Assessing the Impact of COVID-19 on Albemarle County Student Attendance

Rory Black
School of Data Science
University of Virginia
Charlottesville, United States
qmn9tb@virginia.edu

David Siamon
School of Data Science
University of Virginia
Charlottesville, United States
dws3qd@virginia.edu

***Abstract*—General student physical and mental well-being has always been a major concern though it has significantly increased since the beginning of the SARS-COVID-19 (COVID-19) pandemic. Evidence of student well-being may be reflected in increased student absences from school and poorer standardized test performance. The present project was undertaken to identify patterns in middle and high school student attendance in the Albemarle County Public School (ACPS) system to make inferences about student well-being. Specifically, it seeks to reveal differences in trends in the years before, during, and after the pandemic as it relates to educational engagement and assessment outcome levels. Variables included daily period-by-period information on a student's presence in class status, demographic information accompanying each student (e.g., grade level, male/female, racial/ethnic identity, etc.), and the Virginia Department of Education Standards of Learning (SOL) limited testing results. To confirm suspected differences in various student subpopulations, two-proportion z-tests were utilized to confirm patterns among subpopulations. Beginning with exploratory data analysis, and continuing with survival analysis and logistic regression, we identified specific subpopulations of students more at risk of unexcused absenteeism. The number of absences increases for all students from the pre-COVID era to the remote school year and then recovers in the 2021-2022 school year yet not quite reaching the levels seen in the 2018-2019 school year. Results from logistic regression modeling were predictive of student absenteeism with many variables returning a p-value less than 0.001, which may serve as a tool for school administration to identify at-risk students in near real-time. Moreover, an increased association between the proportion of students failing SOL tests, used to assess student success, and the number of students recording unexcused absences in the SOL subject in which the tests are taken. Though precise causal relationships concerning the influence of COVID-19 were unclear, from the perspective of student well-being, results from these analyses highlight where ACPS leadership might put their focus toward maintaining student engagement and academic success in the post-pandemic period.**

***Index Terms*—Student well-being; absenteeism; education; COVID-19**

I. Introduction and Background

The COVID-19 pandemic has resulted in significant disruptions globally and education has not been left out [1]. In the Commonwealth of Virginia's Albemarle County Public School (ACPS) system, the way students learned changed dramatically as the remote learning system was adopted. This article seeks to track the change in ACPS student attendance before and after the COVID-19 pandemic to identify the impact of the pandemic on student behavior. The relevant work will be synthesized with similar projects tracking student grades, disciplinary infractions, and UVA Hospital data to glean insight into how student mental health has been affected by the pandemic.

As this serves as a beginning to a larger project, research questions have been shaped around gaining familiarity with student attendance behavior and how said behavior differs between demographic groups. A primary goal is to identify pockets of students that have proportionally worse attendance after the pandemic than their peers, potentially denoting alarming trends to examine more closely in future work.

II. Related Works

Many studies have been devoted to student mental health since the beginning of the COVID-19 pandemic [2]. Although there has yet to exist a study into the specific student behavior occurring among Albemarle County Public School students, it may be assumed that there exists some overlap and value in the studies of other students around the world [3].

In their attempt to identify why student attendance plummeted during the COVID-19 pandemic, a team from Australia answers questions about what factors are driving this pattern and the relevancy of attendance in student performance [4]. Specifically, the answer to the question about the relevancy of attendance to student performance emphasizes the importance of uncovering the cause for this drop in attendance among ACPS students.

In previous attempts to use Survival Analysis to evaluate student drop-out rates, information on students that extended beyond just their recorded number of absences was used to predict the time of drop-out [5].

III. Data Description

Data concerning school attendance was collected by ACPS over four different school years: *2018-2019*, *2019-2020*, *2020-2021*, and *2021-2022*. There exists both demographic and attendance data for all middle and high school students that were

Thank you to our sponsors, Tara Hofkens (UVA School of Education and Human Development) and Brian Wright (UVA School of Data Science), as well as our faculty advisor, John D. Van Horn.

enrolled during any of the four listed school years. Although the data is presented in a de-identified format, there are still plenty of important pieces of information that remain. First, the demographic data contains information on each student's grade level, ethnicity, special education status (SPED), English language program enrollment (EL), the number of Standards of Learning (SOL) tests taken, and the number of SOL tests failed in one single school year.

In addition to demographic data for each student, there is class-by-class attendance data. Each absence is linked to a student ID and contains information on the term, academic department, specific course, and period in which this absence occurred in addition to the type of absence: *excused*, *unexcused*, or *tardy*.

Variable	**Description**
Grade Level	Integer value: Between 6-12
Ethnicity	Hispanic/Latino, White, Mixed Race, Black, Asian, American Indian, Native Hawaiian, Other
SPED	Boolean value: Yes, No
EL	Boolean value: Yes, No
SOLs Taken	Integer value
SOLs Failed	Integer value, "–" if no SOLs were taken

For the purposes of analysis, the 2019-2020 school year was excluded. Data collected during that year is not consistent nor does it exist for almost half of the school year. In the figure below, we see the distribution of unexcused absences for the three school years included in the analysis. This pattern alone sets the foundation for all suspicion that the COVID-19 pandemic disrupted academics and school attendance. This large increase in unexcused absences across all high schoolers and middle schoolers for the 2020-2021, remote instruction, school year confirms that suspicion that there is a problem among students.

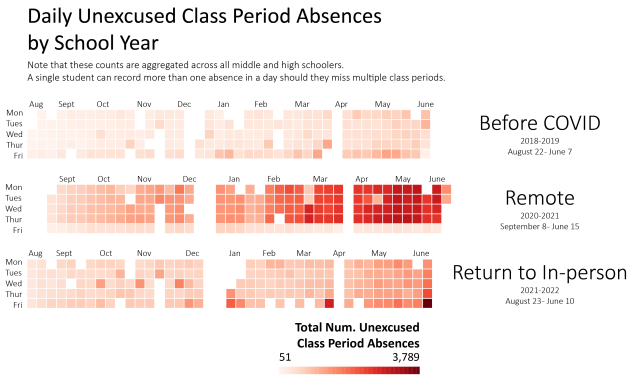

Fig. 1. Students noticeably recorded the most unexcused absences during the remote schooling year.

IV. Methodology

A. Exploratory Data Analysis

A core piece of understanding student attendance data has been exploring general trends to contextualize more sophisticated modeling techniques. This required a range of analyses, including comparisons between middle and high schoolers, grade-by-grade comparisons, analyses of EL versus non-EL students, SPED versus non-SPED student comparisons, and comparisons between different ethnic groups (white and non-white). All of these comparisons were made by looking at trends over the three school years seen in **Fig. 1**. The study also investigated interaction effects between the different demographic groups to identify potential correlations.

In addition to these analyses, aggregated Standards of Learning (SOL) [6] failure rates across demographic groups were assessed to identify any correlations with absence rates. This analysis had limitations due to unreliable sample sizes explained further in **Section V-A**. Overall, this study aimed to gain a comprehensive understanding of the various factors that may impact student performance and attendance rates, and identify potential areas for improvement in educational outcomes for different groups of students.

B. Logistic Regression

To examine what factors are important when identifying sub-populations of students that display some concern in terms of attendance, a logistic regression model was built. With the ability to evaluate a student's probability of being an "at-risk" student at the beginning of the school year, the hope is that school administration can intervene earlier before absences become an issue for the student's overall academic performance and general well-being. This model was built with the full disclaimer that the probability of being "at-risk" is based solely on a student's demographic data. Although it is difficult to make definitive conclusions solely from the data provided, this model can be used as a supplemental tool in assisting teachers and administrators to proactively identify problems to help slow the growth of increased student disengagement with school [7].

To begin, a binary response variable needed to be produced from the number of absences. Since the aim of this model was to flag students displaying concerning behavior, absences were filtered down to only include unexcused absences. Each student's absences were tallied up to give each student a count for each of their courses for the school year. The binary response was built to answer the question, "Did a student record at least 8 absences in any one course in a school year?". This cutoff of 8 absences derives from ACPS policy stating that missing 8 class periods will result in a discussion with administration. Although this policy may vary across individual schools, the choice to observe 8 absences was made to support model simplicity. The school year was transformed from a categorical to a discrete numerical variable by calculating the number of years since the start of the COVID-19 pandemic. This was done to allow the model to

be applied to future school years without the addition of a new school year being recognized as an addition to a level of categories within the variable.

To further distance ourselves from the ethical concerns we had regarding determining a student's attendance risk using race as a factor, we chose to create a final logistic regression model that excluded race. Even after reducing race to the binary of "white" or "non-white", we still saw a lot of interaction between race and other factors. This only supported our decision to remove race from the model. The *SPED* predictor was also removed because its p-value of 0.274 indicated that it was insignificant. The equation below is drawn from the final logistic regression model and displays how a student's probability of being at risk for failing a course due to absences is calculated.

$$ln(\frac{\hat{p}}{(1-\hat{p})}) = b_0 + \sum_{i=1}^{11} b_i X_i \tag{1}$$

In equation 1, each b value represents 1 of 10 coefficients included in the model. Corresponding coefficient values can be seen in the table below.

Variable	Estimate	Std. Error	p-value
Intercept	-1.437	0.064	<0.001
Years Since COVID	0.318	0.017	<0.001
Grade Level 7	0.265	0.079	0.001
Grade Level 8	0.724	0.081	<0.001
Grade Level 9	0.847	0.080	<0.001
Grade Level 10	0.181	0.076	<0.001
Grade Level 11	0.571	0.075	<0.001
Grade Level 12	0.322	0.077	<0.001
EL – Yes	0.862	0.060	<0.001
SOL Tests Taken	-0.753	0.025	<0.001
SOL Tests Failed	0.753	0.029	<0.001

The coefficient estimates for each variable indicate whether we can expect this predictor to increase the log likelihood of a student being at risk of reaching the attendance threshold or decrease it. The only variable in which we see a negative coefficient is *SOL Tests Taken* because as the number of tests a student takes in a year increases, we expect the probability of the student being a risk to decrease. All predictors except for *SPED* reveal very small p-values indicating that those predictors are significant in determining the probability of a student reaching a concerning number of unexcused absences.

C. Survival Analysis

1) Kaplan-Meier Method:

After this initial attempt to determine which factors tend to be contributing to student absence counts and should raise possible flags of concern, the interest shifted more to how the time of year was affecting student attendance. Although commonly used in the field of health to model how disease affects a population, survival analysis served as a way to model the point of the year students tend to be reaching a concerning number of absences in any one course.

The date of the 8^{th} unexcused absence in any one course in the student's record was used to mark that a student has "failed" or "died" out of the population of students that remain in good standing. The difference in this metaphorical disease of chronic absenteeism is that students cannot "recover" from the disease within a school year until the last day of school. A student is labeled with the status "censored", meaning they remain in good standing in regards to attendance, if that student never reached the threshold of 8 unexcused absences in any one course.

We used the Kaplan-Meier method, a popular non-parametric method, to model the pattern of student attendance drop off throughout any one school year. Using Python's *lifelines* package, we were able to explore both survival analysis mentioned within this paper. The method utilizes the survival probability function displayed below [8].

$$S_{KM}(t) = \prod_{t_j<t} (1 - \frac{d_i}{n_i}) \tag{2}$$

In this equation, $S_{KM}(t)$ represents the probability of survival given the time. The probability is calculated by taking the product of 1 minus the result of the number of attendance "failures" at the time over the size of the population still at risk at the time.

2) Cox Proportional-Hazards Model:

Another way of approaching survival analysis is using the Cox Proportional-Hazards model. This method differs from the Kaplan-Meier estimator because it is a semi-parametric method that makes assumptions on density and utilizes covariates to determine hazard [8]. The hazard function below is very dependent on covariate weights.

Much like the logistic regression model, we chose to remove ethnicity as a predictor because the sample sizes across groups were not evenly distributed, and ethnicity did not appear as a significant covariate when constructing the model.

$$h_0(t) = \sum_t \frac{c}{\sum_n e^{\hat{\beta} x_i}} \tag{3}$$

In the equation above, we are able to calculate the hazard at any given time. The c in the numerator represents the attendance "failure" status of each student where 1 indicates the student reached the 8-absence count and 0 indicates the student did not. The $\hat{\beta}$ term in the denominator is found from the MLE of each β, or covariate, weight.

V. Results

A. Exploratory Data Analysis

Before the pandemic, there was a strong linear trend between grade level and absences, with higher grades having more absences ($r = 0.992$, $p < 0.001$). This trend was not observed during the remote year, and a weaker albeit definite positive trend ($r = 0.87$, $p < 0.001$) was noted during the post-COVID school year as can be seen in **Fig. 2**.

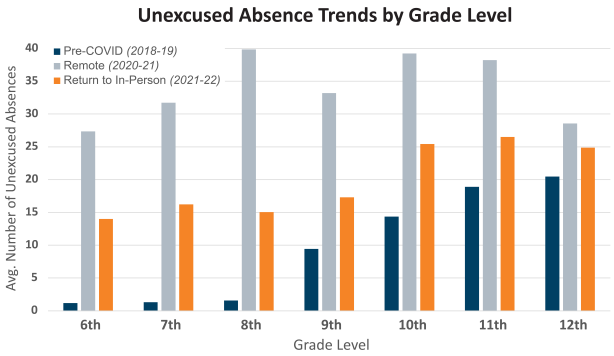

Fig. 2. The relationship between grade level and unexcused absences varied noticeably across the three school years shown.

The increases and decreases observed in the absence patterns had a trend that varied by grade, with larger gaps between school years for lower grades.

Furthermore, non-white, English Language Learners (EL), and Special Education (SPED) students missed significantly more classes each year than their control groups.

To better visualize the differences in absences between non-white and white students, **Fig. 3** shows the proportion of students absent by race/ethnicity for each academic year.

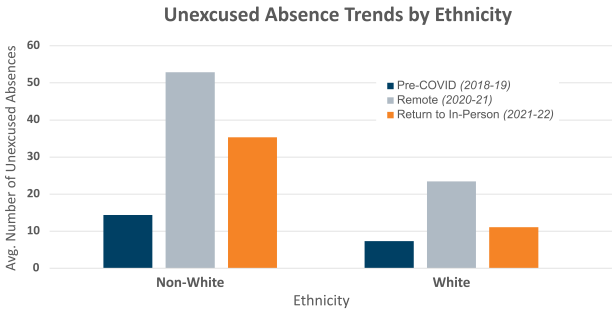

Fig. 3. Non-white students recorded more unexcused absences across all three school years.

One key finding is that non-white students had a significantly higher proportion of absences post-COVID than before, with non-white students missing approximately 1.96 times as many classes as white students pre-COVID (14.36 vs. 7.34), 2.26 times during the remote year (52.89 vs. 23.45), and 3.19 times post-COVID (35.31 vs. 11.08).

B. Logistic Regression

The logistic regression model revealed that every factor included is significant in determining the probability of a student reaching the threshold of concerning absences. Those factors are the years since the start of the COVID-19 pandemic, grade level (6-12), English language program participation (Y/N), number of SOL tests taken, and number of SOL tests failed. The confusion matrix below displays the performance of the model in accurately predicting whether a student should be labeled as a risk:

	Predicted Risk	**Predicted No Risk**
Actual Risk	74.68%	1.92%
Actual No Risk	20.67%	2.73%

To further assess model capabilities, we include cross-validation folds into the training and testing process. With 10 cross-validation folds, we receive an average accuracy of 77.23% with a probability threshold of 0.55. This means that if the probability of a student being "at-risk" is less than 0.55, they will be classified as not a risk. Again, this probability can be evaluated on its own on its continuous scale instead of forcing students to reside in one of the two risk categories based on a threshold to help administration assess students more objectively.

C. Survival Analysis

1) Kaplan-Meier Method:

The Kaplan-Meier estimator can be visualized as a plotted curve. In **Fig. 4** below, we see a clear difference in student attendance behavior. This figure shows displays the increase in students reaching their 8^{th} unexcused absence in a school year in any one course. As previously stated, this number of absences indicates that the student may be required to engage in discussion with administration regarding attendance.

The remote instruction year (2020-2021) reveals a serious decrease in the percentage of students that are remaining in good standing regarding attendance throughout the year. Not only are a higher percentage of students recording 8 unexcused absences in at least one single course, the rate at which students are reaching that day is much faster than the pre-COVID school year. In the return to in-person instruction school year, we still do not see attendance levels reach the point at which they once were. Students are still posing risks of failure in at least one of their courses at a higher overall percentage than the pre-COVID year.

When we separate students into their school level groupings, high schoolers and middle schoolers, we tend to see the same overall pattern of a severe increase in percentage of students tagged as "at-risk" from before COVID to remote instruction and then a slight recovery into the return to in-person school year. Even though that same pattern is observed, we notice in **Fig. 5** and **Fig. 6** that high schoolers are responsible for a majority of the "at-risk" students in the 2018-2019 school year. Overall, it appears that high schoolers tend to have more students recording at least 8 absences in at least one course than middle schoolers.

As seen in **Fig. 5**, middle schoolers experienced a very significant increase in the percentage of students that were approaching the risk of failing a course from before the pandemic going into the remote instruction school year. Like

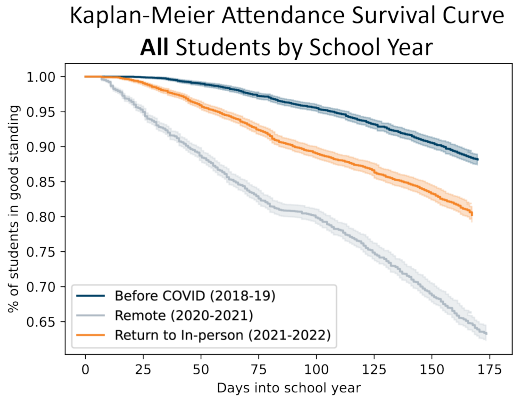

Fig. 4. Attendance levels improve but do not reach previous levels after the COVID-19 pandemic.

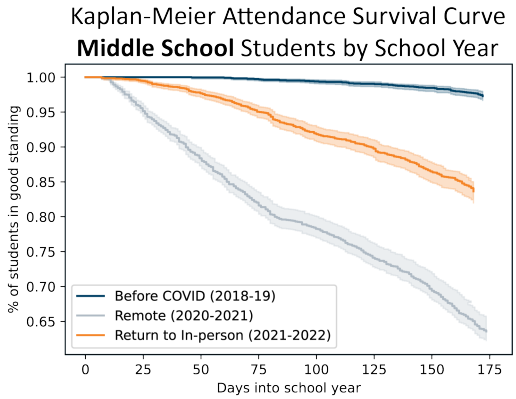

Fig. 5. Middle schoolers experience a large increase in percentage of students posing as a risk for failing a course during the COVID-19 pandemic.

in **Fig. 4**, we see attendance recover during the return to in-person instruction school year. Even so, the 2021-2022 school year still differs greatly from the 2018-2019 school year.

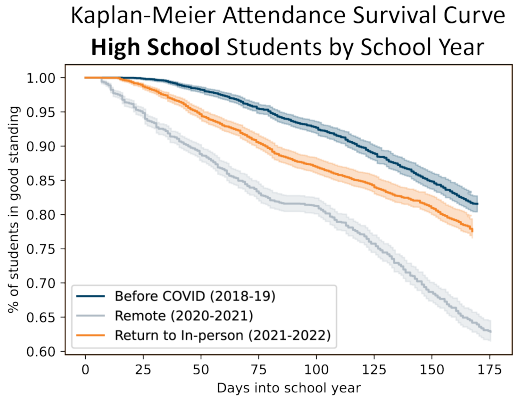

Fig. 6. High schoolers display an overall higher percentage of students at risk of failing a course than middle schoolers in the same school years.

2) Cox Proportional-Hazards Model:

The Cox Proportional-Hazards model revealed very similar patterns to the Kaplan-Meier attendance survival curves. The general pattern of percentage of students at risk of failure increasing during the remote year and decreasing again in the return to in-person school year was seen. The advantage of the Cox Proportional-Hazards model is that we are able to observe covariate coefficients as they relate to the hazard function. The table below showcases the coefficients for *SchoolYear*, *SchoolLevel*, *SPED*, *EL*, *SOLTestTaken*, and *SOLTestFailed*. The p-values displayed within the table reiterate that each of these variables is significant in predicting a student's risk of failure.

Variable	Coefficient	exp(Coefficient)	p-value
SchoolYear – 2020-2021	1.16	3.20	<0.005
SchoolYear – 2021-2022	0.49	1.63	<0.005
SchoolLevel – High	0.33	1.39	<0.005
SPED – Yes	0.20	1.22	<0.005
EL – Yes	0.75	2.13	<0.005
SOLTestTaken	-0.39	0.68	<0.005
SOLTestFailed	0.51	1.66	<0.005

For the categorical variables listed above, each one has a corresponding control group. For *SchoolYear*, the 2018-2019 serves as the control group. For the *SPED* and *EL* variables, the absence of either serves as the control group. An example of an interpretation of one of these coefficients would be saying that the risk of "failure" for students that are in the special education program is 1.22 more than the students that are not in the special education program. This rate of 1.22 comes from $exp(0.20) = 1.22$. Overall, we see that a lot of our available variables serve as significant indicators for predicting a student's risk of failing a course due to being absent for at least 8 classes.

VI. Conclusion

Overall, we observe a significant difference in student attendance behaviors among the three school years observed. More importantly, we see the same pattern of absences increasing significantly once students enter the remote instruction school year (2020-2021). This pattern is affirmed by retaining about the same structure across different subsets of students. Although we can only speculate, it is likely that this pattern can be attributed to heightened anxiety and lower academic motivation that came along with the introduction of the COVID-19 pandemic. This effect is seen lingering in the 2021-2022 school year as attendance levels do not return to the place they were only a few years prior. With the observance of these patterns, attention turns to leaders in education to come up with a response to the question, "How will we get student attendance levels back to where they were before the interruption of the COVID-19 pandemic while addressing this observed decrease in general student well-being?".

As mentioned, the work presented here may be seen both as a standalone analysis of student attendance and as a beginning to a larger, more cohesive exploration of student mental health changes over the pasts several years. Future work will aim to expand upon the results outlined above. We hope that analysis will be expanded to examine not only the point in the year where the 8^{th} unexcused absence is recorded in a course, but

to further dig into the pattern of time between absences. We believe there is value in observing how closely together a student's absences are.

As UVA Education continues to foster its relationship with ACPS, there is the potential for future research teams to receive data with more demographic factors. This is contingent on legal agreements as this data wouldn't necessarily protect student anonymity. The potential for more extensive work is certainly present with such agreements but needs to be done in a manner that prioritizes student privacy.

One more possible path of future research expansion would be the merging of attendance data with the data on student grades and disciplinary infractions. These three separate files are currently unable to be joined, thus preventing analytical techniques such as tracking the effect of unexcused absences on grades at the student level. Once the merge of data is approved, analysis can be broadened to make even more inferences about student mental health during this unprecedented period in time.

REFERENCES

[1] J. Hoofman and E. Secord, "The Effect of COVID-19 on Education," *Pediatric Clinics*, October 2021.

[2] H. M. et al., "Epidemiology of mental health problems in COVID-19: a review." *F1000Research*, June 2020.

[3] M. Kuhfeld, J. Soland, B. Tarasawa, A. Johnson, E. Ruzek, and K. Lewis, "How is COVID-19 affecting student learning?" *Brookings*, December 2020.

[4] E. S. Rudling, S. Emery, B. Shelley, K. te Riele, J. Woodroffe, and N. Brown, *Education and Equity in Times of Crisis*. New York City, NY: Springer Publishing, 2022.

[5] S. Ameri, M. J. Fard, R. B. Chinnam, and C. K. Reddy, "Survival Analysis Based Framework for Early Prediction of Student Dropouts," in *CIKM '16: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*.

[6] "K-12 Standards & Instruction." [Online]. Available: https://www.doe.virginia.gov/teaching-learning-assessment/instruction

[7] L. Zhang and H. Rangwala, "Early Identification of At-Risk Students Using Iterative Logistic Regression," *George Mason University, Department of Computer Science*, 2018.

[8] A. Nag, *Survival Analysis with Python*. Boca Raton, FL: CRC Press, 2021.