Vaccine Equity Project

Leland Ling, Kristi Li, Brian Tao, Yu Yan

CS506 Data Science Tools and Applications

Professor Lance Galletti

December 5, 2022

I. Background & Motivation

It is no secret that the effects of SARS-COVID-19 have been devastating to communities around the world. However, it isn't as well known how disproportionate the effect across different communities and demographics has been. Dr. Julia Koehler approached us to explore these discrepancies, especially in the context of vaccinations.

To explore these effects, we made our scope of our project to be the cities of Massachusetts, namely Revere, Chelsea, Newton, Wellesley, Everett, and Springfield. Revere, Chelsea, Everett and Springfield were classified as underserved communities by Dr. Koelher, while Newton and Wellesley are classified as better-served communities. Dr. Koehler elaborated that there had been a grassroots organization called La Colaborativa serving the communities in Chelsea by doing outreach, specifically demystifying and providing access to vaccine rollout.

The ultimate goal of this project was to be able to directly visualize the effects of La Colaborativa on Chelsea and to discern a qualitative difference between the communities that have grassroots support (Chelsea specifically in the scope of our project) and those that do not.

II. Previous Work

Since the emergence of SARS-COVID-19, there haven't been any specific attempts at quantifying this specific virus or its vaccine response in the past. The only other similar outbreak was in 2003 with a similarly described SARS outbreak, in which development into a vaccine occurred yet never made it past the first trials(SARS and MERS). Interestingly enough, some academic literature around this time cautioned against the vaccine because of conflicting data on its effectiveness (Caution raised over SARS vaccine - PMC). Due to these past actions, the emergence of COVID-19 and its vaccine response has been extremely novel, and as such its vaccine rollout can't be compared to other events.

III. Data Collection

Dr. Koehler had supplied us with two datasets, both collected by Massachusetts' Bureau of Infectious Disease and Laboratory Sciences. The first one was organized based on zip code while the other had been organized by county and city. Both of the datasets contain weekly logged vaccination statistics since the beginning of COVID. The datasets contained the number of individuals vaccinated with at least one dose of the vaccine, fully vaccinated, vaccinated with Booster one, and vaccinated with Booster two over different demographics, such as age, race, and gender identity. The data within was mostly complete, except for a couple of missing points to which we had cleaned.

During the earlier times, booster two vaccination data is missing. This makes sense as the booster was not out yet during that time period. We did not make any changes to them, and we considered these time periods completely separate when plotting them. During the later times, we figured that some of the partially vaccinated data was missing. We cleaned that part of the data by subtracting the fully vaccinated data from at least one dose data. We also realized that for most times, the infant data is missing. We simply filled them out with 0. Lastly, the partially

vaccinated data of Wellesley city from line 2396 to 2403 seems to be shifted up by one row, so we shifted them down by one row.

To ensure that the different cities had a statistical difference, we used the T test, finding that the data between cities and races is significantly different as p value is lower than 0.05 when comparing them to each other.

IV. Data Visualization & Exploration

We quickly realized that analyzing every demographic contained within the datasets would be much too much for the timeline we were given, so we decreased the scope of our project to only race data. Furthermore, out of the races present in the datasets (AI/AN, Asian, Black, Hispanic, Multi, NH/PI, White, and Other/Unknown), we decided to only look at data for Asian, Black, Hispanic, and Caucasian, as the other race data was too small in scale – only having data for on average ~150 data points. Asian data somewhat exhibits this issue as well, having a sample size of 810 individuals on average.

We normalized our data by changing each value to be a function of population, or an overall percentage. For example, the number of fully vaccinated Caucasian individuals of date 03/09/2021 living in the city of Chelsea was divided by the population of Caucasian living in Chelsea observed on the same date. This was done to be able to compare city data at the same scale, as the different cities had different populations for different demographics.

We began data visualization by plotting each city's data against time, separating the demographic data by race as shown below in Figure 1. We find that some of the city data have distributions similar to each other; Hispanic data shows that Chelsea matches Revere distribution during the initial stages of vaccination rollout, i.e. 03/2021 to 07/2021 and then surpasses the other distributions all together. Black data also seems to follow this trend as well. Caucasian data shows that the distribution has a much higher vaccination percentage on average than each of the other cities. Asians in Chelsea were shown to have much higher values overall, even surpassing 100%. While we initially thought that this was an issue on our part, perhaps tampered with during data cleaning, the race data for smaller population races, such as the ones we deliberately decided to omit because of small sample sizes, sometimes would have populations less than the number of individuals vaccinated of that demographic.

To quantify how similar each city's vaccination data is to one another, we used Hypothesis Testing to compare corresponding race data. Our hypothesis was that the two cities' corresponding race data are similar - using the p-value generated from t-test as our metric. Our hypothesis was the data being compared are similar distributions, while the null hypothesis was that the distributions were not. The code to this can be found in the github at https://github.com/BU-Spark/ds-vaccine-equity/blob/dev/fall22-team-1/HypothesisTesting.ipynb . With the T test, if the resulting p value is above 0.05, we can reject the null hypothesis and conclude that the data distributions being compared are statistically similar. Below we show

significant observations, or T test comparisons that result in p values above 0.05, and omit the ones that result in a p value under 0.05.

When comparing Chelsea and Newton, we found that the Hispanic demographic data of both cities had a p value of 0.911, meaning that they statistically are probably similar distributions. White demographic data when compared had a p value of 0.111, meaning that they probably are somewhat similar as well.

Black and Hispanic data distributions of Chelsea and Revere result in p values of 0.470 and 0.535 respectively when performing the T test.

Chelsea and Springfield's Black data reveal a p value of .463 when performing the T test on the two distributions.

While there are more significant comparisons between the rest of the data, we are more interested in comparisons between Chelsea and the other cities specifically, looking for some kind of statistical way to show the effects of La Colaborativa. The rest of these T test comparisons are within the same github link above.

To further understand the similarities and differences between Chelsea and the other city's racial data, we explored the rates of change in percent of fully vaccinated data of each race. We calculated the rate of change of each demographic data by applying the python package Pandas' dataframe differential function to the data points. This function essentially calculates the difference between a data point and its previous neighboring data point and divides it by two, as we are only analyzing the rate of change over two points. Initially, when graphing rate of change in percent fully vaccinated for each race over time, we find that rate of change starts from values ranging -0.8 to 0.0 and tapers off to 0 on average, having a very minor positive rate of change in times 03/2021 to 07/2021 over all data distributions. (NEED TO SHWO THESE PLOTS)

However, plotting the data on the timeline from 03/2021 to 11/2022 made it difficult to observe these minor increases and decreases in rate of change. Therefore we split up the timeline into 5 parts — early, middle, late vaccine rollout, booster one rollout, booster two rollout — making the data easier to view and analyze. Early, middle, and late vaccine rollout time intervals were calculated by splitting the timeline from the beginning of the data to booster one rollout into thirds. Booster one rollout ends when booster two rollout begins.

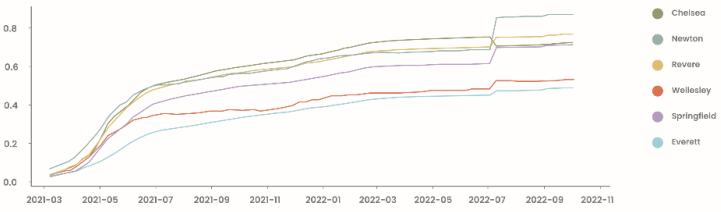
We then graphed the following four graphs for all 5 time periods. In the first two graphs (Data *vs.* Rate of At Least 1 Dose by Region, Data *vs.* Rate of Fully Vaccinated by Region), each line corresponds to a region and each graph corresponds to a race. Now working on a smaller scale, we find there are multiple points in the data that could possibly correlate with real events – to which we will explore these possibilities more in the next section.

However, when we applied the same hypothesis testing that we did above to the rate of change, we found that there are not significant findings. Each application of the T test as a comparison results in a p value of above 0.05, ranging from 0.21 to 0.98.

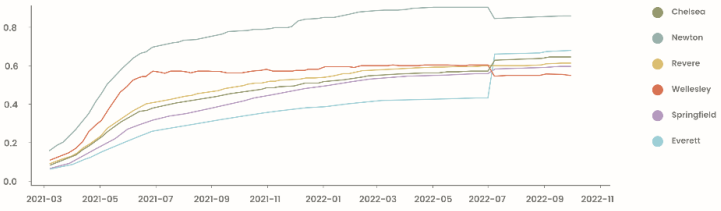
CITY DATA *VS*
% OF FULLY VACCINATED DATA

In the following graphs, each line corresponds to a region,
and each graph corresponds to a race.

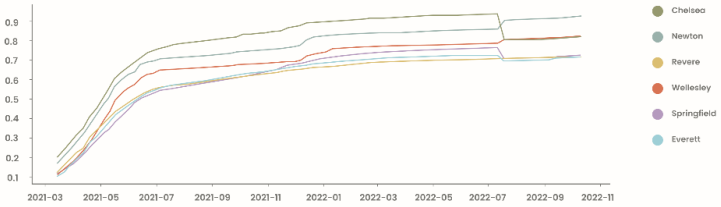
HISPANIC



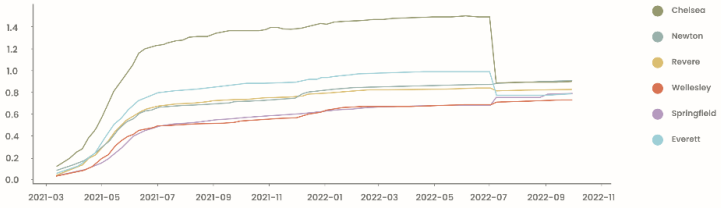
BLACK



WHITE



ASIAN



VI. Interpretation & Limitations – incomplete

Using the results from the hypothesis testing to examine the similarity between Chelsea and its influence from La Colaborativa and the other cities of Massachusetts, we found many correlations.

The cities that we chose to analyze fall into three categories - underserved, underserved with grassroots efforts, and better served. While there could be another category of Better Served with Grassroots efforts, to our knowledge we do not know of any of these non-governmental organizations (NGO) that do so in the better served cities that we are analyzing. However, an NGO serving a better served city would be included in the definition of better served regardless.

Newton and Wellesley were selected as examples of better served communities. Revere, Everett and Springfield were selected as underserved communities. Chelsea has La Colaborativa as an NGO and therefore would fall into the category of underserved with grassroots efforts.

As mentioned above, the fact that Chelsea's Hispanic data when compared to Newton's results in a p value of 0.911, suggests that Chelsea's Hispanic vaccination data follows similar trends as Newton's. This correlation possibly could be the result of La Colaborativa's influence, as their work was done mostly in spanish.

Chelsea and Revere's Hispanic and Black populations also have correlation, as indicated above. Chelsea and Revere are both underserved communities, and as such it is no surprise that they have such a statistical similarity – having p values of 0.470 and 0.535. Springfield also falls in this category and its black population exhibits similarity with Chelsea as well, with a p value of 0.463.

While Chelsea's hispanic population data has similarities with both Newton and Revere, the p value resulting from comparing the Chelsea and Newton hispanic populations with the T test versus that of comparing Chelsea and Revere suggests that the distributions of Chelsea and Newton are more similar than the Chelsea and Revere distributions.

It is important to note that Chelsea does not exhibit similarity with Everett, another city that supposedly falls into the same category.

Qualitatively, there are a number of points within the rate of change graphs that could be correlated to some real world event.

We are not sure what the initial spikes displayed in the rate graphs represent. However, immediately after, it appears that there is a gradual increase in the rate of change of percentage of individuals that are fully vaccinated. This trend seems to begin around early april 2021, peaking

at mid may 2021, and gradually decreasing to zero after, until july 2021. This could be explained by the gradual increase in availability of the vaccine.

It seems before the holiday season in mid late november 2021 and after in january 2022 until gradually tapering off until march 2022, there also appears to be a gradual increase in the rate of change for percentage of individuals that are fully vaccinated for all demographics.

V. Questions Answered – incomplete

Based on the data visualizations we made, we were able to answer the following questions outlined in the project document:

1. **Do grassroots organizations have a statistically significant effect on vaccination rates in hispanic populations?**
    - I think we might have to run the T test on individual points.
2. **What are the factors that affect vaccination rates in the data that we are seeing**
    - Race seems to be a factor that affects vaccination rates, and the effect varies from city to city.
3. **Does the rate of vaccination rate change as la colaborativa was able to do more work in the community?**
    - We cannot tell because there is no timeline of La Colaborativa's actions.
4. **Are there changes in the vaccination rate?**
    - Yes. They will be explained in the next paragraph.

VII. Challenges Faced – incomplete

While we can discern that there is a qualitative difference between Revere and the other graphs, we quickly realized that without a timeline to La Colaborativa's efforts, we cannot pin these differences to specific events in the timeline we observed.

VIII. Suggestions for Future – incomplete

Some suggestions for the future of the project is to look at other aspects of neighborhoods in which residents are in, such as analyzing the effect of public transportation and economic resources. This could be possible if provided with a more detailed dataset (such as neighborhood ridership, resident's annual income, etc.)

Exploration and Findings (Part II）-- already integrated

Using this to examine the similarity between Chelsea and its influence from La Colaborativa and the other neighborhoods of Boston, we uncovered multiple findings.

When comparing Chelsea's race data with Newton's race data, we found that the Hispanic populations and Caucasian populations are statistically similar. Performing the T test to compare the hispanic population data between these two cities results in a p value of 0.910, meaning that these two data distributions are very statistically similar. Comparing the caucasian population data results in a p value of 0.111. While being above 0.05 (which still proves significance), the p value is substantially lower than the p value of the T test for hispanic data. There also appears to be some similarity between Chelsea and Revere's data, as comparing their black populations with T test results in a p value of 0.469 and comparing their hispanic populations results in a p value of 0.534. Chelsea and Springfield's black population also show signs of similarity with a T test comparison resulting in a p value of 0.113.

We attempted to figure out if there is a significant difference in rate of change of fully vaccinated population. We took the difference of percentage of the fully vaccinated population between two consecutive times and then used hypothesis testing to examine how similar they are.

When comparing Chelsea and other cities, surprisingly we found that the rates of change for hispanic, black, asisn and white population are all statistically significant because their p-value are greater than the significance level (0.05).

not yet integrated - not sure how to integrate

After constructing all the graphs for vaccination rates of "at least one dose" and "fully vaccinated", we are able to compare the vaccination rates across cities and races. Based on the plots of early vaccine rollout corresponding to regions, we can see that Chelsea has the highest vaccination rate of "at least one dose" for each race, but the trend does not decrease or increase over time. For the middle vaccine rollout, the graph begins to fluctuate. For hispanic data, Revere and Chelsea have generally the highest increasing vaccination rates. Similarly, for African American data, Revere and Chelsea have the highest rate of covid vaccinations. For Causian and Asian data, the vaccination rate of Chelsea is much higher than that of Revere, Wellesley and Newton. By observing the plots of late vaccine rollout, we can see that there is a significant spike appearing between 11/01/2021 and 12/01/2021 for each race, which could be contributed by flu season. Also, we found out that the rates of death tend to be increasing in this time period, which could be the reason that people attempt to get vaccinated. For booster 1 rollout, we discovered that the trend of vaccination rate is similar across cities, but there is significant drop for Causian data at the time interval 01/2022 to 02/2022. We need to do further

investigation on this. For booster 2 rollout, Chelsea city has the lowest vaccination rate at the middle of July in 2022 except for the African American population. Then, the vaccination rates of “at least one dose” become the same for all cities after the latter half of July. (9:53)