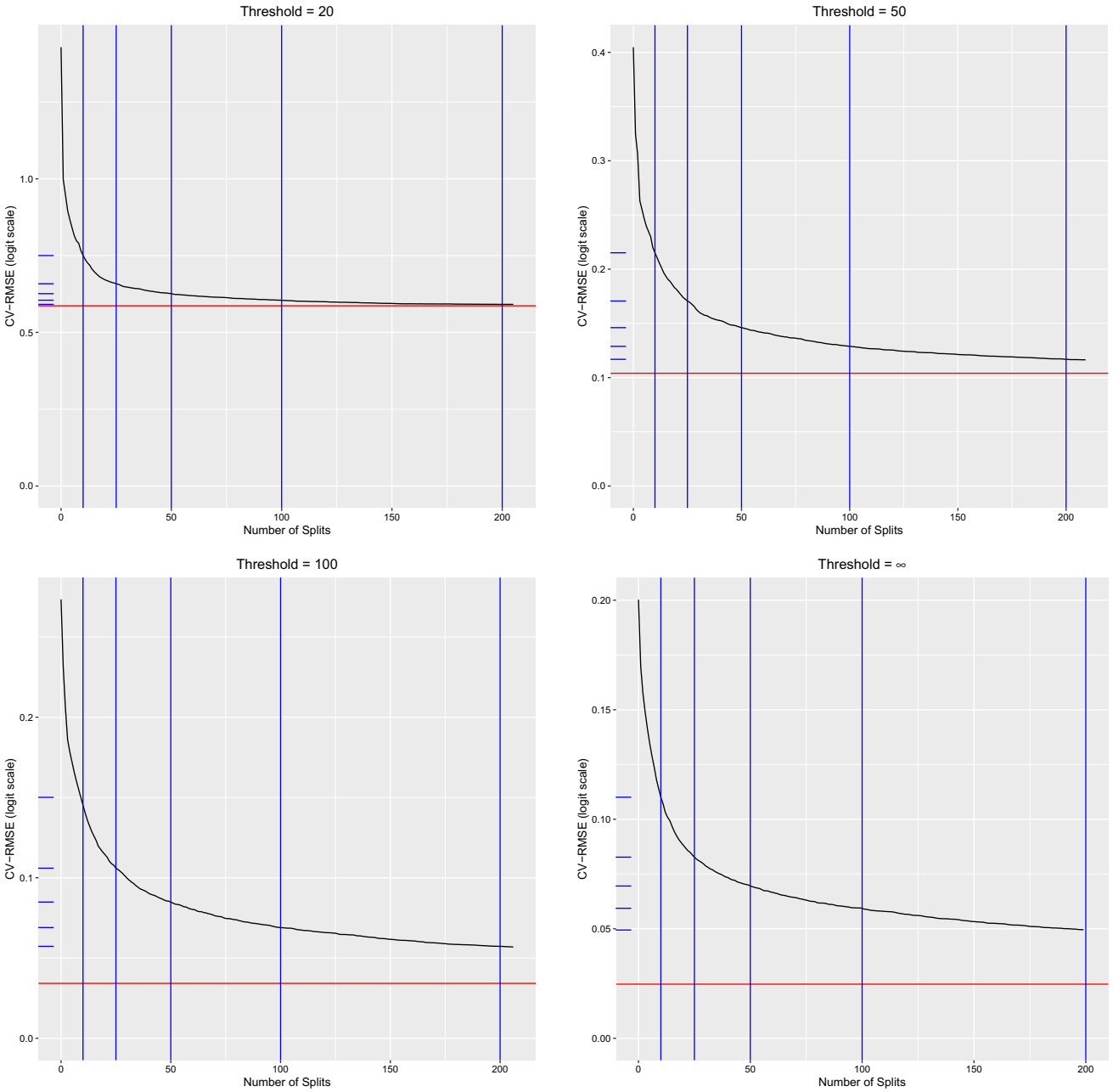

Fig. 9 CV-RMSE for predicting logit-peak outbreak size across tree sizes for each class size threshold. Vertical lines correspond to trees with 10, 25, 50, 100 and 200 splits (All trees with 10 splits are worse than the best tree with 9 splits with respect to the criterion used for tuning. As such, the optimal 9-split tree is used in place of a 10-split tree. For consistency, we still refer to this as the 10-split tree.), with ticks on the *Y*-axis at these trees' CV-RMSE values. The horizontal line is the global minimum

student but before removing isolated components. Removing isolated components both before and after thinning leads to an excess of removed enrollments. Finally, we remove any isolated components and repeat the simulation as described in the rest of this section. The only difference is that here the parameter ϕ only takes values 20, 50 and 100, and is thought of as indexing the degree of thinning rather than as an explicit threshold (smaller values of ϕ correspond to smaller networks after thinning). Results of these simulations (not shown) suggest that our thresholding strategy is considerably more effective than thinning at random.

Table 7 Variable importance measures for selected trees of interest in each class size threshold for predicting logit-peak outbreak size. Values of ≈ 0 round to 0. Blank cells indicate that no splits were made on that variable by that tree

Threshold	Tree	ρ_A	ρ_{I1}	θ_{I2}	q_E	q_A	q_{I1}	q_{I2}	q_{EA}
20	10		0.08	0.77			0.02	0.13	
	25		0.09	0.71			0.06	0.14	0.01
	50	≈ 0	0.09	0.69			0.06	0.14	0.02
	100	0.01	0.09	0.68		≈ 0	0.06	0.14	0.03
	200	0.01	0.09	0.67	≈ 0	≈ 0	0.06	0.14	0.03
	CV-1se	0.01	0.09	0.67	≈ 0	≈ 0	0.06	0.14	0.03
	CV-min	0.01	0.09	0.66	≈ 0	≈ 0	0.06	0.14	0.03
50	10		0.08	0.50			0.05	0.37	
	25		0.10	0.46			0.08	0.33	0.03
	50	0.01	0.10	0.44			0.09	0.31	0.05
	100	0.01	0.11	0.43		≈ 0	0.10	0.30	0.06
	200	0.01	0.11	0.42	≈ 0	≈ 0	0.10	0.29	0.07
	CV-1se	0.01	0.11	0.41	≈ 0	0.01	0.10	0.29	0.07
	CV-min	0.02	0.10	0.41	0.01	0.01	0.10	0.28	0.06
100	10		0.07	0.40			0.07	0.45	
	25		0.13	0.34			0.12	0.37	0.04
	50		0.12	0.34			0.12	0.35	0.06
	100	0.00	0.12	0.33		0.00	0.13	0.34	0.08
	200	0.00	0.12	0.33	0.00	0.01	0.13	0.33	0.08
	CV-1se	0.01	0.12	0.32	0.01	0.02	0.12	0.32	0.08
	CV-min	0.01	0.12	0.32	0.01	0.02	0.12	0.32	0.08
∞	10		0.16	0.22			0.22	0.40	
	25		0.14	0.20			0.18	0.41	0.07
	50		0.13	0.19		≈ 0	0.18	0.40	0.10
	100		0.13	0.19	≈ 0	0.01	0.18	0.38	0.10
	200	≈ 0	0.13	0.19	0.02	0.01	0.18	0.37	0.10
	CV-1se	0.01	0.12	0.19	0.03	0.02	0.17	0.35	0.10
	CV-min	0.01	0.12	0.19	0.03	0.02	0.17	0.35	0.10

Table 8 CV-RMSE for predicting logit-peak outbreak size using selected trees across class size thresholds. *CV-RMSEs for trees chosen based on this metric are optimistically biased

Threshold	10	25	50	100	200	CV-1se*	CV-min*
20	0.75	0.66	0.63	0.61	0.59	0.59	0.59
50	0.21	0.17	0.15	0.13	0.12	0.11	0.10
100	0.15	0.11	0.08	0.07	0.06	0.03	0.03
∞	0.11	0.08	0.07	0.06	0.05	0.02	0.02

5.1 Limitations

Our study has some limitations which restrict the generalizability of its conclusions. First is the source of the data. Our network is constructed using only data from a single university, SFU. Since different schools will have different enrollment networks, we do not necessarily expect our conclusions to generalize. However, the methodology we use is quite general, and other institutions could repeat our analysis to see whether similar conclusions hold there. Code used to perform our simulation and data analysis, as well as the enrollment data used to generate the network, are available in an accompanying GitHub repository (Ruth and Lockhart 2022).

We now discuss some limitations of our dataset. This is not meant to be an exhaustive list, but rather to illustrate some of the challenges involved with modelling disease spread on a real population. To start, our network only links students through shared classes. As is clear from a cursory inspection of any university campus, classes are not the only way in which students interact. It is conceivable that we could incorporate data on living arrangements for students in residence, but no dataset could account for all the ways in which students meet for coffee, or stand near each other outside a classroom, or on a bus... In short, we cannot account

for all the ways in which a disease can spread throughout the student population, so instead accept that we must limit our study (and, therefore, its conclusions) to the effect of transmission through shared courses.

Another limitation is the implicit assumption that every student who is enrolled in a class attends every meeting of that class. This assumption is clearly not true. In fact, there may be systematic bias toward lower attendance for classes at less popular times (e.g. the earliest classes at SFU start at 8:30 am). SFU does not keep records of class attendance, so the data required to account for attendance in our model do not exist. Some work has been done to study rates of class attendance (Devadoss and Foltz 1996; van Blerkom 1992), but incorporating these models into our study is beyond the scope of this paper.

The last limitation we discuss relates to class scheduling. At SFU, classes meet at the same times each week, typically for one or more hours on one or more days. We were only able to obtain data for the day(s) on which a class meets. This prevents us from accounting for the amount of time actually spent in a room with classmates. Given more detailed information, we could develop a model which more closely reflects real-world behaviour, but data privacy concerns limit the specificity of the data we are able to access.

5.2 Extensions

In the previous section, we discussed some inherent limitations to our study based on the dataset we were provided. Here, we briefly mention some ways our model could be expanded to incorporate other aspects of disease transmission, as well as ideas for related analysis which are of interest but beyond the scope of a single manuscript.

As was discussed in Sect. 5.1, our model does not account for the possibility of infection outside of classes. While it would be impossible to fully model student behaviour, one might introduce a random number of infections at each time step. The addition of random infections from outside the disease model is referred to as a spark term and is discussed by Deardon et al. (2010). These additional infections would represent out-of-class interactions that take place on-campus, as well as the possibility of contracting the disease somewhere off-campus. Random infections could be assigned uniformly across the susceptible population, or a separate model could be developed to describe students' heterogeneous risks of transmission outside classes.

Our simulation uses only three distinct values for each of the disease parameters due to the sharp increase in computational cost as more values are included. Future work might focus on a finer exploration of the parameter space.

There are many possible control measures to limit further spread by infectious individuals. Examples include mask wearing and not coming to class when sick. Masking can be incorporated into an existing model by reducing transmission rates. Other work suggests that mask use is important for reducing transmission risk (Zhou et al. 2021). One can also imagine numerous strategies for keeping sick students out of classes. Examples include quarantining individuals who feel sick, or moving individual classes online if any enrolled students show symptoms. While a more comprehensive control strategy which makes use of any of the methods described here or elsewhere (see, e.g., Gressman and Peck 2020) will be more effective than any one measure in isolation, our work specifically illustrates the benefit to be gained by moving certain classes online.

Our statistical analysis is somewhat limited. While regression trees are interpretable, other methods often have better statistical properties (Hastie et al. 2009). See Section 5 of the Supplemental Material for a parallel analysis using logistic regression. It would be interesting for future work to include a more detailed machine learning analysis of our simulation output focusing on prediction instead of interpretability. Such an analysis may uncover patterns that our tree model is unable to detect.

An important feature of stochastic disease modelling is whether a single infected individual produces a full outbreak, or whether the infected cases all recover before a critical number is reached (see, e.g., Britton and Pardoux 2019). This "extinction probability" is best studied empirically using a single initial infected individual, so our framework is ill-suited to measure this quantity (we always start with ten initial cases). One way this might be studied is to repeat our simulation with a single initial case and investigate the rate at which outbreaks go extinct before infecting many students.

6 Conclusions

It is clear from a cursory analysis of our simulation results that moving large classes online is an important tool in managing the risk of a large outbreak at SFU. More precisely, to ensure that large outbreaks are unlikely (e.g. those in which more than 50% of students are infected over the course of the term), even moderately sized classes must be moved online. Applying a threshold value of 20 gives qualitatively different behaviour than any other level, with most students avoiding infection and the peak number of simultaneous cases being quite small (often below 1% of students). Although the number 20 may not be appropriate for other contexts, our findings do suggest that all but the smallest classes must be moved online to mitigate the chance of a severe outbreak.

We have also identified which disease parameters are most strongly associated with case counts. Fortunately, our models' predictive power is mostly determined by a small

number of parameters. In particular, the total number of infections is driven mostly by the disease's transmissibility rather than by the duration of infectiousness. Conversely, the largest number of simultaneous cases is more heavily influenced by infection duration; particularly when more classes are allowed to meet in-person. This increased influence is especially pronounced for symptomatic cases, which is the stage of the disease when individuals are most infectious. These results suggest that efforts to reduce the duration of infectiousness, by quarantining for example, are best focused on symptomatic individuals. This is comforting since it is much more challenging to detect asymptomatic and pre-symptomatic cases.

Our findings apply specifically to Simon Fraser University. However, other institutions can repeat our analysis on their enrollment networks to provide conclusions more specifically tailored to their unique circumstances. We stress that our findings are meant to be interpreted qualitatively and to be used alongside other analytics in the support of policy decision making.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13721-022-00375-1.

Acknowledgements The authors would like to thank Mallory Gilmore for her thoughtful feedback on an early version of this manuscript, as well as to the staff at SFU who provided us with the data.

Data Availability Statement All data and code used in this study are available from Github in the repository called SARS_CoV-2_Transmission_in_University_Classes (https://github.com/wruth1/SARS-CoV-2_Transmission_in_University_Classes). Raw data are in the folder called "Data". The output from our simulation is too large to store on Github but is available from the authors upon request.

Declarations

Funding This work was funded by the Natural Science and Engineering Research Council of Canada (RGPIN/06099-2014). Computing resources were provided by the Digital Research Alliance of Canada (https://alliancecan.ca/en). The authors have no relevant financial or non-financial interests to disclose.

References

Ambatipudi M, Gonzalez PC, Tasnim K, Daigle JT, Kulyk T, Jeffreys N, Sule N, Trevino R, He EM, Mooney DJ, Koh E (2021) Risk quantification for SARS-CoV-2 infection through airborne transmission in university settings. preprint, Occupational and Environmental Health, April. URL http://medrxiv.org/lookup/doi/10.1101/2021.03.31.21254731

Bahl R, Eikmeier N, Fraser A, Junge M, Keesing F, Nakahata K, Wang LZ (2020) Modeling COVID-19 spread in small colleges

BCCDC (2021) Epidemiology & modelling presentations. http://www.bccdc.ca/health-info/diseases-conditions/covid-19/modelling-projections, January

Bezanson J, Edelman A, Karpinski S, Shah VB (2017) Julia: a fresh approach to numerical computing. SIAM Rev 59:2017

Borowiak M, Ning F, Pei J, Zhao S, Tung HR, Durrett R (2020) Controlling the spread of covid-19 on college campuses

Brauer F (2008) Compartmental models in epidemiology. In: Brauer F, van den Driessche P, Jianhong W (eds) Mathematical Epidemiology, chapter 2. Springer-Verlag, Berlin, pp 19–79

Brauer F, Carlos C-C, Zhilan F (2019) Mathematical models in epidemiology. Springer, Berlin

Breiman L, Jerome F, Stone CJ, Olshen RA (1984) Classification and regression trees. Chapman and Hall/CRC, Boca Raton

Britton PT (2019) Stochastic epidemic models with inference. Springer, Berlin

Buitrago-Garcia D, Egli-Gany D, Counotte MJ, Hossmann S, Imeri H, Ipekci AM, Georgia S, Nicola L (2020) Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: a living systematic review and meta-analysis. PLoS Med 17:9

Byambasuren O, Bell K, McLaws L, Glasziou P (2020) Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. J Assoc Med Microbiol Infect Dis Canada 5:4

Byrne AW, David ME, Collins Aine B, Kevin H, Miriam C, Ann B, Francis B, John G, Lane Elizabeth A, Conor MA, Kirsty OB, Patrick W, Walsh Kieran A, More Simon J (2020) Inferred duration of infectious period of SARS-CoV-2: a rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. BMJ Open 10:2

Chang SL, Harding N, Cameron Z, Cliff Oliver M, Mikhail P (2020) Modelling transmission and control of the COVID-19 pandemic in Australia. Nat Commun 11:5710

Christensen H, Turner K, Trickey A, Booton Ross D, Gibran H, Emily N, Caroline R, Leon D, Matthew H, Ellen B-P (2020) Covid-19 transmission in a university setting: a rapid review of modelling studies covid-19 transmission in a university setting: a rapid review of modelling studies covid-19 transmission in a university setting: a rapid review of modelling studies. MedRxiv. https://doi.org/10.1101/2020.09.07.20189688

Clark J, Holton DA (1991) A first look at graph theory. World Scientific, Berlin

Cowling BJ, Wong JY (2020) The use of seroprevalence data to estimate cumulative incidence of infection. In: Held L, Hens N, O'Neill P, Wallinga J (eds) Handbook of infectious disease data analysis. Chapman and Hall/CRC, Boca Raton

Deardon R, Brooks SP, Grenfell BT, Keeling MJ, Tildesley MJ, Savill NJ, Shaw Darren J, Woolhouse Mark EJ (2010) Inference for individual-level models of infectious diseases in large populations. Stat Sin 20:239–261

Devadoss S, Foltz J (1996) Evaluation of factors influencing student class attendance and performance. Am J Agr Econ 78(3):499–507

Ernesto E (2020) COVID-19 and SARS-CoV-2. Modeling the present, looking at the future. Phys Rep 869:1–51

Frazier P (2020) Reopening cornell during the covid-19 pandemic. URL https://datascience.columbia.edu/event/data-for-good-peter-frazier-cornell-university/

Gressman PT, Peck JR (2020) Simulating COVID-19 in a university environment

Hastie T, Tibshirani R, Jerome F (2009) The elements of statistical learning. Springer, Berlin

Johansson MA, Quandelacy TM, Sarah K, Venkata PP, Molly S, Brooks John T, Slayton Rachel B, Matthew B, Butler Jay C (2021) SARS-Cov-2 transmission from people without COVID-19 symptoms. JAMA Netw Open 4:1

Kharkwal H, Olson D, Huang J, Mohan A, Mani A (2020) University operations during a pandemic, A flexible decision analysis toolkit

Kiss IZ, Miller JC, Simon PL (2017) Mathematics of epidemics on networks. Springer, Berlin

Martcheva M (2015) An introduction to mathematical biology. Springer, Berlin

Naik PA, Yavuz M, Qureshi S, Zu J, Townley S (2020) Modeling and analysis of COVID-19 epidemics with treatment in fractional derivatives using real data from Pakistan. Eur Phys J Plus 135:795

Naik PA, Owolabi KM, Jian Z, Mehraj-ud-din N (2021) Modelling the transmission dynamics of COVID-19 pandemic in Caputo type fractional derivative. J Multiscale Model 12:3

Naik PA, Zu J, Bilal GM, Mehraj-ud-din N (2021) Modeling the effects of the contaminated environments on COVID-19 transmission in India. Results Phys 29:104774

Özköse F, Yavuz M (2022) Investigation of interactions between COVID-19 and diabetes with hereditary traits using real data: A case study in Turkey. Comput Biol Med 141:105044

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/

Rădulescu A, Williams C, Cavanagh K (2020) Management strategies in a SEIR-type model of COVID 19 community spread. Sci Rep 10:21256

Ruth W, Lockhart R (2021) Network analysis of sfu course registrations. URL https://doi.org/10.48550/arXiv.2104.12769

Ruth W, Lockhart R (2022) SARS_CoV-2_Transmission_in_University_Classes. URL https://github.com/wruth1/SARS-CoV-2_Transmission_in_University_Classes. GitHub repository

Therneau T, Atkinson B (2019) rpart: Recursive Partitioning and Regression Trees. URL https://CRAN.R-project.org/package=rpart. R package version 4.1-15

Thompson HA, Mousa A, Dighe Amy F, Han A-PA, Peter B, Juan B-B, Qifang B, Antonio C, Liling C, De ML, Matthias H, Kiran M, Kangqi N, Jagadesan R, Gurpreet S, Biju S, Vicente S, Francesca V, Luigi V, En WL, Justin W, Ghani Azra C, Ferguson Neil M (2021) Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) setting-specific transmission rates: a systematic review and meta-analysis. Clin Infect Dis 73:3

Tuite AR, Fisman DN, Greer AL (2020) Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. CMAJ 192:19

van Blerkom ML (1992) Class attendance in undergraduate courses. J Psychol 126(5):487–494

Vespignani A, Tian H, Dye C, Lloyd-Smith JO, Eggo RM, Shrestha M, Scarpino SV, Gutierrez B, Kraemer MUG, Wu J, Leung K, Leung GM (2020) Modelling COVID-19. Nature Reviews Physics, 2: 279–281. ISSN 25225820. 10.1038/s42254-020-0178-4. URL http://dx.doi.org/10.1038/s42254-020-0178-4

Weeden KA, Cornwell B (2020) The small-world network of college classes: Implications for epidemic spread on a university campus. Sociol Sci 7:222–241

Xin HY, Li PW, Li Z, Lau EHY, Qun Y, Wang L, Cowling BJ, Tsang T, Li C (2019) Estimating the latent period of coronavirus disease COVID-19. Clin Infect Dis 746:2021

Zhou Y, Li L, Ghasemi Y, Kallagudde R, Goyal K, Thakur D (2021) An agent-based model for simulating covid-19 transmissions on university campus and its implications on mitigation interventions: a case study. Inform Discov Deliv 49:216–224. https://doi.org/10.1108/IDD-12-2020-0154