

A5 - Extension Plan
DATA 512
Preston Stringham

Motivation/Problem Statement

COVID-19 has impacted us all in various ways, but I feel there has been no group of people more negatively impacted by COVID-19 than the hospital and healthcare workers who risk their lives to take care of people with severe cases of the virus. My problem statement addresses the impact of hospitalization rates in Middlesex County, New Jersey and how the introduction of vaccination in the United States impacted this. I intend to investigate the change of hospitalization rate in Middlesex County, New Jersey. In particular, I am interested in the correlation of hospitalization rate and vaccination rate and also how the hospitalization rates of particular cities in Middlesex compare to the county as a whole. The reason this is a human-centered problem is that it focuses on the impact of both healthcare and non-healthcare workers. From the perspective of non-healthcare workers, more hospital beds and resources could be used to support those suffering from non-COVID related illnesses. From the perspective of the healthcare worker, their work would be less risky as they would interact with less COVID patients than usual. Both of these aspects are clearly human-centered.

Research Questions

There are many questions that I have about the data. In particular, I want to incorporate the work I did in A4 into my research in A5. So, I will be interested in the relationships between vaccination, virus spread, and hospitalization rate in Middlesex County. Again, I feel that the primary research conducted will be about hospitalizations and will be from a human-centered perspective. I want to look at the vaccination rate over time and see if any significant change occurs and how it correlates with the hospitalization rate. I believe that there will be significant changes to the number of hospitalizations before and after the beginning of vaccination. However, considering that vaccinations did not completely resolve the pandemic, I imagine I will see an uptick in the hospitalization rate even after the introduction of vaccinations. A formal hypothesis for this problem might be, "There is no relationship between vaccination rate and hospitalization rate."

Furthermore, I believe that it would be interesting to compare the mean hospitalization rate in particular cities in Middlesex County, such as Edison city to the rest of the county. A formal hypothesis could be, "There is a significant change in the mean hospitalization rate in Woodbridge Township than the rest of Middlesex County." I think that this is an interesting question since Woodbridge Township is the most densely populated area in the county, so I feel

that the hospitalization there would be much greater. I will need to consider the bias of both the population density as well as the hospital density of this region.

Data

For my analysis, I have found two datasets both provided by United States government agencies. The first dataset I am interested in is the [healthdata.gov](https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capacity-by-Facility)'s "COVID-19 Reported Patient Impact and Hospital Capacity by Facility" dataset which features detailed information about hospital metadata, such as specific types of resources available at each facility and the number of each type of resource that is available. I don't think that there are many ethical concerns about using this data. It seems to come from a reliable source and is provided by the U.S. Department of Health & Human Services. I was unable to find a corresponding license for this dataset, but as the data is available to download at the link provided, I imagine that it is okay for my usage. This dataset contains a 7 day average and sums for the amount of hospital resources (ICU beds, normal beds, staffed beds, etc.) for each hospital/healthcare facility across the United States starting from July 31, 2020 and ending on October 31, 2021. This should span a large majority of the dates we looked at in A4. It expands on A4 as we can now look at how hospitalizations and COVID infection rate are related. A link to the "COVID-19 Reported Patient Impact and Hospital Capacity by Facility" can be found below,

<https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capacity-by-Facility>

Another dataset I am interested in is the CDC's, "COVID-19 Vaccinations in the United States,County" dataset. This dataset contains the number of vaccinations per day for each county in the United States starting from December 13, 2020 to today. What I hope to show with this dataset is that vaccinations did help decrease hospitalizations and the change of the infection rate of the virus. This dataset falls under the Public Domain U.S. Government license and should be perfect for my use. A link to the "COVID-19 Vaccinations in the United States,County" dataset can be found below.

<https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh>

Unknowns and Dependencies

One unknown with respect to this project is a lack of vaccination data prior to December 13, 2020. I will be unable to make any assumptions about the efficacy of vaccines before they are introduced to the general public. This restricts my analysis regarding vaccines to the range of the vaccination dataset. Other than this, I cannot think of any outside factors that could contribute to me not being able to complete my analysis.

Methodology

With regards to my methodology, I plan to first gather the data and join the datasets by their date, similar to what was done in A4. I then want to compare the mean hospitalization rate of the 3 major cities/townships in my area. Namely, Edison city, Woodbridge Township, or New Brunswick city. These areas are more densely populated. I will find which of these cities has the greatest population density and claim that the normalized mean hospitalization rate is greater in this area than the rest of the county. A t-test can be used to do this and I will plan to use that. Additionally, I am interested in the correlation of vaccination rate and hospitalization rate, so I will use a linear regression model to find this correlation. I will take the results of my t-test linear regression and show them in my presentation as some form of visualization. Perhaps show a timeline of the days which were significant. The visualization portion will certainly be something that I will figure out during my analysis.

Timeline to Completion

The timeline to completion is listed as follows:

Week 1: Data Exploration and preparation.

Week 2: Analysis and initial attempts at solving problems stated in the problem statement section

Week 3: Find final results, begin to prepare presentation, and begin working on report

Week 4: Finalize presentation

Week 5: Deliver final report

I feel that this initial timeline gives me ample time to complete all of the tasks required. In particular, I am giving myself extra time to work on the final report as I know this will probably take a substantial amount of time.