Supporting Information for: Higher education responses to COVID-19 in the United States: Evidence for the impacts of university policy

Brennan Klein[†*1,2], Nicholas Generous[†1,2,3], Matteo Chinazzi[1,2], Zarana Bhadricha[1,4], Rishab Gunashekar[1,4], Preeti Kori[1,4], Bodian Li[1,5], Stefan McCabe[1], Jon Green[1,6], David Lazer[1], Christopher R. Marsicano[7,8], Samuel V. Scarpino[1,9,10,11], and Alessandro Vespignani[*1,2]

[1] *Network Science Institute, Northeastern University, Boston, USA*
[2] *Laboratory for the Modeling of Biological and Socio-Technical Systems, Northeastern University, Boston, USA*
[3] *Biosecurity and Public Health Group, Los Alamos National Laboratory, Los Alamos, USA*
[4] *College of Engineering, Northeastern University, Boston, USA*
[5] *College of Professional Studies, Northeastern University, Boston, USA*
[6] *Shorenstein Center on Media, Politics and Public Policy, Harvard University, Boston, USA*
[7] *Educational Studies Department, Davidson College, Davidson, USA*
[8] *College Crisis Initiative, Davidson College, Davidson, USA*
[9] *Vermont Complex Systems Center, University of Vermont, Burlington, USA*
[10] *Santa Fe Institute, Santa Fe, USA*
[11] *Pandemic Prevention Institute, Rockefeller Foundation, Washington, D.C., USA*

S1 Text

The Campus COVID Dataset

Collecting URLs of COVID-19 dashboards

We used the Python library `googlesearch` [1] to automate Google search queries, and we used the `gspread` Python library [2] to aggregate the output into a single Reference Google Sheet. For each of over 4,000 IHEs in the IPEDS 2018 data, we generated the following Google search queries: "`university_name` COVID dashboard", "`university_name` COVID cases", "`university_name` COVID testing", with and without quotation marks to increase specificity or broaden the search yields (if the IHE's name does not include the state it is

[*] Correspondence: b.klein@northeastern.edu, a.vespignani@northeastern.edu; [†] Equal contribution

located in, we append the state name to the search queries to avoid duplicate or similarly-named institutions in different states). This process produced several candidate URLs, and we select the most common URL to include in the Reference Google Sheet; this was usually effective at creating one single URL for each IHE, since "COVID cases" and "COVID testing" are semantically similar terms that are likely to co-occur on the COVID-19 information pages for universities. We repeated this process for thousands of IHEs, and we were left with URLs for the COVID-19 dashboards (or analogous website) of 2,739 IHEs.

Standardized data collection

In order to create a more streamlined data collection process, we created separate pre-formatted Google Sheets for every IHE using the `gspread` Python library. Each of these Google Sheets was indexed by the `ipeds_id` of the IHE and was formatted like Table B. The "total_tests" and "positive_tests" columns were left blank and eventually filled in by research assistants, who would visit the COVID-19 dashboard urls of each IHE and record the data. Many IHEs elected to report data using interactive visualization software, such as Microsoft Power BI or Tableau dashboards; others used a simple data table to report their numbers; others wrote weekly email updates to the campus community that included the testing and case counts for the previous week.

Many IHEs reported none of this information, for a variety of potential reasons. For one, many IHEs were closed for the Fall 2020 semester (or entirely, due to financial hardship), not allowing students on campus at all and therefore not actively recording case counts among the campus population. Some schools were at least partially in-person for the Fall 2020 semester but did not appear to collect data about cases on campus. We categorized these schools as "cannot find data". Other IHEs had passages of text on their websites that suggest they had knowledge of the number of cases on campus (e.g. "case counts on campus are low") but did not publicly report raw numbers. These schools were categorized as "not publishing data". These categorizations can be found in Table A.

Matching counties with broadly similar demographics

Because COVID-19 outcomes correlate with demographic variables (e.g. age [3], among others), extra care must be taken when comparing averages between groups of counties to ensure that any observed differences are not due to differences in population structure of the underlying populations. Here, we try to create counties with as similar as possible distributions of age, race, income, urban-rural code, and population size. (Note: "urban-rural code" is an ordinal variable ranging from 1 ("large central metro") to 6 ("non-core") and is assigned by the National Center for Health Statistics [4].)

In order to compare differences in average case counts and deaths (as in Fig. 2) among counties with IHEs that returned to in-person education compared to those with IHEs that remained online, we rely on data collected by the College Crisis Initiative [6]. This dataset assigns a category for over 1,800 IHEs based on the reopening strategy for the Fall 2020 semester. From these IHEs, we sum together the total full-time enrollment in each reopening

complete time series	971
cumulative data only	477
cannot find data	1,271
total IHEs in data	1,448
total IHEs searched	2,719

Table A: Current status of the Campus COVID Dataset. In total, the Campus COVID Dataset includes data about more than 1,400 IHEs. To collect these data, we searched among over 2,719 IHEs; approximately 40% of these are IHEs with data that we could not find (because the IHE does not collect self-reported positive tests and/or does not conduct campus testing, etc.) or with data that we believe exists but was not being shared publicly by the IHE. There are over 971 IHEs with time series of testing and/or case counts for the Fall 2020 semester. If an IHE reported only cumulative testing or case counts, we classify it as "cumulative only".

date	**total_tests**	**positive_tests**	**college**	**URL**	**ipeds_id**	**notes**
2020-08-01			`university_name`	`dashboard_url`	`ipeds_id`	
2020-08-02			`university_name`	`dashboard_url`	`ipeds_id`	
2020-08-03			`university_name`	`dashboard_url`	`ipeds_id`	
...	...	...	...	...	...	
2020-12-16			`university_name`	`dashboard_url`	`ipeds_id`	

Table B: Example template for inputting data. Each IHEs in the Campus COVID Dataset has a unique URL that leads to a dataframe with this structure. For each date that the IHE reports a number of new cases ("positive_tests" above) or new tests administered ("total_tests" above), we input that value in its corresponding row. For IHEs that report testing and case counts weekly, we insert the data at the first collection date, which makes for more accurate smoothing when performing 7-day averages. If the IHE only reports *cumulative* cases or tests for the Fall 2020 semester, we leave the "total_tests" and "positive_tests" columns blank and report the "cumulative_tests" and "cumulative_cases" in the "notes" column, which we extract later in the analyses.

category (e.g. a county may have three universities—with 250, 800, and 15,000 students enrolled full-time; if the school with 15,000 students is categorized as "primarily online" and the other two are "primarily in-person" we say that specific county has 15,000 primarily online students and 1,050 primarily in-person students). As discussed in the main text, the final piece for constructing groups of counties with similar demographics is the percent of the county population affiliated with the IHE. That is, we define a "college county" based on the what percent of the total population is made up of students enrolled full-time at an IHE in the county.

We create two groups of counties from the 1,238 unique counties in the College Crisis Initiative [6]; we iteratively vary the threshold of inclusion into these two groups, finding the threshold that minimizes the total Jensen-Shannon Divergence (JSD; the JSD between two distributions, P and Q, is $JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$, where $M = \frac{1}{2}(P + Q)$

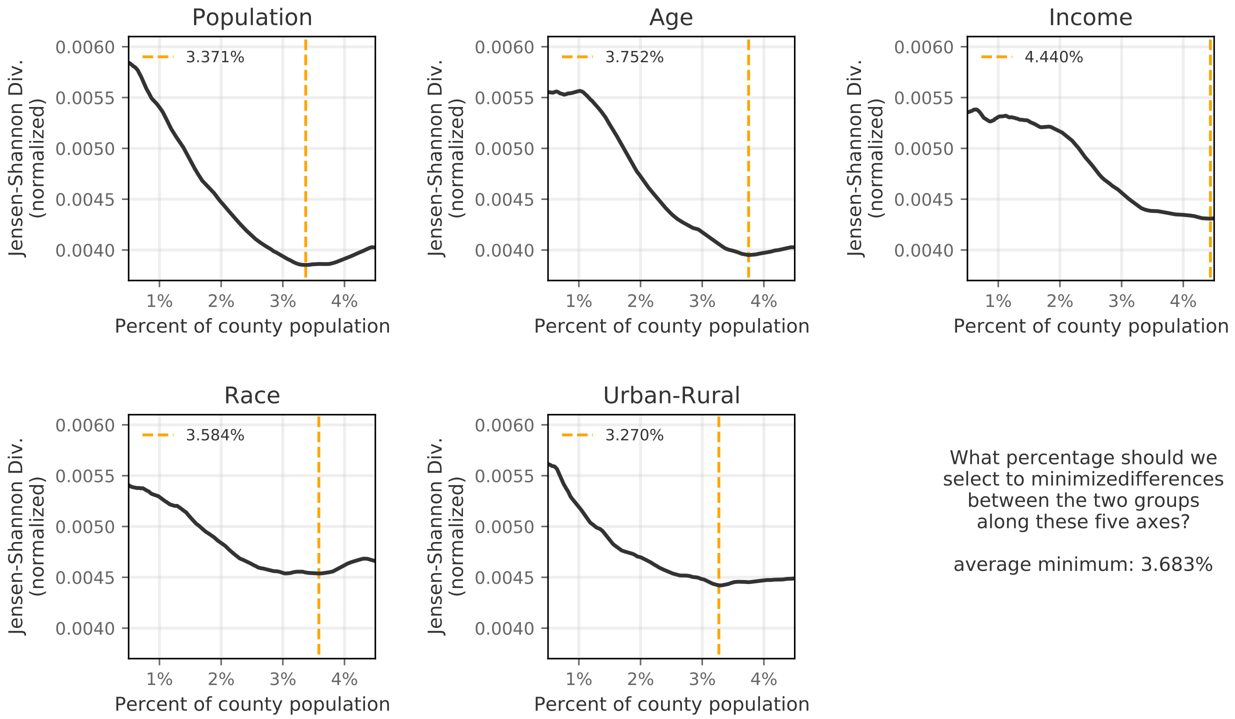


Figure A: Jensen-Shannon Divergence between distributions of demographic variables. As we vary the threshold for inclusion into the two groups—counties with IHEs that returned primarily in-person for Fall 2020 and counties with IHEs that remained primarily online—the Jensen-Shannon Divergence also changes. We want to select the value for this threshold based on whatever minimizes the Jensen-Shannon divergence, on average.

and $D(P||Q) = \sum_{x\in X} P(x)\log(\frac{P(x)}{Q(x)})$, the Kullback-Leibler Divergence) between the average demographic distributions of the counties that comprise the groups. In our case, this threshold corresponds to the percent of the total county population enrolled in IHEs within that county full-time. Intuitively, a "college county" is one with a relatively large percentage of its population being affiliated to an IHE, but the precise value of this percentage is not commonly defined. Here, we test a range of thresholds and select the one that minimizes the JSD between the resulting groups' distributions of the five demographic variables of interest (see Fig. A).

By selecting the threshold that minimizes the average JSD between the demographic distributions of the two groups, we get closer to making sound comparisons along the true axis of interest. Following this procedure, the resulting groups of counties are similar in demographics (Fig. B), spatially spread out across the country, and minimally clustered based on region (Fig. C).

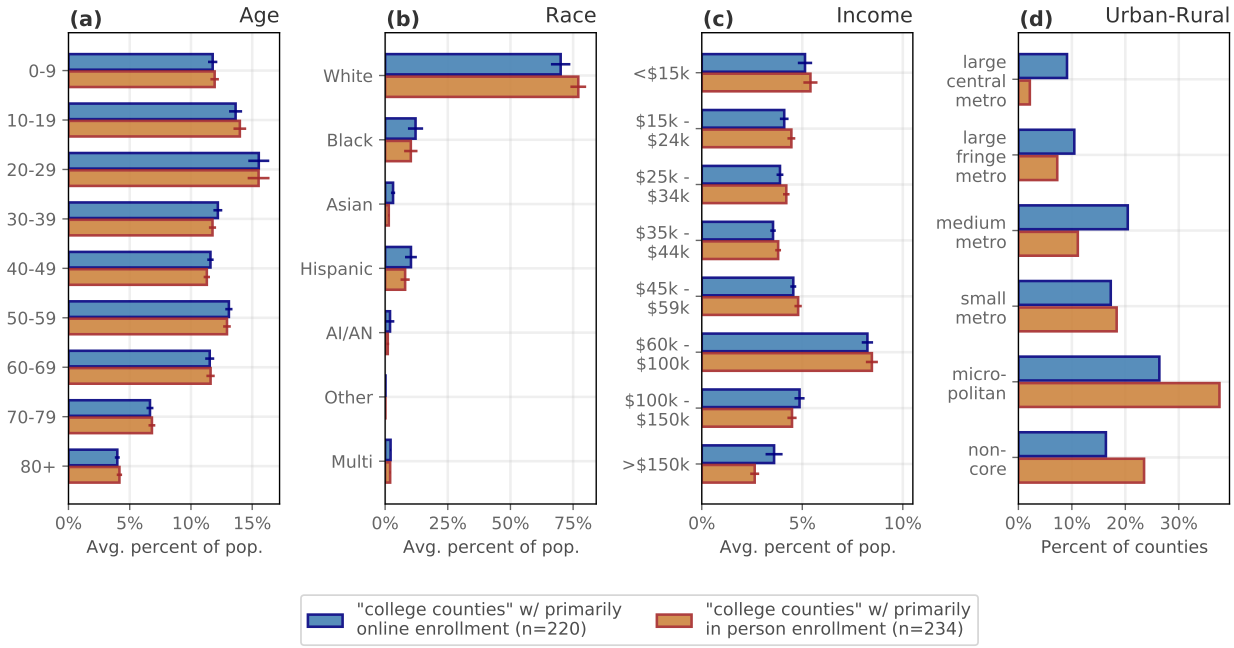
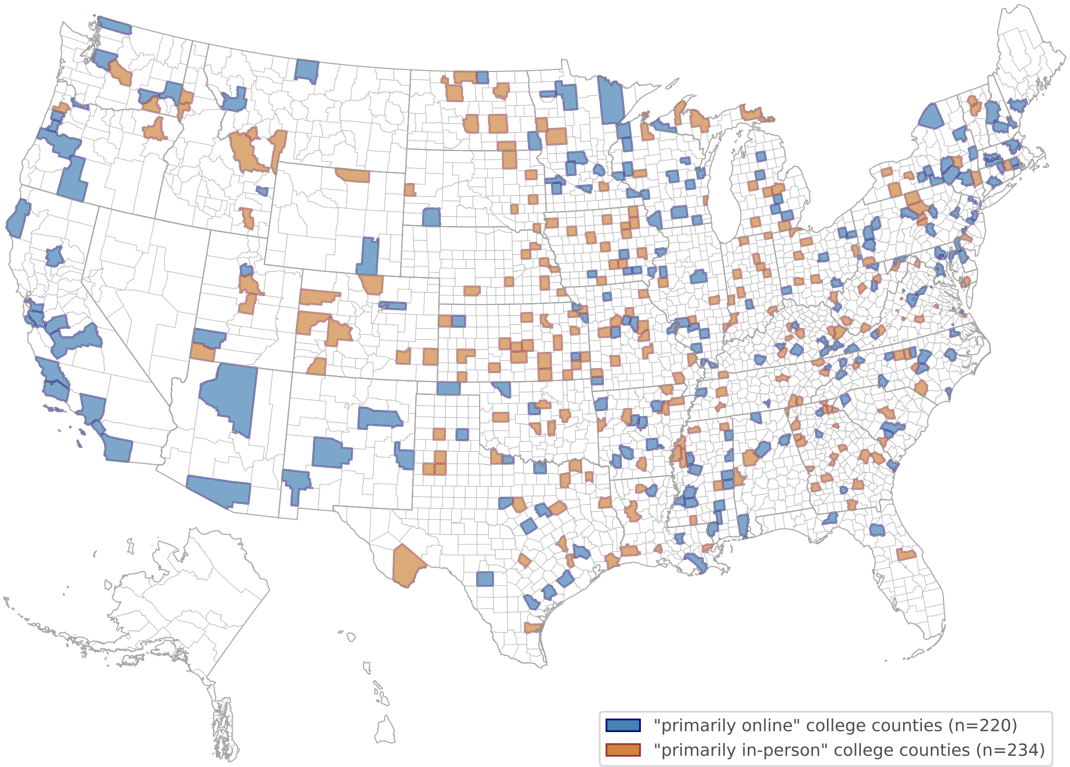


Figure B: Comparison of county-level demographics between groups. Here, we compare the two groups—counties with IHEs that returned primarily in-person for Fall 2020 and counties with IHEs that remained primarily online—based on distributions of **(a)** age, **(b)** race, **(c)** income, and **(d)** urban-rural designation. Error bars: 95% confidence intervals.

Statistical controls for counties with active mitigation policies

To control for potential biases in our construction of county groups—possibly arising due to political affiliations [7], population structure [8], or arise from differences in local mitigation policies—we assigned each county to an "active mitigation policies" score based on policy tracking data from the Oxford COVID-19 Government Response Tracker [9]. These are daily time series data indicating whether or not a number of different policies were active on each day for a given state. Not only does this dataset list the presence or absence of a given policy, it also includes information about the severity (e.g. restrictions on gatherings of 10 people vs. restrictions on gatherings of 100 people, or closing all non-essential workplaces vs. closing specific industries, etc.). From these indicator variables, Hale et al. (2021) define a summary "stringency index" that characterizes the daily intensity of the mitigation policies that a given region is undergoing over time. We include this "stringency index" variable in an Generalized Linear Model regression to quantify the extent to which this time series of policy measures—along with data about IHE testing and enrollment policy, demographic data about the county itself, and average temperature—predicts COVID-19-related deaths (Table 1). The descriptions of the variables used in the regression can be found in Table C and are visualized in Fig. D.

This model predicts deaths per 100,000 county population at a 38-day lead time (i.e., using data from today to predict the number of COVID-19 deaths reported 38 days from


"primarily online" college counties (n=220)
"primarily in-person" college counties (n=234)

https://www.census.gov/geographies/mapping-files.html

Figure C: Map of counties included in matched analysis. With the exception of California, which includes many primarily online IHEs, there are very few regions where the counties are clustered based on campus reopening strategy. County and state boundary maps downloaded from the United States Census TIGER/Line Shapefiles [5].

now), which was selected because it is the lag that maximizes the Pearson χ^2 of the model. Note: this specific value, 38 days, is in line with the CDC's median window of the time between the onset of infection and death [15], and by varying the lead time we do not see any substantial differences in the sign, value, and significance of the coefficients for the different variables. To test for possible multicollinearity, we assigned a Variance Inflation Factor (VIF) to each of the variables in regression. The VIF is defined as $VIF = 1/(1-R_i^2)$; a $VIF = 1$ indicates that a variable is uncorrelated, and VIF between 5-10 suggests that a given variable is highly correlated in the regression. For the regression from Table 1, the VIF values are: "average temperature" $= 1.226$, "urban/rural code" $= 1.951$, "population density" $= 2.244$, "median income" $= 1.422$, "population over 60" $= 1.408$, "2020 voting behavior" $= 1.753$, "IHE fulltime enrollment" $= 1.492$, "IHE fulltime enrollment (online)" $= 4.554$, "IHE fulltime enrollment (in person)" $= 4.042$, "stringency index" $= 1.094$, "county new tests" $= 1.408$, "IHE new tests" $= 1.498$.

There are limitations to the regression used here. For one, the policy data we use is at the state level, while the analyses are at the county level. Part of the reason for this is

variable	**transformation**	**description**	**source**
County demographic variables			
population density	log-scale	number of people per mile^2	[10]
median income	log-scale	median income of residents in the county	[10]
population over 60	per 100,000 county population; log-scale	sum of population in the following age brackets: 60-64, 65-69, 70-74, 75-79, 80-84, 85+	[10]
urban/rural code	code = 4, 5, 6 → 1; code = 1, 2, 3 → 0	Urban-Rural Classification (1: large central metro; 2: large fringe metro; 3: medium metro; 4: small metro; 5: micropolitan; 6: non-core).	[4]
County COVID-19 response			
stringency index (OxCGRT)	none	aggregate indicator of active mitigation policies, see Hale et al., (2021)	[9]
new COVID-19 tests	7-day rolling avg.; per 100,000 county population; log-scale	daily number of COVID-19 diagnostic tests reported in the county	[11]
IHE variables			
fulltime enrollment	per 100,000 county population; log-scale	number of fulltime students enrolled at colleges in our dataset	[12]
fulltime enrollment (online)	per 100,000 county population; log-scale	number of fulltime students enrolled at colleges classified as "primarily online"	[6, 12]
fulltime enrollment (in person)	per 100,000 county population; log-scale	number of fulltime students enrolled at colleges classified as "primarily in-person"	[6, 12]
new COVID-19 tests	7-day rolling avg.; per 100,000 county population; log-scale	sum of daily number of COVID-19 diagnostic tests reported by IHEs within the county	this work
Other			
2020 voting behavior (% Rep.)	$\frac{\text{votes}_{Rep.}}{\text{votes}_{total}} \times 100$	percent of total votes cast for the Republican candidate in the 2020 presidential election	[13]
average temperature (°Celsius)	none	daily average temperature, from the National Oceanic and Atmospheric Administration	[14]
Dependent variable			
new deaths from COVID-19	7-day rolling avg.; per 100,000 county population	daily number of reported deaths from COVID-19 at the county level	[11]

Table C: Description of variables in Table 1. Where appropriate, we use the "per 100k" designation—the variable's value divided by county population, multiplied by 100,000. Here "log" refers to the natural log, which we apply to variables that follow heavy-tailed distributions (e.g. income and population density).

that there remains inconsistent data about county-level policies and more reliable data at the state level. Also, among datasets that do include finer-scale policy data (e.g. [16]), it is often about specific cities, as opposed to counties. Another limitation of using these kinds of policy data in our analysis is that we still do not have information about adherence to

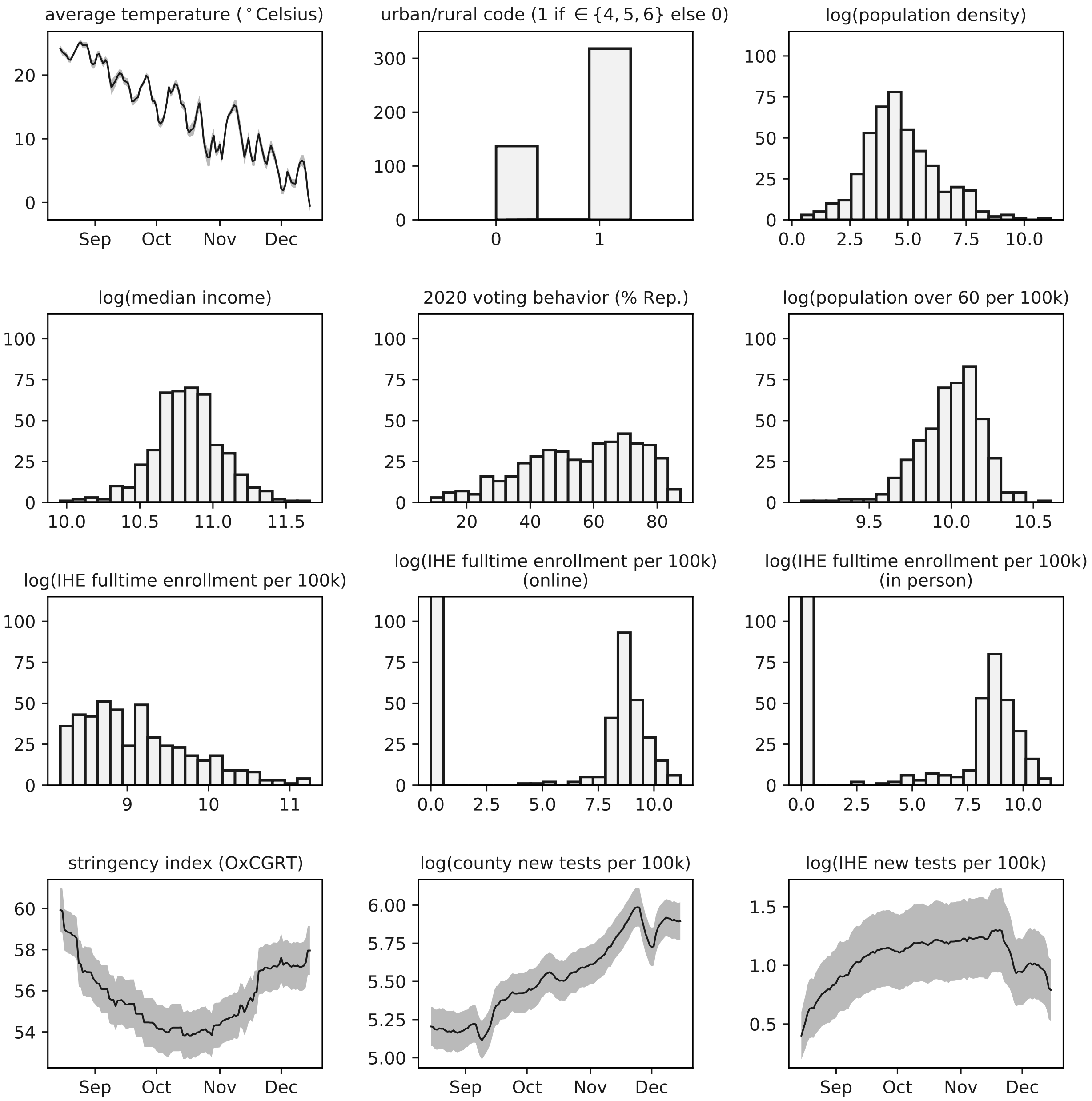


Figure D: Distributions of the variables used in the regression in Table 1.

policies once they are in place. For instance, even in zip codes in states with mask mandates, we do not see full—or consistent—compliance from county to county [17]; similarly, we see differences in social distancing and mobility patterns among counties in the same state [18].

Campus COVID Dataset: Examples of IHE data

The Campus COVID Dataset includes detailed *time series* data about the number of tests and cases reported by the almost 1,000 IHEs. In Figures E, F, and G, we show three examples of IHEs that report daily case counts and tests conducted on campus. The shape of these three IHEs' testing curves exemplifies three broad trends seen in the testing cadence across IHEs during the Fall 2020 semester: For example, Northeastern University (Fig. E) conducts a large number of tests throughout the semester, a rate that remains relatively high throughout September, October, and November. North Carolina State University at Raleigh (Fig. F) reported thousands of "entry tests" early in the semester, followed by a relatively lower volume for the rest of the semester. University of California, Los Angeles (Fig. G) dramatically increased its testing volume during late November, which followed a gradually increasing amount of tests reported throughout the semester. In each of the bar plots in this section, darker color bars indicate accelerating cases/tests.

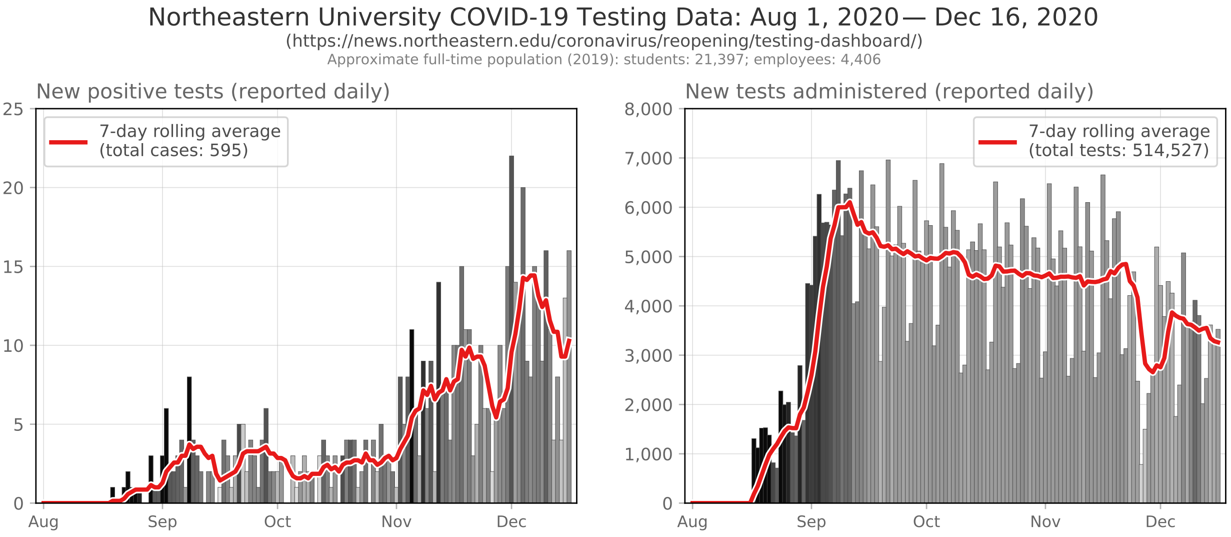


Figure E: Example data: Northeastern University.

Another common trend among the IHEs that conduct and report testing is an increase in testing volume prior to the Thanksgiving holidays (late November), which is when students often return to their homes for several days. For Purdue University, University of Miami, and Georgia Institute of Technology (Figures H, I, and J), we see examples of these this mid-November spike in testing, followed by a sharp decline. This pattern is common across many IHEs, which often requested that students not return following this holiday, as the semester was already almost over and infection rates were beginning to surge in many parts of the United States.

Many IHEs only report cases and test counts in weekly intervals (e.g. Duke University and Ohio State University, Figures K and L), and for hundreds of IHEs, we were only able to collect a summary of the cumulative number of tests conducted and cases reported throughout the semester.

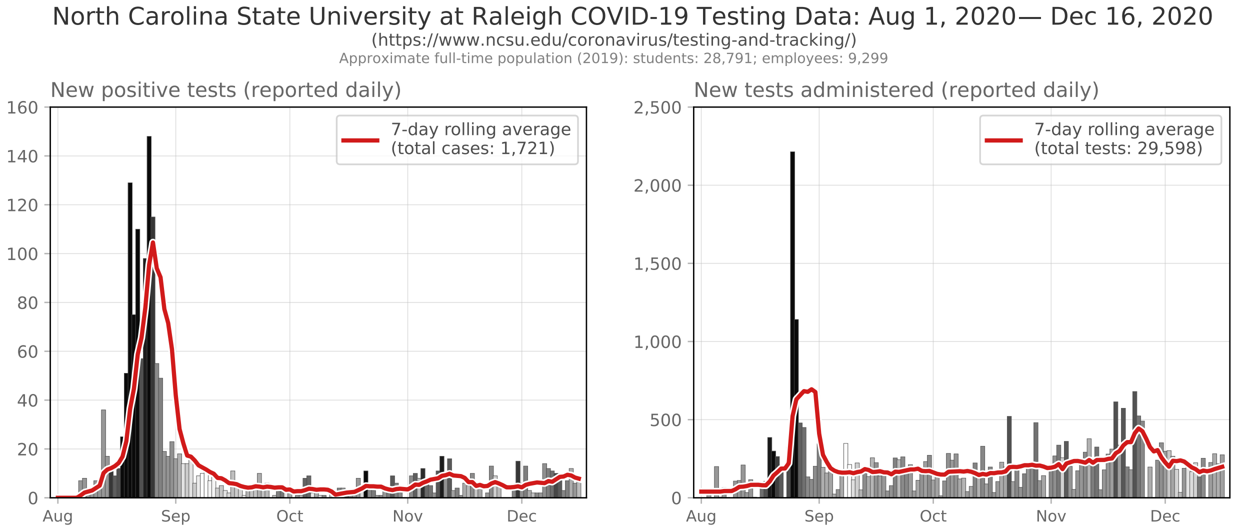


Figure F: Example data: North Carolina State University at Raleigh.

These examples are not meant to be exhaustive—indeed many IHEs in the Campus COVID Dataset do not have time series data for their testing volume or case counts. It is our hope that by releasing these data and analyses that they will motivate other large scale data collection/standardization efforts moving forward. To view examples of time series from other IHEs or to download the entire Campus COVID Dataset, see [19].

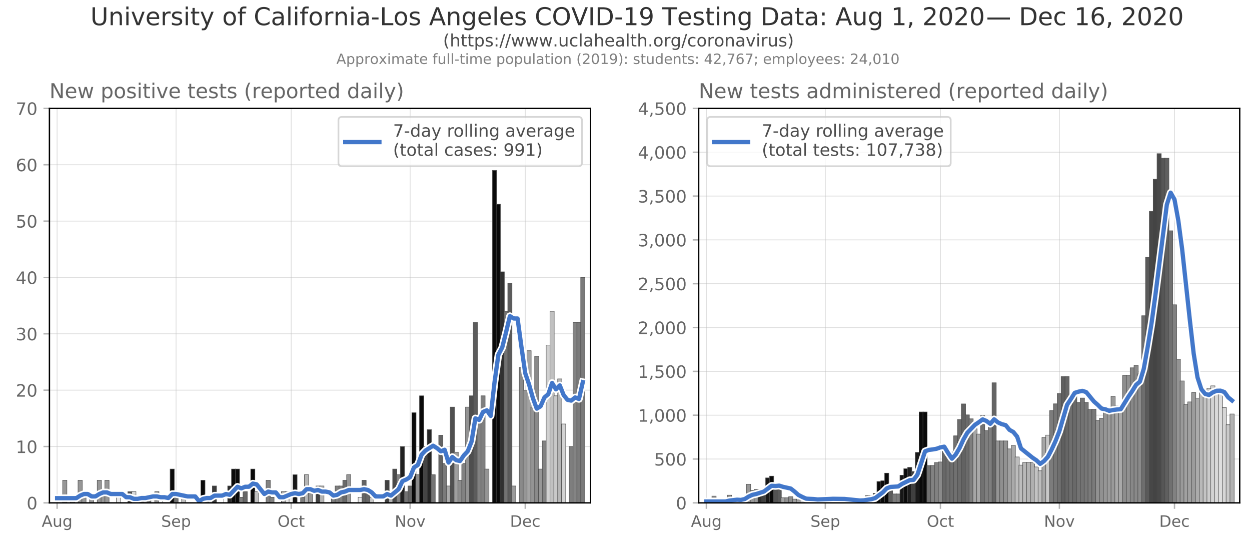


Figure G: Example data: University of California-Los Angeles.

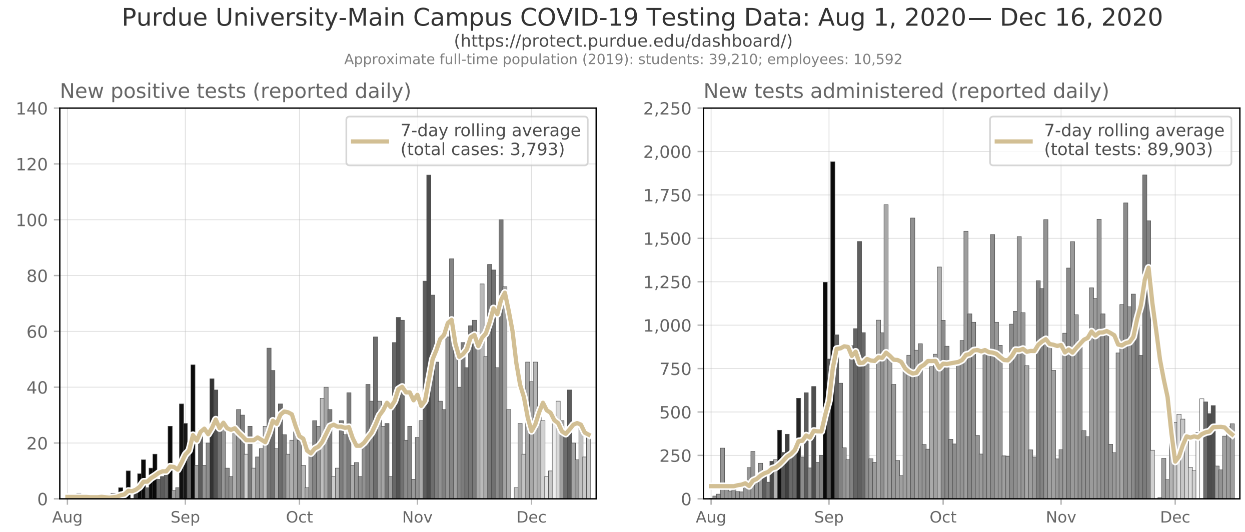


Figure H: Example data: Purdue University-Main Campus.

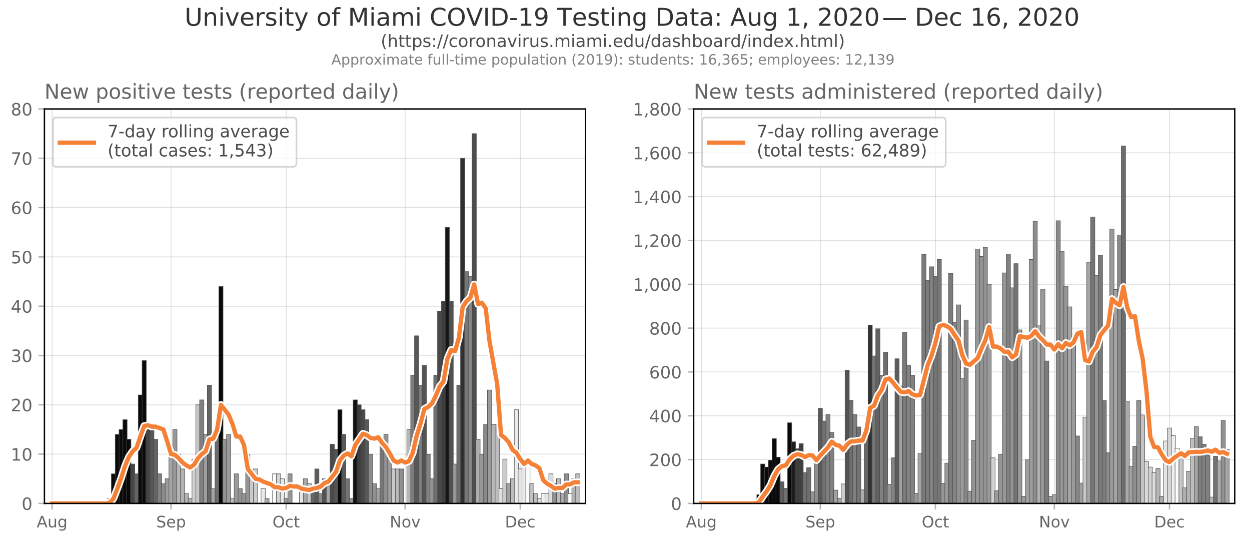


Figure I: Example data: University of Miami.

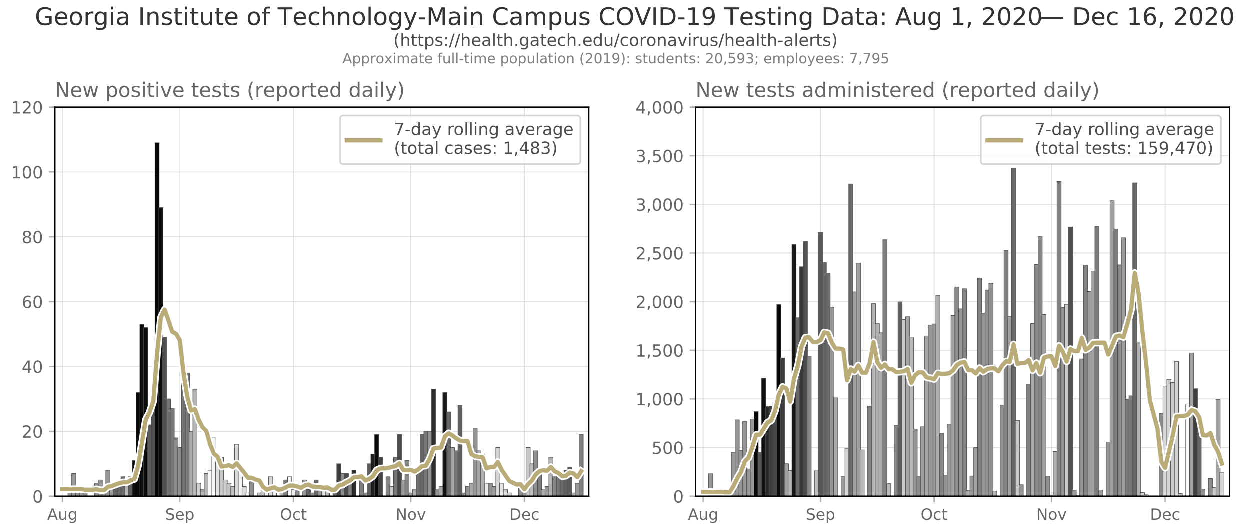


Figure J: Example data: Georgia Institute of Technology-Main Campus.

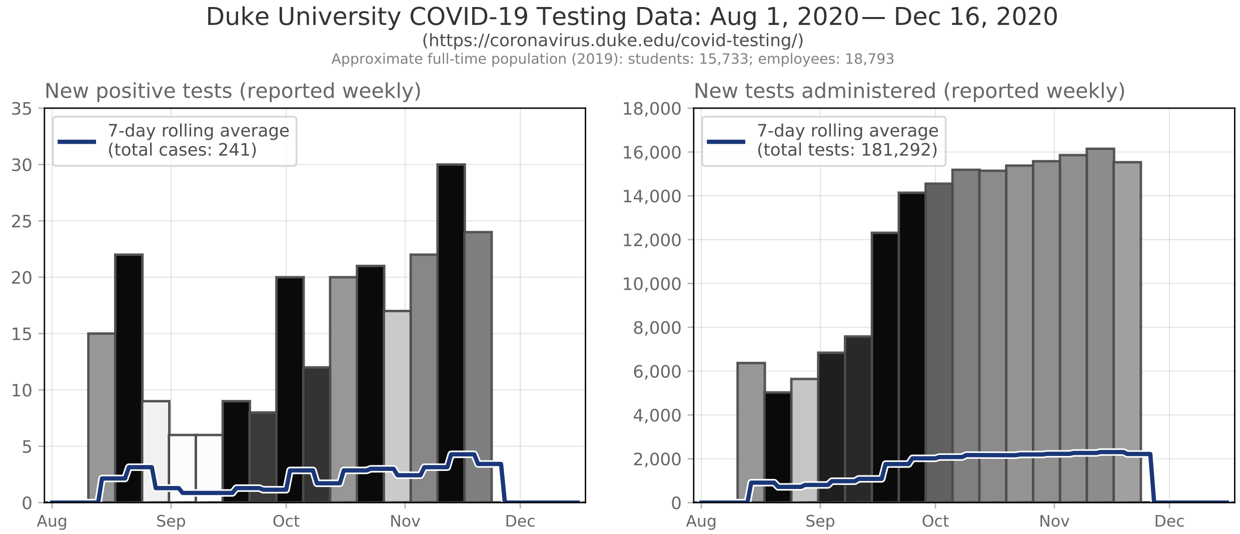


Figure K: Example data: Duke University.

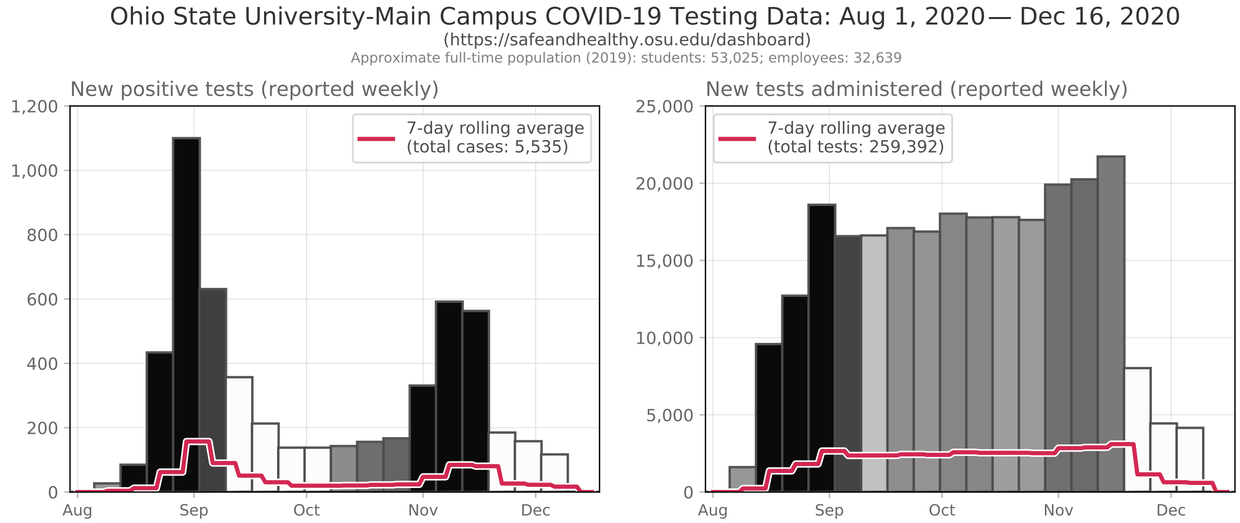


Figure L: Example data: Ohio State University.

References

[1] *googlesearch v.1.0.1*. https://github.com/Nv7-GitHub/googlesearch. 2021.

[2] *gspread v.4.0.1*. https://github.com/burnash/gspread. 2021.

[3] Linda Juel Ahrenfeldt, Martina Otavova, Kaare Christensen, and Rune Lindahl-Jacobsen. “Sex and age differences in COVID-19 mortality in Europe”. In: *Wiener klinische Wochenschrift* 133.7 (2021), pp. 393–398. DOI: 10.1007/s00508-020-01793-9.

[4] National Center for Health Statistics. *NCHS Urban-Rural Classification Scheme for Counties*. https://www.cdc.gov/nchs/data_access/urban_rural.htm. 2013.

[5] United States Census Bureau. *TIGER/Line Shapefiles - 2018*. 2018. URL: https://www.census.gov/geographies/mapping-files.html.

[6] Christopher R. Marsicano, Kathleen Felten, Luis Toledo, and Madeline Buitendorp. “Tracking campus responses to the COVID-19 pandemic”. In: *APSA Preprints* (2020). DOI: 10.33774/apsa-2020-3wvrl.

[7] Jon Green, Jared Edgerton, Daniel Naftel, Kelsey Shoub, and Skyler J. Cranmer. “Elusive consensus: Polarization in elite communication on the COVID-19 pandemic”. In: *Science Advances* 6.28 (2020). DOI: 10.1126/sciadv.abc2717.

[8] Thomas P. Smith, Seth Flaxman, Amanda S. Gallinat, Sylvia P. Kinosian, Michael Stemkovski, H. Juliette T. Unwin, Oliver J. Watson, Charles Whittaker, Lorenzo Cattarino, Ilaria Dorigatti, Michael Tristem, and William D. Pearse. “Temperature and population density influence SARS-CoV-2 transmission in the absence of nonpharmaceutical interventions”. In: *Proceedings of the National Academy of Sciences* 118.25 (2021). DOI: 10.1073/pnas.2019284118.

[9] Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, and Helen Tatlow. “A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)”. In: *Nature Human Behaviour* 5.4 (2021), pp. 529–538. DOI: 10.1038/s41562-021-01079-8.

[10] United States Census Department. *American Community Survey*. https://www.census.gov/programs-surveys/acs. 2018.

[11] Ensheng Dong, Hongru Du, and Lauren Gardner. “An interactive web-based dashboard to track COVID-19 in real time”. In: *The Lancet Infectious Diseases* 20.5 (2020), pp. 533–534. DOI: 10.1016/S1473-3099(20)30120-1.

[12] United States Department of Education. *National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS)*. https://nces.ed.gov/ipeds/use-the-data. 2018.

[13] MIT Election Data and Science Lab. *County Presidential Election Returns 2000-2020*. Version V9. 2018. DOI: 10.7910/DVN/VOQCHQ.

[14] Alisa H. Young, Kenneth R. Knapp, Anand K. Inamdar, William Hankins, and William B. Rossow. "The International Satellite Cloud Climatology Project H-Series climate data record product". In: *Earth System Science Data* 10.1 (2018), pp. 583–593. DOI: 10.5194/essd-10-583-2018.

[15] Centers for Disease Control and Prevention. *COVID-19 pandemic planning scenarios*. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html. 2020.

[16] Georgetown University Center for Global Health Science and Security and Talus Analytics. *The COVID Analysis and Mapping of Policies*. https://www.covidlocal.org/amp. 2020.

[17] Benjamin Rader, Laura F. White, Michael R. Burns, Jack Chen, Joseph Brilliant, Jon Cohen, Jeffrey Shaman, Larry Brilliant, Moritz U.G. Kraemer, Jared B. Hawkins, Samuel V. Scarpino, Christina M. Astley, and John S. Brownstein. "Mask-wearing and control of SARS-CoV-2 transmission in the USA: A cross-sectional study". In: *The Lancet Digital Health* 3.3 (2021), e148–e157. DOI: 10.1016/S2589-7500(20)30293-4.

[18] Brennan Klein, Tim LaRock, Stefan McCabe, Leo Torres, Lisa Friedland, Maciej Kos, Filippo Privitera, Brennan Lake, Moritz U.G. Kraemer, John S. Brownstein, Richard Gonzalez, David Lazer, Tina Eliassi-Rad, Samuel V. Scarpino, Alessandro Vespignani, and Matteo Chinazzi. "Quantifying collective physical distancing during the COVID-19 outbreak". In: *preprint* (2022). URL: https://www.mobs-lab.org/uploads/6/7/8/7/6787877/covid19mobility_report2.pdf.

[19] Brennan Klein. *jkbren/campus-covid: campus-covid*. Version v1.0. 2021. DOI: 10.5281/zenodo.5395340.