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

Cornell COVID-19 Modeling Team
Analysis completed March 2022 and published June 2023

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 recently-diagnosed COVID-positive student is associated with a small increase in a student’s risk of testing positive.

Our statistical approach estimates that an average student with no COVID-positive classmates had a 0.1404% chance (about 1 in 710) of testing positive on a given day during the semester, while an average student with the average number of COVID-positive classmates had a 0.1427% chance (about 1 in 700).

Over the entire semester, our approach estimates that a student without COVID-positive classmates had a 13.6% chance of testing positive and a student with a typical level of positivity among their classmates had a 13.8% chance. This increase of 0.2 percentage points is half of the increase predicted by physical modeling in August 2021 (this predicted increase was 0.4 percentage points).

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.

---

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. A preliminary version of this analysis was posted in February 2023. That preliminary analysis unintentionally permuted outcomes and covariates for some datapoints — the issue arose when pandas dataframe rows were dropped for dates after a student’s confirmed infection, changing the index in the dataframe. Tests and a careful review of the code conducted to finalize this work detected the issue. The current version of the report, published in June 2023, fixes this issue.

2 More precisely, the number of “COVID-positive classmates” experienced by student A in a week is the number of (student B, course) pairs for which (1) students A and B are both enrolled in the course; and (2) student B was diagnosed with COVID-19 within the last week. We estimate that having 0 co-enrolled COVID-positive classmates instead of the average number reduces student A’s chance of testing positive on a given day in the semester from 0.1427% to 0.1404%.
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.

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 see statistical evidence that being enrolled in the same class with a positive case increases one’s risk of infection, but the increase is very small. The estimated increase in infection probability over the semester due to class co-enrollment is about 2 in 1000, which is approximately half of that predicted by physical modeling in August 2021 (4 in 1000) as part of the risk analysis conducted before moving to the new mode of classroom operation. 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 August 26 - December 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 2,891,459 entries that contains, for each of 28,312 students, an entry corresponding to each day leading up to their first notification date of a positive test result (if any). 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.

We assume that a positive infection status is equivalent to being infectious. 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 Approximation 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 $\tau$ as the fraction of students in class $c$, other than $s$, that was positive on day $\tau$:

---

3 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.
\[
\text{class\_prevalence}(s,c,\tau) = \frac{\text{number of positives other than } s \text{ in class } c \text{ on day } \tau}{(\text{number of students enrolled in class } c) - 1}.
\]  

(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 \( \text{class\_positivity} \) as the sum of \( \text{class\_prevalence} \), 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 \( \tau \). Then,

\[
\text{class\_positivity}(s,t) = \sum_{\tau=\max(0,t-T)}^{t} \sum_{c \in C(s, \tau)} \text{class\_prevalence}(s,c,\tau) \times (\text{hours of class } c \text{ on day } \tau),
\]  

(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, \( \text{class\_positivity} \) 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 \( \text{campus\_positivity} \), 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 \( \text{campus\_positivity} \) 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,

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

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

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

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 participation in Greek life and athletic teams explained much of the heterogeneity in infection risk across the student body (Frazier et al. 2022, Wan et al. 2023).

Descriptive statistics of the dataset are given in Table 1.

<table>
<thead>
<tr>
<th>Student group</th>
<th>Description</th>
<th>Population size (%)</th>
<th># person-days in the regression (%)</th>
<th># positive person-days in the regression (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UG_G</td>
<td>Undergraduates in Greek life</td>
<td>2,685 (9.48%)</td>
<td>255,188 (8.83%)</td>
<td>1,350 (32.72%)</td>
</tr>
<tr>
<td>UG_A</td>
<td>Undergraduate athletes not in Greek life</td>
<td>1,124 (3.97%)</td>
<td>112,471 (3.89%)</td>
<td>321 (7.78%)</td>
</tr>
<tr>
<td>UG_other</td>
<td>Undergraduates neither in Greek life nor athletes</td>
<td>12,634 (44.62%)</td>
<td>1,295,444 (44.80%)</td>
<td>1,633 (39.58%)</td>
</tr>
<tr>
<td>GM</td>
<td>Graduate</td>
<td>1,744 (6.16%)</td>
<td>180,414 (6.24%)</td>
<td>108 (2.62%)</td>
</tr>
</tbody>
</table>
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:

$$\ln\left(\frac{p(s, t)}{1 - p(s, t)}\right) \sim 1 + \text{student\_group}(s) + \text{class\_positivity}(s, t) + \text{campus\_positivity}(s, t),$$

where $p(s, t)$ is the probability that infection status$(s, t) = 1$. 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 0.5096 with a p-value of 0.004. The coefficients on student_group are all statistically 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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-7.6584</td>
<td>0.033</td>
<td>&lt; 0.001</td>
<td>[-7.723, -7.594]</td>
</tr>
<tr>
<td>class_positivity</td>
<td>0.5096</td>
<td>0.178</td>
<td>0.004</td>
<td>[0.161, 0.858]</td>
</tr>
<tr>
<td>campus_positivity</td>
<td>0.0069</td>
<td>9.81E-05</td>
<td>&lt; 0.001</td>
<td>[0.007, 0.007]</td>
</tr>
<tr>
<td>Student group (ref = UG_other)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UG_G</td>
<td>1.4235</td>
<td>0.037</td>
<td>&lt; 0.001</td>
<td>[1.351, 1.496]</td>
</tr>
<tr>
<td>UG_A</td>
<td>0.8105</td>
<td>0.061</td>
<td>&lt; 0.001</td>
<td>[0.690, 0.931]</td>
</tr>
<tr>
<td>GM</td>
<td>-0.7277</td>
<td>0.1</td>
<td>&lt; 0.001</td>
<td>[-0.923, -0.532]</td>
</tr>
<tr>
<td>GR</td>
<td>-0.9557</td>
<td>0.055</td>
<td>&lt; 0.001</td>
<td>[-1.063, -0.849]</td>
</tr>
</tbody>
</table>

Table 1: Descriptive statistics of data used in regression.
Table 2: Summary of the regression output.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>p-value</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA</td>
<td>0.5195</td>
<td>0.075</td>
<td>&lt; 0.001</td>
<td>[0.372, 0.667]</td>
</tr>
<tr>
<td>VM</td>
<td>0.2607</td>
<td>0.112</td>
<td>0.02</td>
<td>[0.041, 0.481]</td>
</tr>
</tbody>
</table>

Interpretation of results

From the regression, we conclude that co-enrollment with a COVID-positive person increases the probability of testing positive oneself the following week, but the increase is very small and much smaller than that from participating in social activities. This estimated increase is also 50% lower than predicted by physical modeling in August 2021.

To derive the estimate for the increase in probability of testing positive due to co-enrollment, we first use the regression results to compute a population-average “baseline” probability of testing positive on a day without the effect of co-enrollment. Then we compute the same probability with the effect of co-enrollment, and observe its difference from the baseline probability. Details are given in Appendix B.

We find that co-enrollment increases the baseline infection probability by a relative increase of **1.66%**, from 1.404E-3 (a 1.404 in 1000 chance of testing positive on a given day) to 1.427E-3 (a 1.427 in 1000 chance). Over the whole semester, co-enrollment increases the probability of testing positive by **0.2 percentage points**, from a 13.6% chance to 13.8%. This is half of the increase predicted by physical modeling in August 2021 (this predicted increase was 0.4 percentage points).

Furthermore, we would like to contextualize the effect of co-enrollment by comparing it with other factors that affect student infections. Here, we compare the baseline probability of testing positive on a day for the groups UG_other and UG_G, which, to a certain degree, captures the increase in infection probability due to social activities. We find that the baseline probability for UG_other is 1.24E-3 (a 1.24 in 1000 chance of testing positive on a given day), while that for UG_G is 5.16E-3 (a 5.16 in 1000 chance of testing positive), which is **316%** larger than UG_other. This implies that social activities may increase the risk of infection by 316% / 1.66% = 190 times as much as co-enrolling in classes with positive students does.

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\textunderscore 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,

$$P(\text{transmission}) = 1 - \exp(-\alpha \cdot D),$$

where $\alpha$ is a positive constant.

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

$$\rho \cdot \text{class\_prevalence}(s, c, \tau) \cdot (\text{hours of class } c \text{ on day } \tau).$$

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,

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

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

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

B. Estimating the risk of co-enrollment with a positive student

In this section, we present details on estimating the risk of co-enrollment with a positive student based on the regression outputs.
For data entry $i$, the regression prediction is the natural logarithm of the odds ratio, $\log(\hat{OR}_i)$. The odds ratio can be converted to probability through $
frac{\hat{OR}_i}{1+\hat{OR}_i}$.

We first set `class.positivity` to 0 for all entries and compute for each one the predicted baseline probability, which we denote $\hat{p}_{i,0}$. The average across the population, $\hat{p}_0 = \frac{1}{N}\sum_{i=1}^{N} \hat{p}_{i,0}$, where $N = 2891459$, is calculated to be $1.404\times10^{-3}$.

We then include the `class.positivity` covariate and compute the predicted probability for each entry, which we denote $\hat{p}_{i,1}$. The average across the population, $\hat{p}_1 = \frac{1}{N}\sum_{i=1}^{N} \hat{p}_{i,1}$, is calculated to be $1.427\times10^{-3}$, a $1.66\%$ increase from $\hat{p}_0$.

We then extend these results to estimate the increase in infection risk over the whole semester. The probability of testing positive during the study period (104 days) without the effect of co-enrollment is approximately $1 - (1 - \hat{p}_0)^{104}$, while the same probability with the effect of co-enrollment is $1 - (1 - \hat{p}_1)^{104}$. These two quantities are computed to be 0.136 and 0.138 respectively. The increase, $0.002$, is 50\% less than the 0.004 predicted by physical modeling in August 2021.

Furthermore, we contextualize the effect of co-enrollment by comparing it with other factors that affect student infections. Here, we compare the baseline probability of testing positive on a day for the groups $UG\_other$ and $UG\_G$ separately. The difference between them can be thought of as a proxy for the effect of actively participating in social activities on infection risk.

With slight abuse of notation, we let $i \in UG\_other$ (and $i \in UG\_G$) represent that data entry $i$ is in the group $UG\_other$ (and $UG\_G$), and we let $N(UG\_other)$ and $N(UG\_G)$ denote the number of corresponding entries in the dataset respectively. We separately compute $\hat{p}_{0,UG\_other} = \frac{1}{N(UG\_other)}\sum_{i \in UG\_other} \hat{p}_{i,0}$ and $\hat{p}_{0,UG\_G} = \frac{1}{N(UG\_G)}\sum_{i \in UG\_G} \hat{p}_{i,0}$. We find that $\hat{p}_{0,UG\_other} = 0.00124$ and $\hat{p}_{0,UG\_G} = 0.00516$, $316\%$ more than $\hat{p}_{0,UG\_other}$.

In other words, while co-enrollment increases the infection probability by $1.66\%$, active engagement in social activities increases the infection probability by $316\%$, approximately 190 times the proportional increase due to co-enrollment.