A5: Extension Plan

Sandeep Tiwari

Nov 10, 2021

American prisons, jails, and detention centers are some of the riskiest locations for contracting the coronavirus. The cramped and typically unhygienic conditions of correctional facilities have acted as a cesspool for incubating the virus, as social distancing is simply not possible. Early on in the pandemic, testing prisoners was not a priority, and politicians were slow to change this trend and slow down the spread of the virus within prisons, primarily due to lack of public pressure. Simply put, prisons are not built to handle the spread of disease, especially during a pandemic, and it's difficult to find a worse set of conditions to prevent the spread of coronavirus than those in prisons. Prisons essentially consist of many people stuffed into cramped spaces, interacting with security guards that come and go to the outside world.

Recently a NYT article reported that the rate of infection in United States prisons was more than three times as much as that of the general population. This fact raises a few moral questions about the ethics of prioritizing, or not prioritizing vaccinations among prison populations. This also may give the opportunity to explore what types of statistical bias may persist when collecting this data, as we particularly explore the prisons in Oklahoma County, Oklahoma. This data clearly represents a very vulnerable subset of the population who are at a much higher risk of getting infected than members of the general population. So I hope to look at the impact of the virus within correctional facilities, to get a better understanding of these outbreaks. It might also be useful to leverage the information acquired from the Common Analysis to compare the infection and death rates to those in the prisons, which might give some insight into what is going on.

This data consistently shows the disproportionately high rates of coronavirus within prisons across the country relative to the general population, so I would like to answer the question of why correctional facilities were among the most dangerous places to be during the pandemic. A sub-question that stems off of this regards weighing the ethics of prioritizing prison communities for vaccination as well. Before the vaccines were first developed and available to the United States public, the CDC suggested that prisoners and prison guards be eligible for the early phases of the rollout. This suggestion received mixed reviews, but legislators used several guidelines in determining the allocation of vaccinations, evaluating benefits v.s. harm, mitigating health inequities, promoting justice, and promoting transparency. So it would be good to use these principles to understand the decision-making process for vaccinating prisoners in Oklahoma County. I would also like to examine the potential of some sort of a bias introduced in the data, whether it be via data collection, or aggregation, especially because testing for coronavirus within the prisons was not prioritized.

The New York Times collected data regarding infections, deaths, and testing for COVID-19 in state and federal prisons, immigration detention centers, juvenile detention centers, local, regional, and reservation jails from March 2020 through March 2021. This data was

collected over inmates as well as correctional officers. Because there was no standardized method for reporting national coronavirus cases in correctional facilities, and some prisons decided to stop releasing that information, some of the data was collected from websites that were regulated by state and federal prisons. For data that was not readily available, the Times collected it through direct requests, and other media and news sources covering county and state officials. For statistics about mortality rates, they used coroners' reports, medical records provided by families and reports from investigative agencies. For computing the total number of infections in jails, they used data found over the internet which was then corroborated by jailers, sheriff departments, or local government and health officers. But for missing infection data, they managed to collect data on infection and death rates via public records.

The data that I will be using comes in the form of two CSV files. The first file, `facilities.csv`, contains location information, and infection and death counts for inmate and officer. The second file, `systems.csv`, contains information on the prison and detention systems, themselves. This dataset can be found here. The data is licensed under the same terms as the Coronavirus Data in the United States data of the New York Times. The data was made publicly available for high-level, noncommercial public usage, particularly for medical and public health researchers, policymakers, analysts, and local news media. The data must be attributed to "The New York Times" in any publication. There are few discrepancies between the data reported from the different correctional facilities which may cause an over- or underrepresentation of cases within the prison systems. For one thing, prisoners who contract the virus multiple times are only recorded as having been infected once. Clearly, there is both an ethical dilemma and statistical bias in the way the data is being reported, as we may not be getting the entire story, which could affect the recommendation for prioritizing or not prioritizing prisoners to be vaccinated, given the typically squalid conditions they are subjected to.

But after conducting further research, I have found data from the Marshall project, which aims to compile data on the prevalence of COVID-19 infection in prisons across the country, as well. This data can be found here. However, they provide weekly time series data, which could provide just enough granularity for me to make a comparison of the two populations over time. This contains all the same information, but only at a state-level and gives updated information each week. This pretty much contains the same data as the New York Times dataset over multiple tables, just broken down over a weekly basis.

There appear to be many different unknowns that, while not preventing me from performing this extension of the analysis, might affect the final results and interpretation of it. Infection data across all facilities most probably are undercounted due to the general lack of testing. Most importantly, states have varying testing strategies within prisons and for their general population, indicating that some of these rates will most probably reflect a deceptively low infection rate, with some states' rates being more accurate than others. While many prisons were testing their respective inmates multiple times, many inmates across the country had not been tested after reporting COVID-like symptoms. Now for data more specific to my county, there are some potential setbacks because Oklahoma stopped providing facility-level infection

data in late February. Some state and federal prisons, and ICE did not regularly report facility-level data for inmate infections or tests administered to prison staff. So system wide infection and death totals are usually more than totals of the facility-level data. Many local jails tested fewer inmates than most, and many prisons released inmates who had contracted the virus without including them in the infection count, but were included in The Times' count if identified. But the data does not include facility-level testing counts because the total number of tests was usually not provided by the correctional facilities—however, most provided a cumulative count for their respective facilities. But some states that provided cumulative counts decreased the number of tests for seemingly no reason. Essentially, the running totals through March 2021 are all the data we will have to work with, which might make it difficult to compare to the data from the Common Analysis, as we can only make an inference on a coarser, more aggregated dataset with respect to time. Therefore, all these inconsistencies in reporting infection and death rates within these correctional facilities across the countries could potentially lead to some bias, which is exactly what we'll try to explore in this part of the analysis.

In order to make a reasonable analysis, I will only be able to take the data from the Common Analysis that was collected between March 2020 through March 2021, the same timeframe that this data was collected. I will also probably have to find some way to aggregate the data, either at a facility- or a county-level. Because the two datasets are split on different features, I will not join them, but use them separately to address different parts of the aforementioned questions. Because most of the New York Times prison data isn't broken down by date, I will have to do most of the preprocessing on it so that I can make a valid comparison between the two populations represented by the two sets of data. This will allow me to see what the total effect of the pandemic was through March 2021, specifically in Oklahoma County. From that standpoint, once the raw cases dataset is aggregated accordingly, I can see what the infection and death rates were in each county for each population. Obviously, there might be some bias here due to the lack of reporting in some counties, but I will assume that discrepancy is negligible. Now to drill down deeper into this analysis, I will also look at the Marshall Project data so that I can compare how the rates changed over time within each population. This could be more insightful, as viewing the time series graphs side by side will be useful in seeing how different points of time, and policy changes may have impacted the two communities differently.

The key thing to remember here is that the crux of this analysis is to highlight the discrepancies between the infection rates of prisoners v.s. the infection rates of the entire state. I believe the most instructive way of approaching this is to construct a visualization of the metrics for both populations and place them side-by-side, which is why I have chosen to try out several simple visualizations that make comparisons easy to see. I will begin by visualizing the two populations, perhaps showing how much higher the infection and death rates were in Oklahoma County prisons compared to those of the general public, via a bar chart, or line plot. I am still a bit unclear on which visualizations will be the most informative, and believe this requires some investigation on my part, but I believe that utilizing visualizations will be the best way of communicating the results of my findings while addressing these questions.

The first step in the pipeline for completing this analysis is to collect and consolidate all the data, so that I can explore it effectively. They all come in the form of CSV files, and so do not require any formatting on that front. This should take very minimal time, but once I have it all, I plan to clean and preprocess the data so that it will become usable for any further analysis. I am to spend about 1 week (through November 17) cleaning up the data. Once the data is all cleaned and formatted appropriately, I will then conduct my analysis, which I expect to take about 10 days (through November 28). This analysis will primarily focus on generating various visualizations that will allow me to see how the pandemic affected prison populations, and if there was any bias in the data. I expect about 2 weeks to consolidate my results and formulate a presentation, which I should be ready to give by the end of the quarter.