COVID-19 Hospitalization Rate

Claire Chour, Karina Nguyen, Annie Zheng

ABSTRACT

As the spread of COVID-19 increases, doctors, nurses, and medical workers are prioritizing victims of the epidemic before themselves, hoping to save as many lives as possible. Since the growth of the pandemic is difficult to predict, it is extremely difficult to prepare the right number of protective equipment and supplies for each hospital. According to NBC News, a general problem within all hospitals and health centers right now is the lack of healthcare supplies. Doctors, nurses, and medical workers who contracted COVID-19 were forced to isolate themselves for at least three weeks, leaving the remaining healthcare workers with more patients to take care of.

There were many instances in Italy and New York when hospitals were short of beds and equipment to help COVID-19 patients. In Lombardy, there were 500 beds for those in need of intensive care, but this was not enough. As days progressed, the number of confirmed COVID-19 cases and of deaths increased significantly, but the abundance of supplies remained the same if not a little more. Similarly, New York has a huge overflow of patients, yet not enough resources. At Brookdale Hospital, Dr. Arabia Mollette stated that the COVID-19 has turned the Brooklyn-based hospital into "a war zone" and that the hospital needs more gowns, gloves, masks and ventilators (CNN News). The morgue is also overflowing as well, as it can usually hold about 20 people but the number is significantly higher due to the pandemic.

Ultimately, it is important that hospitals receive as many supplies and resources as possible to help all patients. It is difficult to predict how many patients will be going to each hospital, so it would be helpful for hospitals to have an estimate of the future hospitalization rate.

INTRODUCTION

As of April 18, 2020, COVID-19 was hitting a few hot spots in the United States. The choropleth map below shows the total deaths in each county. While most counties across the country saw fewer than 100 deaths (purple sections), a few areas faced significantly higher death totals (yellow sections). The highest death total of 135,572 was seen in New York.

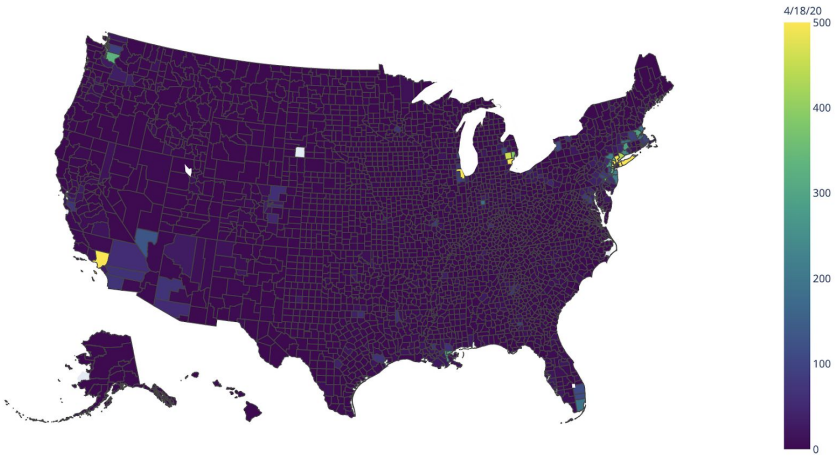


Figure 1. Total deaths per country by 4/18/20.

During the past few weeks of quarantine, news channels have been broadcasting devastating reports of overcrowded hospitals in COVID-19 hot spots. Hospitals in New York and Italy have reached maximum capacity and healthcare workers are running short on resources. During a pandemic, a hospital's lack of preparedness and resources puts its patients at a huge disadvantage. In many cases, healthcare workers had to make the tough decision of choosing which patient to save with the limited ventilators. We hope to explore various factors that impact hospitalization rates in different regions and draw insights that can help hospitals better prepare for future COVID-19 cases.

PROBLEM STATEMENT

Since overcrowded hospitals have been a source of devastation for a number of states, we want to use various factors to create predictions of hospitalization rates. This way, hospitals can better prepare and allocate enough resources for the pandemic and avoid situations in which hospitals are overcrowded and understaffed.

INITIAL EXPLORATION

At first, we attempted to assess the situation in the first and last state that implemented the stay-at-home lockdown to see if that had a huge impact on the number of confirmed cases. However, we realized that it did not have a significant impact and the two states, California and South Carolina likely have other factors that impacted the growth rate.

Instead, we decided to use California and New York because they were more similar as a state than California and South Carolina. We wanted to determine whether the lockdown date for both states reduced the rate of growth of COVID-19. We calculated the slope (growth rate) between each consecutive day and plotted the growth rates below. California's lockdown date was 2020-03-19 and New York's was 2020-03-22; from the graph, California's growth rate significantly decreased after

lockdown, but New York performed the opposite. This may be because New York is more crowded, so it is difficult to contain the pandemic even if people are quarantined.

To account for the NaNs during the first few dates when there were no confirmed cases yet, we filled them with zeros since the rate was 0%. There was an infinite value in the line graph for New York because the rate was above 100%, so we dropped that row and filled it with the average rate of growth during that period of time. The lockdown definitely did not help New York, especially since the number of COVID-19 cases skyrocketed after the lockdown was implemented.

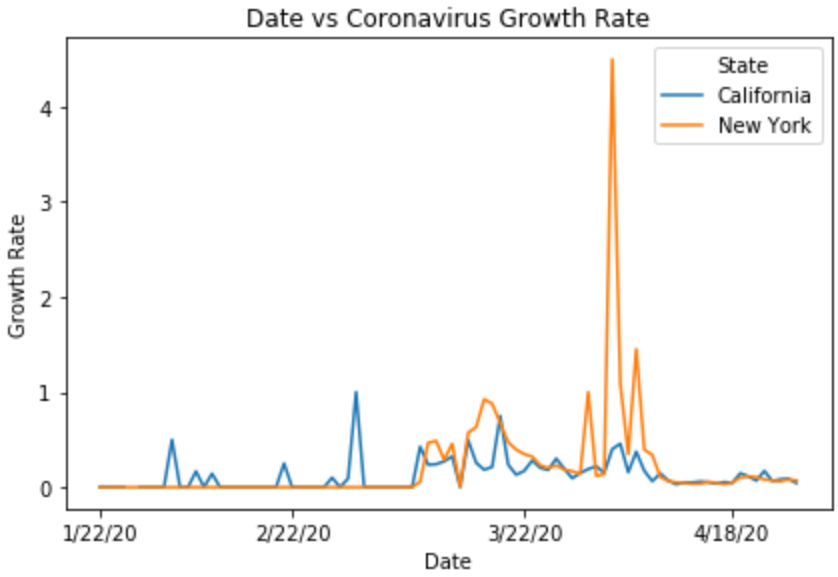


Figure 2. Line graph plotting the growth rate of the pandemic.

DATA SELECTION

In order to create a model that would predict the states' hospitalization rates, we needed to find the features that were most informative.

Access to Healthcare

Initially we had assumed access to healthcare would play a big role in determining hospitalization rates. We examined indicators such as HPSA (Health Professional Shortage Areas) scores, Medicare enrollment, and SVI (Social Vulnerability Index) in the hopes of discovering relationships between healthcare accessibility and hospitalization rates. HPSA shortages and Medicare enrollment were given as county totals, but we converted them to percentages of the county population to draw better comparisons across counties of different sizes. Hospitalization rates were calculated by dividing the hospitalization counts by the number of confirmed cases.

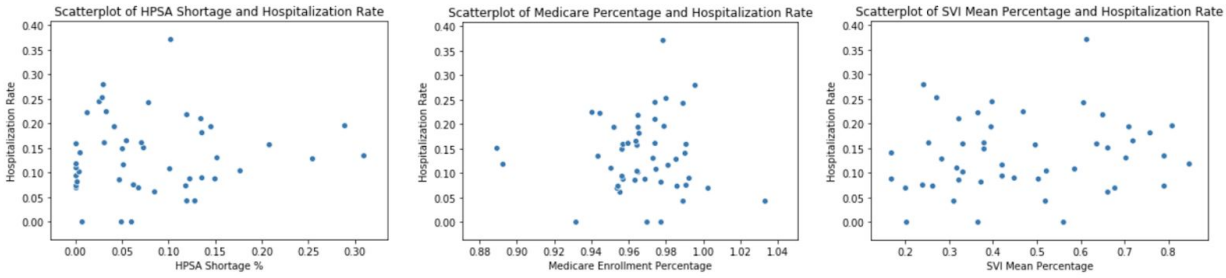


Figure 3. Scatter plots comparing selected features with hospitalization rates.

To our surprise, the points on all three scatter plots appear to be scattered randomly; therefore, there must be little to no correlation between either of the three features--HPSA shortage percentage, percentage of Medicare enrollment, SVI--and hospitalization rates. Upon further reflection, it makes sense that these factors would play less of a role during a pandemic than they would in normal circumstances. Under these special circumstances, there are likely unique factors that have a greater impact on hospitalization rates.

Policy Dates

Factors that are unique to this pandemic are state policies implemented to eliminate crowds and reduce the spread of the virus. Since we had access to the dates that US states implemented certain policies with the intention of decreasing the growth of the pandemic, we wanted to find out whether these policies actually had an effect on the growth rate of COVID-19 cases and hospitalization.

Political Affiliation

In the next steps, we explored how the states' response to the crisis may have improved or worsened the virus' impacts. Since the effectiveness of testing rate might imply policy practicing to handle the situation we were curious if there was any direct impact of political affiliation on hospitalization. Given that our data had the "Democratic: Republican Ratio", we converted that feature to categorical variables of Democrats and Republican. We acknowledge that some states might have been independent, but we categorized them as Democratic. We made the box plot (Fig.8) to see the range of testing rate in republican and democrative states. It seems that higher testing rates are prevalent in democratic states, however, we don't know the total population of those states.

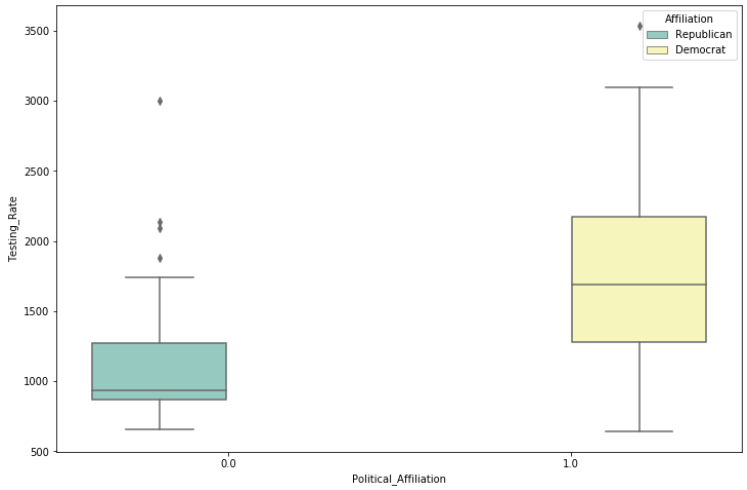


Figure 4. The range of testing rate, categorizing data into political affiliation

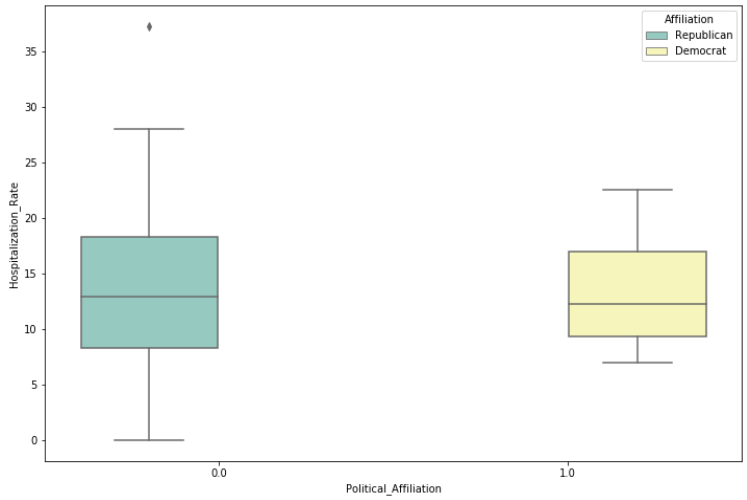


Figure 5. The range of hospitalization rate, categorizing data into political affiliation

We did the same box plot for hospitalization rates (Fig.9), and it seems that they are having almost the same upper bound. The box plots show that overall there was a difference in how democratic states and republican states dealt with the situation and we decided to include political affiliation in our next steps.

Moving forward, we decided to look more closely at state policy enactment dates, state political leanings, testing rates, incident rates, mortality rates, and days since the first case. We were interested in seeing how different states have responded to the crisis, and how their responses affect the hospitalization rates. In particular, we were curious to see if states of different political parties responded to the situation differently. At this point, we had grouped all of our data by county.

CONDUCTING PCA

Although we had eliminated a few features during data selection, we needed to narrow down our dataset even more in order to build a strong prediction model. We decided to apply PCA to reduce the dimensionality of our dataset. This would allow us to see which features have the greatest impact on variance. For PCA, we looked at 12 columns: stay at home, >50 gatherings, >500 gatherings, public schools, restaurant dine-in, entertainment/gym, foreign travel ban, testing rate, incident rate, mortality rate, democratic to republican ratio, and days since first coronavirus case.

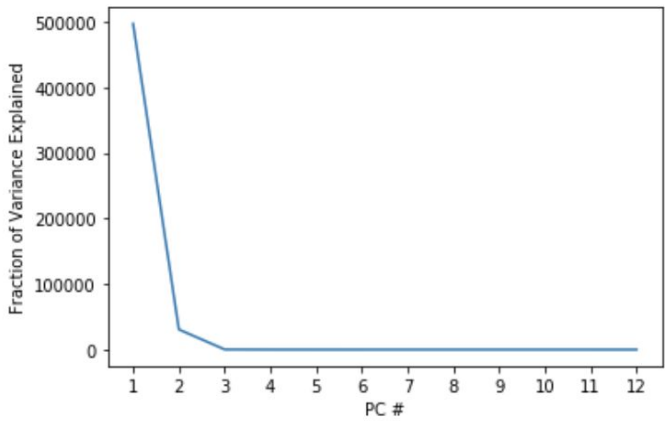


Figure 6. Scree plot of PCs.

As seen in the scree plot above, The PC1 has the greatest significance while PC2 has slight significance. Anything beyond PC2 was pretty much negligible.

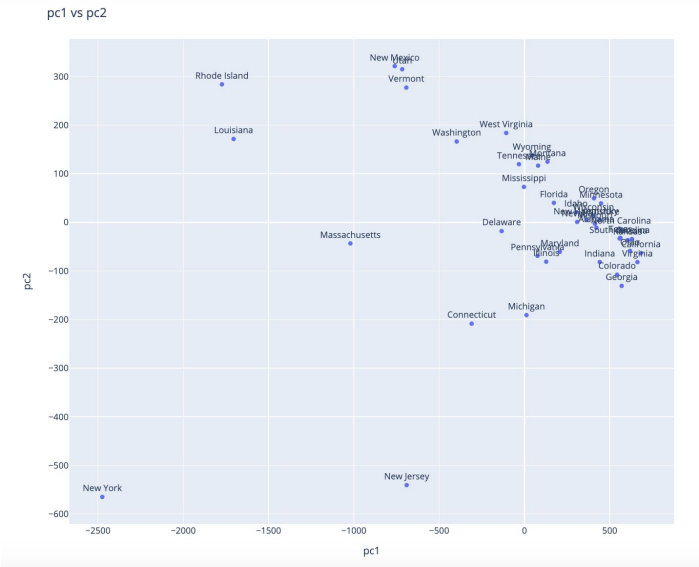


Figure 7. Scatter plot of PC1 and PC2, labeled by state.

The scatterplot shows that there are outliers such as New York and New Jersey. Those are the two states with the highest number of cases.

Next we plotted the first row of V^T to see the contributions that each feature has on the principal components.

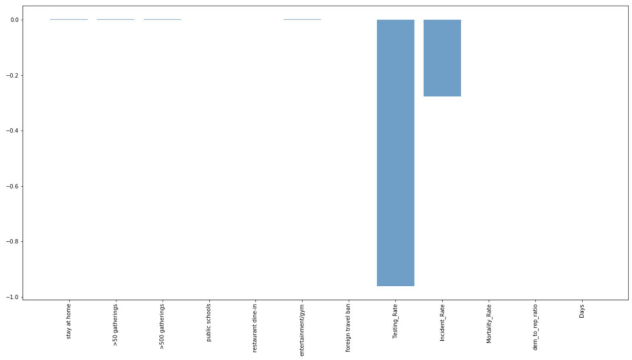


Figure 8. First row of V^T.

After PCA we realized that the two features that had the most variance were Incident Rate and Testing Rate. Testing rate is the total test results per 100,000 personas, where total test results include both positive and negative. Incident rate shows the number of cases per each state. Since those two features showed the highest variances, we decided to use them in our logistic regression model.

RESULTS

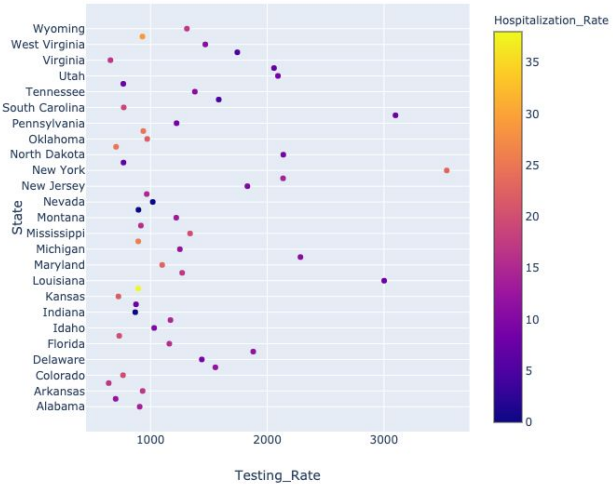


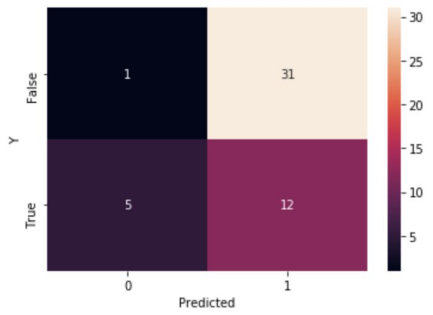
Figure 9. Testing Rate and Hospitalization Rate per State.

We wanted to explore both features in all the states (Fig.10). Assuming that testing rate is how effectively the public health system can conduct testing per 100,000 people in the region, having a high testing rate will indicate high efficiency. However, we couldn't neglect the amount of hospitalization rate too per region since testing much depends on how much people are hospitalized. Having a low hospitalization rate and the highest testing rate will indicate the best practices of handling the pandemic. Moreover, it was important for us to not fall into the trap of considering testing rate = 0 or hospitalization rate = 0 because of NaN (not reported values). Based on that we identified that the threshold for a 'reasonable' hospitalization rate would be less than 10 and maximized testing rate. Therefore, states like Utah, Vermont, Louisiana, MA Washington, Pennsylvania, New Jersey formed the subset of what we can call 'the most effective states.'

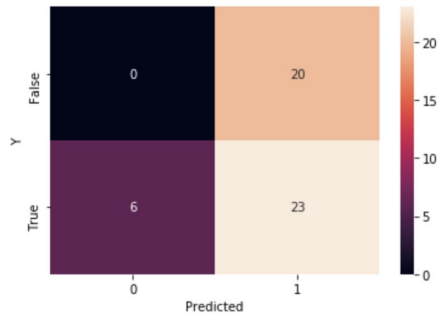
Logistic Regression

After conducting PCA, we realized that the political affiliation ratio and policy enactment dates didn't have much of the variance, and we decided to use two important features - Testing Rate and Incident Rate in our model. In order to understand the hospitalization rate we decided to classify hospitalization rate. If the hospitalization rate is <10 (classified as 0) then we assume that it's a reasonable capacity for health workers to handle the cases, however if it's more than 10, we consider an extreme hospitalization rate (classified as 1). We used Logistic Regression because in this case our hospitalization rate (dependent variable) is binary and we wanted to explore the relationship between hospitalization rate vs testing rate and incident rate. We then explored three different thresholds (10,15,20) of the hospitalization rate to explore the model accuracy because theoretically we couldn't know the actual capacity of the healthcare system.

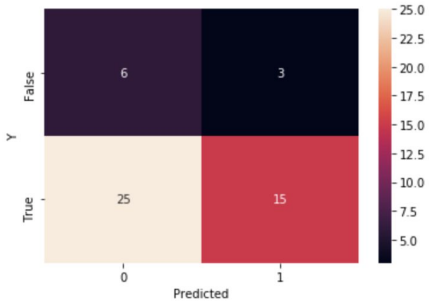
To evaluate the performance of our models we used a confusion matrix to calculate recall and precision:



<u>Threshold of Hospitalization Rate 10</u>
Recall: 41%
Precision: 28%
Training Accuracy: 77%
Prediction: 27%



Threshold of Hospitalization Rate 15
Recall: 79%
Precision: 43%
Training Accuracy: 75%
Prediction: 43%



Threshold of Hospitalization Rate 20
Recall: 51%
Precision: 83%
Training Accuracy: 80%
Prediction: 42%

Figure 10. Confusion Matrices for Different Categorization of Hospitalization Rates.

It's seen that the second model is the best one because it has the highest recall percentage and the prediction accuracy compared to other models.

DISCUSSION / REFLECTION

We thought that the democratic:republican ratio feature would be useful, however, it turned out to be ineffective. With an assumption that the political affiliation of the state is closely tied with the government support to medical assistance and public health support, we thought that the feature could be useful in understanding different policies and hospitalization rate. However, it turned out after EDA this wasn't the case.

One of the challenges that we faced was dealing with NaN (unreported/missing) values. In some situations, we dropped in the rows and columns that had NaN because replacing them with 0s would skew the dataset. However, in other cases we replaced the NaN values with 0 because we did not want to lose the rest of the data tied to that row or column. We were comparing states with one another and we did not want to completely disregard states that did not fully report their data.

We assumed that testing rate and hospitalization rate were indications of the quality of medical/public health in the respective regions. We assumed that the higher the rates, the better resource allocation

and collective effort was. However, it has led us to wrong conclusions due to the lack of contextual data that would better inform us about the public health situation in each state.

Social distancing data would be helpful in exploring whether the introduced policies were effective enough to measure the government responsiveness. Public health historical data in the similar pandemic situations would give us a better picture of the overall policies performance. Data on the quality of equipment and healthcare capacity per region would be very helpful in identifying the major factors that will help to cope with large hospitalization rate, and consequently, to what extent the the government should introduce measurements (i.e. New Zealand's level 2 vs "shelter in place" etc.)

ETHICS

Data collection in itself might have had inaccurate numbers, considering that people who were tested vs. people who were confirmed that they had COVID-19. There might be so many more people who didn't have access to testing and who were not included in the reported data.

It's known that COVID-19 disproportionately affects different communities, mostly affecting low-income communities and regions that don't have enough resources. With an assumption that everyone has the access to insurance and many other factors that might affect an individual's health conditions, hospitalization rates can be very high in marginalized communities. However, we don't have enough data to ensure how much each region is equipped to deal with high rates.

FUTURE WORK

If we had more time and knowledge about the topic, we would build a time series forecasting model since we have dates and the cumulative number of COVID-19 cases. We would also continue from our Logistic Regression model and use the formula of the logistic growth curve to analyze the number of potential cases in the future. The closer the curve is to the carrying capacity, the slower the pandemic should grow.

SOURCES

https://www.nbcnews.com/news/world/medical-workers-spain-italy-overloaded-more-them-catch-coronavirus-n1170721
https://www.cnn.com/2020/03/30/us/brooklyn-hospital-coronavirus-patients-deaths/index.html
https://www.courier-journal.com/story/news/2020/03/31/kentucky-hospitals-prepare-for-covid-19-surge/2923107001/