Supplement to SARS-CoV-2 Transmission in University Classes

William Ruth[1] and Richard Lockhart[2]

[1]Corresponding Author - Department of Statistics and Actuarial Science , Simon Fraser University , Burnaby, BC Canada , wruth@sfu.ca, ORCID: 0000-0002-4975-1572
[2]Department of Statistics and Actuarial Science , Simon Fraser University , Burnaby, BC Canada, ORCID: 0000-0002-7570-1982

This supplement contains five sections. In Section 1, we present a detailed investigation of our dataset and the corresponding enrollment network. In Section 2, we describe the probabalistic model used for our simulation. Section 3 contains a detailed explanation of the values chosen for our parameter values. Section 4 shows the splits made by the regression trees used for our data analysis. Finally, Section 5.2 describes an alternative analysis of our simulation output using logistic regression.

1 Dataset

In this section, we discuss some descriptive summaries of our dataset. In particular, we investigate connectedness properties of the enrollment network.

Weeden and Cornwell (2020) (hereafter WC) studied registration data for students at Cornell University in the fall of 2019 seeking understanding of the potential for the SARS-CoV-2 virus to spread from student to student by classroom contact. Their work considers three groups of students at Cornell. One group consists of undergraduate students

registered in the College of Arts and Sciences (CAS). The second group is all undergraduate students while the largest group adds in almost all graduate students.

Much like in our work, each of the three data sets used by WC is a list of records, with one record for each individual enrollment in a single course. From this data one can construct two networks: a two mode network linking students to their courses and a single mode network of students, created by regarding two students as being connected if they took the same class in the same term. An important difference between the Cornell dataset and ours is that their dataset lacks any information about when or how often classes meet. This makes their dataset conducive to network analysis, but less helpful for a detailed simulation study as in our work.

Interest centres on the connectedness of students. The idea is that students who share a class would come together in reasonably close proximity several times a week; this would be expected to increase transmission of SARS-CoV-2. These students then have other classes in common with other students to whom they could potentially pass the virus. Connectedness is measured using statistics designed for networks (aka graphs).

We perform a similar analysis on the network of students at Simon Fraser University (SFU). Details of our dataset are described in Section 1.2. While WC had access to information at multiple organizational levels (university-wide, undergraduate only and single faculty), our dataset contains no such stratification. We reproduce some of the various tables and summary statistics of WC, then discuss the comparison. In the interest of comparing quantities which are most similar, we compare our results only to those obtained by WC on their university-wide network.

Our main findings are:

1. Compared to Cornell, students at SFU take fewer courses in larger classes.

2. The student body is very interconnected at both institutions. That is, for almost all

pairs of students there is a short chain of students in common courses through which the disease could pass to get from one to the other. This similarity is despite the differences in class sizes and numbers of classes taken.

3. Many of the summary statistics used by Weeden and Cornwell (2020) don't take account of the number of short routes between students; instead they focus on the existence of any short route between two students. If the disease can spread from one student to another by many different routes then transmission from one student to the other will be much more likely than if the only way to get the disease from one to another is by following one of a small number of possible chains. We suggest below that one might compute, as another summary to be considered, the number of paths that lead from one student to another in k steps, for small values of k.

1.1 General Summary Statistics

General summary statistics for the data are in Tables 1 and 2. These tables combine data for the fall of 2019 for the full network at Cornell, taken from Table 1 in Weeden and Cornwell (2020), with the network at SFU for spring 2019, fall 2019, and spring 2020. Table 1 describes the students in these groups and Table 2 describes the courses.

The numbers of students at SFU are similar to, if slightly higher than, the number at Cornell. We have fewer courses per student and substantially lower numbers of co-enrolled students. For clarity, co-enrollment numbers are computed as follows. For each student i, count the number of students sharing a class with i; count each other student at most once. This is exactly the total of the ith row of the matrix M. Note, however, that if i and j share several classes together and i becomes infectious then the disease is more likely to be transmitted to j than if they shared only one class together. The numbers recorded in the table are the mean and median of these co-enrollment counts over all students i.

Table 1: Characteristics of students at Cornell University and at SFU. Cornell data are from Weeden and Cornwell (2020, Table 1). SD is standard deviation.

	Cornell	SFU		
	Fall 2019	Spring 2019	Fall 2019	Spring 2020
# students (n)	22,051	24,071	25,089	23,836
Courses per student				
Mean (SD)	5.4 (1.8)	4.5 (2.0)	4.7 (2.1)	4.5 (2.0)
Median	5	4	5	4
Co-enrolled students				
Mean (SD)	612 (425)	309 (241)	325 (252)	294 (217)
Median	542	258	266	256

Table 2: Characteristics of courses at Cornell University and at SFU. Cornell data is from Weeden and Cornwell (2020, Table 1). SD is standard deviation.

	Cornell	SFU		
	Fall 2019	Spring 2019	Fall 2019	Spring 2020
# courses (m)	6,072	3,546	3,789	3,552
Mean enrollment (SD)	19.5 (42.9)	30.6 (40.1)	31.0 (41.0)	30.4 (38.4)
Median enrollment	8	18	18	18
90^{th} percentile	45	62	64	65
# w/ 100-199 students	110	158	157	165
# w/ 200+ students	64	41	51	39

Cornell has considerably more courses than SFU, but SFU's courses tend to be larger. The only exception to this size difference is in the largest of courses (those with 200+ students), of which Cornell consistently has the most.

We now construct the formal networks. Specifically, the records in our dataset are used to create a bi-adjacency matrix, which we call A, linking classes and students: 1 row for each unique student and 1 column for each unique class. If student i took class j then there would be a 1 in the i, jth entry in the matrix; if not the entry is 0.

The network described by a bi-adjacency matrix can also be represented as a two-mode network (also called a bipartite graph). A two-mode network is a graph where the nodes can be divided into two groups, and edges only exist between nodes in different groups. In our case, the two groups are classes and students, and edges only exist between a student and a course that student is enrolled in (not between students or between courses).

1.2 The 2-mode network

In Table 3 we provide various summary statistics for the bipartite graphs of the Cornell data set and the three SFU terms. Cornell numbers are drawn from Table 2 of (Weeden and Cornwell, 2020). The bipartite or two-mode graph for spring 2019 at SFU could have a maximum of $n \times m = 24,071 \times 3,546 = 85,355,766$ edges. In fact, there are only $108,554$ edges: an edge density of $108,554/85,355,766 = 0.0013$. Other terms at SFU are similar, and all densities at SFU are higher than at Cornell[1]. The formula we use is not the only choice for density. If the two mode structure is ignored there are a total of $n + m = 27,617$ nodes in the graph and the maximum number of undirected edges between any two distinct nodes would be $27,617 \times 27,616/2$. WC used this second normalization in their work; we have adjusted their results to match our definition of density. We provide 1 more digit in

Table 3: Social Network Measures for the 2-mode (student-to-course) graph. Data from part of Table 2 in Weeden and Cornwell (2020) with SFU added. Network density is given as $l/(nm)$, which compares the number of edges to the maximum possible for a bipartite graph.

	Cornell	SFU		
	Fall 2019	Spring 2019	Fall 2019	Spring 2020
# students (n)	22,051	24,071	25,089	23,836
# courses (m)	6,072	3,546	3,789	3,552
# edges (l)	118,314	108,554	117,588	107,902
Network density	0.0009	0.0013	0.0012	0.0013
Largest component				
Proportion: students	0.991	0.971	0.966	0.969
Proportion: courses	0.976	0.970	0.969	0.968

our calculations than did WC because the densities are so small.

A popular concept in network analysis, and one which is relevant for thinking about chains of disease transmission, is connectedness. Consider any pair of nodes in the 2-mode graph (students, classes or one of each), and ask whether it is possible to find a chain of nodes starting with i and ending with j so that each two consecutive nodes in the chain are linked. If so then we say i and j are in the same component of M. Note that such a chain would necessarily consist of an alternating sequence of students and classes. Conceptually, the students in this chain form a sequence of potential transmissions and the classes in this

[1] As an aside, these bipartite graph densities tend to be low, and to be lower in schools with lots of students. Consider a network of n students and m courses. Suppose the average course load for a student is $\bar{L}$ courses and the average class size of $\bar{C}$. Then the number of edges in the graph is $n\bar{L} = m\bar{C}$ because these are two different ways of counting the number of rows in our original data set. The number of possible edges in the bipartite graph is nm–this would arise if every student took every course! So the actual edge density is

$$\frac{n\bar{L}}{nm} = \frac{\bar{L}}{m} = \frac{m\bar{C}}{nm} = \frac{\bar{C}}{n}.$$

That is, if class sizes are the same at two institutions then the one with more students will have a lower edge density. Or you might note that $\bar{L}$ is likely to be in the range 3 to 5 no matter what the institutional size is. So those with higher numbers of courses will likely have lower densities.

chain are vulnerable to a chain reaction of outbreaks. For the SFU network in spring 2019, we find 149 components of which one is very large. The other 148 components represent various special cases, but the results reported here focus on the single large component.

Also recorded in Table 3 are the proportions of students and courses in the largest component. That is, the number of courses and students in the largest component divided by the corresponding counts in the entire network. These numbers are very close to 1; almost all students and courses are connected to each other.

1.3 The 1-mode network

The bi-adjacency matrix can be used to build a network of students. This network is represented by an adjacency matrix with 1 row and 1 column for each student. The matrix has a 1 in row i and column j if students i and j are in some class together; the other entries are 0. For definiteness, the diagonal elements, where $i = j$, are also set to 0; you cannot transmit the disease to yourself. The graph theory jargon is that our student network / graph has no 'loops'.

This adjacency matrix is computed in three steps. First, multiply the bi-adjacency matrix above by its transpose. The resulting matrix (denoted $AA^\top$) is symmetric; the i, jth entry counts the number of courses these two students, i and j, attend together. This matrix is then replaced by a matrix of the same size which is 1 wherever the corresponding entry in $AA^\top$ is at least 1, and 0 otherwise. Finally, set the diagonal elements to 0 and call the resulting matrix M. This matrix corresponds to a 'one-mode network' or a graph.

It is somewhat more natural to discuss connectedness of the one-mode network. For any two students, say i and j, if we imagine that i is infectious, then we ask whether it is possible to find a chain of students starting with i and ending with j so that each consecutive pair of students in the chain are in a class together. If so then i and j are

Table 4: Social Network Measures for the projected 1-mode (student-to-student) graph. Data from Table 2 in Weeden and Cornwell (2020) with SFU added. SFU data reflects only the largest component of the network. Path counts are not given by Weeden and Cornwell.

	Cornell	SFU		
	Fall 2019	Spring 2019	Fall 2019	Spring 2020
# unique edges (l)	5,832,358	3,706,679	4,064,905	3,500,917
Network density	0.024	0.013	0.013	0.013
Average geodesic	2.466	2.779	2.730	2.722
Network diameter	10	16	16	15
Proportion reachable in k steps				
$k = 1$	0.024	0.014	0.014	0.013
$k = 2$	0.594	0.449	0.460	0.445
$k = 3$	0.921	0.894	0.903	0.906
$k = 4$	0.966	0.938	0.947	0.951
k-step path counts				
$k = 1$	—	7.49×10^6	8.22×10^6	7.00×10^6
$k = 2$	—	3.70×10^9	4.25×10^9	3.18×10^9
$k = 3$	—	2.09×10^{12}	2.52×10^{12}	1.63×10^{12}
$k = 4$	—	1.26×10^{15}	1.58×10^{15}	8.86×10^{14}

in the same component of M. This is equivalent to asking whether students i and j are connected in the two-mode network (the chains in the one-mode network come from those in the two node network by eliminating the courses that connect two students). We define the length of a chain to be 1 less than the number of students in the chain. That is, a chain linking two students via a third has length 2, and a chain linking a student to themself with no other links has length 0. When discussing lengths, it is common to refer to chains as paths.

Table 4 records summary statistics for the network of students; as indicated above,

statistics for SFU were computed only on the largest connected component. A few terms in this table warrant definition. First, a 'geodesic' is any path of shortest length between two nodes. The distance between two nodes is then the length of a geodesic (in our networks there are typically many geodesics between any two nodes). Two students who share a class are at distance 1. If they do not share a class but each share a class with the same third student their distance is two. The average geodesic is obtained by averaging the lengths of the geodesics between every pair of students. The network diameter is the length of the largest geodesic.

If a pair of students share a class, we say that they are reachable in 1-step (as in WC). If two students either share a class or both share classes with a common third student, we say that the original pair is reachable in 2-steps. More generally, a pair of students is called reachable in k steps if there is a path of length at most k which links these two students. Table 4 gives the proportion of student pairs which are reachable in k steps, for $k = 1, ..., 4$.

Finally, we give the number of paths of length k between any pair of students, for $k = 1, ..., 4$. Although this statistic is not given by WC, we feel that it is especially relevant to disease transmission. Any path of length 1 (i.e. any pair of students who share a class) is a (relatively) easy path for the disease to be transmitted. Any path of length 2 (i.e. any pair of students who share a class with a common third student) is a path for disease spread; albeit one along which transmission is more difficult than a 1-step path. This reduced probability of transmission is offset somewhat by the vastly larger numbers of paths of higher order. There are typically around three orders of magnitude more paths of length $k + 1$ in our networks than paths of length k.

Some connectivity properties are invariant to inclusion or exclusion of tutorials and labs, collectively referred to as small sections. Under the assumption that every student

registered in the main course is also registered in at least one of the smaller sections, the presence or absence of short paths connecting students in unaffected by whether we include the small section or not. In particular, in Table 4, this invariance property holds for the average geodesic length, network diameter, and proportion reachable in k steps. However, the number of k-step paths is not invariant to the presence of small sections. Students who share a small section of some course represent 2 paths for disease transmission, whereas students who share only the main course represent only 1 path.

There are a few summaries in which the differences between SFU and Cornell seem potentially important to us. First, the average number of co-enrolled students per student is quite a bit larger at Cornell. Table 1 of WC reports an average co-enrollment of 612. It appears, however, from a calculation in the bottom three lines of page 228 that the number which should be compared to the SFU numbers we report is 529; the difference may arise from counting the same pair of students co-enrolled in two classes as two co-enrollments and so on. Since the average class sizes are so much smaller at Cornell and since this ordering is maintained for median and 90th percentiles it appears that the extra co-enrollments may be due to some extremely large classes.

Network densities in the student to student network show the same pattern. However, recall that mean co-enrollment is network density times number of students (see Footnote 1 in Section 1.2). Since the number of students is fairly similar between the two schools, network densities just reproduce co-enrollment figures.

SFU network diameters are much larger than at Cornell. We think, however, that these diameters are much affected by things like some small graduate programs which are connected to the bulk of graduate students only by chains involving small numbers of graduate students taking single graduate courses outside their department or perhaps senior undergraduates taking a graduate course in their discipline. Patterns of cross-listing

graduate and undergraduate courses may differ between the two schools. Thus we think these differences in diameters are not likely to play an important role in whether or not an infection spreads widely.

2 Disease Model

This is a more detailed explanation than we give in the main paper. Do we need to spell things out with formulas? I find it hard to imagine that this would help with understanding.

In this section, we describe our model for SARS-CoV-2 in detail. Recall that the model has the following compartments: Susceptible, Exposed, Asymptomatic, Presymptomatic, Symptomatic and Recovered. We denote these compartments respectively by **S**, **E**, **A**, **I1**, **I2**and **R**. See Section 2.1 of the main text for a discussion of how infected individuals progress through these compartments.

On each day, the number of individuals progressing out of compartments **E**, **A**, **I1**and **I2** each follows an independent binomial random variable, with a separate parameter governing each transition (e.g. q_A is the probability of a single individial leaving compartment **A** on a particular day), and number of trials equal to the number of students in that compartment. For individuals leaving compartments **A**, **I1**and **I2**, there is only one possible destination. Conditional on the number of individuals leaving compartment **E**, the number going to **A** follows another binomial distribution, with probability q_{EA} and number of trials equal to the number of individuals leaving **E**.

The number of individuals leaving **S** on a particular day is more complicated. During a meeting of a class of size n, a single infectious individual may infect a single **S** with probability $\theta_X/\sqrt{n}$, where X is the compartment to which the infectious student belongs. Thus, the probability of a particular susceptible student becoming infected on a particular day in a single class is $\tau_* = 1 - (1-\tau_A)^{M_A}(1-\tau_{I1})^{M_{I1}}(1-\tau_{I2})^{M_{I2}}$, where M_X is the

number of individuals in the class who are in compartment X and τ_X is the transmission probability from contagious individuals in compartment X. Finally, the probability of a single **S** becoming infected on a particular day is $p = 1 - \prod_c (1 - \tau_*^c)$, where the product is over all classes, c, in which the student is enrolled and which meet on the day being considered. Thus, the total number of individuals leaving compartment **S**on a particular day is a sum over all timetables (i.e. over all observed assignments of classes to students), where each timetable is assigned an independent binomially distributed random variable with number of trials equal to the number of students with exactly that timetable and probability as derived above. Note that only classes which meet on the day in question are incorporated into the calculation of p.

3 Parameter Values

Here, we describe how we chose values for our model parameters.

3.1 Infectiousness Parameters

3.1.1 θ_{I1} - Proportionality constant for symptomatic spreaders

Thompson et al. (2021) perform a meta-analysis of secondary attack rates across various settings, and between different sub-groups. Their pooled secondary attack rate for symptomatic individuals is 0.14, with 95% CI of 0.10 to 0.17. Secondary attack rate is the probability of transmission between a particular pair. In our model, this is analogous to transmission in a class of two people, which has probability $\theta_{I1}/\sqrt{2}$. Thus, we use values for θ_{I1} of 0.141, 0.198 and 0.240.

3.1.2 ρ_A - Infectiousness of asymptomatic relative to symptomatic cases

Johansson et al. (2021) give several estimates across multiple other studies for infectiousness of asymptomatic individuals relative to symptomatic ones. Taking a few of these, we get ρ_A values of 0.4, 0.75 and 1.

3.1.3 ρ_{I2} - Infectiousness of presymptomatic relative to symptomatic cases

Buitrago-Garcia et al. (2020) give an estimate for the relative risk of infection from presymptomatic individuals as 0.63, with a 95% CI of 0.18 to 2.26. While the largest value here is quite extreme, there has been some discussion in the literature of peak transmissibility occurring before symptom onset (see, e.g., He et al., 2020).

3.2 Transition Parameters

3.2.1 q_E - Pre-infectious

Xin et al. (2021) estimate the latent period to be 5.5 days, with a 95% CI of $5.1 - 5.9$ days. Since the mean latent period in our model is $1/q_E$, we use values of 0.168, 0.182 and 0.196 for q_E.

3.2.2 q_A - Asymptomatic

Byrne et al. (2020) give several estimates of the infectiousness period for asymptomatic cases. The most appropriate one for us is from Ma et al. (2020), which is 7.25 days, with a 95% CI of $5.91 - 8.69$ days. This gives values for q_A of 0.115, 0.138 and 0.169.

3.2.3 q_{I1} - Pre-sympomatic

Xin et al. (2021) also estimate the incubation period to be 6.9 days, with a 95% CI of 6.3-7.5 days. Their estimated latent period for symptomatic cases is 5.5 days (5.1-6.0).

Subtracting the latent period for symptomatic cases from the incubation period, we get estimated mean holding times in **I1** of 1.2, 1.4 and 1.5 (subtracting bottom from bottom, middle from middle and top from top respectively). These in turn, give estimates for q_{I1} of 0.667, 0.714 and 0.833.

Alternatively, Byrne et al. (2020) give several estimates for the duration of presymptomatic infectiousness. A nice one due to He et al. (2020) is 2.3 days, with a 95% CI of $0.8 - 3.0$ days. Our model does not allow for holding times of less than one day, so we replace the lower bound from Byrne et al. (2020) with the lower bound of 1.2 from Xin et al. (2021).

Taken the above two studies together, we get values for q_{I1} of 0.333, 0.435 and 0.833.

3.2.4 q_{I2} - Symptomatic

Byambasuren et al. (2020) perform a meta-analysis of studies which report duration of symptomatic infectiousness. Many of these studies focus on hospitalized populations, since recovery is not easily defined (see their paper for more details). Their pooled estimate is 13.4 days, with a 95% CI of 10.9 to 15.8 days. This gives values for q_{I2} of 0.063, 0.075 and 0.092.

3.3 q_{EA} - Proportion of asymptomatic cases

A meta-analysis by Byambasuren et al. (2020) estimate the proportion of asymptomatic cases to be 0.18, with a 95% CI of $0.09-0.26$. We use the estimate from their random effects model as it has a larger range of values. These authors also take care to ensure that studies included in their analysis correctly distinguish asymptomatic cases from presymptomatic ones.

Parameter	Min	Center	Max	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)

Table 5: Chosen values for each parameter, along with relevant reference(s). See text for reasoning.

3.4 Summary

See Table 5 for a summary of our chosen parameter values.

4 Small Trees' Splits

This section contains plots of the first 25 splits of regression trees fit to predict (logit-transformed) CII or peak outbreak size at each class size threshold. This gives a more qualitative description of the variable importance values in Tables 4 and 7 of the main text. Since these plots are produced in `R`, they use slightly different terminology for the parameters. Conversions are given in Table 6. Figures 1-4 give the CII trees, and Figures 5-8 give the corresponding trees for predicting peak outbreak size.

5 Alternative Analysis

In this section, we outline an alternative analysis of our simulation results using logistic regression. We discuss the model used in Section 5.1 and the results of our analysis in Section 5.2. Another alternative investigation of our data with a different compartment model for SARS-CoV-2 is given in an earlier version of our paper (Ruth and Lockhart,

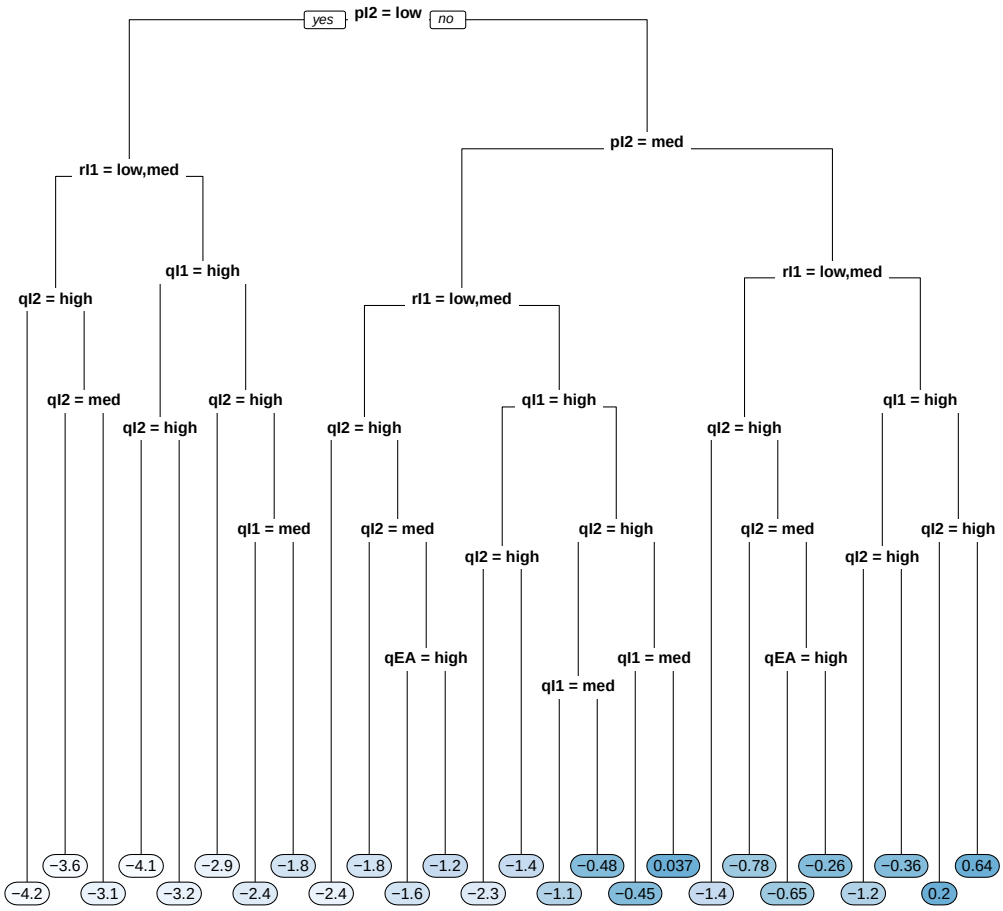

Figure 1: Pruned tree with 25 splits for predicting logit-CII with a class size threshold of 20. At each split, points which satisfy the listed condition move left, while the other points move right.

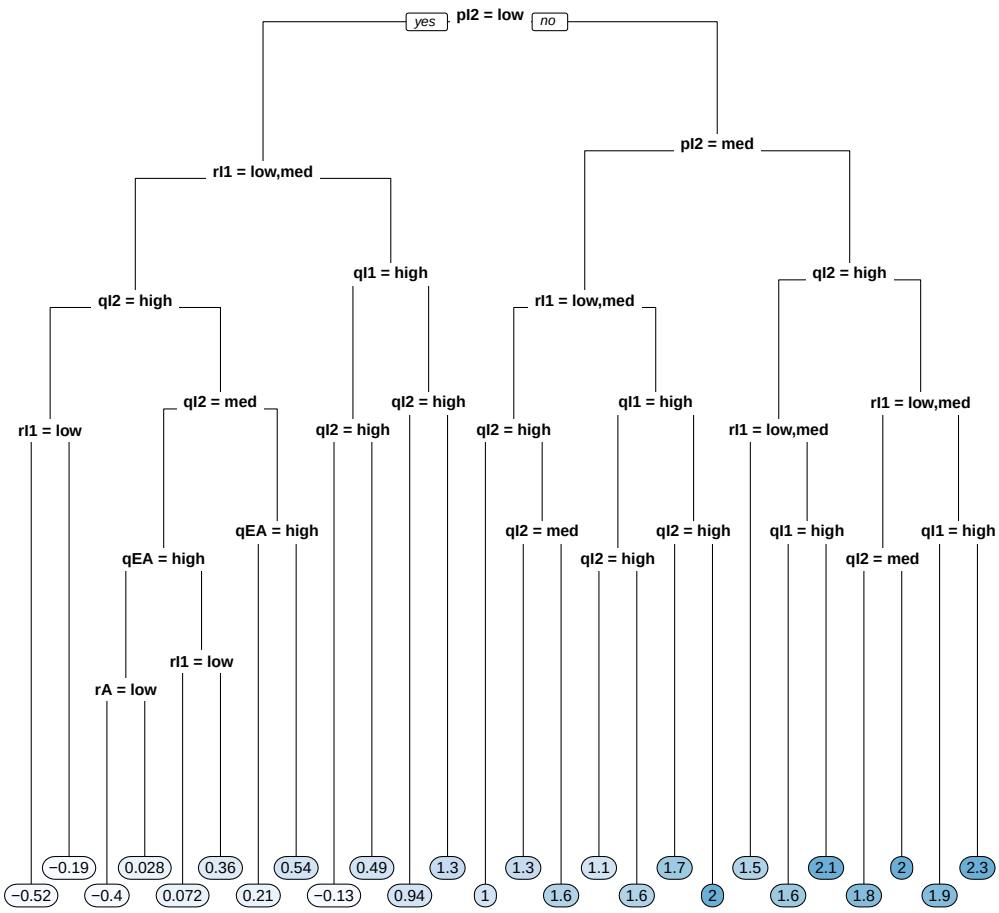

Figure 2: Pruned tree with 25 splits for predicting logit-CII with a class size threshold of 20. At each split, points which satisfy the listed condition move left, while the other points move right.

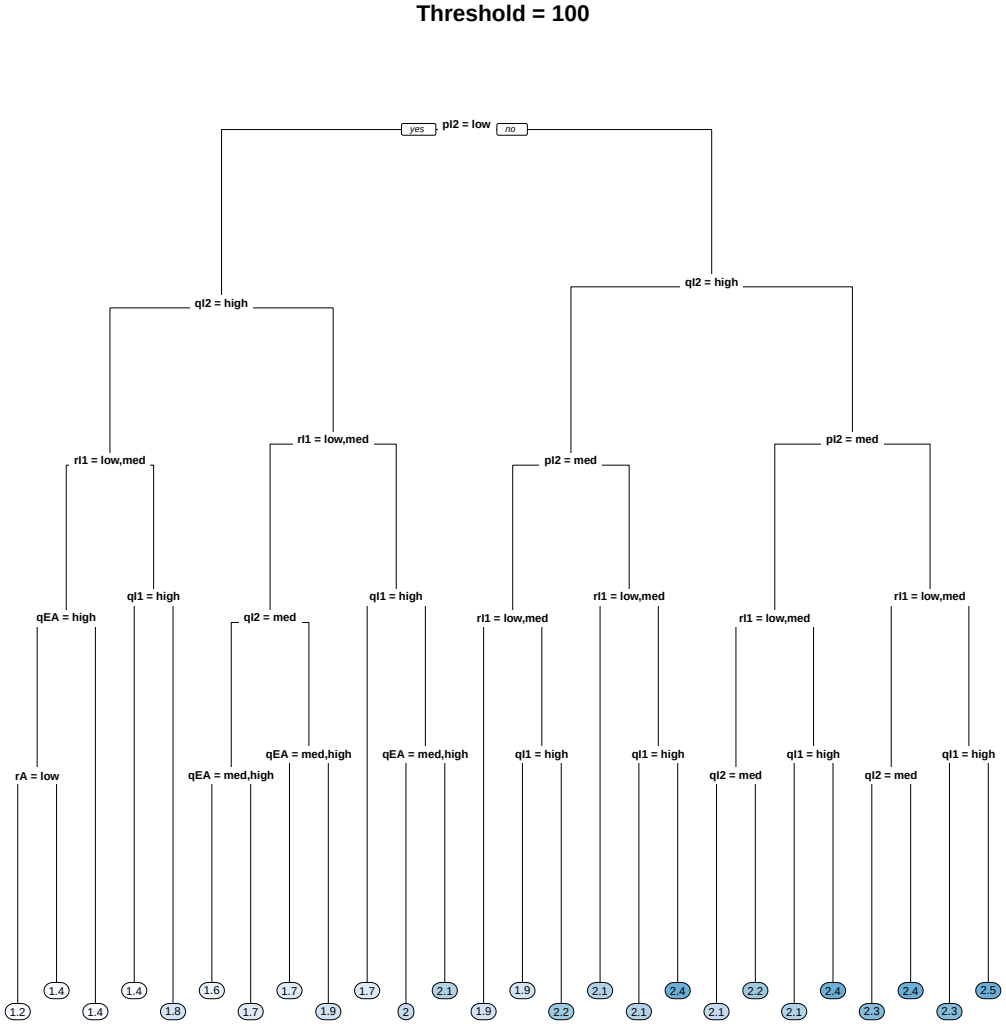

Figure 3: Pruned tree with 25 splits for predicting logit-CII with a class size threshold of 20. At each split, points which satisfy the listed condition move left, while the other points move right.

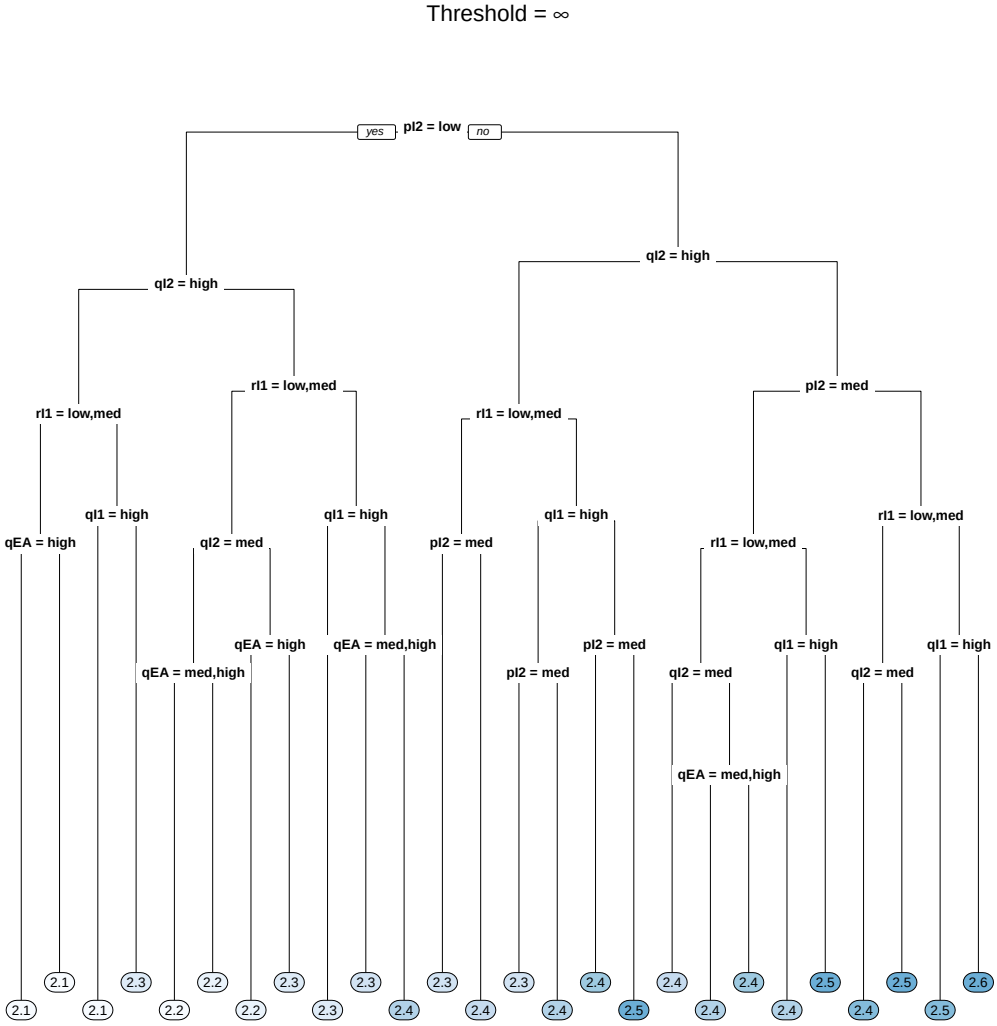

Figure 4: Pruned tree with 25 splits for predicting logit-CII with a class size threshold of 20. At each split, points which satisfy the listed condition move left, while the other points move right.

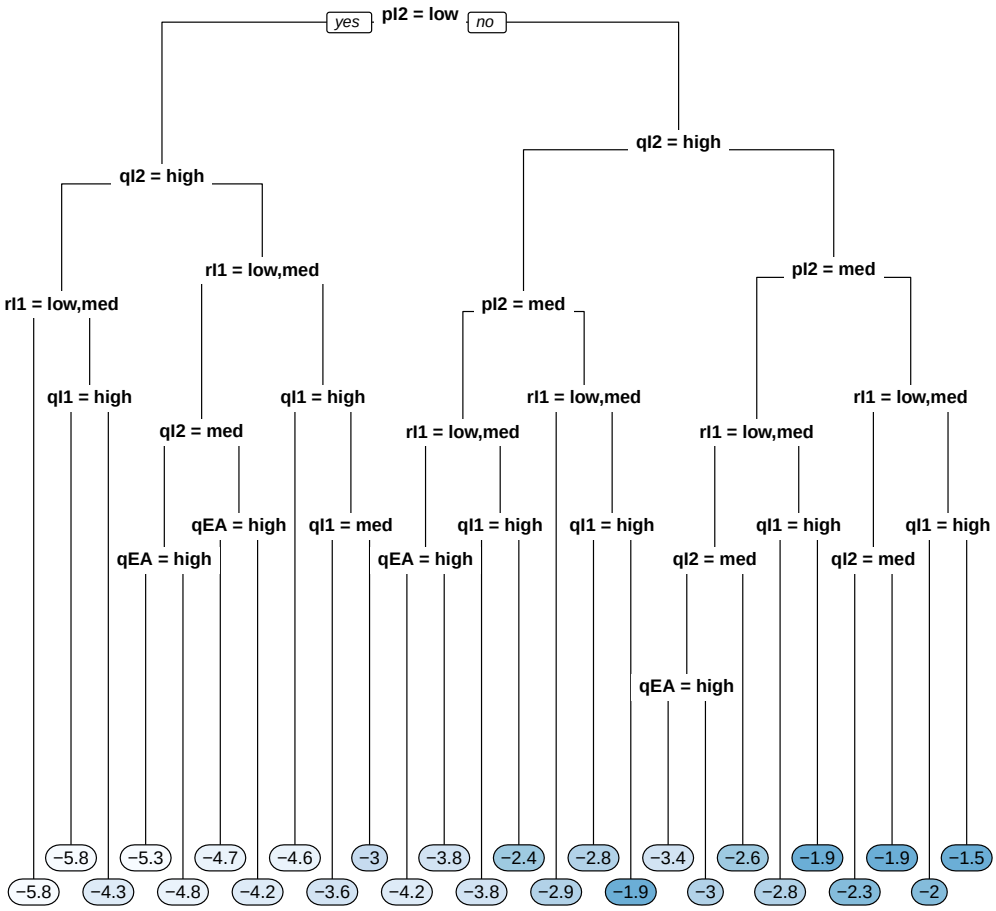

Figure 5: Pruned tree with 25 splits for predicting logit-peak outbreak size with a class size threshold of 20. At each split, points which satisfy the listed condition move left, while the other points move right.

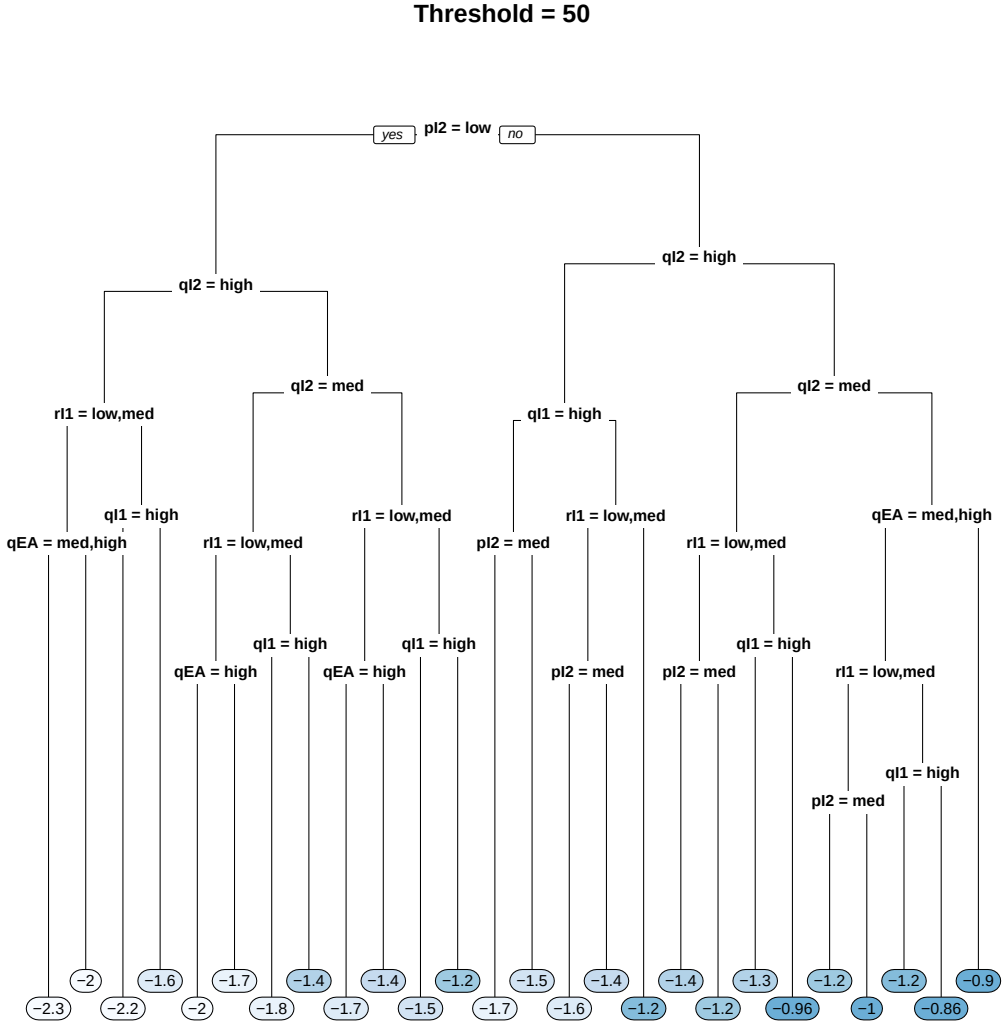

Figure 6: Pruned tree with 25 splits for predicting logit-peak outbreak size with a class size threshold of 20. At each split, points which satisfy the listed condition move left, while the other points move right.

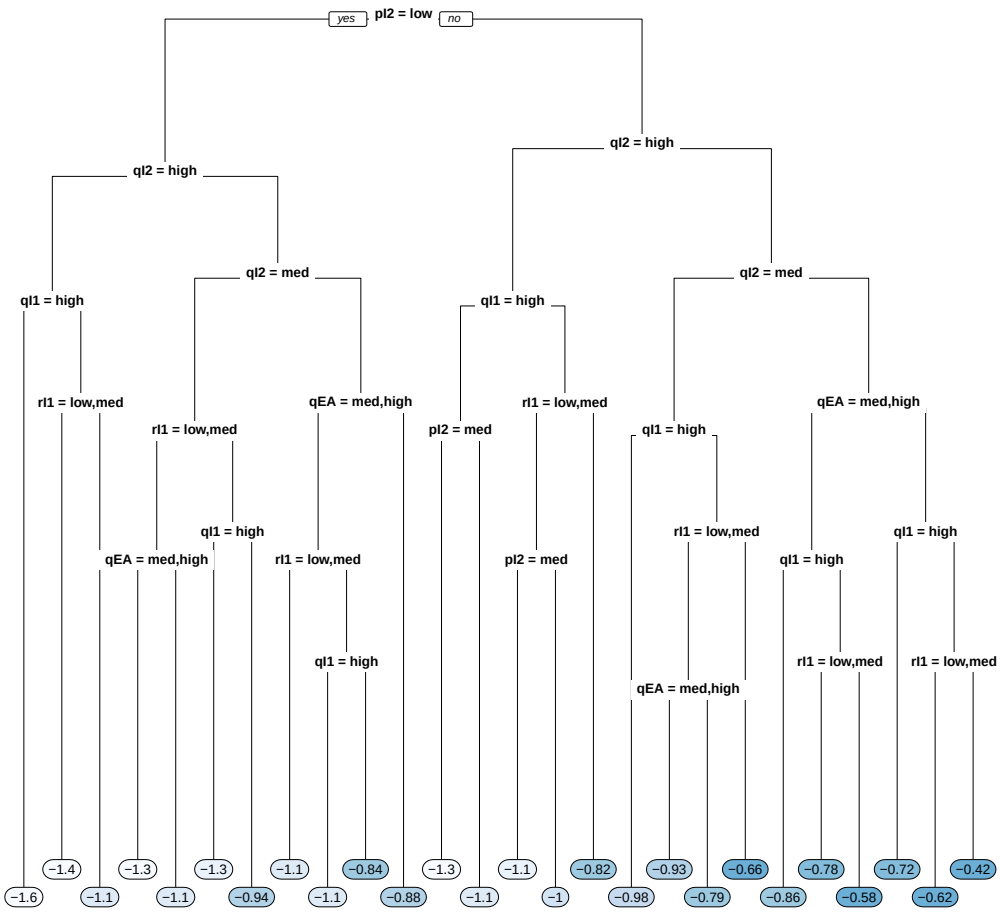

Figure 7: Pruned tree with 25 splits for predicting logit-peak outbreak size with a class size threshold of 20. At each split, points which satisfy the listed condition move left, while the other points move right.

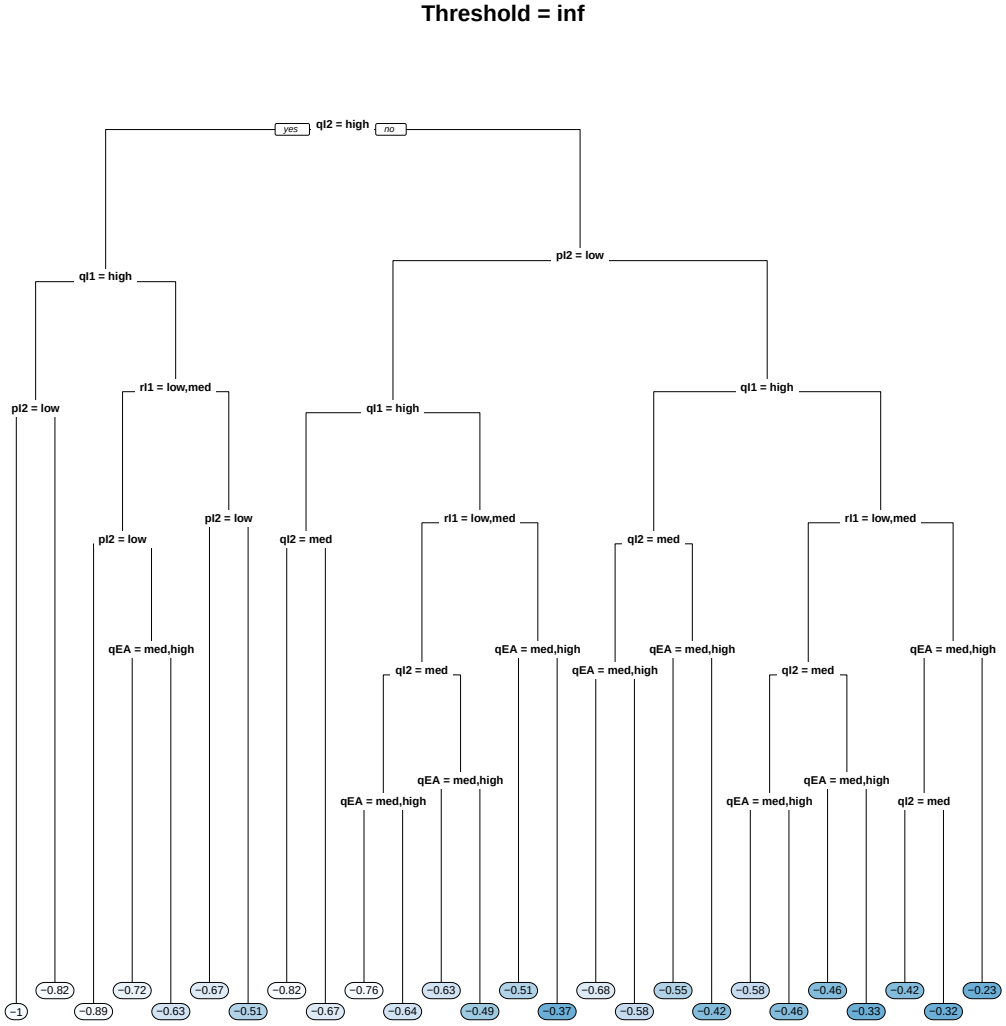

Figure 8: Pruned tree with 25 splits for predicting logit-peak outbreak size with a class size threshold of 20. At each split, points which satisfy the listed condition move left, while the other points move right.

Parameter	R Name
θ_{I2}	pI2
ρ_A	rA
ρ_{I1}	rI1
q_E	qE
q_A	qA
q_{I1}	qI1
q_{I2}	qI2
q_{EA}	qEA

Table 6: Chosen values for each parameter, along with relevant references. See text for reasoning.

2021).

5.1 Analysis

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

The analyses of our two response variables is similar, so we decribe the common methodology here. See the main paper for an investigation into the distribution of the responses across levels of each parameter. 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.

We have 10 replicates for each parameter combination, so we begin by investigating pure replication variability (i.e. the variability of replicates within one level of the covariates). This gives us some information about how to handle dispersion when fitting mean models. Specifically, within each parameter setting, we compute the mean and variance of the response variable (over the 10 observations available within that setting). We then plot this observed variance vs observed mean. It turns out that the relationship between these

two quantities varies greatly across levels of ϕ, the class size threshold, so we produce separate plots for each threshold level.

Next, we fit a logitstic regression model to explain the mean response using our parameters as covariates (see, e.g., McCullagh and Nelder, 1989). Preliminary exploratory analysis suggests that the class size threshold is much more strongly associated with CII than any epidemiological parameters. As such, we divide our data into groups based on the class size threshold, and fit a separate model within each group.

For each threshold level, we fit a sequence of models with interactions of order one through four (first-order interactions here meaning main-effects). After investigating the improvement in deviance obtained by each increase in model complexity, as well as the deviance-per-fitted parameter of each model (not shown), we conclude that it is most appropriate to retain only the main effects for the epidemiological parameters in each class size threshold.

For both summaries, the investigation of pure replication described above shows the presence of overdispersion with respect to the usual binomial variance function[2]. The presence of ovedispersion suggests that we should use a 'quasi-likelihood' model, which includes an overdispersion parameter that is estimated as part of the model fitting process. In fact, the overdispersion varies substantially between class size thresholds. As such, we include a separate overdispersion parameter in each of the class size thresholds' models.

After fitting the models described above, we proceed to examine some diagnostic plots to assess the quality of our model fits. Due to the size of our datasets (26,244 divided evenly

[2]Note that we use a dispersion model obtained from pure replication to inform how we expect the variance to depend on the mean in our model. The former variability is around an unbiased estimate of the mean, while the latter mean estimate is only unbiased if our assumed model is correct. In general, accurate estimation of variance effects in (generalized) linear models is a challenging problem, see, e.g., McCullagh and Nelder (1989); Carroll and Ruppert (1988). We accept this limitation as part of the cost in using an interpretable model, but a valuable extension to our work would be to use a more sophisticated model to predict the mean and variance of our trajectory summaries.

into four threshold levels) and the limited number of predictors being used, all model tests are strongly significant, even after correcting for multiple comparisons (i.e. testing any regression parameter against a null hypothesis of 0 in `R` gives a p-value of $< 2 \cdot 10^{-16}$). Since all differences are statistically significant, we focus on qualitative features of the models. That is, we investigate the deviance contribution of each covariate (both absolute and relative), and measure the covariates' relative importances. This variable importance is best thought of as a measure of how one ought to prioritize available resources for measuring disease parameters if one wants to predict the particular response being modelled.

5.2 Results

We present results separately for cumulative incidence of infection (CII) and peak outbreak size.

5.2.1 Cumulative Incidence of Infection

Figure 9 plots the pure replication variance within each parameter combination against the corresponding mean CII (across the 10 simulation replicates within each combination). Each class size threshold is plotted separately[3]. The red curves correspond to the variance predicted by a binomial model for the total number of cases (i.e. $p(1-p)/n$, where p is the mean CII). We see that the theoretical variance is a poor fit for the obersved variance, although not uniformly so across class size threshold. This lack of fit recommends the use of quasi-likelihood models.

Table 7 gives the fitted overdispersion parameter for each class size threshold.

Next, we present diagnostic plots for the fitted models in each class size threshold. Figure 10 shows the deviance residuals against fitted values (on the observed scale). Pearson

[3] An outlier has been removed from the threshold = 100 group, in which one of the 10 trajectories stalls at a small number of cases. This produces a single point with extremely large observed variance.

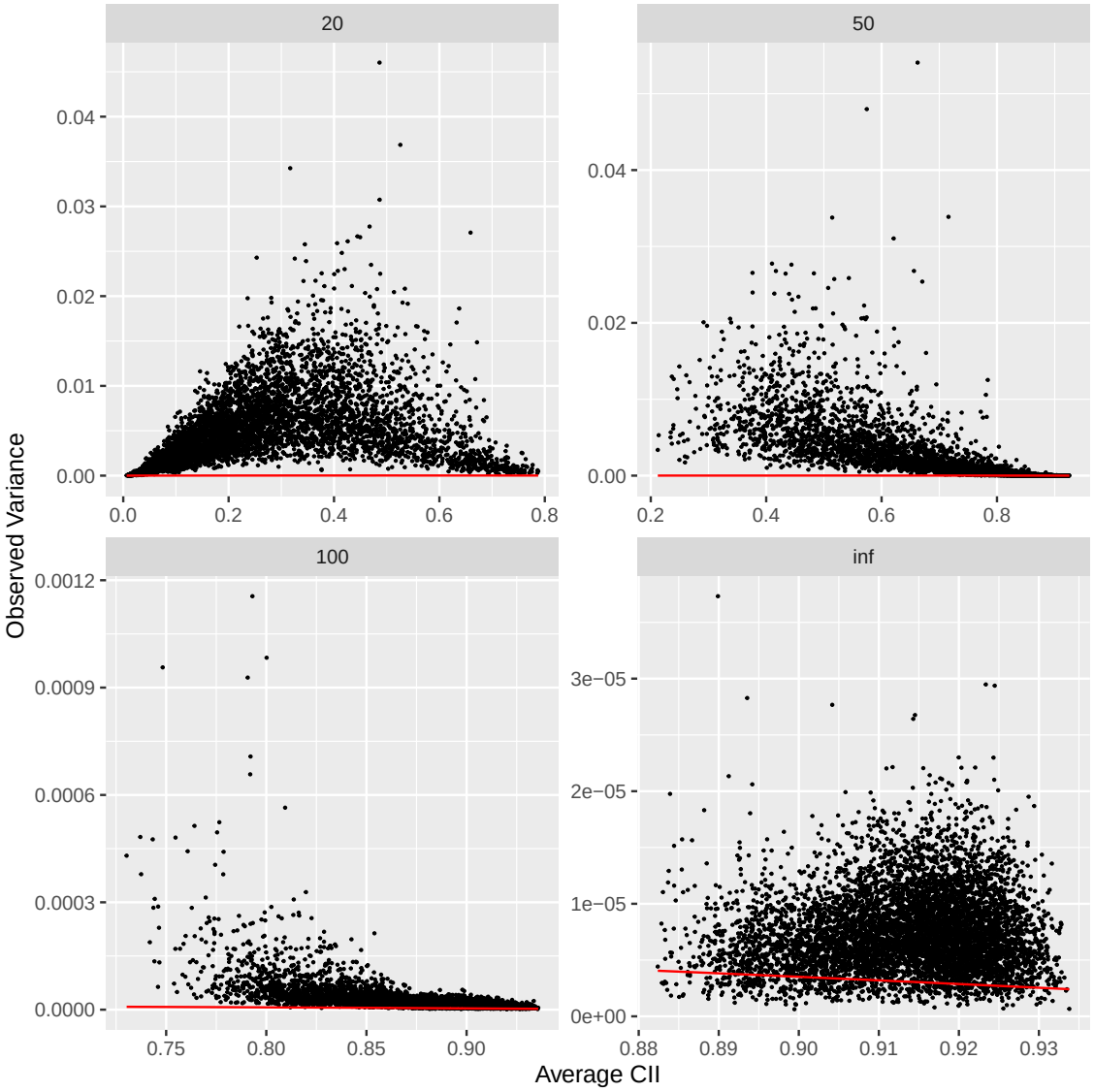

Figure 9: Average CII vs pure replication variance for each class size threshold. Red curve is the theoretical variance from a binomial model.

Threshold:	20	50	100	∞
Overdispersion:	633	374	20	4

Table 7: Overdispersion parameters relative to binomial variance in each class size threshold for CII.

residuals are similar. Figure 11 shows the observed vs fitted values on the observed scale, along with the reference line $Y = X$. Note the presence of some heteroscedasticity and outliers. This suggests that a more sophisticated model may be able to better describe the relationship between CII and our model parameters.

Table 8 gives the deviance increase caused by omitting each parameter from a model containing all others, with a separate column for each class size threshold. We also include the residual deviance of each full model.

	CII			
Threshold:	20	50	100	∞
ρ_A	2.75E+06	2.58E+06	2.67E+05	3.00E+04
ρ_{I1}	3.73E+07	1.99E+07	1.71E+06	2.07E+05
θ_{I2}	1.43E+08	1.27E+08	1.24E+07	1.29E+06
q_E	8.40E+05	7.96E+05	2.76E+04	2.50E+03
q_A	4.19E+05	3.32E+05	4.93E+04	6.34E+03
q_{I1}	8.07E+06	2.08E+06	3.99E+05	6.38E+04
q_{I2}	1.83E+07	1.42E+07	2.65E+06	3.11E+05
q_{EA}	6.15E+06	4.30E+06	6.01E+05	7.27E+04
Residual	4.31E+07	2.51E+07	1.30E+06	2.60E+05

Table 8: Deviance increase caused by omitting each variable from a model containing all others, as well as residual deviance of the full model. Models are fit separately for each class size threshold.

Table 9 gives the relative deviance increases, where the values in each column are divided by the largest increase in that column. The full models' residual deviances are also included in this comparison. Table 10 ranks the deviance increases from largest (rank 1) to smallest (rank 8). The residual deviance has been excluded here. The order of variables given in Table 10 can be thought of as a way to prioritize the use of resources to estimate disease parameters more effectively.

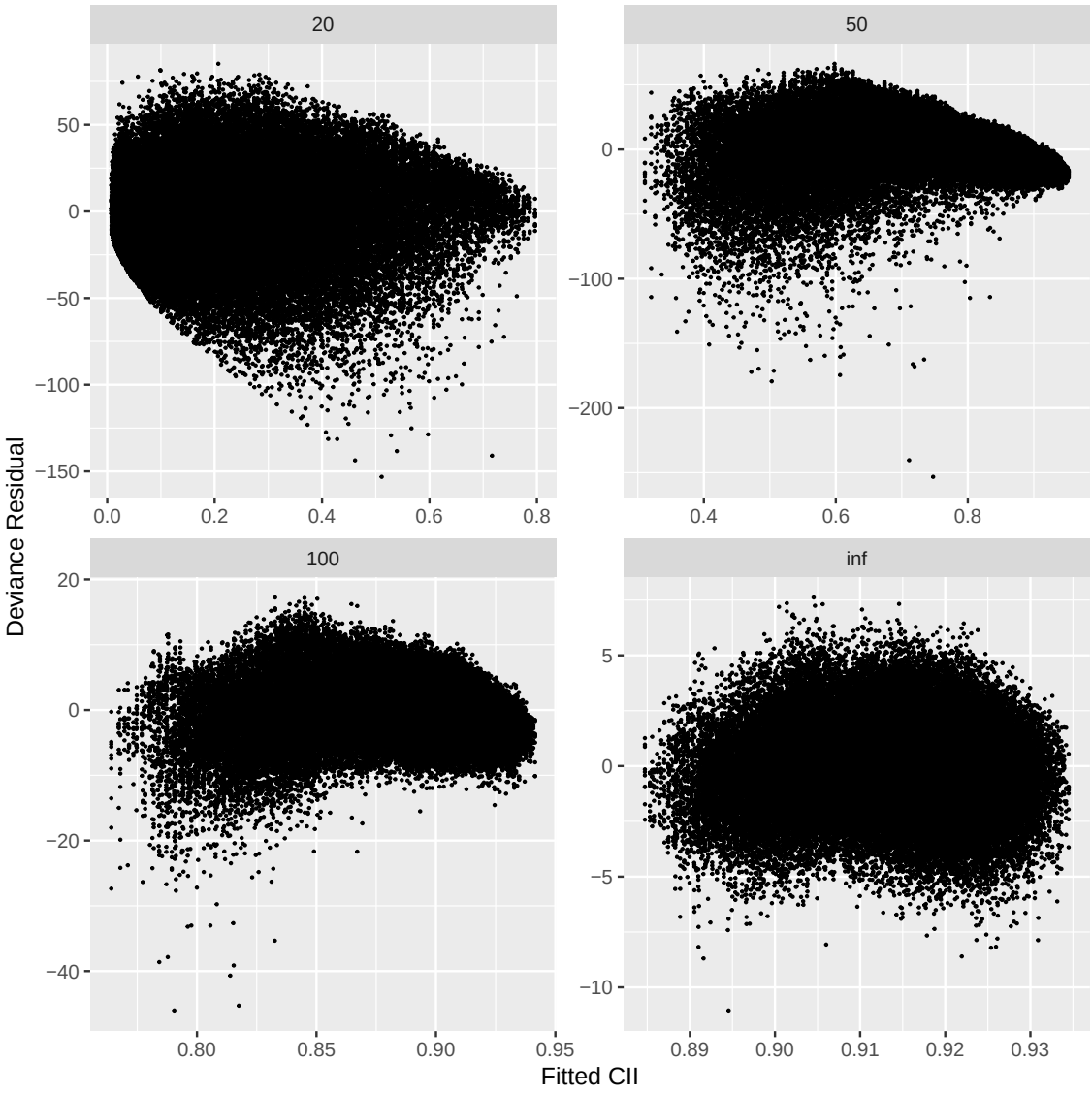

Figure 10: Deviance residuals vs fitted values for CII at each class size threshold.

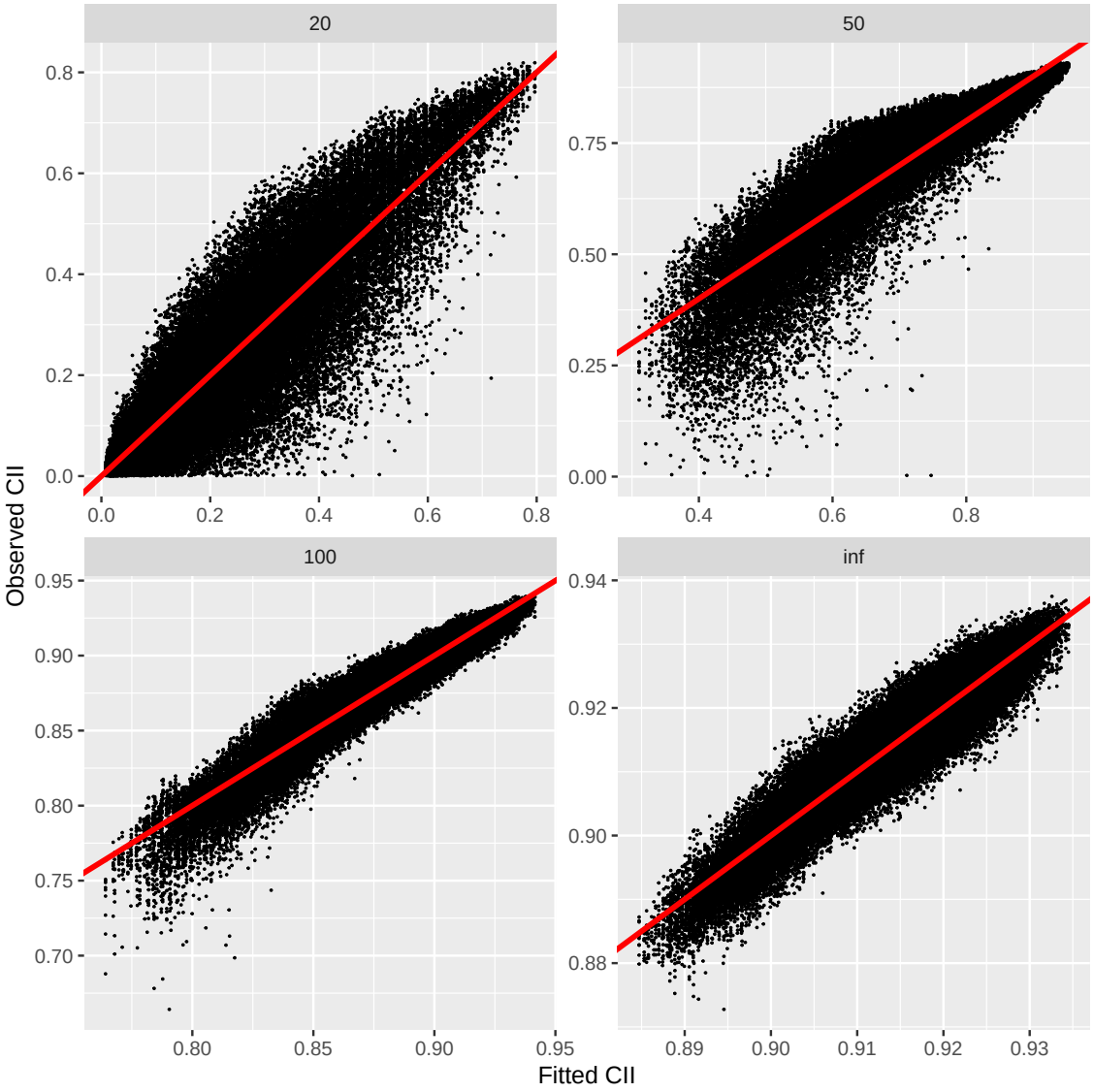

Figure 11: Observed vs fitted CII values at each class size threshold. Red lines are $Y = X$ for reference.

	CII			
Threshold:	20	50	100	∞
ρ_A	0.019	0.02	0.022	0.023
ρ_{I1}	0.26	0.16	0.14	0.16
θ_{I2}	1	1	1	1
q_E	0.0059	0.0063	0.0022	0.0019
q_A	0.0029	0.0026	0.004	0.0049
q_{I1}	0.056	0.016	0.032	0.05
q_{I2}	0.13	0.11	0.21	0.24
q_{EA}	0.043	0.034	0.049	0.056
Residual	0.3	0.2	0.11	0.2

Table 9: Relative deviance increases within each class size threshold, as well as relative size of the full model's residual deviance.

	CII			
Threshold:	20	50	100	∞
ρ_A	6	5	6	6
ρ_{I1}	2	2	3	3
θ_{I2}	1	1	1	1
q_E	7	7	8	8
q_A	8	8	7	7
q_{I1}	4	6	5	5
q_{I2}	3	3	2	2
q_{EA}	5	4	4	4

Table 10: Deviance increase ranks within each class size threshold. Rank 1 has the greatest incease and 8 has the smallest.

5.2.2 Peak Outbreak Size

Figure 12 plots the pure replication variance within each parameter combination against the corresponding mean peak outbreak size (across the 10 simulation replicates within each combination). Each class size threshold is plotted separately[4]. The red curves correspond to the variance predicted by a binomial model for the total number of cases (i.e. $p(1-p)/n$, where p is the mean peak size). We see that the theoretical variance is a poor fit for the obersved variance, although not uniformly so across class size threshold. This lack of fit

Threshold:	20	50	100	∞
Overdispersion:	133	39	20	13

Table 11: Overdispersion parameters relative to binomial variance in each class size threshold for peak outbreak size.

recommends the use of quasi-likelihood models.

Table 11 gives the fitted overdispersion parameter for each class size threshold.

Next, we present diagnostic plots for the fitted models in each class size threshold. Figure 13 shows the deviance residuals against fitted values (on the observed scale). Pearson residuals are similar. Figure 14 shows the observed vs fitted values on the observed scale, along with the reference line $Y = X$. Note the presence of some heteroscedasticity and outliers. This suggests that a more sophisticated model may be able to better describe the relationship between peak outbreak size and our model parameters.

Table 12 gives the deviance increase caused by omitting each parameter from a model containing all others, with a separate column for each class size threshold. We also include the residual deviance of each full model.

Table 13 gives the relative deviance increases, where the values in each column are divided by the largest increase in that column. The full models' residual deviances are also included in this comparison. Table 14 ranks the deviance increases from largest (rank 1) to smallest (rank 8). The residual deviance has been excluded here. The order of variables given in Table 14 can be thought of as a way to prioritize the use of resources to estimate disease parameters more effectively.

[4] As with the analysis of CII, an outlier has been removed from the threshold = 100 group (the same point in both cases). At this setting, one of the 10 trajectories stalls at a small number of cases, which produces a single point with extremely large observed variance.

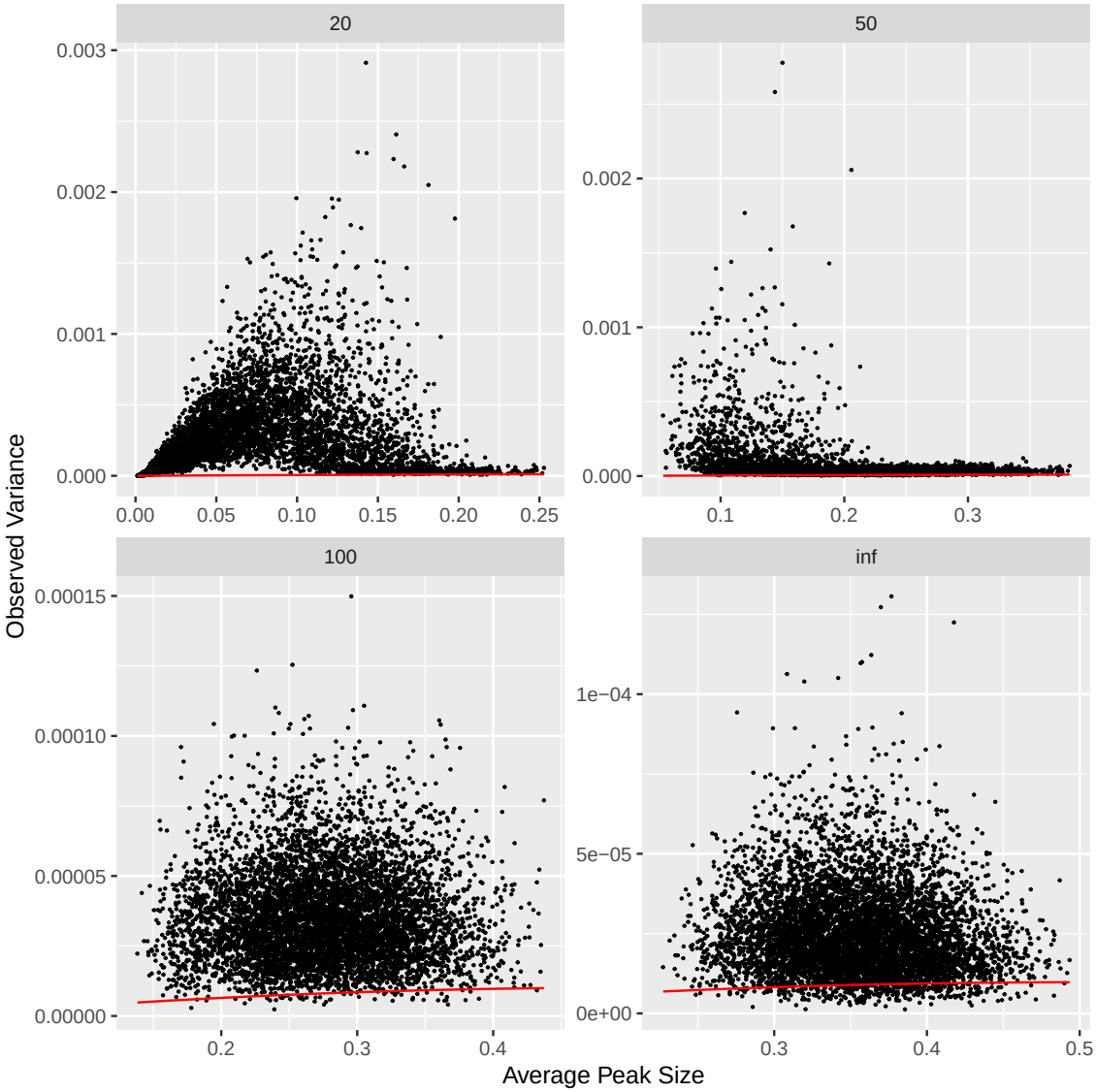

Figure 12: Average peak size vs pure replication variance for each class size threshold. Red curve is the theoretical variance from a binomial model.

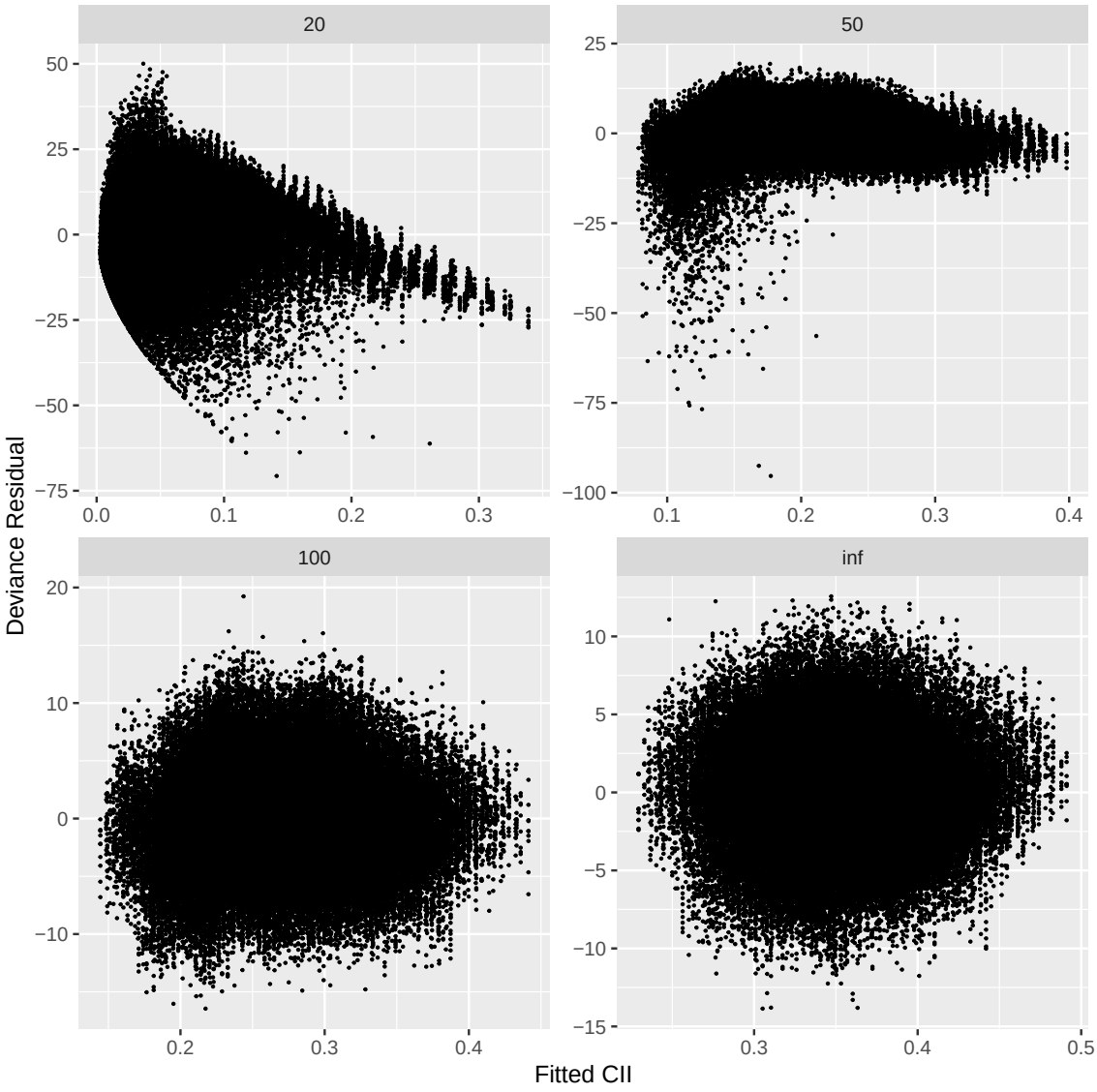

Figure 13: Deviance residuals vs fitted values for CII at each class size threshold.

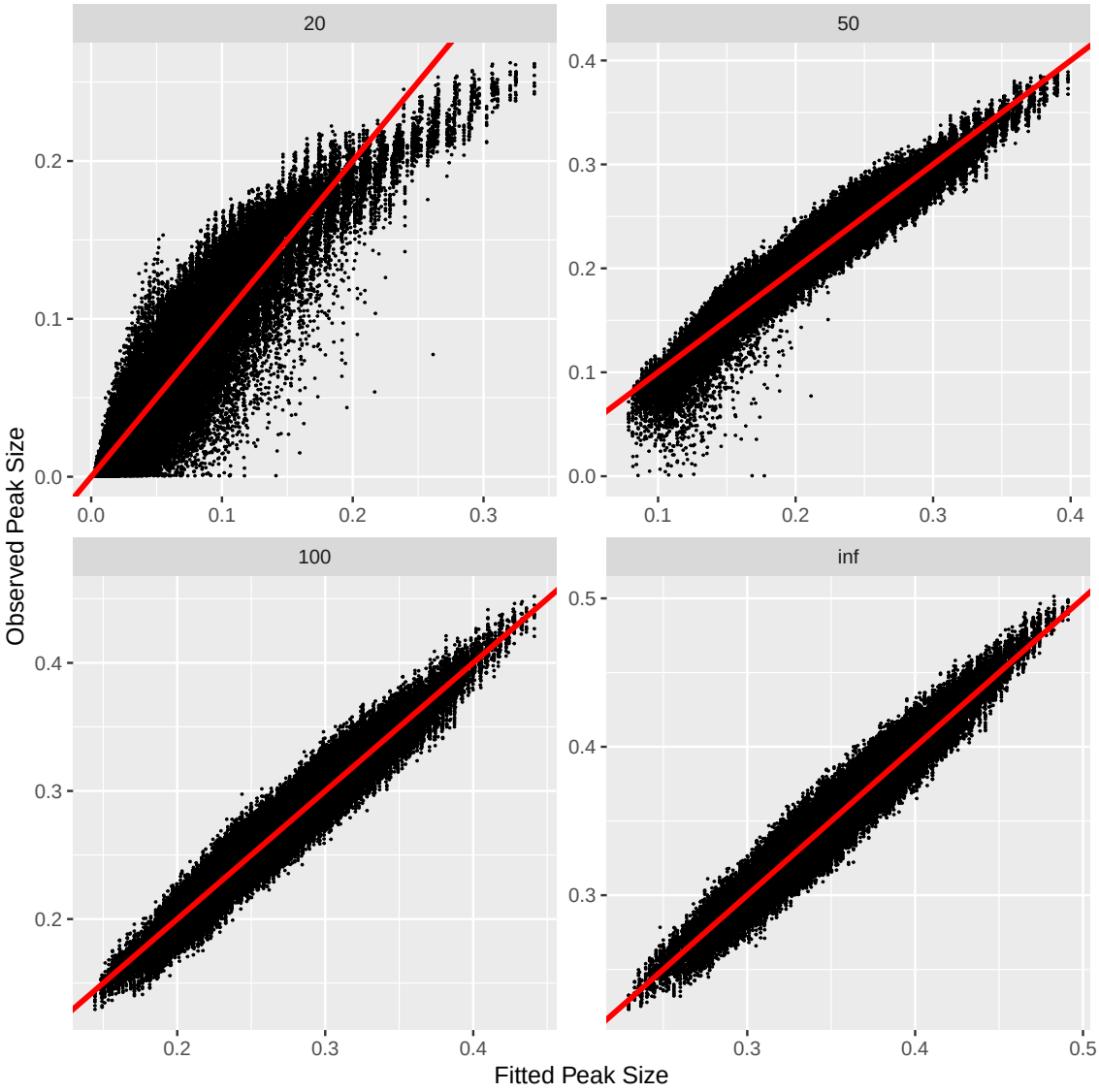

Figure 14: Observed vs fitted CII values at each class size threshold. Red lines are $Y = X$ for reference.

	Peak Size			
Threshold:	20	50	100	∞
ρ_A	5.17E+05	2.97E+05	1.60E+05	6.95E+04
ρ_{I1}	6.63E+06	3.26E+06	2.33E+06	1.47E+06
θ_{I2}	3.30E+07	1.36E+07	7.22E+06	2.75E+06
q_E	2.06E+05	2.11E+05	2.75E+05	4.38E+05
q_A	1.79E+05	2.80E+05	2.91E+05	3.12E+05
q_{I1}	2.45E+06	2.76E+06	2.70E+06	2.56E+06
q_{I2}	7.63E+06	9.46E+06	7.29E+06	5.29E+06
q_{EA}	2.08E+06	2.22E+06	1.83E+06	1.49E+06
Residual	9.22E+06	2.73E+06	1.33E+06	8.63E+05

Table 12: Deviance increase caused by omitting each variable from a model containing all others, as well as residual deviance of the full model. Models are fit separately for each class size threshold.

References

Diana Buitrago-Garcia, Dianne Egli-Gany, Michel J. Counotte, Stefanie Hossmann, Hira Imeri, Aziz Mert Ipekci, Georgia Salanti, and Nicola Low. Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: a living systematic review and meta-analysis. *PLoS Medicine*, 17(9), 2020.

Oyungerel Byambasuren, Katy Bell, Louise McLaws, and Paul Glasziou. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. *Official Journal of the Association of Medical Microbiology and Infectious Disease Canada*, 5(4), 2020.

Andrew William Byrne, David McEvoy, Aine B. Collins, Kevin Hunt, Miriam Casey, Ann Barber, Francis Butler, John Griffin, Elizabeth A. Lane, Conor McAloon, Kirsty O'Brien, Patrick Wall, Kieran A. Walsh, and Simon J. More. 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, 2020.

	Peak Size			
Threshold:	20	50	100	∞
ρ_A	0.016	0.022	0.022	0.013
ρ_{I1}	0.2	0.24	0.32	0.28
θ_{I2}	1	1	0.99	0.52
q_E	0.0062	0.016	0.038	0.083
q_A	0.0054	0.021	0.04	0.059
q_{I1}	0.074	0.2	0.37	0.48
q_{I2}	0.23	0.7	1	1
q_{EA}	0.063	0.16	0.25	0.28
Residual	0.28	0.2	0.18	0.16

Table 13: Relative deviance increases within each class size threshold, as well as relative size of the full model's residual deviance.

	Peak Size			
Threshold:	20	50	100	∞
ρ_A	6	6	8	8
ρ_{I1}	3	3	4	5
θ_{I2}	1	1	2	2
q_E	7	8	7	6
q_A	8	7	6	7
q_{I1}	4	4	3	3
q_{I2}	2	2	1	1
q_{EA}	5	5	5	4

Table 14: Deviance increase ranks within each class size threshold. Rank 1 has the greatest incease and 8 has the smallest.

R.J. Carroll and D. Ruppert. *Transformation and weighting in regression.* CRC Press, 1988.

Xi He, Eric H. Y. Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y. Wong, Yujuan Guan, Xinghua Tan, Xiaoneng Mo, Yanqing Chen, Baolin Liao, Weilie Chen, Fengyu Hu, Qing Zhang, Mingqiu Zhong, Yanrong Wu, Lingzhai Zhao, Fuchun Zhang, Benjamin J. Cowling, Fang Li, and Gabriel M. Leung. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26,

2020.

Michael A. Johansson, Talia M. Quandelacy, Sarah Kada, Pragati Venkata Prasad, Molly Steele, John T. Brooks, Rachel B. Slayton, Matthew Biggerstaff, and Jay C. Butler. SARS-Cov-2 transmission from people without COVID-19 symptoms. *JAMA Network Open*, 4(1), 2021.

Shujuan Ma, Jiayue Zhang, Minyan Zeng, Qingping Yun, Wei Guo, Yixiang Zheng, Shi Zhao, Maggie H. Wang, and Zuyao Yang. Epidemiological parameters of coronavirus disease 2019: a pooled analysis of publicly reported individual data of 1155 cases from seven countries. *medRxiv*, 2020.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, second edition, 1989.

William Ruth and Richard Lockhart. Network analysis of sfu course registrations, 2021. URL `https://doi.org/10.48550/arXiv.2104.12769`.

Hayley A Thompson, Andria Mousa, Amy Dighe, Han Fu, Alberto Arnedo-Pena, Peter Barrett, Juan Bellido-Blasco, Qifang Bi, Antonio Caputi, Liling Chaw, Luigi De Maria, Matthias Hoffmann, Kiran Mahapure, Kangqi Ng, Jagadesan Raghuram, Gurpreet Singh, Biju Soman, Vicente Soriano, Francesca Valent, Luigi Vimercati, Liang En Wee, Justin Wong, Azra C Ghani, and Neil M Ferguson. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) setting-specific transmission rates: a systematic review and meta-analysis. *Clinical Infectious Diseases*, 73(3), 2021.

Kim A. Weeden and Benjamin Cornwell. The small-world network of college classes: Implications for epidemic spread on a university campus. *Sociological Science*, 7:222–241, 2020.

Hualei Xin, Yu Li, Peng Wu, Zhili Li, Eric H. Y. Lau, Ying Qun, Liping Wang, Benjamin J. Cowling, Tim Tsang, and Chongjie Li. Estimating the latent period of coronavirus disease 2019 (COVID-19). *Clinical Infectious Diseases*, 746, 2021.