CITADEL/CORRELATION ONE DATATHON

COVID-19 Data Availability in the US: An Analysis of the Effects of Racial Disparities in Reporting Across the United States

Marlin Figgins[1] | Luca Pistor[2] | Sid Rastogi[3] | Nick Zemtzov[3]

[1]University of Washington, Seattle, WA, 98195, USA

[2]Stanford University, Stanford, CA, 94309, USA

[3]Harvey Mudd College, Claremont, CA, 91711, USA

Correspondence
Email:
mfiggins@uw.edu,
lpistor@stanford.edu
srastogi@hmc.edu
nzemtzov@hmc.edu

The course of the COVID-19 pandemic has largely been shaped by questions of data. In the light of the emerging pathogen, the availability of data becomes a question of life or death. As a way to ask what data has been made available and what effect data has had on the course of the pandemic, we interrogate the public data sets available at the state level in the United States. To this aim, we ask whether the public availability of high quality data affects the downstream outcome of an epidemic. We accomplish this by developing a metric of "public data availability" which takes into account both the presence of various forms of data as well as external checks on the quality of this data based on heuristics about disease spread. Using a Bayesian model, we show that as racial and ethnic data has become more available, racial and ethnic gaps in the growth rate of new cases have declined.

1 | BACKGROUND

The course of the SARS-CoV-2 pandemic has, in large part, been charted by the availability of data and the extent to which this data informs government policy. With COVID-19 being considered by some [1] to be "humanity's first data-driven pandemic," our ability to recover from its effects will be helped or hampered by access to quality data on the individual and institutional levels [2].

Policymakers walk a tightrope when determining the extent to which they ought to prioritize open access to COVID data at the individual level, and this decision making can have tangible effects. For example, the prevalence of cell phone location data has been an enormous boon to governments wishing to track the movement of their citizens to better contain the spread of disease [3]. However, providing access to individual-level location data can present unique challenges in a world in which data anonymization is sometimes viewed as an afterthought. Providing granular data at the individual level has led to discrimination against LGBT communities in South Korea when an individual linked to a super-spreader event was found to have visited a gay bar in Seoul [4], highlighting the potential risks associated with publishing easily de-anonymizable data on individual movements.

Conversely, the U.S. COVID-19 response has largely been hampered by a lack of access to fine-grained data, particularly when it comes to treatment disparities along racial and ethnic lines [5, 6]. Previous studies of epidemics for diseases such as mumps [7] highlight the ways in which knowledge of racial and ethnic breakdowns can be important in understanding localized transmission chains of viruses like SARS-CoV-2. In particular, it has been shown that epidemics can be largely localized to ethnic minority communities and public availability of this data can affect the extent to which individuals can evaluate their personal or community's risk of disease. In particular, we argue that the absence of race-disaggregated data serves to widen the gap in epidemic outcomes between ethnic minority populations and the U.S. population as a whole. We approach this question by analyzing the relationship between state-level racial disparities in COVID-19 deaths over time, and the date from which the state in question began publishing race-disaggregated data on COVID-19 deaths. We correct for the overall trend in deaths on the state level, focusing our analysis on the relationship between state-level data availability and racial disparities in epidemic outcomes.

2 | DATA EXPLORATION

2.1 | Datasets

The majority of our data came from The COVID Tracking Project by *The Atlantic*. In particular, we made use of their state-level time series data on the racial and ethnic composition of case and death counts. We also used data on racial population distributions by state, published by the *Kaiser Family Foundation* [8].

2.2 | Quantifying availability

The COVID Tracking Project assigns every state a score that is intended to serve as a measure of data quality and availability. The score is a categorical letter grade based on whether the state provides data for a number of different, equally-weighted metrics. In our exploratory data analysis, we mapped these categories to numerical values (0 being unavailable, 1 being available) in order to assess whether there was a relationship between data availability and COVID-19 cases and deaths across the country.

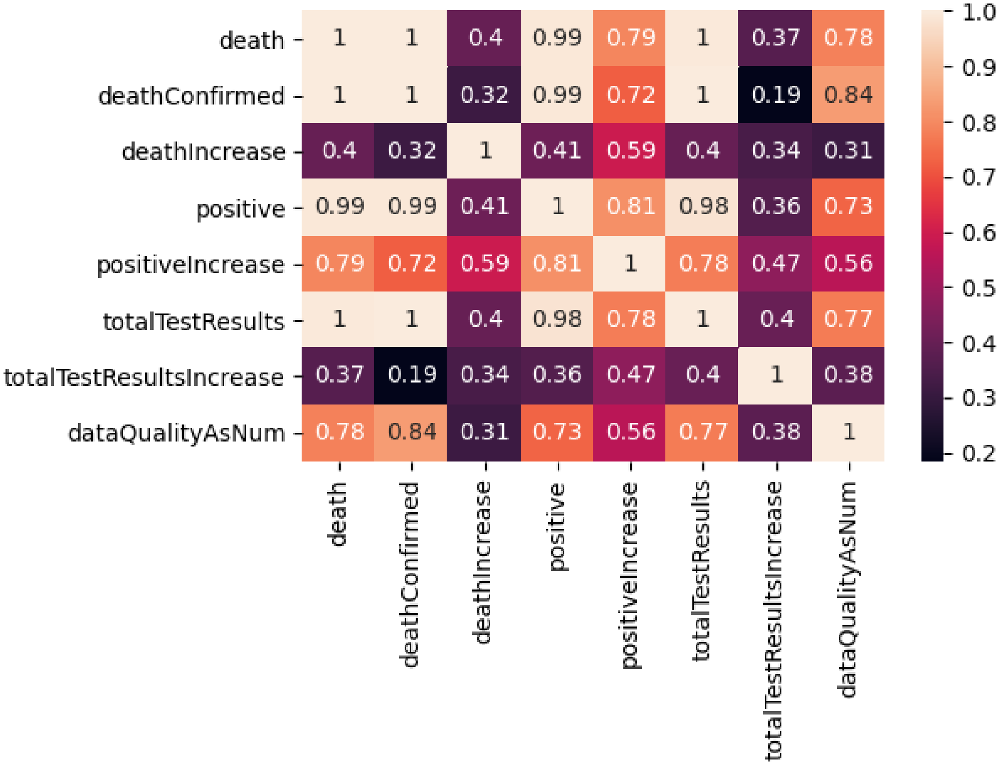

FIGURE 1 Visualizing correlations between the COVID Tracking Project's data quality score and various variables they provide.

The correlation matrix in Fig. 1 showed that there were potentially strong relationships between factors like number of cases/deaths and the amount of data that was made available. However, the data quality grade from the COVID Tracking Project didn't provide historical information on data availability, so we decided to develop our own metric in order to analyze the effect of the availability of data over time.

Using their state-by-state historical data, we constructed a new data set, recording the first date when certain metrics became available. In particular, we focused on the availability of data regarding deaths, current hospitalizations, ICU patient counts and ventilator patient counts, total recoveries, and nursing home deaths. We did not take the availability of testing data into account due to differing standards in testing and reporting across states. Our derived availability score falls within the range $[0, 1]$, and represents the percentage of variables of interest that were made available on a given day, with each individual variable being weighted equally. We then plotted this data availability score for every day dating back to the first confirmed case of COVID-19 reported by each state.

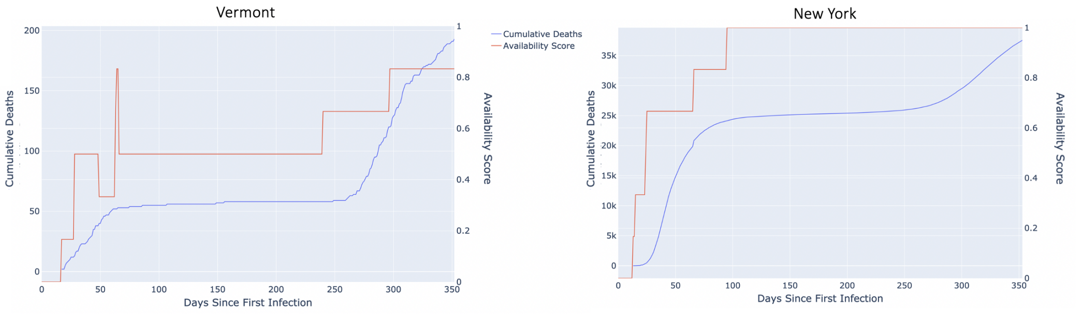

FIGURE 2 Visualizing data availability and cumulative deaths in Vermont and New York. Notice the correlation between the summer intermission and availability of new data.

3 | STATISTICAL ANALYSIS

We focus our analysis on ten states (Florida, Michigan, Montana, Ohio, Colorado, Oklahoma, Arkansas, New York, Vermont, and Washington) which span much of the range of geographical, political, and socioeconomic diversity in the U.S. We chose to focus our analysis to allow for a cross-sectional study of a small number of states employing different strategies against the COVID-19 pandemic. Our ten states were chosen for the balance that they strike between presenting similar categories of data over similar periods of time, while being sufficiently different in terms of testing intensity, political disposition, and mitigation strategies. We then subset to 5 states, carefully chosen based on our results in 3.1.

3.1 | Covariance of data accessibility and epidemic outcomes across states

We analyze the covariance between data accessibility and daily deaths our set of states. The daily deaths are right skew so rather than use mean like a typical covariance calculation, we use median as it better captures the center of the distribution.

$$\text{cov}(x, y) = \frac{\sum(x_i - \tilde{x})(y_i - \tilde{y})}{N - 1}$$

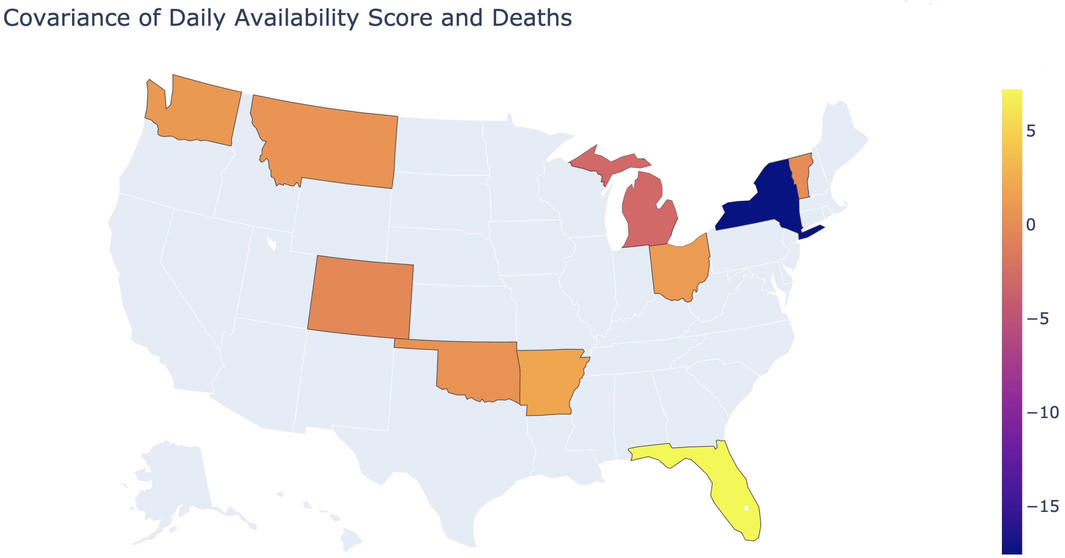

FIGURE 3 Visualizing correlation between data availability and deaths for selected states.

If there were a inverse correlation between deaths and data availability, as hypothesized, we would expect the correlation to be negative. However, the majority of states had positive covariance, and the covariance for all states combined was also positive at just over 2.0.

3.2 | Effective Reproductive Number by race across states

We analyze the effective reproductive number $R(t)$ for different racial groups on the state-level using data provided by the COVID Tracking Project. The effective reproductive number $R(t)$ can be understood as the average number of cases currently observed cases at time t will later produce. Using empirical estimates of the serial interval g i.e. the timing between successive cases in a transmission chain [9], we develop a Bayesian model of this using the race stratified case counts in each state. In short, we can think of modeling the number of confirmed cases $C_X(t)$ of race X on day t as

$$C_X(t) = \sum_{i=1}^{M} R_X(t-i) C_X(t-i) g(i),$$

where $g(i)$ is the probability that a case from i days ago causes a case today. In particular, we estimate $R_X(t-1)$ as a piece-wise constant function which changes at $k = 25$ time points. Assuming a random walk prior on $\log(R_X(t))$, we drew 1000 samples from the posterior distribution for each of race and state combination depicted in Figure 4 using the No-U-Turn Sampler [10]. This analysis was implemented in Julia using the Turing package [11]. The code for our implementation can be found at https://github.com/Luca-Pistor/datathon2021.

Intuitively, an $R(t)$ value greater than one corresponds to a growing epidemic while $R(t) < 1$ corresponds to a declining epidemic. Fitting this model to the case counts in 5 different states to see how observed racial differences in the case reproductive number evolved in time after the point racial and ethnic data became publicly available. We see that in most states the racial differences between case reproductive rate begin to shrink over time.

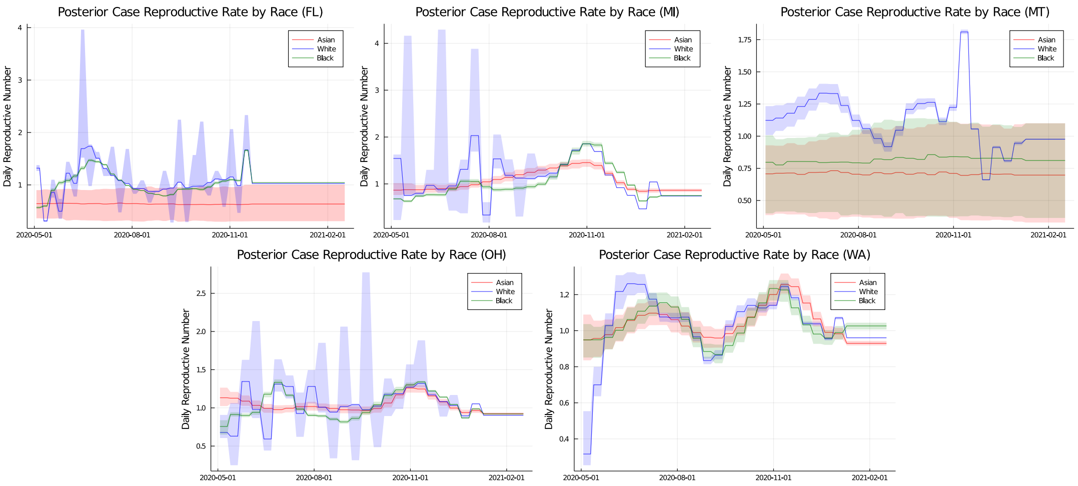

FIGURE 4 Visualizing the effective reproductive number in Florida, Michigan, Montana, Ohio, and Washington state. Notice due to the lack of available data on Asian people in Florida and the low counts of Asian and Black people in Montana, $R(t)$ cannot be reliably inferred. Shaded regions denote the 95 % credible interval for each $R(t)$ value.

3.3 | Discerning over- and under-representation in COVID outcomes by race.

In order to properly consider the implications of our analysis, it was important to consider our results in light of the degree of representation of particular racial groups. For the five states we chose to focus on (Florida, Michigan, Montana, Ohio, and Washington), we began by simply examining the percentage of cases by race and ethnicity (Asian, Black, White; Hispanic, non-Hispanic) over time. Since the proportion of cases whose race/ethnicity was Native American and Alaskan Natives, multiracial, or unknown was considerably lower than other racial groups as well as inconsistency in the usage of these categories between states, we chose to remove them from the analysis.

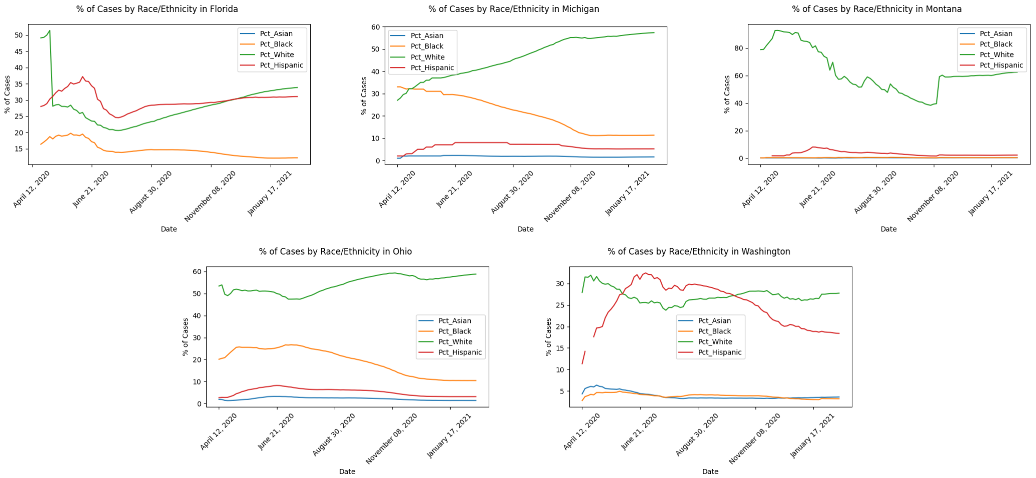

FIGURE 5 Percentage of cases by race and ethnicity (Asian, Black, White; Hispanic, non-Hispanic) over time in Florida, Michigan, Montana, Ohio, and Washington.

However, we then decided to adjust these numbers based on the proportion to see the proportional representation of each racial group according to their fraction in the states population. We call this metric the proportional representation score of racial group X. In terms of deaths due to COVID, this can be represented mathematically as

$$\text{PR}_X = \frac{\text{Deaths of Group X}}{\text{Expected Deaths of Group X}}$$

where we've defined the expected deaths of group X as the total deaths in a population multiplied by the proportion of the population that belongs to group X. Group X having a proportional representation score greater than one would mean that group X is over-represented in the number of deaths due to COVID in a given state. The resulting graphs are shown in Fig. 6. We performed the same analysis on COVID-19 deaths in these states to obtain the the graphs in Fig. 7. It is important to note that there is some missing data in states like Florida and Montana, which prevents us from drawing some conclusions about racial/ethnic breakdowns in these states. However, the racial disparities in health outcomes remain clear. After adjusting for the racial breakdown of the state populations, we find that racial disparities in both cases and deaths appear to decrease as time goes on.

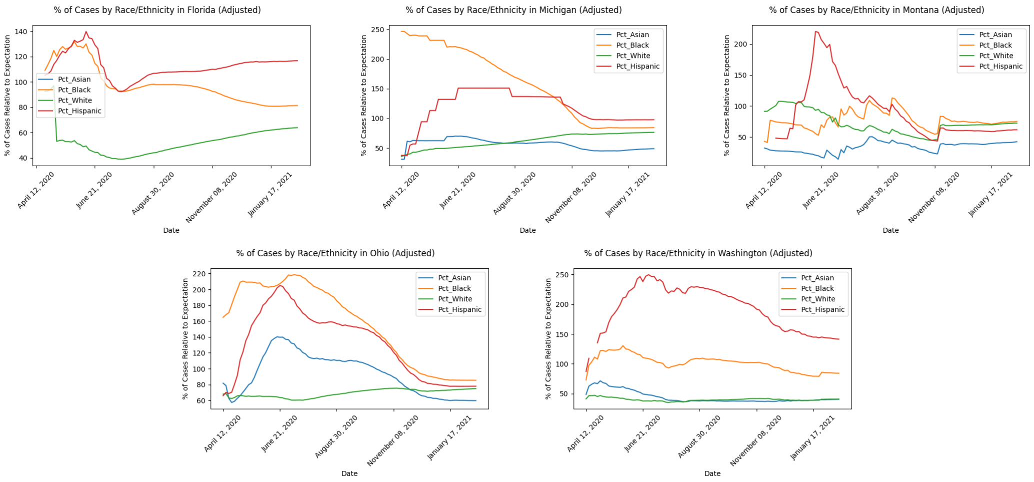

FIGURE 6 Proportional representation score of cases by race and ethnicity (Asian, Black, White; Hispanic, non-Hispanic) over time in Florida, Michigan, Montana, Ohio, and Washington.

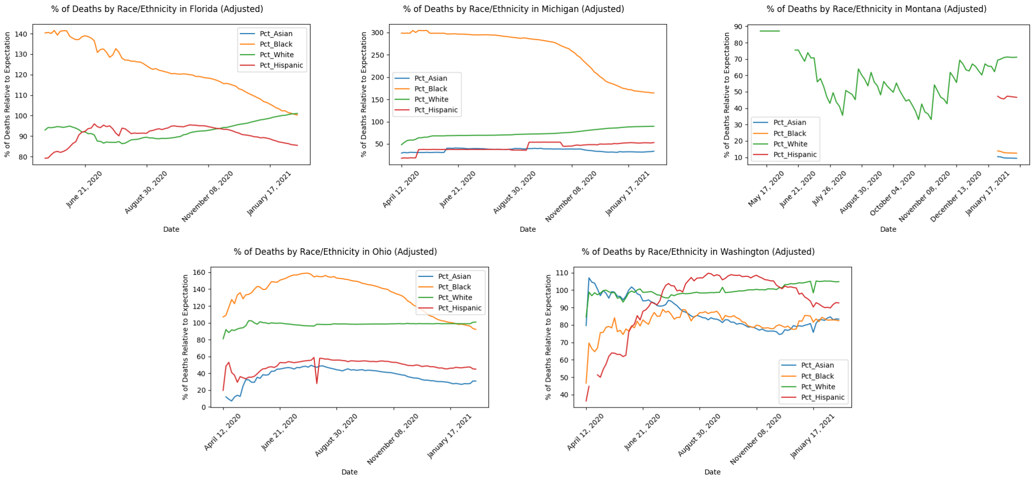

FIGURE 7 Proportional representation score of deaths by race and ethnicity (Asian, Black, White; Hispanic, non-Hispanic) over time in Florida, Michigan, Montana, Ohio, and Washington.

4 | CONCLUSION

We initially hypothesized that as more data regarding the state of the COVID-19 pandemic became available, individuals and institutions would make better informed decisions, resulting in relatively fewer deaths. We found no evidence for this being the case using our generalized metric for data availability. Given the positive covariance between data availability and deaths, we might draw a related conclusion however: that as deaths increase and pandemic-related infrastructure develops, state governments are obliged to (and better able to) provide data to the public. This manifests in the common trend of states to not publish new data metrics over the summer. Looking on the state level, the availability of data stratified by racial and ethnic categories has a strong relationship with the growth rates of cases, so there is evidence for our initial hypothesis when assessing the availability of race and ethnic data.

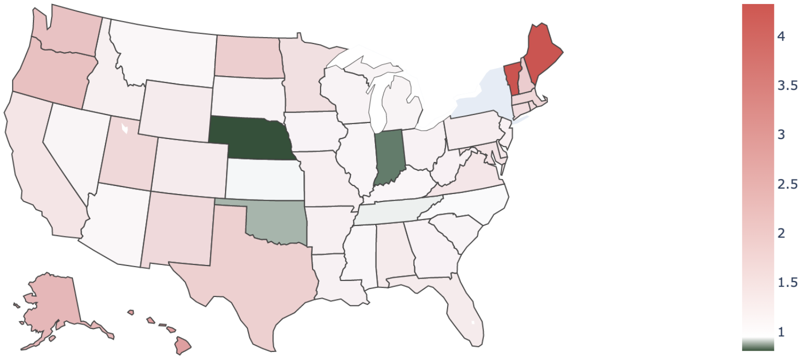

FIGURE 8 This chloropleth map represents the disparities in epidemic outcomes across racial lines using the proportional representation score for Black Americans in each state. Across the United States, there are only six states where Black Americans do not test positive for coronavirus at proportionally high rates. Only 10.9% of Black Americans live in one of these six states.

We find that racial and ethnic gaps in the growth of cases have shrunk as data has become more available over the course of the epidemic. However, we find persistent differences in the number of cases, as well as case outcomes between individuals of different ethnic and racial backgrounds in several states. This is especially pronounced when we consider disparities in case outcomes between Black and non-Black Americans, which are visualized in 8. What is notable is that these gaps appear to be shrinking through time, which holds with the trend that we have observed throughout the states that we've analyzed. Thus, as we strive toward increasing equity of outcomes across different levels of our society, providing race-disaggregated health data is an essential measure to support our understanding of racial substructure in population dynamics and to provide support for communities at the individual level.

references

[1] Rocha R, The data-driven pandemic: Information sharing with COVID-19 is 'unprecedented' | CBC News. CBC/Radio Canada; 2020. https://www.cbc.ca/news/canada/coronavirus-date-information-sharing-1.5500709.

[2] Piller C, Data secrecy is crippling attempts to slow COVID-19's spread in U.S., epidemiologists warn; 2020. https://www.sciencemag.org/news/2020/07/us-epidemiologists-say-data-secrecy-covid-19-cases-cripples-intervention-strategies.

[3] Rader B, Scarpino SV, Nande A, Hill AL, Adlam B, Reiner RC, et al. Crowding and the shape of COVID-19 epidemics. Nature medicine 2020;26(12):1829–1834.

[4] Borowiec S, South Korea's Nightclub Outbreak Shines Unwelcome Light on LGBTQ Community. Time; 2020. https://time.com/5836699/south-korea-coronavirus-lgbtq-itaewon/.

[5] Vestal C, Lack of Public Data Hampers COVID-19 Fight; 2020. https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2020/08/03/lack-of-public-data-hampers-covid-19-fight.

[6] Melamed C, Data sharing to fight COVID-19; 2020. https://www.sustainablegoals.org.uk/data-sharing-to-fight-covid-19/.

[7] Moncla LH, Black A, DeBolt C, Lang M, Graff NR, Pérez-Osorio AC, et al. Repeated introductions and intensive community transmission fueled a mumps virus outbreak in Washington State. medRxiv 2020;https://www.medrxiv.org/content/early/2020/10/21/2020.10.19.20215442.

[8] Population Distribution by Race/Ethnicity; 2020. https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/.

[9] Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. International Journal of Infectious Diseases 2020;93:284–286. https://www.sciencedirect.com/science/article/pii/S1201971220301193.

[10] Hoffman MD, Gelman A, The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo; 2011.

[11] Ge H, Xu K, Ghahramani Z. Turing: a language for flexible probabilistic inference. In: International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain; 2018. p. 1682–1690. http://proceedings.mlr.press/v84/ge18b.html.