Ethical Dilemmas of Analyzing and Predicting Outbreak of Infectious Diseases: A Study of Covid-19

Anika Tabassum°, Berna Oztekin-Gunaydin*, Rahel Tekle*, Shailik Sarkar°, Vasanth Reddy Baddam°
°Department of Computer Science, Virginia Tech
*Department of Urban Planning, Virginia Tech
Email:{anikat1, boztekin, ertekle, shailik, vasanth2608}@vt.edu

ABSTRACT

This study aims to understand the impact of the COVID-19 outbreak in the New York city metropolitan area and ethical dilemmas related to data privacy, contact tracing, mitigation and response to the novel coronavirus. Novel Coronavirus' impact varies not only geographically, but also across different socio-demographic economic status (elderly, low-income groups, Hispanics and African Americans). Data analysis is focused on: 1) **Socio-demographic/economic** factors such as race, ethnicity, age, population density, housing and employment. Further, the analysis highlights how these factors correlate with higher risk of infection and presents mitigation challenges such as staying at home, social distancing and telecommuting because low income populations and African American and Hispanic populations are mostly frontline low income employees at grocery stores and restaurants, and transit as drivers. 2) **Systemic health disparities** including the: lack of access to health insurance and historic health disparities are some of the results of diabetes and cardiovascular diseases associated with African Americans, Hispanics, and Native Americans, which place these populations at higher risk of COVID-19infections. 3) **Mobility**, intervention strategies include, stay at home orders, telecommuting and social distancing. Using publicly available data for the New York city metropolitan area, we were able to analyze data and present results a predictive model, mitigation and response considering historically underserved populations. We reviewed existing literature to understand methods and results. We analyzed data to determine how socio-demographic/economic factors such as race, income, jobs, housing and mobility place different group populations at a different risk. We analyzed the spread of infections in relation to the different measures taken by the New York authorities such as stay home restriction, mobility and hospitalization data. For these two tasks, we aim to look at whether there are any ethical issues with the data, and address them (if there is any) in our research paper. The data may involve various ethical dilemmas with potential impacts on real time decisions. Through this study we hope to gain deep knowledge about societal factors that contribute to ethical dilemmas and provide possible solutions. For example, it is important to note how predictive models can be used to determine the outbreak of COVID-19, and identify high risk populations and areas. In addition, as ethical questions arise, because prediction, response and preventative measures affect individuals, communities, stakeholders and regions differently. This study will provide insights to strategically predict, employ resources, and preventative measures for high risk areas and vulnerable populations.

All codes and datasets are uploaded in Github[1]

KEYWORDS

datasets, prediction, ethical concerns, bias, coronavirus, Covid-19, health outbreak

1 INTRODUCTION

The recent outbreak of the novel coronavirus in the United States and throughout the world has created major social and economic disruption. It has become incredibly important for the governments globally to come up with effective ways to contain the virus, mitigate effects of such spread and come up with strategies to deploy resources, predict and forecast to minimize high risk areas. In the United States, the concentration and impact of the coronavirus spread varies not only geographically, but also demographically and socioeconomically. In this study we analyzed COVID-19 data for the New York City metropolitan area to understand the impact of this infectious disease, how it correlates with high population density, housing, jobs and mobility and how these socio-economic aspects minimize or put individuals at higher risk.. We compared the state-of-the-art analysis to understand and explore ethical issues that may arise from data privacy, contact tracing, building

2 DATASETS

Socio-economic: The data on COVID-19(2019-nCoV) has been imported from GitHub repository managed by by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)[1]. Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL). This dataset has information on the number of infected cases, recovered and deaths from 2019 novel corona virus. This is a time-series data and has the cumulative value of the above mentioned attributes for each each from 22nd Jan 2019.

Dataset has the following attributes and their description given below:

- Sno - Serial Number
- Province/State - The specific place where cases are reported
- Country/Region - Country of given region
- Lat - Latitude of specific region
- Long - Longitude of specific region
- Observation date - Date of the observation in MM/DD/YYYY

Another dataset that we use for analyzing socio-economic aspect of the pandemic spread is the us census data that we got from

[1]https://github.com/vbaddam/Ethics-Project—COVID19

the zipatlas website. This dataset describes the average household income of each of the city of New York. These are the following attributes of the dataset:

- Zip Code
- Location
- City
- Population
- Avg. Income/H/hold

We aim to combine the CDC released data and the aforementioned data for our eventual analysis of the socio-economic aspect.
Mobility: We collect 2 types of mobility data, DescretaLabs [5] and Apple [3]. Each value is a daily incidence mobility from Jan-early May. of a random of group of people at states (county) level. The value is the median distance travelled by a person of the group. For each region, the group is selected randomly for recording the incidence.

3 METHODOLOGY

3.1 Data Pre-processing

3.1.1 Cleaning Data: As we extract the raw data from GitHub repository, we needed to clean the data and fill out the missing values in the table. For the income related dataset we collect the data from zipatlas.com and then we perform the same pre-processing steps.

3.1.2 Grouping Data: After we have gathered data from these data sources and cleaned them we combine them pairwise according to the analysis we want to perform. 1. We combine the income data with the CDC data that gives us zipcode wise information about the number of positive cases and tested cases, into one single table.

3.2 Exploratory Data Analysis

Exploratory Data Analysis alludes to the basic procedure of performing beginning examinations on data in order to find patterns, to spot anomalies, to test the theory, and to check suppositions with the assistance of outline measurements and graphical representations.

3.3 Exploring Socio-economic Disparity

After getting the table having the combined information of health and income data, we have grouped all the zipcodes in the city to 8 clusters. To decide on the optimal number of clusters we take the help of K-means elbow plot(shown in figure 1). After getting the clusters we try to visualize the average number of positive cases in each one of them to find out if there is any disparity in the number of positive cases depending on their economic strength.

3.4 Exploring health disparity

In this, we will look the hospitalisation data for different section of people based on their sex, race, age. Then we will analyse the data and look the bias or disparity towards certain section of people.

3.5 Analyzing Mobility

With the mobility data, we aim to explore several questions: (i) Does mobility differs in different region in terms of population density and race? (ii) How much interventions are affecting different types of mobility (e.g., transit, driving, parking)? (iii) How much mobility in different utilities such as, groceries, park, home, etc., has changed from standard level due to covid cases?

3.6 A study on Contact-Tracing technology

We also conduct a study on the contact tracing technologies in US. We explore several questions, such as, (i) Types of different contact-tracing technologies, (ii) How much the data anonymized while sharing information, (iii) Is there any privacy risk of an individual or business due to using such technology? Popular article 'Wired' [7] points different views on using contact tracing technologies, mobile network to track user contact history, and facial recognition app to track outbreak from social media contents. They point out different context of ethics and how US government is planning to maintain that. Further, during emergency govt has decided to broaden its power for contact tracing app.

The most popular contact-tracing technology have been proposed by joint project of Google and Apple [2]. Apple provides mobility data for different regions based on county level. They claimed their data is maintaining anonymity and not tracking or disclosing any user information. They select random users on a rolling basis for tracking mobility and their objective is to help policy makers for social interventions decision-making with this data. Their tool is a 2-phased plan where first phase needs to download an app to record the close proximity anonymously. For maintaining privacy, they will keep an anonymous individual and broadcast token which will be changed in an interval. The 2nd phases requires no app download and can directly get the proximity and broadcasting in OS level.

4 EXPERIMENTAL RESULTS

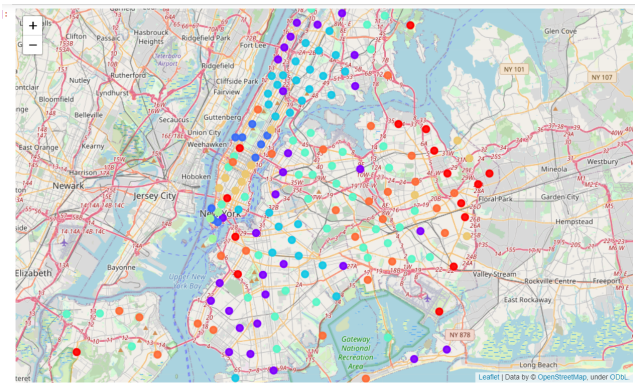

Figure 1: NYC Zipcodes visualized according to average household income

4.1 Socio-economic Disparity

In our experiment we grouped all the zipcodes into 8 clusters according to the average household income. To get a better understanding of the geographic location we visulaized them in the actual map as shown in fig-1 . During this experimentation we managed to find that even though the number of positive cases does not vary

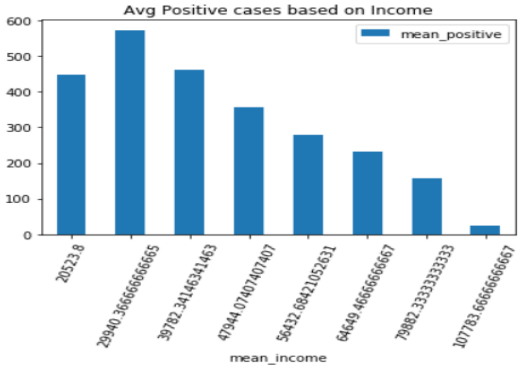

Figure 2: No of Positive cases in each income based clusters

much for first few of these clusters, as soon as the average income crosses a certain thresh hold the number of affected cases drastically decreases. The findings are displayed in fig-2

4.2 Analyzing Mobility

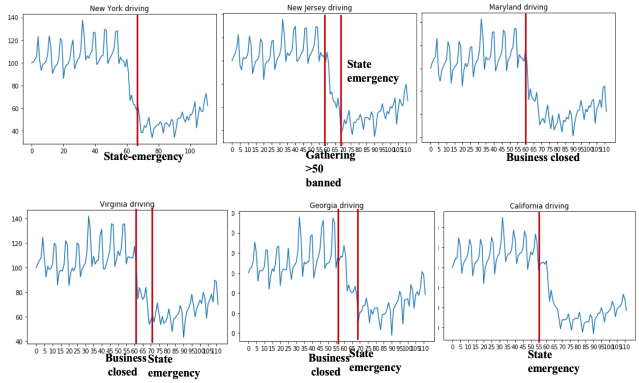

Figure 3: Intervention effects on mobility on major cities in US

We use three analysis in terms of mobility-

4.3 Effect of mobility on Covid-cases

Using DescartesLabs [5] we analyze how the mobility is changing in five different boroughs in NYC and if they have any indirect impact on covid-case counts in those regions. Fig. **??** shows that although mobility in all five boroughs decreased after March 21 (with the declaration of state emergency), it was a lot higher in Queens and Staten Island than Manhattan, Brox, and Brooklyn. Fig. 5 evidents that covid cases are also higher in Queens and Staten Island much more than other boroughs. Further cases in Manhattan is lowest where mobility was also lower before March 20. From this we observe that mobility might have an indirect impact on increase of covid cases.

4.4 Effect of interventions on mobility

We plot mobility of six major cities, New York, New Jersey, Maryland, Virginia, Georgia, and California. Also, we label some major

Figure 4: Mobility Incidence in different POI in NY and NJ.

interventions (state-of-emergency, closure of business) on every states considering them as an inducing changepoint on mobility. We observe that in most of the states mobility starts changing (has a sharp decrease) due to business closure at the very first and after that the mobility pattern becomes stable but with lower magnitude. CA has only one intervention (state emergency) and they reacted to sharp decrease just after that. On the other hand, although NY had several interventions before state-emergency, it only decrease with the state-of-emergency enactment. These findings can give some keen observation and helpful on type of policies to take for different states.

4.5 Change of Mobility in different point of interest

using Google Mobility survey reports we analyze with Fig. 4 how mobility has changed in different points of interest (POI) in NY and NJ since the state emergency (March 21). We observe that other than residence, in every POI mobility decreased from standard incidence (in both the regions). However, the interesting obeservation is at the end of April, just after 1 month of emergency mobility in parks increase very high ($60 - 120\%$), although in other POIs mobility remain stable.

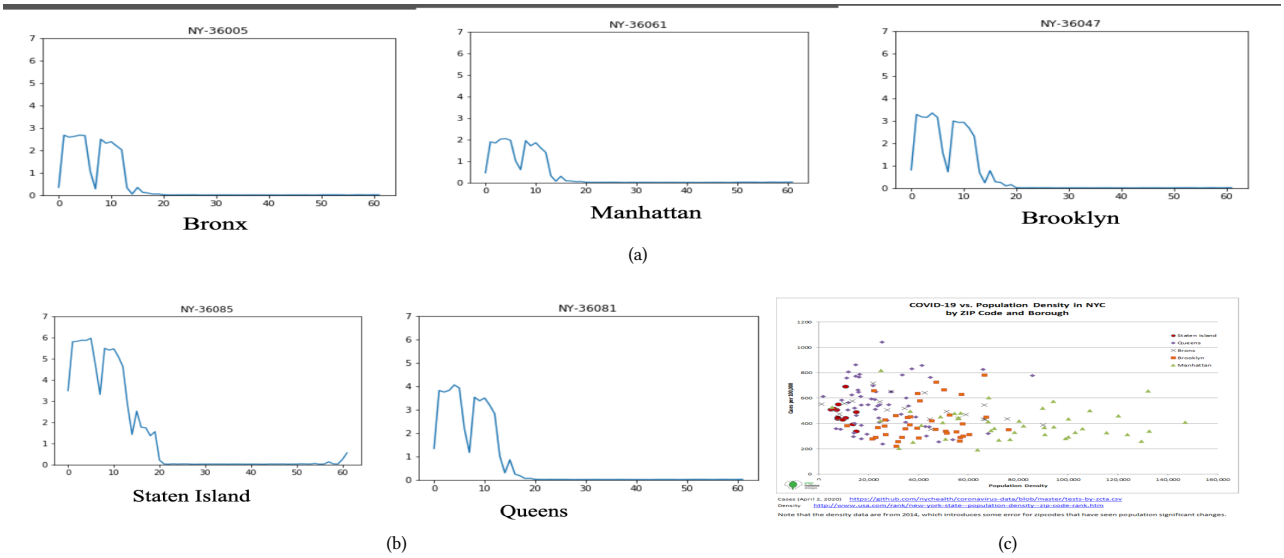

Figure 5: Mobility from March 1-May 2 in (a) Bronx, Brooklyn, Manhattan (b) Staten Island, Queens (c) Covid cases in all five boroughs with population density

5 ETHICAL CONCERNS

The ethical issues that we are looking at in this study focus on the potential issues with data privacy and how the pandemic and mitigation strategies impact different groups from different aspects; socio-demographic/economic, healthcare and mobility. To address these issues the following questions are analyzed.

5.1 Data Collection

5.1.1 Problem 1: Are there any biases/ concerns related to Covid-19 data collection? Data on infected individuals come from tested individuals and there are significant problems with testing in many countries. There is selection bias in testing, which means certain individuals (high-risk individuals such as elderly, or individuals showing symptoms) are tested [4]. Also, infected individual data includes hospitalized individuals but doesn't include people who are not hospitalized (i.e. showing mild or no symptoms). Similarly, data doesn't include people who died outside of hospitals (e.g. individuals who died at their houses). In this regard, we look at potential selection bias in the data that we use.

5.1.2 Problem 2: How is data privacy of the individuals that are infected with Covid-19 virus as well as the general public is protected? Personal information of infected individuals including their health records, demographic, socio-economic and address information is recorded and shared between various agencies (federal, state, local level), healthcare facilities, private companies, and researchers. It is crucial to protect the privacy of individuals through various anonymizing methods (e.g. differential privacy).

5.2 Socio-demographic/economic status

5.2.1 Problem 3: Are there disparities in mitigation strategies (e.g. social distancing) based on socio-demographic variables and socioeconomic status (such as race/ethnicity, age, population density, income)?

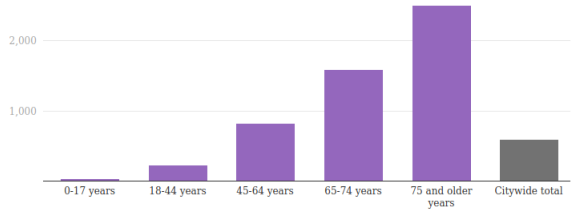

Figure 6: Hospitalisations vs Age

Mitigation efforts affect individuals differently based on their ethnicity, race, age, income, even location. For instance, elderly are more vulnerable to Covid-19 and they have to implement social distancing measures more severely, therefore, they need easy access to basic resources such as food, healthcare facilities, etc. The plot for elderly can be seen in fig 6. Similarly given in [8][9], African Americans and Hispanics are less likely to telework due to the nature of jobs that they have (e.g. construction jobs, service sector jobs). Moreover, African American, Hispanics and Native Americans have a higher proportion of existing conditions such as obesity, diabetes, which makes them vulnerable to Covid-19 compared to White and Asian populations. These groups are more prone to food and income insecurities (higher risk of loss of income among African American and Hispanic populations) due to Covid-19 mitigation efforts. Thus, we analyze Covid-19 data to see any discrepancies based on socio-demographic variables and socioeconomic status.

5.2.2 Problem 4: How do surveillance efforts impact minorities and vulnerable groups (e.g. low income groups)? It is argued in recent reports about Covid-19 that African Americans are referred less for testing[6] and there are fewer testing centers where minority groups reside, and most testing facilities are located in areas with high white populations. Furthermore, there is a need to add ethnicity/race information to infected individual data to study higher

Table 1: Findings on ethical concerns.

Concern	Factors	Findings
Problem 1	Bias on socio-demographic/economic	Yes, observed discrepancy based on on race/ethnicity and income
Problem 1	Bias on health disparity	Yes, only hospitalized cases are included in the data.
Problem 1	Bias on mobility	No bias, mobility is decreased in all groups
Problem 2	Privacy with socio-demographic/economic	No bias, data is anonymized
Problem 2	Privacy with health disparity	No bias, data is anonymized
Problem 3	Disparities on socio-demographic/economic	Yes, observed discrepancies based on socio-demographic variables and socioeconomic status
Problem 4	Surveillance	No bias, data is anonymized
Problem 5	Disparities on hospitalization	Yes, observed discrepancy based on race/ethnicity.
Problem 6	Data sharing restrictions on mobility	No, data is public and on a random group of people. Maintaining anonymity Maintaining anonymity
Problem 7	Anonymity on locations/travel frequency	(i) No, the locations of both DecratesLabs data is also given in terms of latitude, longitude. (ii) Travel frequencies are not anonymized, but individual identities are fully anonymized.
Problem 8	Privacy of infected individuals in mobility	No bias because data is anonymized, but there is risk if individual data isn't protected well. Contact tracing technology in the US is still work-in-progress and Apple proposed for self-reporting for data sharing and each user to have unique ID, which would be anonymous.
Problem 9	Mobility rate on socio-economic conditions	No bias in mobility data but there are discrepancies in impact of mobility based on ethnicity/race.

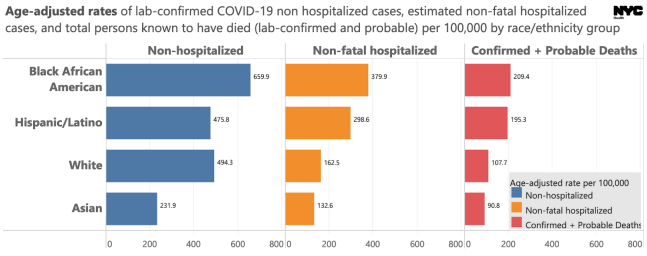

Figure 7: Non-hospitalisations vs ethnicity

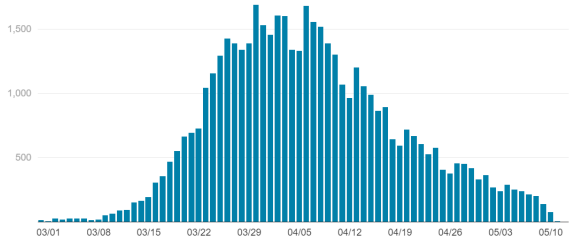

Figure 8: Number of hospitalisations over the period

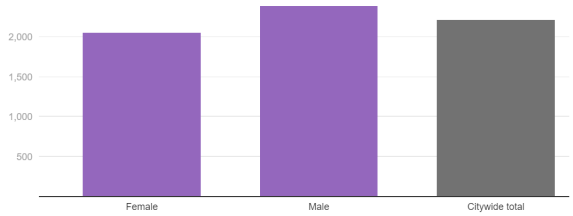

Figure 9: Hospitalisation vs Sex

rates of infection and death among African American, Hispanic and Native American populations. This raises data privacy and potential deanonymization of personal information for infected individuals from these ethnic/racial groups. We look at this issue through NYC hospitalization data. The analysis can be seen in fig 8and fig 9

5.3 Systemic Health Disparities

5.3.1 Problem 5: Are there disparities in terms of hospitalization and health care facilities in low income areas or for different racial/ethnic groups? In addition to having fewer testing centers in African American neighborhoods, it is crucial to check whether hospitalization rates vary between different ethnic/racial groups. We look at the hospitalization rates to see if there are any discrepancies based on ethnicity/race in the fig 7

5.3.2 Problem 6: Does data sharing restrictions pose ethical concerns among stakeholders involved in the emergency response? For example, do government public health officials, private healthcare providers or state/local governments protect individuals' data and privacy? It is crucial to share data during emergencies. During Covid-19 pandemic, many agencies at local, state and federal levels have been sharing data among themselves and with external parties including healthcare and communication providers, researchers, and even with agencies/ institutions abroad. So, it is crucial for these agencies/companies to protect the privacy of the individuals and make sure that the data remains anonymous when data is shared and used. To address this issue, we look at whether the healthcare data remained anonymous.

5.4 Mobility

5.4.1 Problem 7: What are the ethical concerns around mobility data? Is mobility/ travel frequency data to different locations anonymized? How is the privacy of individuals whose travel information is collected protected? Starting March many US states and local governments took measures to decrease mobility to control the spread of Covid-19. The measures include shelter-in-place orders, closing of

businesses and public schools and daycares. As mentioned under Problem 3, there are racial and ethnic inequalities in the impacts of Covid-19, and these measures limiting mobility affected different income, ethnic and racial groups differently. In this regard, we analyze whether there are differences in mobility based on socio-demographic and racial factors as well as population density and location (i.e. Boroughs of NYC).

5.4.2 Problem 8: Privacy of infected individuals is violated when their social contact history is traced. Digital surveillance can be an effective tool to monitor and control the spread of Covid-19 and contact tracing infected individuals is a common method to do that. However, there are significant privacy concerns around contact tracing, especially when it is done pro-actively by using a mobile app and tracing mobility of individuals. Tracing mobility of individuals and combining it with additional personal data can be ethically challenging and require transparency in how the data is being used and for how long it is collected and kept. Furthermore, it is concerning that infected individuals' identities are revealed to their social network about them being infected with Covid-19. The privacy of these individuals are violated since this doesn't happen when people have other illnesses. This may lead to social bias against these infected individuals in the future. In our analysis, we look at what data privacy measures are taken when contact tracing is implemented.

5.4.3 Problem 9: How are low mobility rates impacting different socio-demographic/economic groups? Are there disparities in access to food in low income areas where mobility is low and residents don't own vehicles and mostly dependent on public transportation? This problem is at the intersection of mobility and socio-demographic aspects.

Implementing social distancing and stay at home orders can be challenging especially for low income individuals, who rely on public transportation and individuals living areas where access to basic needs such as food, medication requires traveling by a vehicle. Due to stay-at-home orders and social distancing measures, many stores limit their operating hours and use of public transportation is limited to essential workers. So, low income people who don't have cars but need to travel to purchase food, medication, etc. are affected more compared (negatively) by these measures compared to people who own a car. We look at this issue in our mobility analysis.

6 DISCUSSIONS ON ETHICAL CONCERNS

The table 1 shows our observations and findings on ethical concerns. We observed bias based on socio-demographic/economic status particularly for African Americans, Hispanics and Native American populations due to their pre-existing conditions and existing biases against these groups. In our analysis of the data, we didn't find any ethical concerns with data privacy issues because data is anonymized and individuals are able to opt in or out to share their information. That said, the privacy issues around data sharing and use should be monitored closely. Further research on data sharing is needed.

7 CONTRIBUTIONS

Contribution for the course project is given in Table 2.

Table 2: Contributions from each individual.

Checkpoint	Task	Contribution
Abstract		Rahel
Introduction		Rahel
Literature survey		All
Data collection	Socio-demographic/economic Health disparity Mobility	Shailik Vashanth Anika
Pre-processing & Analysis	Socio-demographic/economic Health disparity Mobility	Shailik Vasanth Anika
Exploratory Analysis hline Visualization	Contact-tracing Socio-demographic/economic Health disparity Mobility	Anika Shailik Rahel, Vasanth Anika, Berna
Ethical Concern	Questions (All)	Berna
Findings	Socio-demographic/economic	Rahel, Shailik
Discussion	Health disparity	Rahel, Vasanth
	Mobility	Anika, Berna

REFERENCES

[1] [n.d.]. Covid-19 datasets. https://github.com/CSSEGISandData/COVID-19.
[2] 2020. Apple and Google are launching a joint COVID-19 tracing tool for iOS and Android. https://techcrunch.com/2020/04/10/apple-and-google-are-launching-a-joint-covid-19-tracing-tool/.
[3] 2020. Apple opens access to mobility data, offering insight into how COVID-19 is changing cities. https://techcrunch.com/2020/04/14/apple-opens-access-to-mobility-data-offering-insight-into-how-covid-19-is-changing-cities.
[4] 2020. Jason Oke, Carl Heneghan (2020), Global Covid-19 Case Fatality Rates, The Oxford COVID-19 Evidence Service. https://www.cebm.net/covid-19/global-covid-19-case-fatality- rates/.
[5] 2020. Mobility descratesLabs. https://github.com/descarteslabs/DL-COVID-19.
[6] 2020. Rubix Life Science (2020), COVID-19 Minority Health Access, March. https://rubixls.com/wp- content/uploads/2020/04/COVID-19-Minority-Health-Access-7-1.pdf.
[7] 2020. The Value and Ethics of Using Phone Data to Monitor Covid-19. https://www.wired.com/story/value-ethics-using-phone-data-monitor-covid-19/.
[8] Taylor Chin, Rebecca Kahn, Ruoran Li, Jarvis T Chen, Nancy Krieger, Caroline O Buckee, Satchit Balsari, and Mathew V Kiang. 2020. US county-level characteristics to inform equitable COVID-19 response. *medRxiv* (2020).
[9] Clyde W Yancy. 2020. COVID-19 and African Americans. *Jama* (2020).