Fall 2021 Classroom COVID-19 Transmission: Follow-up Analysis

Cornell COVID-19 Modeling Team
Analysis completed March 2022 and published February 2023[1]

Summary

Using data from Cornell's Fall 2021 semester, the statistical analysis in this report finds that being enrolled in the same class with a COVID-positive student did not measurably increase a student's risk of testing positive for SARS-CoV-2.

The data used in this report comes from a time period when masks were mandatory in classrooms and seating was at full density, two factors that influence the risk of SARS-CoV-2 transmission. While it is also important to quantify the risk of transmission when masks are optional, the available data do not permit doing so easily. In periods with mask-optional classrooms at Cornell, most COVID-testing data was self-provided and came from optional tests that individuals chose to take on their own. This self-provided data gives an incomplete view of who was infected when from which it is hard to assess the impact of classrooms.

This analysis complements and confirms an earlier analysis done in August 2021, which used physical simulation and data published in the scientific literature to predict the risk of classroom transmission for students and instructors in classrooms with full density and mandatory masking. The two analyses confirm each others' findings — that classrooms with mandatory masking are a low-risk environment, even with full-density seating.

Introduction

In transitioning to dense in-person instruction from online or socially-distanced classes, understanding the transmission risk of COVID-19 in dense classrooms is crucial. Our previous analysis in August 2021 used physical modeling and parameters from the scientific literature to estimate the risk of classroom transmission. The average risk was projected to be low for Cornell students and instructors, which supported the decision to provide dense in-person instruction in the Fall 2021 semester. Using retrospective data now available from Fall 2021, this report studies the same question and finds a similar answer — the risk associated with in-person instruction using fully dense seating and mandatory masking is low.

[1] The technical analysis was completed over the period from October 2021 to March 2022. Because the results simply confirmed earlier predictions and did not affect decisions, the authors prioritized completing other time-critical analyses needed to help Cornell respond to the pandemic above finishing this report. The authors returned to writing this report later in 2022 once no more time-critical pandemic-related analyses were needed. The report was finished and published in February 2023.

We study all the classes in which Cornell students were registered in Ithaca in Fall 2021, and estimate the effect that being enrolled in the same class with a positive case has on a person's infection status. We do not see statistical evidence that being enrolled in the same class with a positive case increases one's risk of infection. The most pessimistic estimate of the probability that an undergraduate student catches the infection due to classroom transmission over the entire semester is an order of magnitude lower (1e-4) than that predicted by physical modeling (1e-3). Indeed, the risk of transmission in classrooms is low compared to both social activities, where distancing and masking requirements may not be obeyed, and travel, which have been the dominant types of events associated with past student infections.

This analysis is based on a period where Delta is the dominant variant. Omicron is estimated to be 2 to 5 times more transmissible than Delta (Lyngse et al. 2021 UK HSA 2021, Sofonea et al. 2021). Multiplying the risk estimates for Delta by a factor of 2 to 5 therefore gives a rough estimate of the corresponding risk for Omicron. When multiplied by this factor, even pessimistic estimates of risk remain low.

Analysis details

Data and assumptions

We investigate data on students and the courses they are registered for in Fall 2021. The date range is Aug 26 - Dec 7 2021 (104 days). Each positive case has a unique notification date, when the student was first notified by the health department of their positive test result within the study's date range. The courses considered are lectures, discussions, labs, and seminars in Ithaca.

We construct a dataset with 2,839,592 entries that contains, for each student, an entry corresponding to each day leading up to their first notification date of a positive test result (if any)[2]. Each data entry contains the de-identified ID of the student, the date, the student's academic career, the student's infection status on this date, a covariate named $class\_positivity$ that captures a student's amount of exposure due to co-enrollment in classes with positive students in the previous week, and a covariate named $campus\_positivity$ that captures the risk of exposure through general interactions on campus. The period studied does not include the spike in Omicron cases, which occurred after instruction ended in Fall 2021.

We now explain the modeling details, based on which we derive the infection status, $class\_positivity$, and $campus\_positivity$ in each data entry.

[2] In principle, the data should exclude the no-action-positives (NAPs) at the start of the Fall 2021 semester (i.e, positive cases with notification date within 90 days prior to the semester start date), because they would be unlikely to be reinfected within a short time due to naturally acquired immunity. However, there are only 56 such student NAPs, so including them has a negligible effect on results.

We assume that a positive infection status is equivalent to being infectious. Hereafter, "positive case", "positive infection status", and "infectious case" are used interchangeably. We say a positive case is "active" if they participate in classes and other activities normally.

Inferring the infection status of a student on each day, i.e, the response variable, involves making a few simplifying approximations, which we detail below.

Approximation 1: *Each positive case is active and infectious for a week before being notified by the health department and isolated.*

Based on Assumption 1, for a positive case notified on day t, we set their infection status to "positive" on days $max(0, t-6)$ through day t, and to "negative" on days before $t-6$. Days beyond t are not included in the dataset. For a student that never tested positive during the study period, their infection status is negative for all 104 days.

Next, we describe our approach for computing the covariate $class\_positivity$, a measure of exposure due to co-enrollment with positive students. Hereafter, we use the word "class" to denote one instance of meeting (e.g., a lecture on Aug 26, 9-9:50am), and we use "course" to denote the collection of classes under the same name and catalog number, e.g., MATH 1110.

Approximation 1 enables us to construct a history of the presence of positive cases in each course throughout the semester. For each positive case, we get the classes in which they are enrolled during the week before the notification date and accumulate one positive case in that class. Based on this history of positive cases in each class, we compute $class\_positivity$ for each student on each day based on the courses in which they are enrolled.

We first define $class\_prevalence$ for a class c that student s takes on day τ as the fraction of students in class c, other than s, that was positive on day τ:

$$class\_prevalence(s, c, \tau) = \frac{number\ of\ positives\ other\ than\ s\ in\ class\ c\ on\ day\ \tau}{(number\ of\ students\ enrolled\ in\ class\ c) - 1}. \quad (1)$$

We use the fraction rather than the number of positive cases in the classes, because the former has a larger effect on a student's probability of being close to a positive case (assuming that seating is reasonably randomized). For example, we expect the risk to be higher in a 20-person class with 10 positives than in a 200-person class with 10 positives. Though the number of positive cases is the same, a student in the former class is more likely to be close to a positive than a student in the latter.

Approximation 2*: A student's infection status on day t may be affected by exposure within a week preceding t but does not depend on exposure that happened more than a week ago.*

Then, for a student s on a given day t, we define $class\_positivity$ as the sum of $class\_prevalence$, multiplied by the class duration, over all the classes in which they are

enrolled during the week preceding t. Let $C(s,\tau)$ denote the set of classes that student s takes on day τ. Then,

$$class\_positivity(s,t) = \sum_{\tau=max(0,t-T)}^{t} \sum_{c \in C(s,\tau)} class\_prevalence(s,c,\tau) * (hours\ of\ class\ c\ on\ day\ \tau), \quad (2)$$

where T is set to be 7 days.

For small amounts of exposure, the probability of infection is approximately linear in the amount of exposure time, where the probability of being exposed to a positive person is proportional to the prevalence in the classroom. According to this reasoning, $class\_positivity$ is approximately proportional to the probability of transmission. A more detailed discussion is given in Appendix A.

Finally, for each student s on a given day t, we also include a covariate named $campus\_positivity$, to account for the risk due to general interactions with other students on campus outside of the classroom. These include, but are not limited to, eating together, living together, and participating in social gatherings, study groups, and extracurricular activities.

Consistent with Approximation 2, we define $campus\_positivity$ for a student s on day t as the number of (distinct) active infectious students, other than student s, in the week preceding day t. (Here we count the distinct number of positive students to approximate the amount of exposure due to general interactions with other students. Unlike classes with known schedules, the duration and intensity of these general interactions are impossible to observe, so a more refined quantification of exposure is hard to obtain.)

We now explain how $campus\_positivity$ is calculated.

First, based on Assumption 1 that a positive case is active and infectious for a week before being notified by the health department,

$$\#\ active\ infectious\ students\ on\ day\ t = \sum_{\tau=t}^{t+6} \#\ positives\ notified\ on\ day\ \tau.$$

Then, $campus\_positivity(s,t)$, the total number of active infectious students other than s in the week preceding day t, is obtained by setting the limits of the sum on the right-hand-side as $t-6$ and $min(104, t+6)$:

$$campus\_positivity(s,t) = \sum_{\tau=t-6}^{min(104,t+6)} \#\ positives\ notified\ on\ day\ \tau - 1\{s\ is\ notified\ on\ day\ t\}.$$

Statistical analysis

To understand whether co-enrollment with positive cases poses a risk of infection to students, we use regression to assess whether $class\_positivity$ has a significant effect on a student's infection status.

We divide students into groups based on academic career, since academic career affects the amount and types of courses a student enrolls in and is also correlated with their level of exposure outside of class. We further segment the undergraduate students into three groups: those in Greek life, those not in Greek life but are varsity athletes, and those that are neither in Greek life nor varsity athletes. We do so because it has been observed in past analyses that membership in Greek life or sports teams are associated with higher risk of infection. Descriptive statistics of the dataset are given in Table 1.

Table 1: Descriptive statistics of data used in regression.

Student group	**Description**	**Population size**	**Number of person-days in the regression**	**Number of positive person-days in the regression**	**Average $class\_positivity$**
UG_G	Undergraduates, in Greek life	2,395	226,610	1,253	0.0152
UG_A	Undergraduate athletes not in Greek life	1,098	109,765	319	0.0149
UG_other	Undergraduates neither in Greek life nor athletes	12,668	1,298,368	1,681	0.0149
GM	Graduate Management	1,716	177,586	101	0.0150
GR	Graduate Research	8,498	880,845	422	0.0149
LA	Law School	928	95,150	197	0.0148
VM	Veterinary Medicine	499	51,268	84	0.0151
Total		27,802	2,839,592	4,057	0.0149

We regress the infection status of each student s on each day t on their $student\_group$, $class\_positivity$, and $campus\_positivity$. Here, $infection\ status(s,t)$ is binary,

$student\_group(s)$ is categorical, and $class\_positivity(s,t)$ and $campus\_positivity(s,t)$ are numerical. The regression is specified as follows:

$$logit(infection\ status(s,t)) \sim 1 + student\_group(s) + class\_positivity(s,t) + campus\_positivity(s,t),$$

where $logit(p) = ln\left(\frac{p}{1-p}\right)$, and the reference category for $student\_group$ is set to UG_other, undergraduates that are neither in Greek life nor athletes.

The regression output is given in Table 2. The coefficient for $class\_positivity$ is negative (-0.2459) and not statistically significant (p-value = 0.395). Most coefficients on $student\_group$ (except for the one on VM) are significant, indicating a significant difference in infection risk from the baseline group UG_other. The coefficient for $campus\_positivity$ is positive and significant, suggesting that general interactions on campus are influential for one's infection status, and more influential than classes. We also ran the regression without the $campus\_positivity$ variable and $class\_positivity$ is also not statistically significant in that regression.

Table 2: Summary of the regression output.

Variable	***β***	**S.E.**	***p*-value**	**95% CI**
Intercept	-7.6178	0.033	< 0.001	[-7.682, -7.554]
$class\_positivity$	-0.2459	0.289	0.395	[-0.813, 0.321]
$campus\_positivity$	0.0069	9.63E-05	< 0.001	[0.007, 0.007]
Student group (ref = UG_other)				
GM	-0.8923	0.106	< 0.001	[-1.099, -0.685]
GR	-0.9734	0.054	< 0.001	[-1.079, -0.868]
LA	0.3882	0.078	< 0.001	[0.235, 0.541]
UG_A	0.8516	0.06	< 0.001	[0.734, 0.969]
UG_G	1.4344	0.038	< 0.001	[1.361, 1.508]
VM	0.0192	0.124	0.877	[-0.223, 0.262]

Interpretation of results

From the regression, we conclude that co-enrollment with a COVID-positive person does not increase the risk of testing positive oneself the following week in a statistically significant way.

It is, however, possible that co-enrollment with a positive student *does* elevate the risk of testing positive for COVID by an amount so small that it can't be detected with the amount of data available.

How large could the risk be and remain undetected? Our statistical approach allows us to quantify this: we use the lower and upper bounds of the 95% confidence interval of the

coefficient for $class\_positivity$. This approach says that being co-enrolled with a COVID positive student has an impact on the risk of testing positive that is between -0.11% and +0.04% for undergraduates who are not affiliated with Greek-life social organizations or varsity athletics. (Because of noise in our data, there is a possibility that the true risk lies outside this range — 95% confidence intervals, like the one we use here, are designed to contain the value estimated with probability 95%.)

The same approach also produces ranges of possible risks consistent with our data for other groups, listed in Table 3. Details of the calculation are given in Appendix B. Even if the true risks are at the upper bounds of these ranges, they are still low.

Table 3: Plausible range of risk of testing positive for COVID-19 over the semester associated with being in the same classes with positive students.

Student group	**Lower bound of risk**	**Upper bound of risk**
Undergraduates neither in Greek life nor athletes	-0.11%	0.04%
Undergraduates in Greek life	-0.45%	0.18%
Undergraduate athletes not in Greek life	-0.25%	0.1%
Graduate Research	-0.04%	0.02%
Law School	-0.16%	0.06%
Graduate Management	-0.04%	0.02%
Veterinary Medicine	-0.11%	0.04%

While we study the effect of being *enrolled* in the same class as a COVID-positive student, students do not attend all classes. Thus, the risk of being enrolled in the same course as a COVID-positive student is not the same as the risk of *attending* classes in that course. However, many of the students enrolled in a class *do* attend, and so our observation that the risk of co-enrollment with a COVID-positive student is near 0 suggests that the risk of *attending* a class in which a COVID-positive student is enrolled is also near 0. For example, if ⅔ of enrolled students attend class, then the risk of attending a class in which a COVID-positive student is enrolled would be roughly a factor 1/(⅔) = 1.5x higher than the risk of simply being enrolled in that class.

Appendix

A. Class positivity as a valid approximation of transmission probability

We describe in more detail why $class\_positivity$ is approximately proportional to the probability of transmission due to course enrollment.

Physical modeling of COVID-19 transmission typically adopts the *exponential dose-response model* (see Equation 4 in Buonanno et al. 2020 and references therein) to compute the transmission probability as a function of dose, where "dose" is defined as the amount of virus particles a susceptible person is exposed to. Mathematically, the transmission probability given dose D takes the following form,

$$P(transmission) = 1 - exp(-c \cdot D),$$

where c is a positive constant.

In our classroom context, under the assumption that any infectious person emits virus at a constant rate, and a fraction ρ of infectious individuals attend class, the dose that a person s is exposed to in one class c on day τ is proportional to the product

$$\rho * class\_prevalence(s, c, \tau) * (hours\ of\ class\ c\ on\ day\ \tau).$$

The transmission probability in the above equation is approximately linear in the regime we are operating in, so this product is approximately proportional to the transmission probability in a single class.

Next we argue that the linearity approximation applies to estimating the transmission probability during multiple classes. Let C_1 denote the set of classes a student takes in a week before a given day. We observe that the transmission probability during this set of classes can be approximated as the sum over transmission probabilities in individual classes (since the probabilities for individual classes are low). Mathematically,

$$P(transmission\ in\ classes\ C_1) = 1 - \prod_{c \in C_1} (1 - P(transmission\ in\ class\ c))$$

$$\approx 1 - (1 - \sum_{c \in C_1} P(transmission\ in\ class\ c)) = \sum_{c \in C_1} P(transmission\ in\ class\ c).$$

Thus, using the $class\_positivity$ expression in Equation (2) as a proxy for transmission probability due to classroom exposure is justified.

B. Calculating the risk of co-enrollment with a positive student from regression outputs

We first estimate a baseline odds of infection without any exposure due to course enrollment and then characterize how much co-enrollment with a positive person increases the odds of infection.

For an undergraduate student who is not in Greek life or varsity athletics (the reference category for $student\_group$) when $campus\_positivity = 78.99$ (the empirical mean of $campus\_positivity$) and $class\_positivity = 0$, the odds of infection is estimated to be $exp(-7.6178 + 0.0069 * 78.99) = 0.0008$. This quantity can similarly be computed for students in other groups. Among them, the expected odds is the highest for an undergraduate student in Greek life at $exp(-7.6178 + 0.0069 * 78.99 + 1.4344) = 0.0036$. The expected odds is the lowest for a graduate research student at $exp(-7.6178 + 0.0069 * 78.99 - 0.9734) = 0.0003$.

Next we estimate the effect of $class\_positivity$. Because the prevalence among students was low during the semester, the probability that one is in the same class with a positive case is low in the first place. In fact, the empirical average value of $class\_positivity$ is 0.015 across the dataset and does not differ strongly between student groups (recall from Equation (2) that $class\_positivity$ is computed by summing over all classes in a week). Thus, on average, being enrolled in the same class with a positive student changes the odds of infection by a factor of $exp(0.015 * (-0.813)) = 0.988$ to $exp(0.015 * 0.321) = 1.005$, where we take the lower and upper 95% confidence interval bounds on the coefficient of $class\_positivity$ in Table 2.

Finally, we translate this into an estimate for the risk of testing positive for COVID-19 due to co-enrollment with a positive student over the entire semester. Above, we calculated the baseline odds of infection on any given day for different student groups. Without the effect of course enrollment, the probability of infection on a day for odds r is given by $r/(1 + r)$. With the factor of 0.988 to 1.005 due classroom transmission, the probability of infection on a day now ranges from $r * 0.988/(1 + r * 0.988)$ to $r * 1.005/(1 + r * 1.005)$. The difference between these two values is the increase in the probability of testing positive due to co-enrollment with a positive student on a day. As there are 104 days in the semester, the overall increase in infection probability due to co-enrollment with positives over the entire semester is approximately[3] given by

$$[104 * r * (0.988 - 1), 104 * r * (1.005 - 1)].$$

This is calculated for students in each group in Table 3.

Therefore, given that the effect of $class\_positivity$ is found to be insignificant, and that the risk is found to be low under even the pessimistic parameter estimates, the risk of transmission in Fall 2021 classrooms seems to have been either nonexistent or very low.

[3] This approximation is derived using the series $\frac{r}{1+r} = r - r^2 + r^3$ ... and neglecting quadratic and higher order terms.