Statistical Analysis of COVID-19 Case Data from U.S. Colleges

By

Catherine Pommersheim, Undergraduate Department of Biomedical Engineering

Word count: __4291____

Number of Figures: __2___

Number of Tables: ___3___

Number of Equations: ___0___

Number of Supplements: ___0___

Number of References: ___12___

Approved: Brent A. French Date: _____5.10.2021_____

Dr. Brent French, Department of Biomedical Engineering

Statistical Analysis of COVID-19 Case Data from U.S. Colleges

Catherine Pommersheim[a]

[a] Fourth Year Biomedical Engineering Undergraduate at the University of Virginia
[1] Correspondance: cmp4nd@virginia.edu

Abstract

The COVID-19 pandemic, caused by the virus SARS-CoV-2, began as an outbreak in Wuhan, China in late 2019 and made its way to the United States in early 2020. Since then, restrictions have been enacted to limit disease transmission and protect the public from infection. Initial confusion surrounding the most effective policies led to widespread distrust of the scientific community which manifested as the anti-mask and anti-vaccine movements that gained momentum throughout the pandemic. In order to more quickly assemble a response against a disease to limit infections and establish trust in public health officials and institutions, COVID-19 transmission will be analyzed to identify which policies are most effective in limiting its spread. Colleges in the United States will serve as the subjects of this study as they mimic semi-urban population dynamics with the added ability to monitor positive COVID-19 cases among the student body. To perform this investigation, regional variance in mask compliance was analyzed, trends in COVID-19 positive case data were identified, and a review of individual schools' preventative measures were compared to provide insight into how to best limit COVID-19 transmission. The results unexpectedly showed little association between strict COVID-19 policies and the proportion of positive cases out of undergraduate populations. In fact, schools with high proportions of COVID-19 cases report implementing thorough policies which is seen at both UVA and Johns Hopkins. The most significant confounding variable revealed through further investigation is policy compliance levels from individual student bodies which must be investigated and accounted for to fully understand COVID-19 transmission patterns.

Keywords:
COVID-19, transmission, compliance, public health policy

Introduction

The novel coronavirus, COVID-19, was initially identified in late 2019 in Wuhan, China and continued to spread until it became a global pandemic in 2020. This disease, caused by the virus SARS-CoV-2, made its' way to the United States (U.S) in early 2020 which initiated the beginning of a national quarantine and implementation of other public health measures in order to limit disease transmission. The progression of the COVID-19 pandemic has been unparalleled in recent history by its ability to bring social and economic systems to a halt. The effects of the pandemic have been felt by those who have lost loved ones, lost jobs, or been infected themselves and will continue to endure lasting consequences. At the onset of the global pandemic, the lack of information on this novel virus led to initial confusion on how to best protect the public from infection.

The speed with which COVID-19 spread across the world and to the U.S. left public health officials struggling to implement effective protective policies in such a short amount of time. As a result, the initial policies set forth and enforced by the Centers for Disease Control and Prevention (CDC) were later proven to be ineffective and were subsequently revised. For example, the CDC announced that masks were not necessary to be worn by all non-medical personnel and that all surfaces should be regularly disinfected to best protect oneself from infection with SARS-CoV-2. However, in April 2020 the CDC revised this policy to recommend non-medical face masks to be worn in public which later became a policy mandated by over half of the U.S. states and proved to be one of the more effective ways to prevent against COVID-19 transmission. Additionally, the CDC has since released statements that further research on the novel coronavirus reveals that there is a low risk of contracting this disease from contact on surfaces which contradicts the initial platform outlined at the beginning of the pandemic. While the revision of national COVID-19 policies and recommendations has led to more effective prevention measures being put into place, it has also fostered a sense of distrust among some communities who view the changes in policy as uncertainty about the importance of following these guidelines.

This general sense of distrust has manifested throughout the course of the pandemic as an increase in anti-mask sentiment which was so intense that concerned citizens had organized public protests against this policy. Furthermore, anti-vaccine sentiment has seen a recent rise in support since the onset of the COVID-19 pandemic which can be attributed to concerns about the safety and efficacy of a vaccine developed on an accelerated timeline.

The onset and progression of both anti-vaccine and anti-mask sentiment during the pandemic indicates an underlying distrust of the scientific community and public health officials. In order to address this issue, the development of a comprehensive strategy to best limit disease transmission is important so that in the event of a novel disease outbreak, a plan can be more quickly assembled and disseminated. The speed with which a preventative plan is organized will hopefully strengthen trust in the policies outlined by public health officials and emphasize the importance of compliance which has proven to be an issue during the current pandemic, as seen with the resistance to mask mandates.

To identify a selection of COVID-19 policies that best protect against disease transmission, U.S. colleges were chosen as the subjects for this study as they can be considered a microcosm for densely populated environments with the unique ability to monitor their students and how COVID-19 moves throughout their communities. At the point of completion of this work, there are no similar studies that have investigated the relationship between positive COVID-19 cases on college campuses and the associated public health policies that have been implemented. As a result, this work also aims to serve as a preliminary attempt to understand the effect of policy on disease transmission patterns using the methodology developed in this study.

In addition to being able to report the number of cases from their student bodies, colleges are able to implement their own COVID-19 policies and theoretically enforce them to produce a high level of compliance. By understanding COVID-19 transmission patterns on college campuses and examining the associated preventative measures, a clearer picture of the most effective policies can be developed. From there, a more thorough strategy to prevent against disease transmission can be assembled and disseminated in the event of a novel disease outbreak. This will not only limit resulting illness and death due to viral transmission, but will also increase the trust and compliance of the general public with policies set forth by public health officials and related organizations.

Results

Examining Regional Trends in Mask Compliance

To best understand COVID-19 transmission on college campuses, regional trends in mask compliance were considered to determine if they exhibited a significant effect on transmission patterns due to varying levels of compliance. Mask compliance is known to vary regionally due to a few factors that include but are not limited to cultural norms, population density,

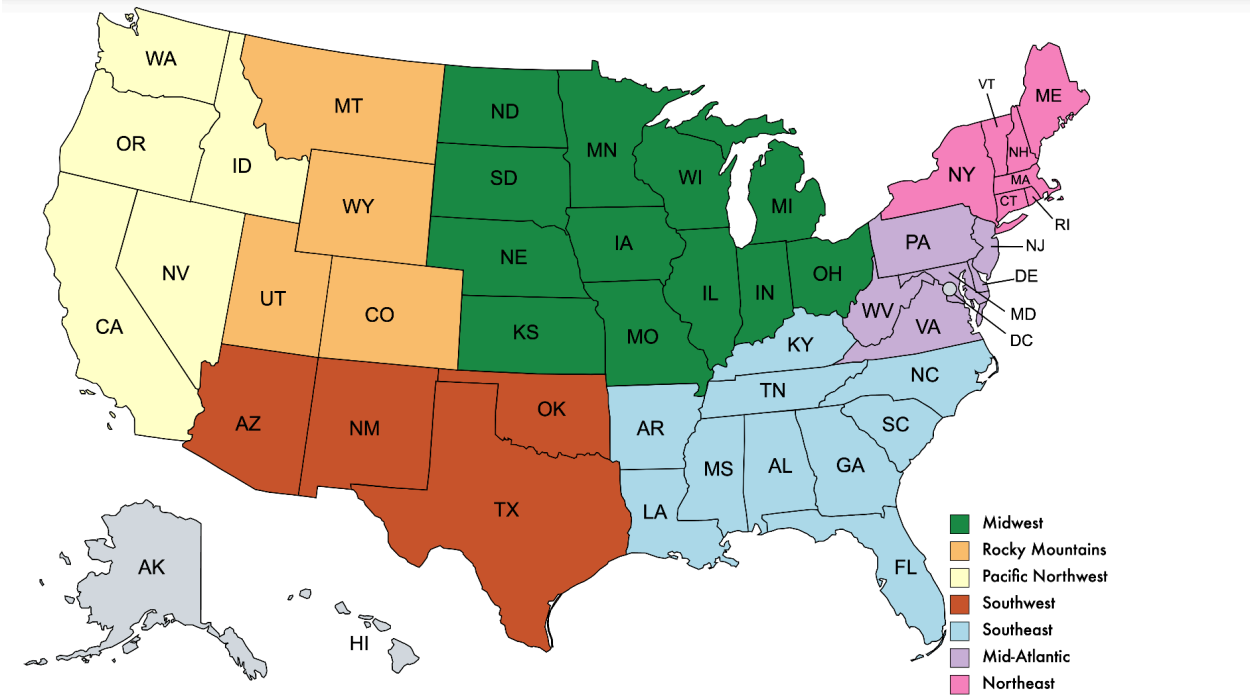

Figure 1: Geographic Breakdown of the US

and perceived risk of infection. To identify trends in mask compliance, the contiguous United States was divided into seven major geographic regions to subsequently identify regional differences in compliance levels. It is important to understand regional demographics and incorporate this information into public health policy to fight COVID-19 as different age, health, and ideological groups may react differently to restrictions imposed for public health.

The demarcations used to divided the contiguous U.S. into seven regions are outlined in Figure 1 as well as the resulting terms used to identify each region throughout the remainder of this study. Mask compliance data was then assembled from the New York Times (NYT) database that was acquired through a national survey asking individuals about their likelihood to wear a mask when required by local mandates. It is important to consider a few sources of bias when examine data from a voluntary survey in order to consider how results may be skewed to make respondents look more favorable. Two main types of bias exist when considering this data. The first is response bias which is created due to the tendency for individuals to respond to a survey when they already possess a strong preference for one perspective over another. Similarly, social desirability bias could affect the collection of responses to this survey as individuals tend to self-report perspectives that appear favorable to the general public and existing social structures. In the case of mask compliance data, this bias might manifest as a probability to over-report one's levels of mask compliance which could skew the data to report higher levels of compliance than one would observe in reality. As a result, it is important to consider that these levels of mask compliance are claims rather than fact.

The responses to the mask complicate survey collected by the New York Times were originally sorted by county. However, for the purposes of this study, the responses were aggregated to form a general level of compliance for each state by organizing the results of each county within a state and finding the highest proportion of responses. The levels of compliance that were included in the survey are as follows: Always, Frequently, Sometimes, Rarely, and Never. A large proportion of respondents self-reported that they would always wear a mask when required by local mandates so it was then straightforward to identify the proportion of

Geographic Regions in the Contiguous United States	Mid-Atlantic	Southeast	Northeast	Midwest	Rocky Mountains	Pacific Northwest	Southwest
Claimed Likelihood to Always Wear a Mask (%)	100	99.5	100	87.7	83.7	96.3	97.8

Table 1: Regional Mask Compliance Claims

individuals within each region who reported that they would always comply with mask mandates compared to lower levels of compliance.

This analysis was performed in R and serves as a general approximation for mask compliance levels in each region after responses were organized by state and then by geographic region. The results of these analyses are reported in Table 1 as the percentage of the respondents in each region that reported they would always wear a mask when required. As one can tell, the Mid-Atlantic and Northeast regions report the highest levels of compliance and are the more urban, densely populated regions. On the other hand, lower levels of compliance are reported for the Rocky Mountain and Midwest regions which tend to be more rural or sparsely populated regions with a lower level of perceived risk of infection from COVID-19.

School Selection

The dataset used to analyze COVID-19 case data was collected from the New York Times, similarly to the mask compliance survey responses. This dataset includes positive COVID-19 case numbers reported from individual colleges and universities to the New York Times as well as their location. As a result, schools that did not contribute their case data to this repository were not considered for this study. For the purposes of this study, 20 schools were selected from each of the seven main geographic regions to be investigated out of the total 1949 schools that contributed COVID-19 case data. To produce the most comprehensive results, schools that had not yet reported case numbers for the Spring 2021 semester by the date of collection, February 26th, were not considered for this study. After removing these schools, there were 1508 schools eligible to be selected. Using a random sampler in R, 20 schools were randomly selected from each region to produce a total of 140 schools whose case data was examined. On a national scale, these samples would be considered to be stratified by region.

School size was not considered when selecting schools for this study as the selection was random. However, school size was later controlled for by identifying the proportion of positive COVID-19 cases given the undergraduate enrollment numbers at each school. The proportions of positive cases were then used for comparisons rather than the raw number of positive reports. This consideration accounts for the first assumption necessary to this study which is that COVID-19 cases are reported for the entire undergraduate population rather than including graduate students or leaving out groups of undergraduate students. The data collected from the New York Times does not indicate whether each school is reporting cases for their undergraduate population or has decided to include cases from post-graduate departments so this assumption aids in standardizing our results.

Another assumption that must be considered when examining the data from these schools is that each school is accurate in their reporting of positive COVID-19 cases. This assumption contains two sub-hypotheses that must be considered valid to establish accurate reporting of case numbers. First, we assume that schools report cases for their entire undergraduate student body and do not omit any groups. This assumption is important to validate our use of undergraduate population as a metric to standardize case numbers whereas we know that this assumption may not hold in reality due to some students not returning to their respective campuses due to the risk of infection with SARS-CoV-2. Additionally, this assumption implies that students who get tested through avenues not

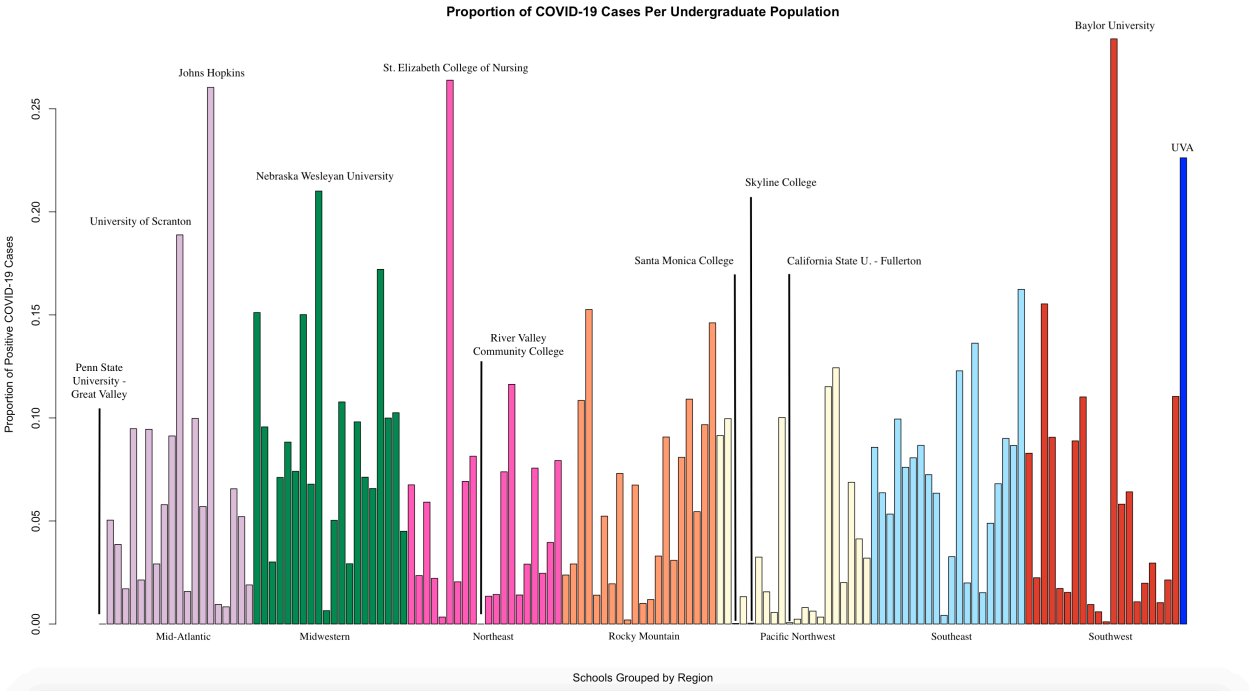

Figure 2: Regional Proportions of Positive COVID-19 cases with the Schools of Interest Identified

directly associated with their schools will still have any positive cases accounted for and reported to the NYT database. Second, we assume that schools possess an accurate testing and reporting system that correctly identifies positive COVID-19 cases with little to no false positives or negatives being reported.

Trends in COVID-19 Case Data

Once all 140 schools were selected for this study, the COVID-19 positive case data was cleaned and prepared for analysis. The first step in preparing this data, was to identify the undergraduate populations for each school. This investigation was preformed manually by accessing publicly available information on undergraduate enrollment numbers. By dividing the total number of COVID-19 cases for each school by its' respective undergraduate population, the proportion of COVID-19 cases at each school was identified. Nationally, the distribution of proportions exhibited a minimum value of 0.00 and a maximum value of 0.284 while the mean was found to be 0.064. The point estimate for the national average proportion of positive COVID-19 cases per school was found by summing the total number of cases for schools selected for the study and dividing it by the sum of total undergraduate enrollment numbers. This point estimate was found to be 0.0713 which means that one could expect 7.13% of students at any school to have tested positive for COVID-19 over the course of the 2020-2021 school year.

There was significant variance exhibited between the proportion of positive COVID-19 cases at schools both regionally and nationally. To test for effects of mask compliance levels on COVID-19 case proportions, the point estimate for each region was calculated and tested against the national average point estimate of 0.0713. No significant difference was found between regions or nationally when testing for similarity between each subgroup of this study as each exhibited such a wide amount of variation between schools. In fact, a proportion test in R was used to perform these tests as well as to test for similarity between schools within each region. When testing for similarity within each region, significant variation was found where the p-value reported for these

tests approached zero to indicate that the variation is so strong that these results case proportions likely did not originate from the same population, which is the case.

To identify what accounts for such large differences in COVID-19 case proportions, the five schools with the largest proportion of positive cases to undergraduate population as well as the most significant p-values when tested against regional and national point estimates were chosen to explore their COVID-19 policies. Similarly, the five schools with the lowest proportions of positive cases to undergraduate population as well as the most significant p-values were chosen as well. These schools have been identified in Figure 2 to better visualize how significant the differences in proportions appear. Figure 2 presents a bar plot of the proportions of COVID-19 cases as separated by region with the labels for each given across the x-axis. These regions are colored in the same manner as shown on the map to maintain consistency. Along the y-axis of this plot is the scale to identify the proportions of positive cases per undergraduate population which is helpful in qualitatively confirming which schools have the highest and lowest proportions of cases.

The proportion of cases for the University of Virginia (UVA) has been included in the right-most side of Figure 2 labeled in bright blue and, while technically belonging in the Mid-Atlantic region, is presented separately due to our special interest in and affiliation with UVA. The proportion of cases per undergraduate population was found to be 0.226 meaning that around 22.6% of the undergraduate population tested positive for COVID-19 during the 2020-2021 academic year, as of February 26th, 2021.

Association of Public Health Policy and College Cases

To begin the investigation into COVID-19 policies and their effect on case count, we first had to identify the policies implemented at each of the ten most significantly varied schools that were chosen for investigation. To accomplish this goal, a thorough review was performed on each school to identify the preventative measures that the school claims to use. This information was collected from each school's website and is all available to the public.

The five schools with the lowest proportion of COVID-19 cases per undergraduate population all shared characteristics of being either a community college or secondary location of a state school. Secondary locations of state schools tend to have characteristically smaller campuses and student bodies. For an example, UVA's secondary location at Wise is a significantly smaller, liberal arts college compared to the much larger university in Charlottesville. Similarly, community colleges tend to be commuter schools which eliminates the aspect of COVID-19 transmission that arises from living and learning in shared spaces which is a common feature of most other traditional four year institutions. Furthermore, all five schools shared the policies of having almost fully online courses with only some laboratory classes meeting in person. Based on the overview of policy used at these schools, it it unsurprising that they exhibited such low levels of COVID-19 in their communities. A full overview of policies

School	Policy Overview
Penn State University - Great Valley	Online and hybrid instruction modes, Requires a negative test 72 in advance of returning to campus
River Valley Community College	All classes online except labs, Temperature checks upon entering buildings, Masks mandatory indoors
Santa Monica College	All classes online except labs
Skylin College	All classes online and will remain online through Fall 2021
California State University - Fullerton	All classes online except labs

Table 2: COVID-19 Policies for Schools with Low Proportions of Cases

School	Policy Overview
University of Scranton	Mandatory masks indoors and outdoors, No visitors, Asymptomatic testing (3x per semester)
Johns Hopkins University	Mandatory masks indoors and outdoors, Asymptomatic testing (3x per week), No gatherings of 5+ indoors and 10+ outdoors, Hybrid and online courses
Nebraska Wesleyan University	Mandatory masks indoors, No gatherings, No visitors, Hybrid and online courses (with weekly small group meetings on campus)
Baylor University	Mandatory masks indoors and outdoors, Asymptomatic testing (1x per week), Hybrid and online courses
St. Elizabeth College of Nursing	COVID-19 policy not publicly available but requires clinical experience in the community

Table 3: COVID-19 Policies for Schools with High Proportions of Cases

implemented at schools with the lowest COVID-19 case proportions is found in Table 2.

On the other hand, the five schools with the highest recorded proportions of COVID-19 cases per undergraduate population reported detailed preventative measures against disease transmission. It it important to note that one school with a high proportion of cases is in fact a nursing college whose students perform clinical rotations in local hospitals. This routine exposure to high-risk environments accounts for the significant proportion of COVID-19 cases amongst the small population at this school. The other four schools selected for this policy review are all traditional four year institutions who boast thorough COVID-19 policies. Preventative measures at these schools include, but are not limited to, indoor and outdoor mask mandates on campus, routine prevalence testing, prohibited gatherings or gathering limits, and optional online learning. A full overview of the policies implemented at each of these schools can be found in Table 3.

Discussion

The analysis of COVID-19 policies used at the schools in this study yielded unexpected results. Schools with low proportions of cases given their undergraduate populations exhibited policies that explain how disease transmission was so limited, primarily due to significantly restricting in-person education and living arrangements. However, schools with the highest proportions of cases indicated that they employed strict COVID-19 policies, as outline in Table 3. There are multiple reasons that could account for the lack of significant association between public health policy and case numbers. One could suggest that the preventative restrictions were actually introduced retroactively after a spike in cases and therefore were not implemented in time to limit transmission. Similarly, one could argue that increased asymptomatic testing, which is a policy implemented across three out of five of these schools, is causing the heightened number of cases as more positive cases are being identified through more rigorous testing. Despite both of these theories, the most significant confounding variable identified through field research is compliance.

Compliance was found to be a significant limitation to these analyses through first hand experience and testimonials. The two schools at which compliance level appears to play a significant role, based on qualitative reports from current, average students, are UVA and Johns Hopkins University. Monitoring COVID-19 trends at UVA is of particular importance to this investigation in order to protect the health and wellness of students and faculty at the university, although this school was examined separately from the rest of the Mid-Atlantic schools.

The similarities between UVA and Johns Hopkins responses to the current pandemic are the

development of thorough COVID-19 policies as well as a case proportion per undergraduate population above 0.22. This high proportion of cases per student body would lead one to believe that these schools did not develop a robust prevention plan; however, both schools imposed gathering limits, regular asymptomatic prevalence testing, mask mandates for indoors and outdoors, and a majority of course being delivered remotely. When comparing the proportions of positive cases and the rigor of COVID-19 prevention measures, the discrepancy is obvious. However, first hand knowledge of student dynamics at UVA combined with insight from a contact at Johns Hopkins indicate that both schools are unable to fully enforce their policies yet have no system to account for compliance. Additionally, first hand accounts identify the avenues through which COVID-19 transmission occurs both on and off college campuses that administrators may not keep track of, let alone be able to report. Without incorporating compliance levels into these analyses, a crucial factor in COVID-19 transmission patterns will be ignored and limit the conclusiveness of our results.

While access to raw data on COVID-19 case numbers as well as the policies implemented at each school can paint a general picture of disease transmission, the most important factor needed to complete this investigation is compliance with public health policies. For the purpose of this study, compliance can be perceived synonymously with administrative enforcement of COVID-19 policies. A key assumption to our analyses is the belief that school administrators enforce institutional policies to the fullest extent, which we assume to be one hundred percent. In reality, it is improbable that an institution is able to monitor its' student body to the degree necessary to eliminate all non-compliant behavior. The varying levels of compliance with policy between schools acts as a confounding variable by masking the effects of public health policies and making them seem less effective than in reality.

Future work must address levels of policy compliance in order to generate more conclusive results that can better inform and influence future policy makers and administrators. One method to account for compliance in future analyses is by testing for the effects of COVID-19 policies on groups of students that have identified as being almost fully compliant with policy compared to groups that report non-compliance. The efficacy of this method would depend on honest, anonymous self-reporting from students which could be difficult to obtain considering the effects of social desirability bias. Another avenue to address compliance is to gather first hand accounts from students at schools of interest that identify perceived levels of compliance among the student body. These results could then be quantified to include compliance as a metric that can be tested in future analyses.

In addition to compliance, limitations exist that hinder our ability to produce definitive conclusions about the efficacy of COVID-19 policies. We assume that mask compliance data reflects current views on mask regulations and is representative of the population within a region; however, we know that this may not be the case since the surrey was collected in July, 2020. Additionally, we assume that all undergraduate students are at equal risk of infection from SARS-CoV-2 which may not be case if students chose to live off-campus or remained at home throughout this academic year. The vaccine rollout, which began in mid-December, 2020, has also affected the likelihood of individuals to contract COVID-19. While this is a general, public good, it does contradict the assumption used for this investigation as student populations are no longer at equal risks of infection.

A final source of error that should be considered in future work is the inclusion of community colleges in analysis on college policies. While community colleges are valid higher education institutions, they lack the same qualities that make traditional four-year institutions good predictors of transmission dynamics. Most importantly, community colleges are typically commuter schools which eliminates the communal living aspect of traditional universities that creates a densely populated community. In future work, one might consider identifying and eliminating community colleges from the investigation. This perspective is backed by the fact that three out of the five schools with the lowest proportions of COVID-19 cases were community colleges which delivered almost completely remote education. The inclusion of these schools could have skewed the results with their characteristically lower case proportions and

should be removed in future studies to test this hypothesis.

The methodology of this study serves as a starting point for further investigation into COVID-19 transmission patterns on college campuses. The procedure outlined in this paper could be applied to monitor transmission patterns of other diseases as well to identify effective preventative measures. While this study was unable to form conclusive results on the efficacy of COVID-19 policies, compliance with policy was identified as a significant confounding variable that must be addressed in future work to produce meaningful results.

Materials and Methods

Mask Compliance Data

Mask compliance data was compiled by aggregating responses by county into their respective states. From there, responses were grouped into the seven regions outlined in Figure 1 by each state. The The dominant response from the entire survey was that individuals claim to Always wear a mask when required by law. As a result, the frequency of counties that reported to Always wear a mask within each region was calculated and reported as a percentage in Table 1.

School Selection

Schools were grouped by their location and sorted into the seven geographic regions outlined in Figure 1. From there, a random sampling function in R was used to randomly select 20 schools from these groupings.

Public Policy Identification

COVID-19 policy identification was done manually by consulting the main websites for the schools of interest within the study. This qualitative data was aggregated in Table 2 and Table 3.

COVID-19 Case Data Analysis

The primary test used to compare the proportions of COVID-19 cases per undergraduate population is the proportion test function in R. This function was combined with the sapply() function to iterate across al rows of schools in the study to individually test their case proportion against regional and national point estimates to identify significance.

References

Times, T. N. Y. (2021, February 26). Tracking Coronavirus Cases at U.S. Colleges and Universities. *The New York Times*. https://www.nytimes.com/interactive/2021/us/college-covid-tracker.html

Katz, J., Sanger-Katz, M., & Quealy, K. (2020, July 17). A Detailed Map of Who Is Wearing Masks in the U.S. *The New York Times*. https://www.nytimes.com/interactive/2020/07/17/upshot/coronavirus-face-mask-map.html

Coronavirus Campus Information. (n.d.). Penn State Great Valley. Retrieved May 10, 2021, from https://greatvalley.psu.edu/coronavirus-campus-information

COVID-19 Information – River Valley Community College. (n.d.). Retrieved May 10, 2021, from https://www.rivervalley.edu/student-support/information-on-coronavirus/

COVID-19 Updates. (n.d.). Retrieved May 10, 2021, from https://www.smc.edu/news/coronavirus/index.php

Campus Re-entry Plan | Titans Return: COVID-19 Recovery. (n.d.). Retrieved May 10, 2021, from http://coronavirus.fullerton.edu/titans-return/

COVID-19 information | Health & Wellness Services | Skyline College. (n.d.). Retrieved May 10, 2021, from http://www.skylinecollege.edu/healthandwellness/video.php

A Safe, Healthy Campus | Royals Safe Together | The University of Scranton. (n.d.). Retrieved May 10, 2021, from https://www.scranton.edu/royals-safe-together/safe-healthy-campus.shtml

COVID-19 (coronavirus) Updates | Nebraska Wesleyan University. (n.d.). Retrieved May 10, 2021, from https://www.nebrwesleyan.edu/about-nwu/covid-19-coronavirus-updates

Curriculum Plans. (n.d.). *St. Elizabeth College of Nursing | Utica, NY*. Retrieved May 10, 2021, from https://www.secon.edu/student/prospective/curriculum/

COVID-19 Student Guide | Coronavirus (COVID-19) Information | Baylor University. (n.d.). Retrieved May 10, 2021, from https://www.baylor.edu/coronavirus/index.php?id=971766

COVID-19-Related Policies. (n.d.). Coronavirus Information. Retrieved May 10, 2021, from https://covidinfo.jhu.edu/health-safety/covid-19-policies/