School teaching posture is not related to increases in transmission of COVID-19 in Ohio

By: Daniel Nason, Wei-Yu Tseng, Julia Keating, Hongsheng (Daniel) Xie, Ziyan (Olivia) Wang
Advisor: Valerie Ventura
Department of Statistics and Data Science, Carnegie Mellon University

Abstract

In this paper, we investigate the relationship between school teaching posture and COVID-19-related outcomes in the state of Ohio during the Fall of 2020. We use data from the Johns Hopkins open source data API for COVID-19 cases and deaths as well as MCH Strategic data to include information about teaching posture across the various school districts. Due to validity issues with reporting for the time series of COVID-19 confirmed cases, we used deaths by county over time to calculate the effective reproductive rate (R_t) using the methods outlined in Cori et al. (2013). Our findings suggest that at the start of the school year, there is no relationship between school teaching posture and changes in R_t. These results could be of interest to researchers, public health officials and policymakers for guidance on K-12 teaching policies; however, it should be noted that the results may have limited applicability as the analysis was performed on data that predates pharmaceutical interventions such as vaccinations.

Introduction & Importance

The purpose of the PHIGHT COVID (Public Health Interventions aGainst Human-to-Human Transmission of COVID-19) project is to explore the relationship between K-12 teaching posture and COVID-19. This analysis can help to guide further investigation into non-pharmaceutical interventions (NPIs) for COVID-19 and be utilized as support for policymakers and public health officials regarding recommendations for public health policy. Extensive public health research has been focused on understanding the transmission of COVID-19 since the start of the pandemic, and various NPIs have been employed by policymakers and public health officials based on findings by researchers. However, despite the attention and resources allocated to the pandemic, there has been limited investigation into the relationship between school teaching posture and COVID-19-related outcomes such as transmissions and death. Schools are specifically a focus because prior research related to influenza has shown how children act as vectors for virus transmission in the community, and this phenomenon may similarly be occurring with COVID-19.

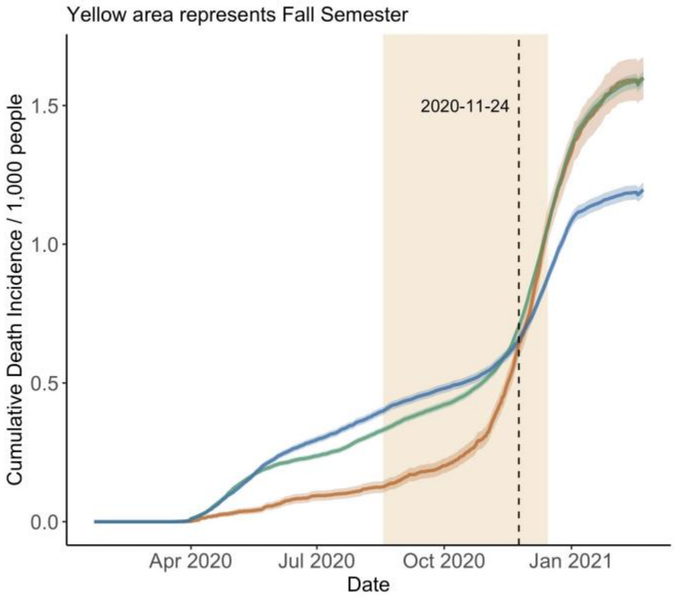


Figure 1: Cumulative Death Incidence per 1000 people

Our project hopes to fill this gap by expanding upon the work previously completed by the MSP Class of 2021 project and Lakdawala Lab that identified a relationship between school teaching posture and COVID-19 in the Fall of 2020, as shown in Figure 1. Our work contributes to this investigation by utilizing modeling techniques that best capture the time-varying effective reproductive number to measure disease progression as described in Cori et al. (2013) to compare trends by teaching posture category to try to identify a link between in-person instruction and COVID-19 death rates.

Methods

Data

We utilized the data from the work previously completed by the MSP Class of 2021 project and in Ehman et al. (2021) which focuses on the state of Ohio from January 2020 to February 2021. As described in Ehman et al., Ohio was selected because of the uniform statewide public health interventions at the start of the pandemic as well as the variation in school teaching posture between school districts in each of the state's counties. The data for COVID-19 cases and deaths by county over time utilized for the analysis was collected from the Johns Hopkins open source data API (Ehman et al. 2021). This includes information about the population of each county as well. Data about the K-12 public school districts was collected from MCH Strategic data and includes information on teaching posture (in-person, online only, or hybrid) by school district at the beginning of the semester (Ehman et al. 2021). To calculate percentage teaching posture, we take the number of students in schools with a given teaching posture relative to all the students in the district. Since there are multiple districts in each county, the district values are aggregated to get the percentage for each of the postures. The majority teaching posture at the start of the semester is then defined as the largest percentage of each of the postures.

Other data we include in the analysis consists of demographic information at the county level from MCH Strategic data and the Ohio state government website. This consists of information such as the population of the county, percentage of the population that is uninsured or over the age of 65, and the NCHS Urban Rural status as defined by the CDC (Ehman et al. 2021). The NCHS Urban Rural status consists of 6 different levels that depend on the population size of the county (Noncore, Small metro, Micropolitan, Large fringe metro, Medium metro, and Large central metro). Additionally, we include information collected by GISAID about different variants in the state during the time frame of the analysis (Ehman et al. 2021). Table 1 shows the majority teaching posture at the start of the semester stratified by the NCHS Urban Rural status:

	Hybrid	On Premises	Online Only	Total
Large central metro	2	0	1	3
Large fringe metro	16	0	1	17
Medium metro	9	2	2	13
Micropolitan	24	8	1	33
Noncore	9	6	0	15
Small metro	5	0	0	5
Total	65	16	5	86

Table 1: NCHS Status by Starting Majority Teaching Posture across counties

Processing the Cases Data

Challenges arise with using daily COVID-19 data, such as large positive and negative fluctuations in the daily reporting numbers by county. These could likely be due to corrections from over- or under-counting cases. Therefore, we processed the data in order to smooth out these spikes and get a more accurate picture of the trend.

The case time series was processed as follows:

For each county:

1. Calculate the rolling medians (window=7 days) and rolling IQR (window=15 days) of cases (we used median and IQR since they are robust to outliers)
2. Estimate standard deviation as IQR/1.35
3. Calculate the acceptance region as: [max(0,median-2.5*sd), median+2.5*sd] (using max(0,ylower) makes sure we flag all negative case numbers)
4. Flag dates with a case number outside of the acceptance region
5. Create a new column where the values are:
    a. For flagged dates, the rolling median on those dates
    b. For non-flagged dates, the number of cases on those dates
6. For each flagged date:
    a. Calculate the number of cases to spread backwards, which equals (# of cases)-median
    b. Calculate weights for each day starting 60 days before the flagged date and up to and including the flagged date using the medians (days with higher/lower median have higher/lower weight)
    c. Multiply these weights by the number of cases to spread backwards and add to the new column

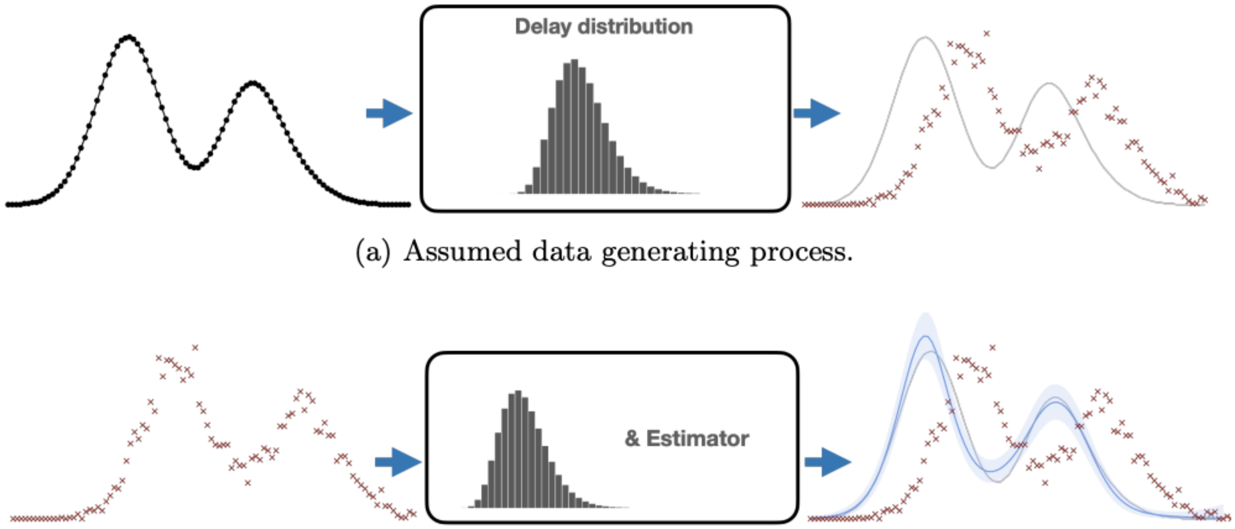
Going forward, we use the processed cases time series for our analysis and refer to this as the cases time series.

Effective Reproductive Number Estimation

In order to investigate the relationship between school posture and COVID-19-related outcomes, we used the reproductive number as a measure of virus transmission. This is the average number of secondary cases of disease caused by an infected individual while they are contagious (Cori et al. 2013). The effective reproductive number R_t as defined by Cori et al. (2013) is estimated as the ratio of the number of new infections generated at a given time t (I_t) to the total infectiousness of infected individuals at the same time. The incidence of cases at a given time on average is estimated as the product of R_t and the total infectiousness of infected individuals at the same time (Cori et al. 2013). Note that the model utilizes Bayesian statistical inference and the expression reduces to a simple analytical expression for the posterior distribution when the gamma prior distribution is assumed (Cori et al. 2013).

However, the expression we used to estimate Rt depends on the time series of infections, which is an issue because we cannot observe this series. We instead attempt to infer the infection time series from the series we can observe, such as cases or deaths. Figure 2 from Miller et al.

(2020) depicts the general approach used to infer infections from cases or deaths. Part (a) of the figure shows that the time series of infections and the delay distribution of time from infection to case or death can be used to generate the time series of cases or deaths. Alternatively, part (b) of the figure shows that a time series of cases or deaths and an estimate of the delay distribution of time from infection to case or death can be used to estimate the time series of infections. Thus, we can use an observable series, such as cases or deaths, to come up with an estimate of the time series of infections which is needed to estimate R_t.



(a) Assumed data generating process.

(b) Estimation procedure

Figure 2: Delay distribution from Miller et al. (2020)

While our data consists of COVID-19 cases and deaths starting from January 2020, there are some inherent issues with reporting during the initial months of the pandemic. For instance, spikes in COVID-19 cases and deaths occur around April in some counties, and this could be due to insufficient testing infrastructure and a backlog of cases that were all reported on the same day. Therefore, we focus our analysis from June of 2020 onward to avoid some of these data validity issues.

We employed the methods outlined in Cori et al., since this is the standard method utilized in the field of epidemiology and is also implemented in the R software programming language as part of the EpiEstim package. Using the model parameters from Ehman et al., we specify the gamma distribution with a mean of 23.9 and coefficient of variation of 0.4 to estimate the probability that someone will die at a specific time *t* given that they are infected at an early time *s*. Specifically, we conducted the following analysis:

1. Calculate R_t with estimation 14 days time windows for Top 10 largest counties in Ohio with the analysis of time series of death
2. Calculate R_t with estimation 14 days time windows for Top 10 largest counties in Ohio with from the analysis of time series of case
3. Compare R_t for different teaching posture in Ohio
4. Compare estimated R_t with scaled time series of death

Determining the Validity of Cases Data

The cases and deaths time series provide information about COVID-19 transmission by county. While the main benefit for using the death time series to estimate R_t is the validity of the data, there are two advantages in using the cases time series: 1) cases happen sooner after infection than deaths, so this reduces the error when measuring R_t; 2) there are more observations for cases, so the estimate of variance is smaller compared to the deaths time series. However, demographics and resources available (i.e. testing capacity) to track COVID-19 varied across counties, so these observations require further validity checks before they can be used to estimate R_t. Different counties in Ohio had varied demographics and resources available to track the spread of COVID-19 cases throughout their communities. Therefore, since as shown in Figure 10 (see appendix) the dominant variant remains constant in Ohio, to detect issues in the validity of the cases time series we examine the ratio of deaths to cases (death rate).

Because of the reliability of the deaths time series and no changes in the dominant variant, we would expect that the death rate should be relatively constant over time across counties in order to use it to calculate R_t. The deaths time series is shifted 14 days in order to better align with the case time series due to the delay between infection and positive test case and infection and death (Bonvini et al. 2021). A Gaussian filter was also applied to both time series before the ratio was calculated in order to smooth some of the noise inherent in the daily reporting of deaths and cases and to provide insight about the relationship between these variables.

Starting School Teaching Posture Comparisons

In order to investigate the impact of school teaching posture on COVID-19 spread in a county, we focus on the start of the semester since this is the data we have available. Since each school had a varied start date for the fall semester ranging from August to December, we chose the median of these start dates (9/15/2020) as the starting point and compared how R_t changed in the communities with varied teaching postures. We define Period 1 as the date range from 9/15/2020 to 10/5/2020 and Period 2 as the date range from 10/6/2020 to 10/26/2020. For each county and Period, we linearly regress R_t on the time period and extract the slope to use as an estimate for the rate of change in R_t in these two periods of time. Comparing these rates while accounting for percentage of teaching posture and the majority teaching posture in the county will allow us to determine if there is any impact of teaching posture on the changes in R_t at the start of the school year.

We also narrow the scope of the analysis to just Micropolitan counties, which the NCHS defines as urban areas with populations between 10,000 and 50,000. This additional blocking has the benefit of making more valid comparisons across counties with different teaching postures at the start of the semester since it better controls for potential differences across county demographics. It should be noted that since only one Micropolitan county has a majority teaching posture during this time, the comparisons are instead made by looking at the percentage of a specific teaching posture for each county compared to the change in R_t.

Results

Processing the Cases Data

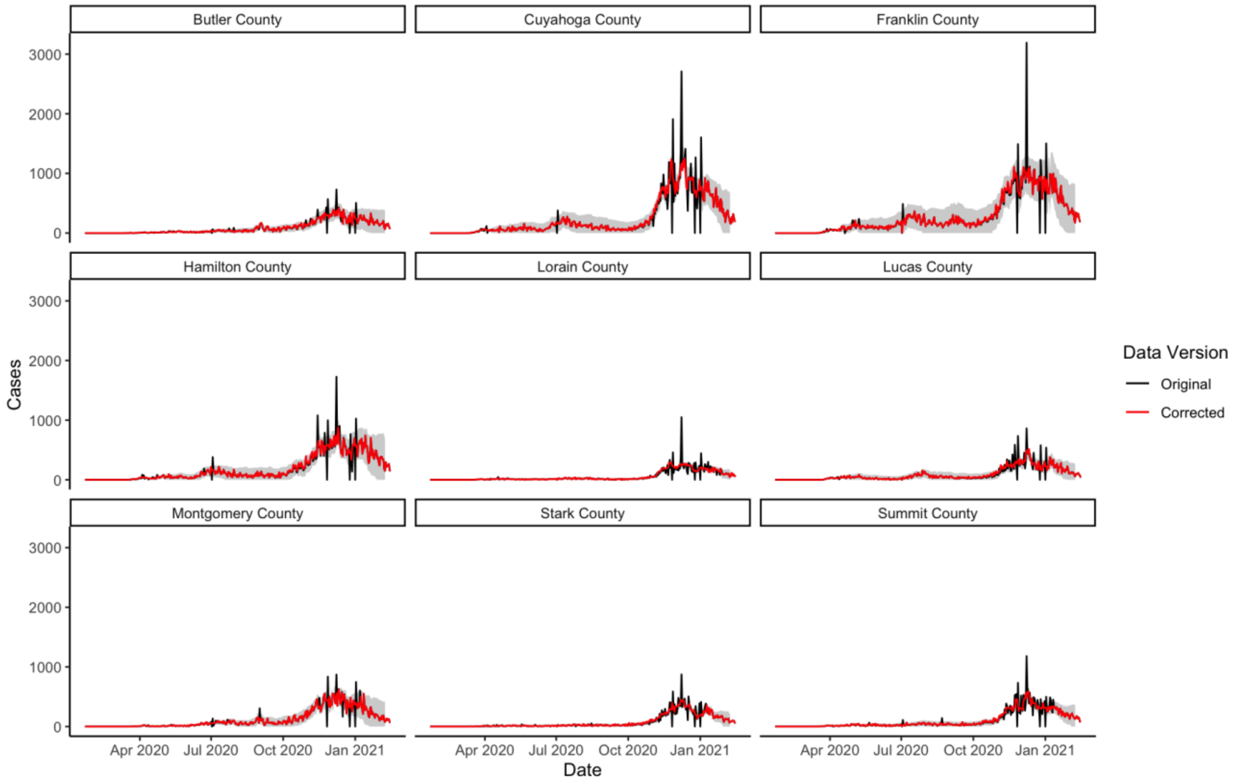


Figure 3: Corrected Cases Data by counties

For the nine largest counties (Butler, Cuyahoga, Franklin, Hamilton, Lorain, Lucas, Montgomery, Stark and Summit), the plots display the interval that was used to identify points that required smoothing as well as the corrected data after the smoothing method was applied.

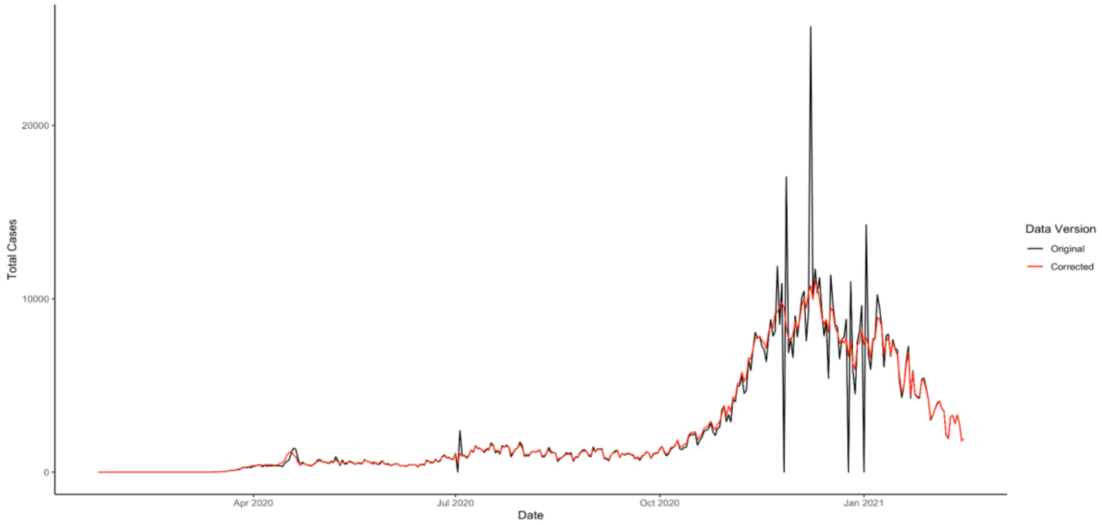


Figure 4: Corrected Cases Data over the entire Ohio State

Aggregating all the data over time, we see that the correction smooths out the spikes well overall. While there are some instances of negative case counts, these are all between -1 and 0 and therefore rounded up to 0. Any decimal values that resulted were rounded before estimating R_t. Some unusual behavior occurs in the earlier data for smaller counties (i.e. large spikes), but since the focus of the analysis is on later dates and aggregated by teaching posture this is less of a concern.

Determining the Validity of Cases Data

The two most populous counties in Ohio, Cuyahoga and Franklin are compared to check whether the case time series during the semester is usable for the analysis.

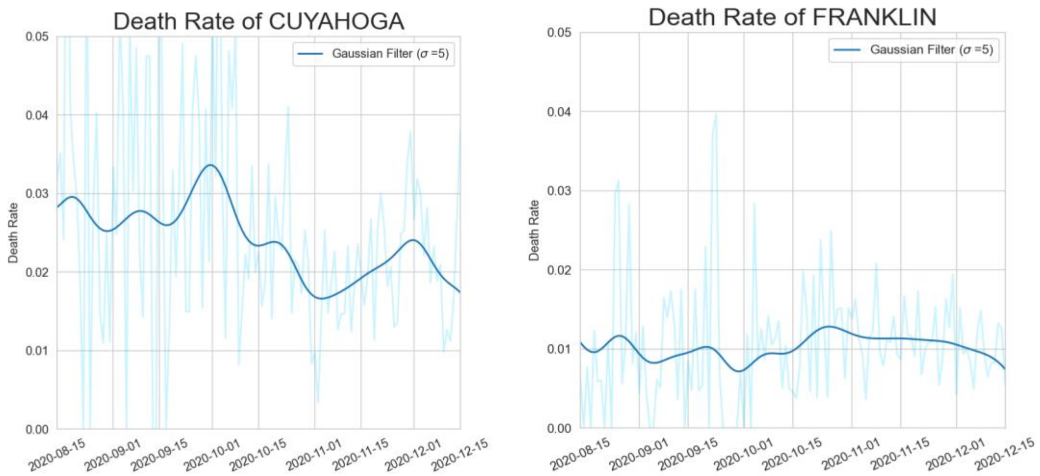


Figure 5: Smoothed death rate comparison for the two largest counties in Ohio

The death rate in Cuyahoga wasn't constant during the 2020 Fall, fluctuating between 1.5% and 3.5%. Although the death rate in Franklin was constant during the time, the 1% death rate was noticeably different from values we observed in Cuyahoga. This suggests that Cases time series may not be correctly reported even after processing the data. An additional check on the validity of the processed data is comparing the estimated R_t over time for the cases and deaths time series. As shown in Figure 11 (see appendix), the estimated R_t plots do not overlap during the semester, and there is a drift between the two series that could be the result of delayed reporting. The peak around Thanksgiving is also smaller in cases compared to deaths, which could be the consequence of underreporting. Because the validity checks show that there are discrepancies between reportings for cases and deaths, we therefore utilize the deaths time series to perform our analysis.

Starting School Teaching Posture Comparisons

Figure 6 plots the estimated R_t calculated after aggregating deaths across the counties by majority teaching posture. The gray region highlighted identifies the fall semester of the school year, and Periods 1 and 2 are labeled on the plots.

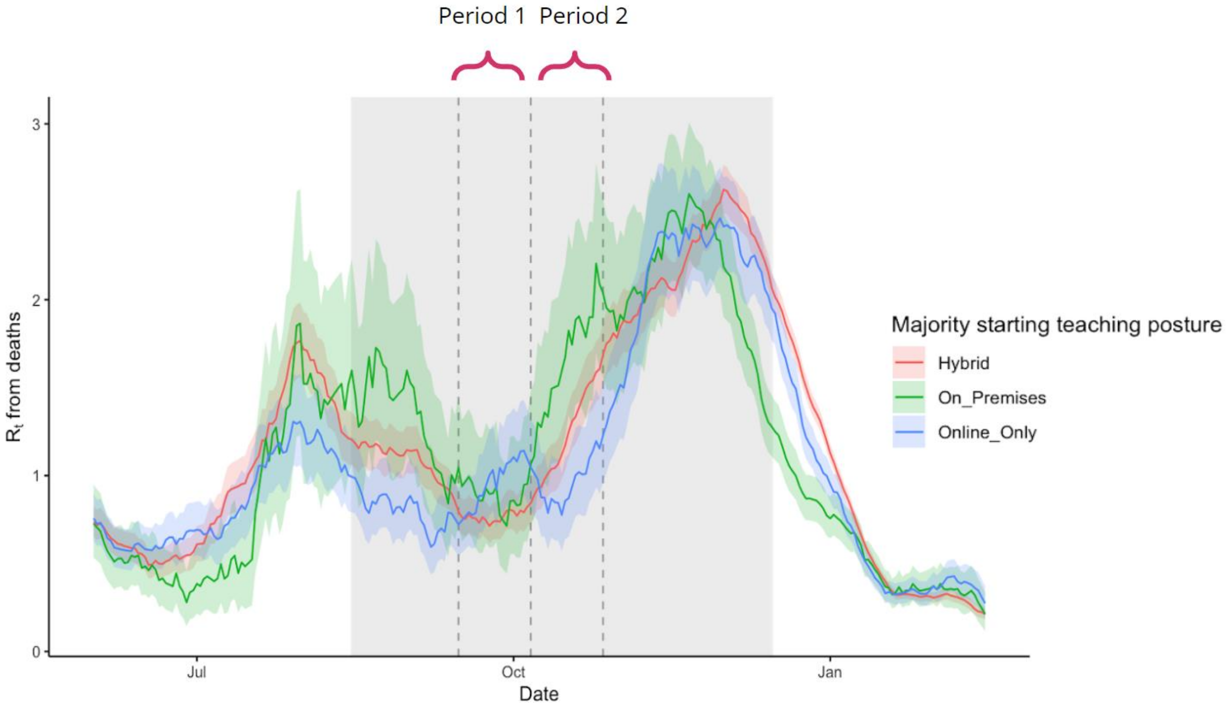


Figure 6: Calculated R_t by teaching postures

When aggregated by teaching posture, the estimated R_t appears to be higher during period 2 in the counties with an on-premises majority instruction. However, the large estimates of variation as indicated by the shaded regions around the line indicates that there may not be real difference across teaching posture. This is due to the count of non-zero observations in the death time series each day.

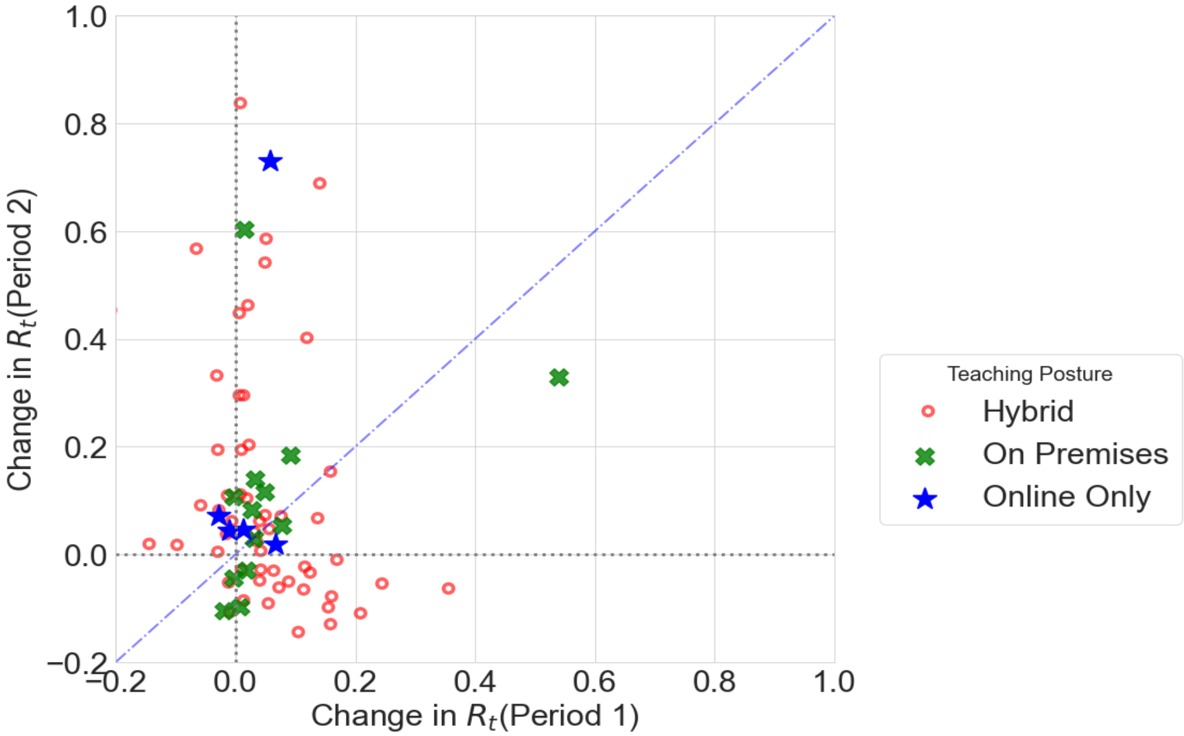


Figure 7: Change of R_t over Periods 1 and 2

The points in Figure 7 represent the change in R_t for a county in Period 1 relative to Period 2, and the diagonal line provides a visual representation of constant change between the two periods. Points to the right of the diagonal represent counties with a larger change in R_t in Period 1 compared to Period 2, while the reverse is true for points to the left of the diagonal. Most of the points are clustered around the origin, and there are no obvious trends for schools with a majority teaching posture of on premises versus online only; that is, the same proportion of counties with these majority teaching postures at the start of the semester lie to the left and right of the diagonal. This suggests that teaching posture is not related to more changes in R_t in a county after the semester started.

Figures 12 and 13 (see appendix) account for additional potential confounders at the county level such as population density and mobility. However, the side-by-side scatterplots show that, comparing across the majority teaching postures, the change in R_t is roughly similar in Period 1 compared to Period 2. The lines drawn on the plots are least squares estimates for the given majority teaching posture and serve as a visual aide to show that the change is roughly constant for R_t even as we account for population density and mobility. Interestingly, the least squares line is downward sloping for on premises counties, though this is likely driven by the influential point that represents Fulton county, which has a change in R_t in Period 2 of -1.61.However, in general we find that the majority of the points are clustered around 0, indicating that there is no change in R_t between the two periods and that these changes are also unrelated to the majority teaching posture.

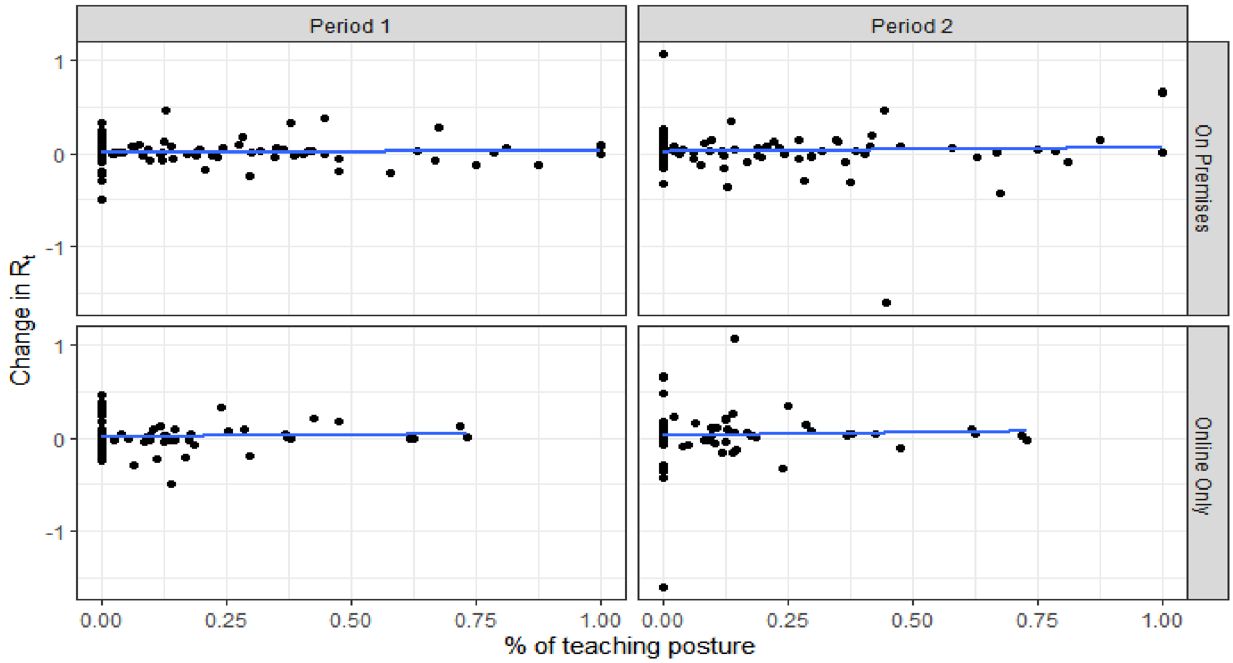


Figure 8: Change of R_t across periods as a function of teaching posture

The issue with the majority teaching posture approach is the limitation in the number of observations across each of the posture types, as indicated in Table 1; that is, there are only 5 counties with an online only majority teaching posture and 16 with an on premises majority teaching posture. Figure 8 accounts for this by showing the change in R_t for a given period as a function of the county's percentage of a specific teaching posture. For example, the upper left plot shows the variation in the change in R_t as percentage of students in the on premises teaching posture increases during Period 1, while the upper right plot displays the same except for Period 2. However, in both cases we see that as the percentage of a given teaching posture increases, the change in R_t appears to be relatively constant and centered around 0. Interestingly, there is slightly more variation in Period 2 compared to Period 1 for both teaching postures, but this does not impact the general trend of the change in R_t.

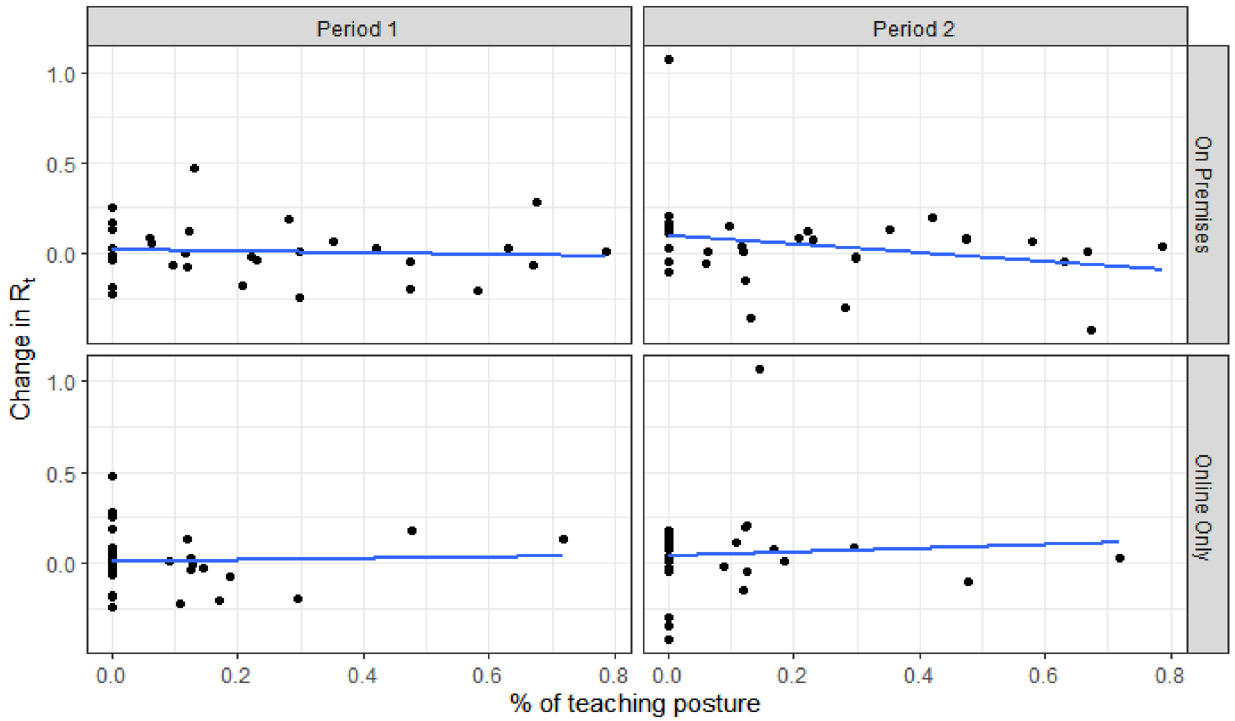


Figure 9: Change of R_t across periods as a function of teaching posture for micropolitan counties

Figure 9 displays similar plots as Figure 8 except only for Micropolitan counties. However, even after blocking on these counties to better control for potential demographic differences across the state, we see a similar result. The percentage of a specific teaching posture does not appear to be related with a change in R_t and the changes are very similar in both Periods.

Discussion

After investigating the death rate, we discovered that there could be some reporting issues in cases, so we chose to use deaths in our analysis. However, the estimated errors using deaths are larger due to the small sample size, and therefore resulting in less precise estimates of R_t. The analysis of the change in R_t at the beginning of the fall semester suggests that there does not appear to be a relationship between teaching posture and COVID-19 disease spread in a county. This lack of a relationship holds even when we account for potential confounders including population density, average mobility, and make comparisons across counties with similar demographics i.e. Micropolitan counties.

However, the analysis is not without limitations that need to be addressed. For instance, the definition of on-premises teaching posture does not account for potential mitigation measures taken by the schools to limit the spread of COVID-19. This includes but is not limited to mask requirements, reduced extracurricular activities, improved ventilation systems, and other social distancing measures. These actions could have sufficiently reduced the risk for students with in-person instruction. Additionally, we assumed a fixed delay distribution when calculating R_t for the deaths and cases time series. However, this may not be appropriate as Jahja et al. (2022)

find that the delay distribution from the onset of symptoms to the reporting changes over time, as seen in Figure 14 (see appendix). This could imply that the estimates of R_t are invalid due to the inappropriately specified distribution.

Additionally, our uncertainty regarding the delay distribution of cases may be a large source of error in our analysis. Documentation of the EpiEstim package used to estimate R_t suggests that we may have been using the incorrect distribution all together, and should use the distribution of time from primary incidence event to secondary incidence event as opposed to the distribution of time from infection to incidence event. Further research on the EpiEstim package is needed to determine the correct distribution to input into the function within the EpiEstim package used to estimate R_t. Since we are also uncertain about the distribution parameters, we might also conduct a sensitivity analysis to determine the effect of changing the distribution parameters on our estimates of R_t.

There are multiple reasonable extensions that could be considered to improve the analysis. A more in-depth index for counties with on premises or hybrid teaching postures to account for potential mitigation measures implemented by the schools to reduce the risk of disease transmission. Differences across these measures could yield insights about the lack of change in R_t for these counties. Additionally, adjusting the parameters of the delay distribution according to the findings from Jahja et al. (2022) could increase the validity of the cases time series for estimating R_t. The main benefit of this would be the increase in sample size and reduced variance estimates when calculating R_t.

More data about how the teaching posture could also benefit the analysis. We make the simplifying assumption that the starting teaching posture by school district is constant for the 6 weeks; however, school districts may have altered their teaching posture after a spike in COVID-19 cases and/or deaths in the community, which could have reduced transmission of the virus. Finally, a multiple linear regression analysis could be performed to quantify the effect of teaching posture on R_t while controlling for relevant confounders such as mobility or population density. This analysis could also be extended to other states with a similarly non-uniform teaching posture status across counties to validate the conclusions found in Ohio.

Interestingly, our findings suggest that since there is no difference in the estimated R_t and teaching posture, on-campus instruction is safe for children. It should be noted that since this data was collected, there have been numerous advances in pharmaceutical interventions to combat COVID-19 infections such as vaccinations; therefore, these findings may be out of date due to advances in medical technology. While further research is required to verify the results of the analysis, such findings build upon Ehman et al. (2021) and provide guidance for policymakers and public health officials in deciding school teaching postures during the pandemic.

References

Anne Cori, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez. (2013) *New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics.*

Ehman, C., Luo, Y., Yang, Z., Zhu, Z., Donovan, S., Avery, A. J., ... & Lakdawala, S. S. (2021). *K-12 School Teaching Posture Correlates with COVID-19 Disease Outcomes in Ohio*.

Katelyn M. Gostic, Lauren McGough, Edward B. Baskerville, Sam Abbott, ... & Sarah Cobey (2020) *Practical considerations for measuring the effective reproductive number, R_t*

Maria Jahja, Andrew Chin, and Ryan J. Tibshirani. (2022). *Real-Time Estimation of COVID-19 Infections: Deconvolution and Sensor Fusion.*

Matteo Bonvini, Edward H. Kennedy, Valerie Ventura, and Larry Wasserman. (2021) *Causal Inference in the time of COVID-19.*

Miller, A. C., Hannah, L., Futoma, J., Foti, N. J., Fox, E. B., D’Amour, A., Sandler, M., Saurous, R. A., & Lewnard, J. A. (2020). Statistical deconvolution for inference of Infection Time Series.

Unwin, H. J. T., Mishra, S., Bradley, V. C., Gandy, A., Mellan, T. A., Coupland, H., Ish-Horowicz, J., Vollmer, M. A., Whittaker, C., Filippi, S. L. et al. (2020). *State-level tracking of COVID-19 in the United States.*

Technical Appendix

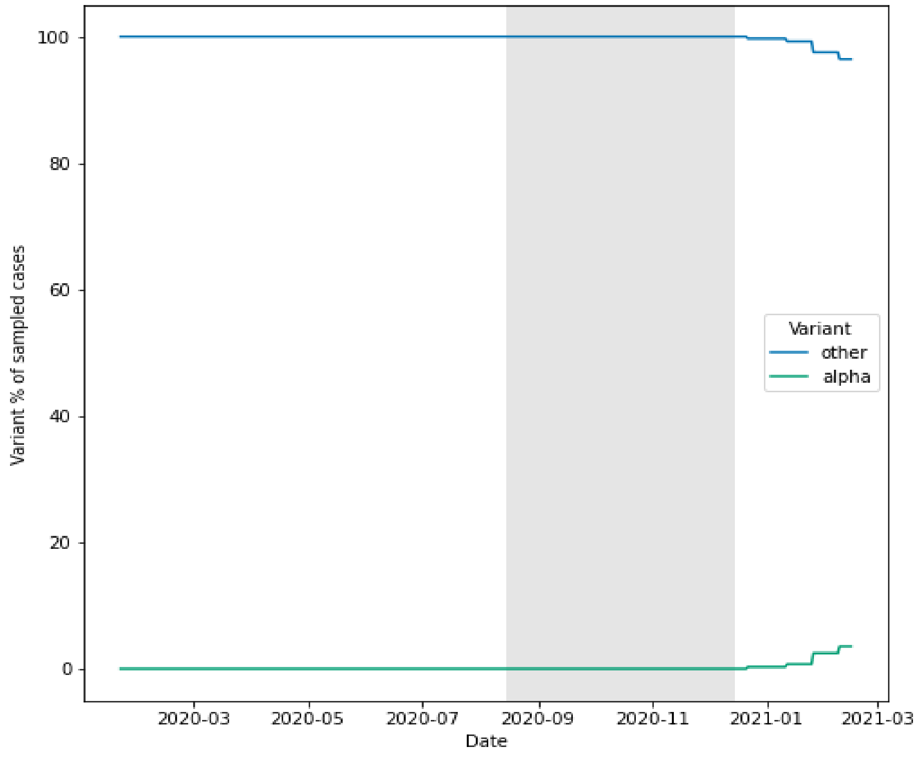


Figure 10: Percentage of Alpha variant presented in Ohio

The alpha variant didn't show up until the Spring of 2021 in Ohio State, so that prior to the time, the virus that dominated the United States was the original COVID-19, and we could assume that the death rate should remain within a certain level .

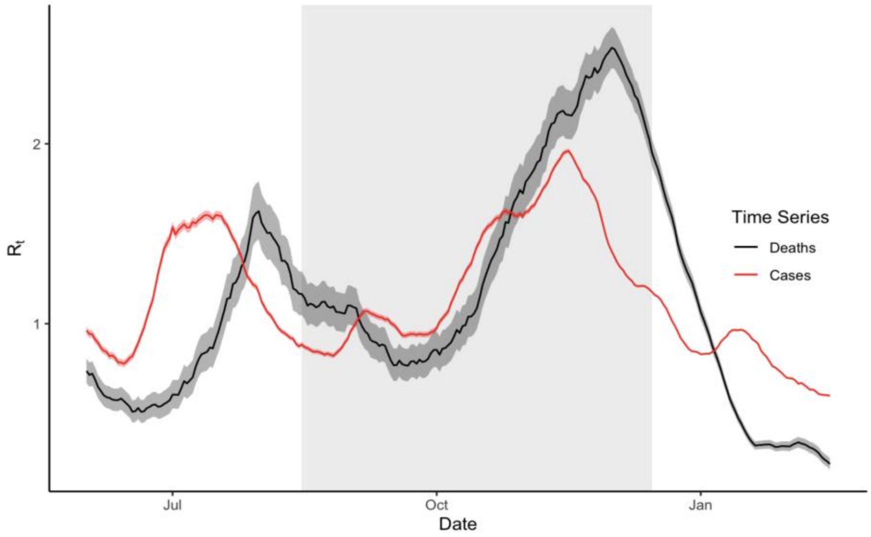


Figure 11: Comparison Cases calculated R_t and Deaths calculated R_t

The graph shows a drift between two series, which could be possibly caused by delayed reporting. The second peak of the cases is smaller, which is likely the result of underreporting.

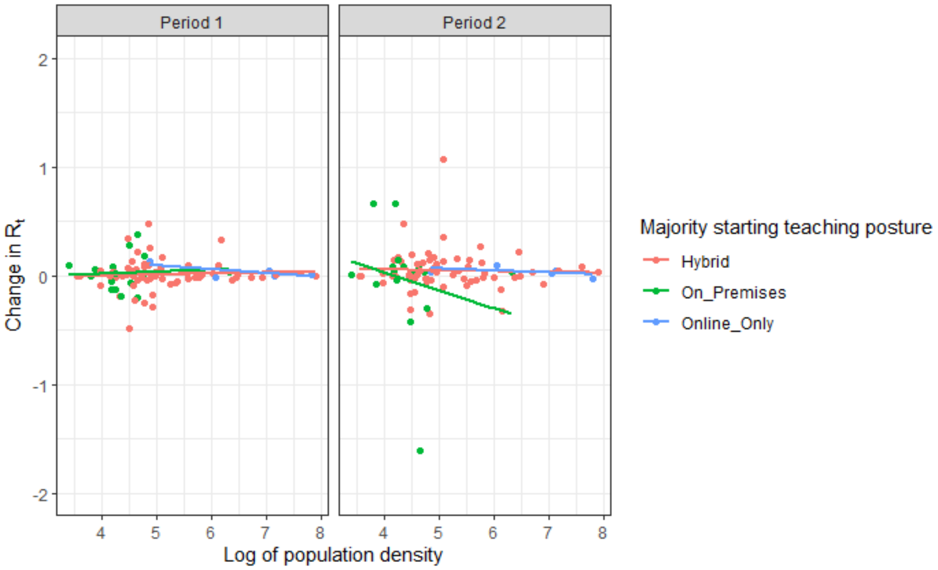


Figure 12: Change of R_t across periods as a function of log population density

This graphic illustrates that as population density varies, there is no change in R_t from period 1 to period 2 across each of the majority teaching postures at the start of the school year.

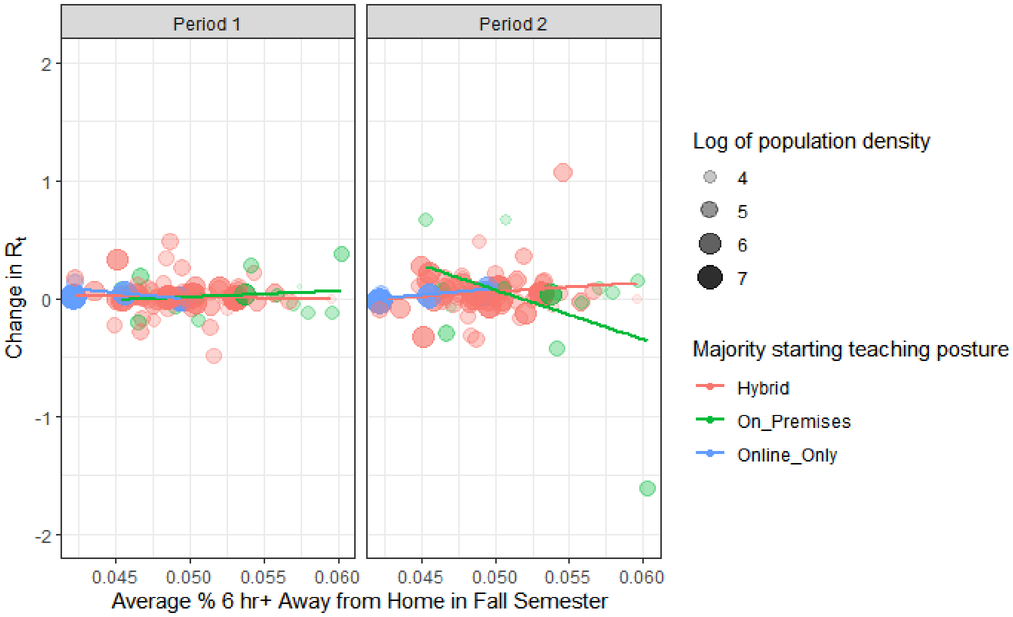


Figure 13: Change of R_t across periods as a function of mobility

Figure 13 shows that there is no change in R_t from period 1 to period 2 across each of the majority teaching postures at the start of the school year even for counties with higher mobility.

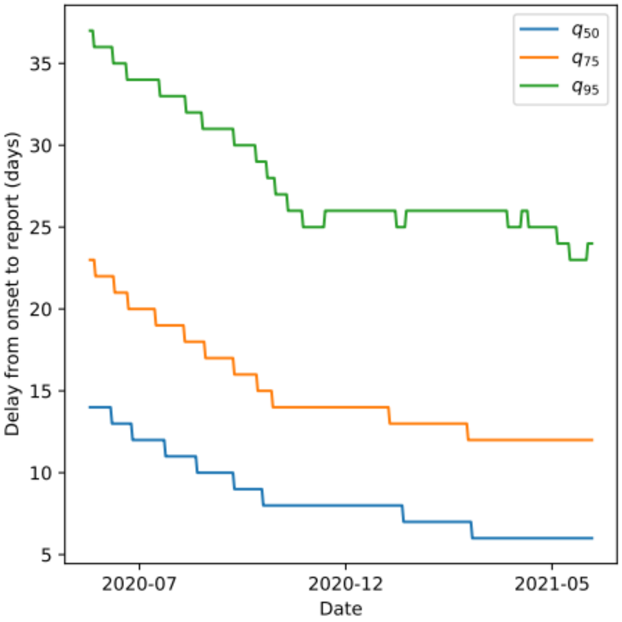


Figure 14: Delay from onset to report changes over time

Taken from Figure 3 of Jahja et al. (2022), the delay from onset to report decreases over time, this could be the consequence of increase in testing capacity.