COVID-19 clusters in schools: frequency, size, and transmission rates from crowdsourced exposure reports

Paul Tupper[1*], Shraddha Pai[2,3], COVID Schools Canada[4], Caroline Colijn[1]
[1]Department of Mathematics, Simon Fraser University, Burnaby, Canada
[2]The Donnelly Centre, University of Toronto, Toronto, Canada
[3]Ontario Institute for Cancer Research, Toronto, Canada
[4]https://covidschoolscanada.org/

[*]To whom correspondence should be addressed; E-mail: pft3@sfu.ca.

The role of schools in the spread of the COVID-19 pandemic is controversial, with some claiming they are an important driver of the pandemic and others arguing that transmission in schools is negligible. School cluster reports that have been collected in various jurisdictions are a source of data about transmission in schools. These reports consist of the name of a school, a date, and the number of students known to be infected. We provide a simple model for the frequency and size of clusters in this data, based on random arrivals of index cases at schools who then infect their classmates with a highly variable rate, fitting the overdispersion evident in the data. We fit our model to reports for several jurisdictions in the US and Canada, providing estimates of mean and dispersion for cluster size, whilst factoring in imperfect ascertainment. Our parameter estimates are robust to variations in ascertainment fraction. We use these estimates in three ways: i) to explore how uneven the distribution of cases is among different clusters in different jurisdictions (that is, what fraction of cases are in the 20% largest clusters), ii) to estimate how long it will be until we see a cluster a given size in jurisdiction, and iii) to determine the distribution of instantaneous transmission rate β among different index case. We show how these latter distribution can be used in simulations of school transmission where we explore the effect of different interventions, in the context of highly variable transmission rates.

1 Background

In the management of the COVID-19 pandemic an important consideration is the role of children and in particular schools. In most jurisdictions rates of COVID infection among children are similar to those in the adult population (*8*). But severity is much lower in children; the infection fatality rate (IFR) of COVID for at age 10 is estimated to be 0.002% versus an IFR of 0.01% at age 25, and 0.4% at age 55 (*24*). Cases are more often asymptomatic among children, less likely to require hospitalization and ICU care (*8*), and less likely to be classified as long COVID (*34*). On the other hand, not all the news is good for children, MIS-C is a serious condition sometimes resulting from COVID infection (*6*), and myocarditis happens more frequently as a side effect of infection among younger individuals (*31*).

Jurisdictions have had to make a choice between closing schools, with all the attendant social, economic, and psychological costs (*9*), and leaving schools open, allowing possible transmission of COVID in that setting (*8*). The direct downside of transmission in schools if it occurs is that children may be infected there, risking the low but non-negligible harms of COVID-19 in that age range, but also adult teachers and staff are put at risk. Transmission in schools may also contribute to overall community transmission, indirectly jeopardizing more vulnerable individuals (*41*). As a concrete example, if a child contracts COVID at school, they may then go on to transmit it to an elderly relative they live with, for whom the consequences are more severe (*21*). Estimating the magnitude of these two kinds of harm and making the decision as to what choice to make involves many sources of uncertainty and value judgements, which helps explain why different jurisdictions have taken different approaches (*18*). In some jurisdictions schools were open for the 2020 to 2021 school year, though many measures were put into place in order to reduce the risk of COVID transmission (*28*). Measures include cohorting, staggered entrance and exit times, masks, improvements in ventilation, extra sanitization measures. In other jurisdictions schools were closed for large portions of the year (*30*).

Studies that have looked at the effect of school closures on the overall rate of COVID transmission find mixed results: some find substantial reduction in community transmission when schools are closed, and others small or no effect (*11, 41*). Given that schools involve many children all sharing a room for many hours a day, it may be surprising that there is not a clearer evidence of significant transmission in schools. One explanation is that children may be less likely to transmit COVID to each other, either by being less infectious or by being less susceptible (*14, 38*). But transmission in schools does occur, and it's worthwhile to estimate the magnitude and characterize the variation in it.

One source of evidence for transmission in schools are school exposure reports. Throughout the pandemic organizations have collected data submitted by volunteers about COVID cases in schools, and such data has subsequently been published online (*4, 5*). Data consists of reports of exposures or clusters in schools, either submitted by parents, or determined from reading newspaper reports. Several such websites exist, though many ceased due to excessive workload after the 2020–21 school year. In some

jurisdictions there are also similar sources of data provided by local government (*17, 29*) or Public Health Agencies (*19, 20*).

Here we propose a simple model of transmission in schools, and we use these data on cluster sizes to estimate parameters of the model for various jurisdictions in the US and Canada. Our model allows for heterogeneity in transmission rate, which is able to capture the considerable variability in the sizes of the clusters, with most exposures leading to no further cases (and so a cluster of size 1) but with few having a large number of cases (*36*). We estimate the the mean and overdispersion parameters for different jurisdictions. We then use our parameter estimates in a few different ways: firstly, we explore the overdispersion of cluster sizes in different jurisdictions, giving estimates of what fraction of all cases are in the 20% largest of all clusters. Secondly, we are able to estimate how long we can expect to wait until we see clusters of a given size in a jurisdiction. Thirdly, we can get an estimate of the distribution of transmission rate β, the rate at which a single infected individual infects a susceptible person when they are in contact. This parameter, in turn, could be used to simulate school transmission and explore the impacts of interventions (*37*) as we explore for several parameter choices.

Finally, two important changes have occurred in 2021 that we expect to impact cluster sizes in schools. On the one hand, in many jurisdictions, large portions of children aged 12 and up have been vaccinated with the Pfizer/BIONTech vaccine (*35*). According to the extent to which the vaccine protects against infection, we expect cluster size will be reduced, as fewer students will be infected if they have been immunized. Observed cluster size may be reduced further even than this, if the vaccine allows harder-to-detect infections to occur. On the other hand, now more infectious variants of the coronavirus have emerged, most noticeably the Alpha variant and subsequently the Delta variant (*7*). Increased transmissibilty would suggest larger cluster sizes, certainly among unvaccinated ages, but the relative impact of vaccination and the new variants together is difficult to gauge. Furthermore, both changes may lead to a change in the variability in cluster sizes.

2 Methods

Our data consist of lists of clusters of confirmed cases among students in schools for 8 US states and 4 Canadian provinces during the 2020–2021 school year. The US data was gathered from the National Educational Association website (*4*) (originally started by Alisha Morris, an educator at a Kansas high school) which collected data from news media and from reports submitted by volunteers (*39*). We selected the 8 states with the most data available, and covering the period between August and November 2020. Canadian data was contributed by Dr Shraddha Pai from COVID Schools Canada (*5*), an initiative of the group Masks for Canada (*1*). We included the 4 provinces from this data set with the most clusters recorded (omitting provinces where there was only the total cases by school recorded); entries ranged over the whole 2020–21 school year For both sources of data, each entry is a single cluster with number

of confirmed infections among students in the school, the school's name, the district, and the date of the report. For the US data we used confirmed student cases for the cluster size, excluding teachers and staff; for the Canadian data we used total cases. Information is usually not available about whether the students were in just one classroom, and some of the clusters exceed 30 students, so are likely not restricted to just one class. Accordingly, we interpret clusters as capturing all linked cases at a given school, and not just one classroom. Figure 1 shows histograms of the observed cluster size in the 12 jurisdictions.

In Table 1 we show, for each jurisdiction, the number of clusters, the number of schools appearing, the number of schools with more than one reported cluster, the fraction of clusters that have only one observed case, and the average number of observed cases in the clusters. There are several striking differences between the US and Canadian data. There are substantially more clusters reported in Canada than in the US, despite the US states having greater population on average. This may partly be explained by the Canadian data being collected over a longer period than the US data, but this is likely not the full explanation: in Figure 2 (top) we show the rate (in clusters per day) that cases appear in the dataset over time. In Figure 2 (bottom) we show rate of COVID incidence in the jurisdiction over the same period of time. Even at times when both US and Canadian datasets record clusters, Canadian rates are higher than US rates. This suggests that the method used for gathering cluster reports in different jurisdictions varied substantially between jurisdictions, especially when we look at daily incident cases in each. Statewide incident cases were generally higher per capita in US states than Canadian provinces. In the US data, cluster reports do not track well with incidence, but in the Canadian data there is a clear correspondence. Furthermore, the majority of schools in the US datasets only report one cluster, whereas the opposite is true of the Canadian data.

There is also substantial differences in statistics of cluster sizes. Mean observed cluster sizes were without exception larger in the US states than Canadian provinces, and Canadian provinces tended to have a higher fraction of clusters with only one case. Given the incomplete nature of the US data, we cannot determine whether these differences are due to real differences in transmission in the jurisdictions, or because smaller clusters were less likely to be reported in the US states.

There is substantial uncertainty in whether each reported cluster of cases accurately represents a set of cases linked by transmission. For any cluster of 2 or more cases, it may be that two independent sets of cases are incorrectly included in the same cluster. This may lead us to overestimate the size of clusters. Likewise, any two cluster in the the same school in the same week or adjacent weeks may actually be a single cluster misidentified as two. Both these factors may occur in our data, but we neglect both of them, taking the observed cluster size as given in the dataset. We are also unable to distinguish between transmission occurring in a school and in social activities with classmates outside of school.

In a given jurisdiction, we assume exposure events occur according to a Poisson process with variable rate. Figure 2 (top) gives an estimate of the rate of exposure as a function of time for the dataset of each jurisdiction. Independently of this process, once an exposure event occurs at a school, we say Z

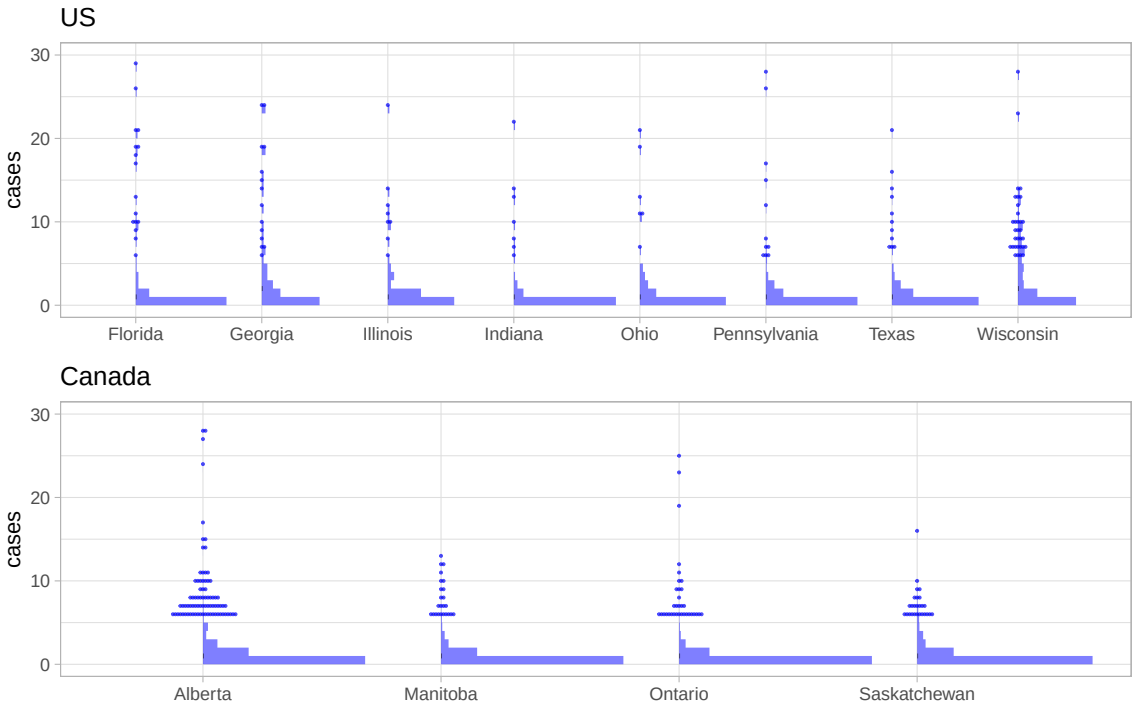

Figure 1: Histograms of observed cluster sizes in 8 US states (top) and 4 Canadian provinces (bottom). We only show clusters of size 30 or fewer. Dots indicate clusters of size 6 or larger, and indicate the presence of (more rare) larger clusters.

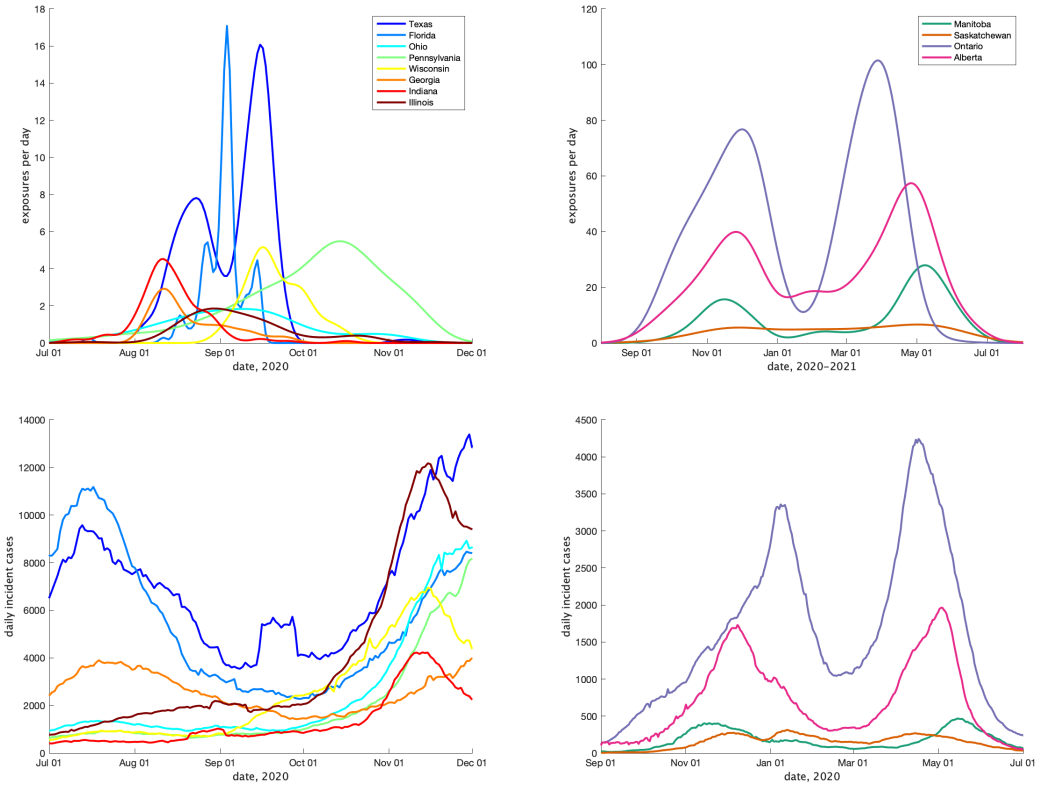

Figure 2: (Top) Estimates of the rate of new clusters being reported in each jurisdiction as a function of time, in 8 US states (left), and 4 Canadian provinces (right). (Bottom) Incident cases per day in the same jurisdictions over the corresponding time interval.

Jurisdiction	Number of Clusters	Number of Schools	Schools with Multiple Clusters	Fraction with One Case	Mean observed Cluster size
Texas	369	326	30	0.69	1.73
Florida	147	134	10	0.69	5.31
Ohio	122	95	12	0.69	2.11
Pennsylvania	322	247	46	0.73	1.75
Wisconsin	130	118	11	0.46	4.02
Georgia	68	53	10	0.46	4.76
Indiana	105	84	13	0.81	2.19
Illinois	81	76	5	0.52	3.47
Manitoba	2806	542	430	0.78	1.35
Saskatchewan	1434	466	321	0.75	1.46
Ontario	12559	3337	2391	0.83	1.24
Alberta	7971	1536	1242	0.70	1.53

Table 1: For each of 8 American states and 4 Canadian provinces, number of clusters in the data, number of schools reported, number of schools with multiple clusters, number of schools with multiple clusters separated by a week or more of no reported cases.

additional people are infected by the index case, for a total of $Z + 1$ individuals in the cluster. The variable Z includes individuals directly infected by the index case, as well as any subsequent infected individuals that are included in the same cluster. Following (*25*), we model Z as a Poisson random variable with parameter ν, where ν itself is a Gamma-distributed random variable. As described by (*25*), Z is then a negative binomial random variable. Rather than the usual parametrization of a negative binomial distribution we use parameters R_c and k. The parameter R_c is the expected number of additional infections in a cluster, and k is the dispersion: a measure of how far the distribution of Z is from being Poisson. As $k \to \infty$, the distribution of Z approaches that of a Poisson distribution with mean R_c. The variance of Z is $R_c(1 + R_c/k)$ and so for smaller values of k we expect more of the secondary cases to occur in rare large clusters rather than in frequent small clusters (*25*).

There are then a total of $Z + 1$ infected individuals in the school. The number that are actually observed X depends on the ascertainment model. We consider a model where each case is observed and contributes to the reported cluster size with probability q, so that the observed cluster size X (conditioned on Z) is binomial with parameters $n = Z + 1$ and probability q. The index case is treated the same as the infectees, so X may or may not include the index case. If none of the cases in a cluster are observed, we assume the cluster is not reported, so our model factors in the effect that smaller clusters are more likely to be missed.

For each collection of cluster sizes in our datasets we estimate the mean R_c and dispersion k using the ascertainment model with $q = 0.75$. We use maximum likelihood estimation to obtain estimates of R and k, and we use the Hessian of the log-likelihood to obtain 95% confidence ellipses for the

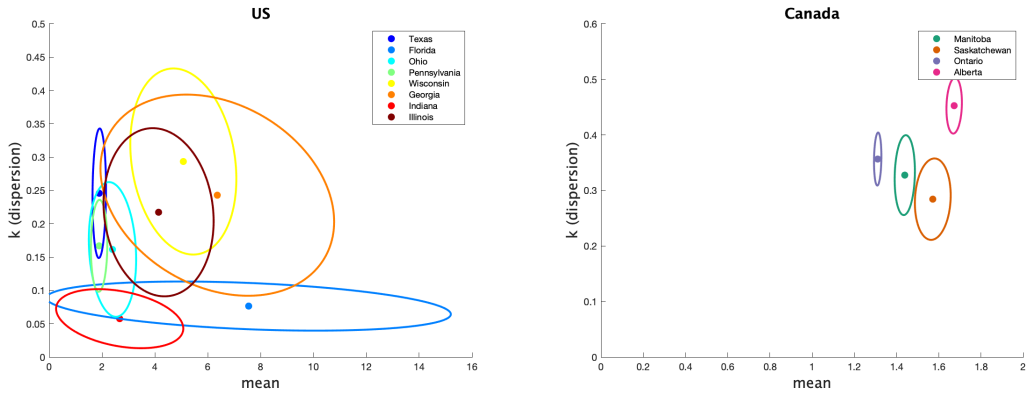

Figure 3: Estimates of mean and dispersion of cluster size for 8 US states (left) and 4 Canadians provinces (right) using the individual ascertainment model with ascertainment rate 0.75. Estimate of mean includes index case.

parameters (*42, Sec. 9.10*). See the Supplementary Information for an explicit statement of the likelihood function.

3 Results

Figure 3 (left) shows the estimated mean cluster size ($= R_c + 1$) and dispersion k for the 8 US states. Figure 3 (right) shows the same for 4 Canadian provinces. In the US data, mean cluster size was estimated to range from about 2 in Texas to almost 8 in Florida. Dispersions ranged from 0.05 to 0.3, showing considerable overdispersion compared to the Poisson distribution. In the Canadian provinces mean cluster sizes ranged between 1.3 and 1.7 cases, and dispersion ranged from 0.28 to 0.45, meaning that observed dispersion was lower than in the US data (recalling that no overdispersion corresponds to $k \to \infty$.)

In the Supplementary Information we explore varying the ascertainment fraction between 0.2 and 1. Though lower ascertainment fractions yield bigger values of R_c and smaller values of k, we see that the parameter estimates are relatively insensitive to values of q between 0.5 and 1. The reason for this is that though a given cluster with multiple cases will look smaller with fewer cases detected, and therefore biasing observed size downwards, many single-case clusters will not be detected at all, biasing the observed cluster size upwards again. We also consider an alternate model of ascertainment, where the chance of a cluster being reported at all depends on the size of the cluster, and vary the rate of ascertainment in it; see the Supplementary Information.

As a way of interpreting dispersion values and what they mean for cluster size, we consider the fraction of all cases that occur in the largest 20% of all clusters. (If the distribution of cases follows the

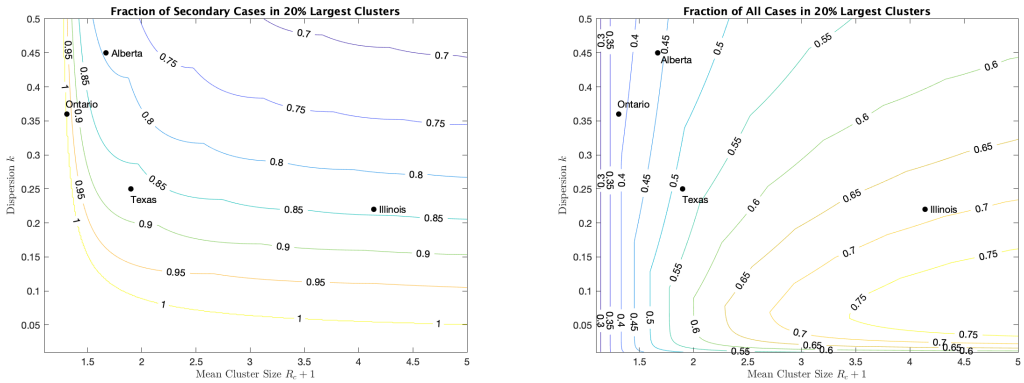

Figure 4: For a range of mean cluster size and dispersion k, the fraction of cases in the 20% largest clusters, counting only secondary cases (left), or all cases, index and secondary (right). Dots indicate the location of selected jurisdictions in the plots.

Pareto principle (*43*) then 80% of the cases will be in the top 20% largest clusters.) If we consider only secondary cases (not including the index case) we see from Figure 4 (right) the fraction that are due to the 20% largest clusters for various values of mean cluster size and k. For example, for Alberta with a mean cluster size of 1.6 and a dispersion k of 0.3, 86% of the secondary cases are in the top 20% of the clusters by size. For Florida, with a mean cluster size of 8 and $k = 0.075$, 97% of secondary cases are in the top 20% of clusters by size. When we include index cases, the fractions are correspondingly lower, as we see in Figure 4 (right).

Another way to visualize the variability of transmission we have inferred from the data is to show the distribution of the Poisson parameter ν, of which R_c is just the mean. In our model ν is the index case-specific expected number of further cases in a cluster, and is a gamma distributed random variable. Figure 5 shows the estimated distribution of ν for each jurisdiction, and Table 3 shows some key properties of the jurisdictions.

As another application of our results, we can estimate the time to observe a cluster of a given size or larger in a given jurisdiction. We used the estimate of rate of cluster occurrence shown in Figure 2 for the day September 15th, 2020. Summing the tail of the cluster size distribution past a certain point with a given R_c and k gives the fraction of clusters observed that are that size or larger. Multiplying these quantities and taking the inverse gives the expected number of days until we would see a cluster this size for larger in the dataset. We show the results of this calculation in Figure 6. These vary widely: in our estimates for Florida, it would take under 5 days for there to be a cluster of 10 cases, whereas in Indiana this would not happen until 11 or 12 days, and in Alberta it would take over 50 days.

Our model does not consider the details of transmission at the individual level, and so does not make use of an instantaneous transmission rate per contact pair. However, by making some simple assumptions

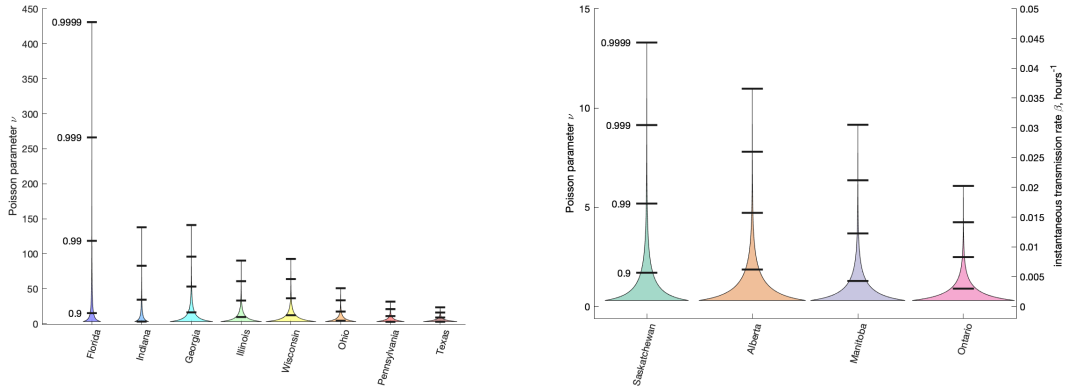

Figure 5: Estimated distribution of ν for different jurisdictions. Left, US states. Right, Canadian provinces. Right axis shows distribution of instantaneous transmission rate β.

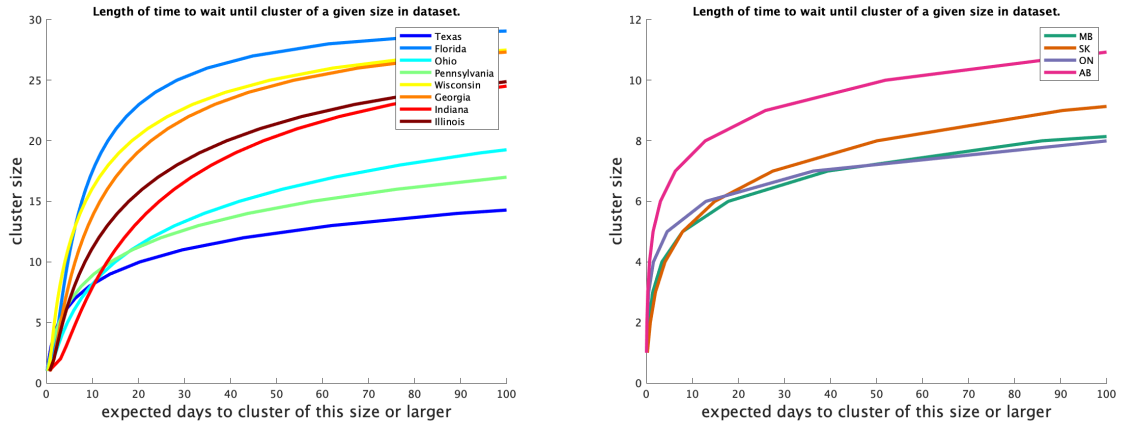

Figure 6: For each jurisdiction, expected number of days to observe a cluster of a given size or larger in the data set.

Jurisdiction	Mean	Standard Deviation
Florida	6.5 (-12.6,25.7)	23.7 (-271.0,318.3)
Indiana	1.7 (-0.2,3.6)	7.0 (-34.1,48.0)
Georgia	5.4 (-1.0,11.8)	10.9 (-24.4,46.1)
Illinois	3.1 (1.7,4.6)	6.7 (-1.6,15.1)
Wisconsin	4.1 (2.8,5.4)	7.5 (1.2,13.9)
Ohio	1.4 (1.1,1.7)	3.5 (1.2,5.7)
Pennsylvania	0.9 (0.9,0.9)	2.2 (1.9,2.4)
Texas	0.9 (0.9,0.9)	1.8 (1.7,1.9)
Saskatchewan	0.57 (0.57,0.57)	1.07 (1.06,1.09)
Alberta	0.67 (0.67,0.67)	1.00 (1.00,1.00)
Manitoba	0.44 (0.44,0.44)	0.77 (0.76,0.77)
Ontario	0.31 (0.31,0.31)	0.52 (0.52,0.52)

Jurisdiction	Median	90th percentile	99th percentile
Florida	6.0e-03	15.1	118.4
Indiana	1.0e-04	3.0	34.3
Georgia	8.8e-01	16.1	53.0
Illinois	4.0e-01	9.5	33.0
Wisconsin	9.6e-01	12.1	36.3
Ohio	7.6e-02	4.2	17.3
Pennsylvania	5.4e-02	2.7	10.8
Texas	1.5e-01	2.7	8.8
Saskatchewan	1.3e-01	1.7	5.2
Alberta	2.8e-01	1.9	4.7
Manitoba	1.2e-01	1.3	3.7
Ontario	9.8e-02	0.9	2.5

Table 2: Properties of the estimated distribution for the Poisson parameter ν, the index-specific expected number of further cases in a cluster. The expected value of ν is R_c and its distribution gives important information about overdispersion of clusters. In units of hours^{-1}, with 95% credible intervals for the mean and standard deviation.

about COVID transmission, we can infer a distribution of transmission rate β from our estimate of the distribution of the parameter ν. Recall that ν is a Gamma-distributed random variable that gives mean number of secondary cases. Another way to estimate mean cluster size is to use an individual contact model where when an infectious person is in contact with a susceptible person, the susceptible person is infected with rate β. In such a model we assume that infected individuals are in a classroom for 2 days before isolating (when they develop symptoms), and that the total contact time with their classmates is $T = 12$ hours. Assuming that all individuals are in the same class, the infected individual is in contact with $n = 25$ other susceptible students for that time period. Then the infected individual will on average infect βnT other students. So we estimate $\beta = \nu/(nT)$. Since ν is Gamma-distributed, our estimate of β is too. The approximations here are only reasonable when it's plausible that all secondary infections

are in the same classroom; this is generally not the case with the US data sets, since there are occasional clusters of more than 30 cases. So we estimate the distribution of β this way only for the Canadian data sets. Figure 5 (right) shows with the right y-axis the distribution of β for the different Canadian provinces. Table 3 shows some of the features of the estimated distribution for β.

Jurisdiction	Mean	Standard Deviation
Saskatchewan	1.9e-03 (1.9e-03,1.9e-03)	3.6e-03 (3.5e-03,3.6e-03)
Alberta	2.2e-03 (2.2e-03,2.2e-03)	3.3e-03 (3.3e-03,3.3e-03)
Manitoba	1.5e-03 (1.5e-03,1.5e-03)	2.6e-03 (2.5e-03,2.6e-03)
Ontario	1.0e-03 (1.0e-03,1.0e-03)	1.7e-03 (1.7e-03,1.7e-03)

Jurisdiction	Median	90th percentile	99th percentile
Saskatchewan	4.2e-04	5.7e-03	1.7e-02
Alberta	9.3e-04	6.2e-03	1.6e-02
Manitoba	4.1e-04	4.3e-03	1.2e-02
Ontario	3.3e-04	3.0e-03	8.3e-03

Table 3: Properties of the estimated distribution for the instantaneous transmission rate β. In units of hours^{-1} , with 95% credible intervals for the mean and standard deviation.

One application of these estimates of the distribution of β is that we can explore the consequences of different types of interventions in the classroom setting. In (*37*) the authors consider a simple model of COVID-19 transmission among a group of contacts and investigate the quantity R_{event}, the average number of secondary infections due to the presence of a single infectious individual. R_{event} is determined by T, the total length of time the infectious individual is with others; n_{contact}, the number of contacts at any point in time, τ the length of time the individual is with a fixed set of contacts; and β, the instantaneous transmission rate. Interventions can be classified according to which of these parameters they modify: reducing transmission reduces β, social distancing reduces n_{contact}, and "bubbling" (staying with the same small group rather than mingling) increases τ to T. If use our distributions for β with the model of (*37*) we can estimate how the distribution of cluster sizes is changed with different interventions under different values of the parameters R_c and k.

In Figure 7 we show estimated size distributions of clusters under different interventions. Our baseline simulation settings intend to capture a pre-COVID high school classroom: $T = 12$ hours (two days of exposure before the index case isolates), $\tau = 3$ hours (each student has four different classes that they attend, $n_{\text{class}} = 25$, and β is sampled from our estimated distribution for a given choice of R_c and k. We consider three interventions: reducing transmission (for example, by introducing masks) reduces β by a factor of 2; social distancing cuts the size of a class in half; strict bubbling sets τ to T. Figure 7(left) shows the distributions of cluster size with the baseline and the different interventions for choices of R_c and k appropriate to Saskatchewan ($R_c = 0.5$, $k = 0.3$). We see that both reducing transmission and social distancing are effective, whereas bubbling does not contribute much to reduce cluster sizes. This is

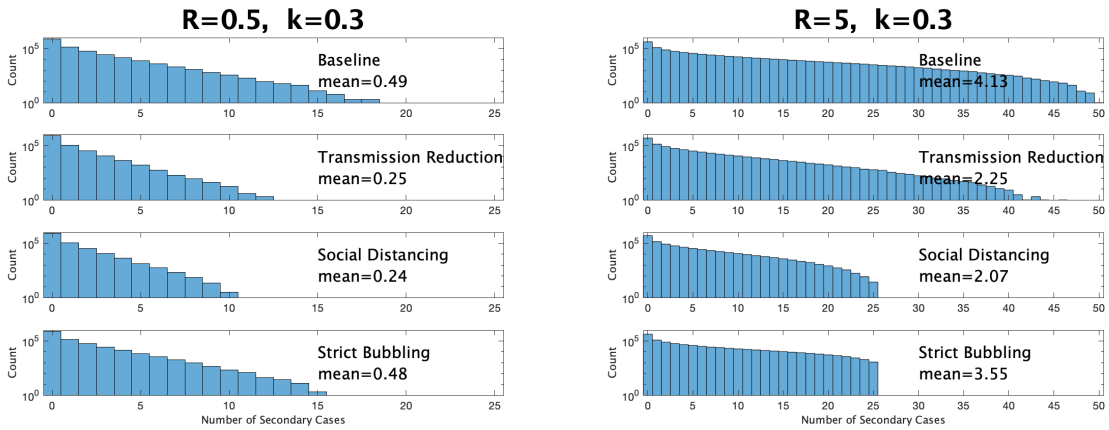

Figure 7: Distribution of cluster sizes under baselines conditions and under three interventions. Left: under parameter choice $R = 0.5$ and $k = 0.3$. Right: with 10 times as large R_c.

characteristic of what (*37*) call the linear regime: the number of secondary infections depends linearly on the time the infectious person is present with others. Figure 7 (right) shows the results in a hypothetical setting where R is much larger ($R_c = 0.5$, $k = 0.3$). Here transmission reduction is less effective than in the linear regime, and strict bubbling more so; these parameters have moved us closer to the so-called saturating regime.

4 Discussion

We have used cluster size data to estimate the mean and dispersion in cluster sizes, accounting for imperfect case detection. We have found that in all jurisdictions in our data, the majority of school transmission occurs in a small number of classrooms, with the top 20% of clusters containing between 80% and 100% of the secondary cases in school settings. We developed a method to estimate the transmission rate per contact per unit time, with reference to a simple model of classroom transmission. Having a direct estimate of the transmission rate allows us to compare the benefits of different control measures. We find that with parameters estimated from Canadian jurisdictions, interventions that reduce transmission rates (such as masking) and reduce number of contacts at any one time (class size reduction), are more effective than strategies aimed at keeping sets of contacts consistent (such as bubbling).

Overdispersion in transmission of COVID-19 and other infectious diseases is well documented (eg (*45*)) and is often described with reference to the 20/80 rule: that 20% of the infected individuals account for 80% of the transmission. Naturally, if the more infectious 20% can be identified, interventions targeting that portion of the population are likely to have a high impact. For SARS in 2003, Lloyd-Smith et al (*25*) estimated that 20% of the cases were responsible for almost 90% of the transmission. Estimates

for COVID-19 also find considerable overdispersion, with the parameter k between 0.1 (*16*) and 0.5 (*22*) (with $R_0 = 2.5$ this gives the top 20% of cases causing 69% - 96% of the transmissions; see (*32*) for a survey). These estimates focus on the distribution of the number of people an infectious person infects directly during the whole course of infection (with mean R_0), which is of obvious epidemiological importance, but for which it is difficult to obtain high-quality data. When a case is identified, we are not always able to determine who they infected, and indirect methods must be used. We may miss cases, and others may be wrongly attributed to a given index case.

In our present study, we examined a different random quantity, the number of additional cases Z infected, either directly or through intermediaries, by a given index case in a given setting. We denoted the mean of Z by R_c. Including the index case means that the cluster size is $Z + 1$, with mean $R_c + 1$. Compared to estimates of R_0, R_c does not count people infected at other sites, but it does include additional cases, because it includes both direct and indirect transmission. Z and its mean R_c are therefore more focused on the particular setting (in this case a school) than R_0 is. In general it will depend on the infectiousness of the index case, as well as how conducive the environment is to transmission, and what activities are undertaken there. Determining the distribution of Z, as we have done here, provides an alternative means of investigating transmission.

However, these two measures of transmissibility (R_0 and R_c, the mean of Z) may be close enough that it is instructive to compare our estimates for Z with the traditional R_0, and our dispersion estimates with dispersion estimates for the number of secondary infections. Our R_c ranges from 0.3 in Ontario to nearly 7 in Florida. The lower range of R_c is inconsistent with R_0 estimates (which range from 2 to 6 (*3*)), and indicate that in these jurisdictions schools are unlikely to be a major contributor to COVID-19 spread. Our medium to higher estimates of R_c are consistent with estimates of R_0 and suggest that school transmissions were an important factor in COVID transmission in general in these jurisdictions at these times. Our estimates for k range from 0.06 (in Indiana) to 0.5 (in Alberta), corresponding closely to earlier estimates of dispersion.

Overdispersion has consequences for controlling transmission and for estimation. Estimating the average transmission rate from a small number of clusters will be difficult, and will result in a high variability. Most likely what will be observed in a small number of sampled clusters will be little to no onward transmission, which would lead to underestimates of the transmission rate. But if one or more larger clusters are included in a sample by chance, then this could lead to an overestimate of the transmission rate.

If we could identify the conditions under which the rare larger clusters occur (high-risk individuals, activities and settings) we could achieve a disproportionately large effect on reducing transmission by applying new measures in these settings. There are myriad possible reasons for overdispersion of transmission for SARS-CoV-2, including variation in viral load (*10*), behaviour, and number of contacts. Properties of the setting may be very important, with some settings (cramped, poor ventilation) being

especially conducive to transmission. It would be good to identify classrooms or schools where there is a high risk of larger clusters. For example, if data were available on the occupancy, ventilation standards, mask use, classroom size, distancing behaviour and other features of classrooms, we could investigate how this related to the cluster size. Rapid tests may be especially good for identifying the most infectious individuals, given that they are sensitive to viral loads (*27*), but additional data collection is likely needed to quantify setting-level risks.

Two important changes have happened since the majority of the data here was collected. Firstly, in the jurisdictions studied, effective vaccines have been developed and deployed for those aged 12 and up (*35*). There are several ways in which this may effect cluster sizes in the school setting. To the extent that the general population (including adults) being vaccinated reduces incidence of COVID (*23, 26, 44*), there will be fewer introductions of COVID into the classroom, and so fewer exposures will occur leading to fewer clusters. This effect may be dampened by relaxation of distancing and other measures that were keeping COVID-19 at bay and are no longer necessary in the context of vaccination. The distribution of cluster sizes when clusters do occur will also change: many students who might otherwise be infected will be protected by the vaccine, others who are vaccinated but infected (breakthrough infection) may have reduced symptoms and therefore may not be detected. We therefore expect the mean cluster size to be reduced by vaccination, in age ranges where vaccination has been deployed. It is unclear what the consequences will be for the dispersion.

Secondly, new, more transmissible variants of SARS-CoV-2 have emerged (*6*), most notably the Alpha variant with a transmissibility approximately 1.5 times higher than that of the original variant, and the Delta variant estimated to have a transmissibility 1.5 times higher than that (*15*). A natural way to implement this change in our model is to multiply R_c by a factor of 2.25, boosting the size of clusters, without changing the dispersion parameter k. This would result in mean cluster sizes ranging from 1.7 in Ontario to 16 in Florida according to our estimates. Data from the period in which Delta was the prevalent strain is not available, but schools in the Canada and the US saw resurgences in clusters in schools around school openings (*2, 13, 33*).

Our data and model have some limitations. The data rely on crowdsourcing, and there is reason to believe that reporting is incomplete. Larger clusters may be more likely to be reported. The rate at which exposures were reported in these data varies a great deal from state to state, and so these results cannot be used to estimate the true rate of exposures in schools in the jurisdictions. For example, California and New York are some of the most populous states in the US, both of which were severely hit by COVID during this period, and yet neither of them had enough reports to make it into the top 8 most reported states. However, the consistency of the Canadian estimates, despite the data being from different jurisdictions with different reporting, lend them credibility. In the modelling, we assumed a Poisson random variable for the cluster size, with an underlying gamma-distributed rate variable. This is a flexible model allowing for considerable overdispersion, but it is simple and does not explicitly handle complexities

such as the differences between elementary and high schools. Our estimates of the transmission rate were derived (where feasible) from a model with a fixed number of hours that the index case would be infectious in the classroom, and fixed class sizes. Accounting for variation in these would result in more variability in the estimates. We also assumed that reported clusters in the Canadian data occurred in the same class. Errors where distinct clusters are incorrectly put together and errors where one larger cluster is incorrectly split and assumed to be two distinct clusters probably both occur in the data. Without additional information, however, we were not able to account for them. Finally, our illustrative modelling of the impact of interventions was simple, and used simple assumptions for the impacts of masking, distancing and cohorts (bubbling). Our estimates of the per-contact transmission rate per unit time could, however, be used in more sophisticated simulation modelling to compare interventions.

Despite these limitations, our approach has distinct advantages. We have developed estimates of the person-to-person transmission rate derived directly from data. The data we use (cluster sizes) are relatively easy to access. This approach does not require individual-level data or contact tracing information, which are often not available; individuals may be identifiable and data are held within public health institutions. Our estimation approach, together with cluster size data, offers a high-resolution view of transmission: we can estimate the distribution of cluster sizes in specific settings, accounting for reporting and overdispersion, and in some contexts we can estimate the transmission rate, all without requiring either individual-level data or assumptions on transmission parameters such as the serial interval (see, in contrast, (*12, 40*) which require serial interval estimates). The results offer context-specific tools to simulate interventions in particular settings (here, schools). The methods are readily generalizable to other structured settings, such as workplace outbreaks where workplaces are similar in size and structure. Our results also suggest the need for data collection activities that can relate cluster sizes to setting variables such as occupancy, density, ventilation, activity and distancing behaviour. Ultimately this would provide the data needed to design interventions that best reduce school and/or workplace transmission.

Acknowledgments

We thank Alisha Morris and the other volunteers at the National Education Association for providing the US data that was used in this study. PT and CC were supported by Natural Science and Engineering Research Council (Canada) Discovery Grants (RGPIN-2019-06911 and RGPIN-2019-06624). CC receives funding from the Federal Government of Canada's Canada 150 Research Chair program.

References

1. Masks4Canada – canadian doctors, professionals, & citizens for masks. https://masks4canada.org/. Accessed: 2021-10-26.

2. Kids are facing covid-19 risks. here's what parents can do - CNN video. *CNN*, August 2021.

3. Yousef Alimohamadi, Maryam Taghdir, and Mojtaba Sepandi. Estimate of the basic reproduction number for COVID-19: A systematic review and meta-analysis. *J. Prev. Med. Public Health*, 53(3):151–157, May 2020.

4. National Education Association. Nea school and campus covid-19 reporting site, https://app.smartsheet.com/b/publish?eqbct=00a2d3fbe4184e75b06f392fc66dca13, accessed: 12.04.2020.

5. Covid Schools Canada. Canada covid-19 school tracker, https://covidschoolscanada.org/.

6. CDC. For parents: Multisystem inflammatory syndrome in children (MIS-C) associated with COVID-19. https://www.cdc.gov/mis/mis-c.html, September 2021. Accessed: 2021-10-26.

7. CDC. SARS-CoV-2 variant classifications and definitions. https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html, June 2021. Accessed: 2021-6-15.

8. Centers for disease control and prevention. Science brief: Transmission of sars-cov-2 in k-12 schools and early care and education programs – updated. https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/transmission_k_12_schools.html, September 2021. Accessed: 2021-10-26.

9. Sonia Chaabane, Sathyanarayanan Doraiswamy, Karima Chaabna, Ravinder Mamtani, and Sohaila Cheema. The impact of COVID-19 school closure on child and adolescent health: A rapid systematic review. *Children*, 8(5), May 2021.

10. Paul Z Chen, Niklas Bobrovitz, Zahra Premji, Marion Koopmans, David N Fisman, and Frank X Gu. Heterogeneity in transmissibility and shedding SARS-CoV-2 via droplets and aerosols. October 2020.

11. Victor Chernozhukov, Hiroyuki Kasahara, and Paul Schrimpf. The association of opening K-12 schools with the spread of COVID-19 in the united states: County-level panel data analysis. *Proc. Natl. Acad. Sci. U. S. A.*, 118(42), October 2021.

12. Anne Cori, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.*, 178(9):1505–1512, November 2013.

13. Beth Reese Cravey. Duval schools report almost 1,500 COVID-19 cases; 2 schools remain closed due to outbreaks. *Florida Times-Union*, August 2021.

14. Itai Dattner, Yair Goldberg, Guy Katriel, Rami Yaari, Nurit Gal, Yoav Miron, Arnona Ziv, Rivka Sheffer, Yoram Hamo, and Amit Huppert. The role of children in the spread of COVID-19: Using household data from bnei brak, israel, to estimate the relative susceptibility and infectivity of children. *PLoS Comput. Biol.*, 17(2):e1008559, February 2021.

15. Michaeleen Doucleff. The delta variant isn't as contagious as chickenpox. but it's still highly contagious. *NPR*, August 2021.

16. Akira Endo, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Sam Abbott, Adam J Kucharski, and Sebastian Funk. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res*, 5:67, July 2020.

17. Ontario Government. Schools COVID-19 data - ontario data catalogue. https://data.ontario.ca/dataset/summary-of-cases-in-schools. Accessed: 2021-10-27.

18. Douglas N Harris. When should schools reopen fully in-person? https://www.brookings.edu/blog/brown-center-chalkboard/2020/09/29/when-should-schools-reopen-fully-in-person/, September 2020. Accessed: 2021-10-26.

19. Fraser Health. School exposures. https://www.fraserhealth.ca/schoolexposures. Accessed: 2021-10-26.

20. Vancouver Coastal Health. COVID-19 potential exposure events in VCH schools. http://www.vch.ca/covid-19/school-exposures. Accessed: 2021-10-27.

21. Rebecca L Laws, Rebecca J Chancey, Elizabeth M Rabold, Victoria T Chu, Nathaniel M Lewis, Mark Fajans, Hannah E Reses, Lindsey M Duca, Patrick Dawson, Erin E Conners, Radhika Gharpure, Sherry Yin, Sean Buono, Mary Pomeroy, Anna R Yousaf, Daniel Owusu, Ashutosh Wadhwa, Eric Pevzner, Katherine A Battey, Henry Njuguna, Victoria L Fields, Phillip Salvatore, Michelle O'Hegarty, Jeni Vuong, Christopher J Gregory, Michelle Banks, Jared Rispens, Elizabeth Dietrich, Perrine Marcenac, Almea Matanock, Ian Pray, Ryan Westergaard, Trivikram Dasu, Sanjib Bhattacharyya, Ann Christiansen, Lindsey Page, Angela Dunn, Robyn Atkinson-Dunn, Kim Christensen, Tair Kiphibane, Sarah Willardson, Garrett Fox, Dongni Ye, Scott A Nabity, Alison Binder, Brandi D Freeman, Sandra Lester, Lisa Mills, Natalie Thornburg, Aron J Hall, Alicia M Fry, Jacqueline E Tate, Cuc H Tran, and Hannah L Kirking. Symptoms and transmission of SARS-CoV-2 among children - utah and wisconsin, March-May 2020. *Pediatrics*, 147(1), January 2021.

22. Ramanan Laxminarayan, Brian Wahl, Shankar Reddy Dudala, K Gopal, Chandra Mohan, S Neelima, K S Jawahar Reddy, J Radhakrishnan, and Joseph A Lewnard. Epidemiology and transmission dynamics of COVID-19 in two Indian states. *Science*, September 2020.

23. Eyal Leshem and Annelies Wilder-Smith. COVID-19 vaccine impact in israel and a way out of the pandemic. *Lancet*, 397(10287):1783–1785, May 2021.

24. Andrew T Levin, William P Hanage, Nana Owusu-Boaitey, Kensington B Cochran, Seamus P Walsh, and Gideon Meyerowitz-Katz. Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. *Eur. J. Epidemiol.*, 35(12):1123–1138, December 2020.

25. James O Lloyd-Smith, Sebastian J Schreiber, P Ekkehard Kopp, and Wayne M Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005.

26. Smriti Mallapaty. China COVID vaccine reports mixed results - what does that mean for the pandemic? *Nature*, January 2021.

27. Michael J Mina and Steven Phillips. Opinion. *The New York Times*, October 2021.

28. British Columbia Ministry of Education. K-12 education restart plan. https://www2.gov.bc.ca/assets/gov/education/administration/kindergarten-to-grade-12/safe-caring-orderly/k-12-education-restart-plan.pdf, July 2020. Accessed: 2021-10-26.

29. State of Michigan. School-Related cluster and outbreak reporting. https://www.michigan.gov/coronavirus/0,9753,7-406-98163_98173_102480---,00.html. Accessed: 2021-10-27.

30. Bellwether Education Partners. Missing in the margins 2021: Revisiting the COVID-19 attendance crisis. https://bellwethereducation.org/publication/missing-margins-estimating-scale-covid-19-attendance-crisis, October 2021. Accessed: 2021-10-26.

31. Mendel E. Singer, Ira B. Taub, and David C Kaelber. Risk of myocarditis from covid-19 infection in people under age 20: A population-based analysis. *medRxiv*, 2021.

32. Kim Sneppen, Bjarke Frost Nielsen, Robert J. Taylor, and Lone Simonsen. Overdispersion in covid-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control. *Proceedings of the National Academy of Sciences*, 118(14), 2021.

33. Star staff and wire services. Today's coronavirus news: Ontario reports 748 cases; province still in fourth wave as cases rise, says ontario's top doctor. *The Toronto Star*, November 2021.

34. Carole H Sudre, Benjamin Murray, Thomas Varsavsky, Mark S Graham, Rose S Penfold, Ruth C Bowyer, Joan Capdevila Pujol, Kerstin Klaser, Michela Antonelli, Liane S Canas, Erika Molteni,

Marc Modat, M Jorge Cardoso, Anna May, Sajaysurya Ganesh, Richard Davies, Long H Nguyen, David A Drew, Christina M Astley, Amit D Joshi, Jordi Merino, Neli Tsereteli, Tove Fall, Maria F Gomez, Emma L Duncan, Cristina Menni, Frances M K Williams, Paul W Franks, Andrew T Chan, Jonathan Wolf, Sebastien Ourselin, Tim Spector, and Claire J Steves. Attributes and predictors of long COVID. *Nat. Med.*, 27(4):626–631, April 2021.

35. The New York Times. See how vaccinations are going in your county and state. *The New York Times*, December 2020.

36. Zeynep Tufekci. This overlooked variable is the key to the pandemic. *Atl. Mon.*, September 2020.

37. Paul Tupper, Himani Boury, Madi Yerlanov, and Caroline Colijn. Event-specific interventions to minimize COVID-19 transmission. *Proc. Natl. Acad. Sci. U. S. A.*, 117(50):32038–32045, December 2020.

38. Russell M Viner, Oliver T Mytton, Chris Bonell, G J Melendez-Torres, Joseph Ward, Lee Hudson, Claire Waddington, James Thomas, Simon Russell, Fiona van der Klis, Archana Koirala, Shamez Ladhani, Jasmina Panovska-Griffiths, Nicholas G Davies, Robert Booy, and Rosalind M Eggo. Susceptibility to SARS-CoV-2 infection among children and adolescents compared with adults: A systematic review and meta-analysis. *JAMA Pediatr.*, 175(2):143–156, February 2021.

39. Tim Walker. Teacher creates national database tracking COVID-19 outbreaks in schools. https://www.nea.org/advocating-for-change/new-from-nea/teacher-creates-national-database-tracking-covid-19-outbreaks. Accessed: 2021-10-27.

40. Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.*, 160(6):509–516, September 2004.

41. Sebastian Walsh, Avirup Chowdhury, Vickie Braithwaite, Simon Russell, Jack Michael Birch, Joseph L Ward, Claire Waddington, Carol Brayne, Chris Bonell, Russell M Viner, and Oliver T Mytton. Do school closures and school reopenings affect community transmission of covid-19? a systematic review of observational studies. *BMJ Open*, 11(8), 2021.

42. Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

43. Wikipedia contributors. Pareto principle. https://en.wikipedia.org/w/index.php?title=Pareto_principle&oldid=1052101893, October 2021. Accessed: NA-NA-NA.

44. Annelies Wilder-Smith and Kim Mulholland. Effectiveness of an inactivated SARS-CoV-2 vaccine. *N. Engl. J. Med.*, 385(10):946–948, September 2021.

45. M E Woolhouse, C Dye, J F Etard, T Smith, J D Charlwood, G P Garnett, P Hagan, J L Hii, P D Ndhlovu, R J Quinnell, C H Watts, S K Chandiwana, and R M Anderson. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc. Natl. Acad. Sci. U. S. A.*, 94(1):338–342, January 1997.

Supplementary Information

Model for Cluster Size

We consider two models for ascertainment (whether a case is actually detected), though we only consider the first in the main text.

In the first ascertainment model (**Individual Ascertainment**) each of the infected individuals is detected with a probability q_1. So X, the total number of infected individuals is binomial (n, p) with parameters $n = Z + 1$ and $p = q_1$. If by chance none of the individuals are observed, we do not observe the cluster. This is meant to model a situation where cases are detected independently of each other, and one detected case does not lead to further tests or screening.

In the second ascertainment model (**Cluster Ascertainment**), at first each case is identified with probability q_2, but then if any of the students are identified they are all subsequently identified. This is intended to capture a situation where a single detected case triggers testing for the whole class. Again, if no cases are detected we do not observe the cluster. This is equivalent to saying that clusters of size m are detected in their entirety with probability $1 - (1 - q_2)^m$.

The number Z of new cases given the presence of one infectious case is a Poisson distributed random variable with a rate ν that is itself a Gamma-distributed random variable with a shape k and scale θ. This means Z has a negative binomial distribution $\mathrm{NB}(r, p)$ where $r = k$ and $p = \theta/(1 + \theta)$. Letting $\Theta = (k, \theta)$, the cdf of Y is

$$P(Z = j) = V_\Theta(j) = \frac{\Gamma(j + r)}{j!\Gamma(r)}(1 - p)^r p^j.$$

Under the individual ascertainment model with ascertainment probability q_1, X, the number of observed cases is binomial (n, p) with $n = Z + 1$ and $p = q_1$. So, the probability that i individuals are observed is

$$W_{\Theta,q_1}(i) = \sum_{j=0}^{\infty} \binom{j+1}{i} V_\Theta(j) q_1^i (1 - q_1)^{j+1-i}$$

for $i = 0, 1, \ldots$. Since we do not observe clusters with no observed cases the probability of observing a cluster of size i is actually $P(X = i) = C_{\Theta,q_1}^{-1} W_{\Theta,q_1}(i)$ for $i = 1, 2, \ldots$, where $C_{\Theta,q_1} = \sum_{i=0}^{\infty} W_{\Theta,q_1}$.

If the observed cluster sizes are X_i, $i = 1, \ldots, n$, the log-likelihood function for $\Theta = (k, \theta)$ under the

individual ascertainment model is then

$$\sum_{i=1}^{n} \log \left[C_{\Theta,q_1}^{-1} W_{\Theta,q_1}(X_i) \right]$$

Under the cluster ascertainment model, the cluster is observed or not with probability $1-(1-q_2)^{Y+1}$. So the probability of observing $X=j$ in a cluster is

$$U_{\Theta,q_2}(i) = [1-(1-q_2)^i] V_\Theta(i)$$

for $i = 0, 1, \ldots$ but then since we cannot observe clusters of size 0, an observed cluster has size i with probability

$$P(X=i) = D_{\Theta,q_2}^{-1} U_{\Theta,q_2}(i)$$

for $i = 1, 2, \ldots$, where $D_{\Theta,q_1} = \sum_{i=0}^{\infty} U_{\Theta,q_2}(i)$.
If the observed cluster sizes are X_i, $i = 1, \ldots, n$, the log-likelihood function for $\Theta = (k, \theta)$ under the cluster ascertainment model is then

$$\sum_{i=1}^{n} \log \left[D_{\Theta,q_2}^{-1} U_{\Theta,q_2}(X_i) \right]$$

Under both ascertainment models, we then go from our estimates of k and Θ to estimates of R_c via the formula

$$R_c = \frac{pr}{1-p} = k\theta.$$

We use the Delta method to obtain confidence intervals for R_c from confidence intervals on k and θ.

Varying the rate and model of ascertainment

In the main text we estimated parameters with the assumption of the individual ascertainment model with an ascertainment probability of 0.75. Here we investigate how our main parameters $R_c + 1$ (expected cluster size) and k (dispersion) vary with this ascertainment probability. We also consider the alternate ascertainment model discussed in the previous section **Model for Cluster Size**.
Figure 8 shows the parameter estimates for the two models. On the left we show the results for the US states, and on the right for the Canadian provinces. The top plots show results for the individual ascertainment model where we set $q_2 = 1$ and vary q_1 from 0.2 to 1. The bottom plots show results for the group ascertainment model with $q_1 = 1$ and q_2 varying from 0.2 to 1. We see that the parameters do vary with the model and the ascertainment fraction, but relative magnitudes of the parameters in different jurisdictions do not change. In particular, for a range of q_1 greater than 0.5 in the individual ascertainment model, there is little change in the parameter estimates.

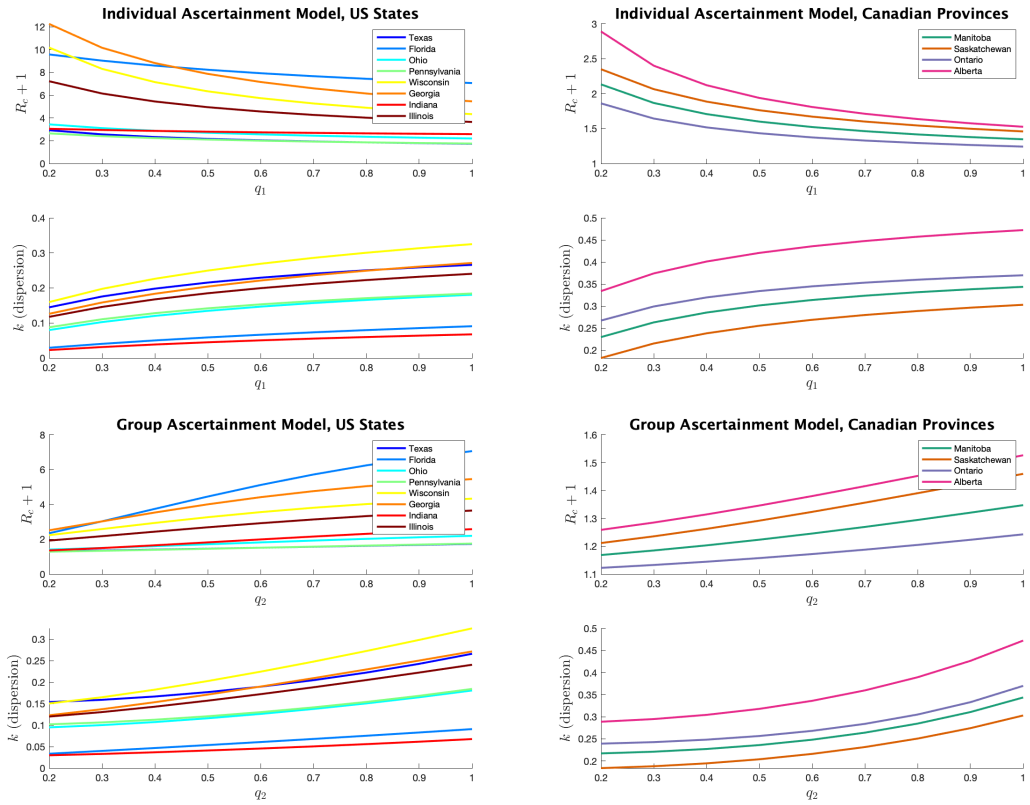

Figure 8: Estimates of mean and dispersion of cluster size for 8 US states (left) and 4 Canadians provinces (right) using the individual ascertainment model (top) and the group ascertainment model (bottom) with varying ascertainment rate. Estimate of mean includes index case.