The COVID-19 Pandemic: How Health(care) and Economic Factors Influence Death Rate

Shauna Hannani, Eshani Mehta, Emily Wang

Abstract

The COVID-19 crisis has quickly escalated into a pandemic in the last 5 months since the first case was reported. In this study, we analyze the existing data on the COVID-19 crisis (confirmed cases and deaths), along with external data regarding the well-being of counties in the United States economic and health wise) to gain a better understanding of how these different factors affect the death rate of people with COVID-19. We analyze three specific factors:

1. General health of the population (looking at how much of the county is high-risk to dying due to COVID)
2. Healthcare system of the county
3. Economic state of the county (income level, poverty percentage, unemployment rate)

Ultimately, this study aims to find a connection between different societal factors to see if one can predict how at-risk people are to be severely impacted by the coronavirus, to better understand and prepare for future impacts of the virus.

Introduction

The first case of COVID-19 was confirmed on December 31, 2019 in China. In the past five months since December 31, the virus has spread to 219 countries and there are over four million total confirmed cases, and over 297,000 deaths worldwide. In the United States alone, there are over 1.4 million cases and over 84,000 deaths; the United States has over 34% of the total cases worldwide, even though the virus originated in China. In light of this, we wanted to understand how different factors affected the death rate different counties had from the coronavirus. We limited our focus to just the United States, instead of worldwide.

For each county, we looked at three different aspects to see how each facet affected the death rate due to COVID-19. The first item we looked at was the general health of the population. The "health" of the population was determined by how much of the population was deemed high-risk, whether they be older than 65 or a smoker. We also looked at how high-risk the county itself is, by factoring in the mortality due to heart disease and respiratory illness, as the lungs and the heart are the two most impacted organs from the coronavirus. We predicted that counties with a higher health risk would have a higher death rate due to COVID-19.

The second aspect we looked at was the state of healthcare of the county. Here, we looked at the SVI percentile of the county, which (by the CDC's standards) shows the overall social vulnerability of that county in terms of an index. We also took into account the access to healthcare in each county, from looking at the number of hospitals and ICU beds to the number of MDs and hospital workers. We predicted that counties with more access to healthcare would have lower death rates due to COVID-19.

The third and final factor we looked at was the economic state of the county. This included the median income of the county, the unemployment rate of the county, and the percentage of poverty in the county. We predicted that counties with higher poverty rates would have higher death rates due to COVID-19.

Data

To begin our exploratory data analysis, we generated several graphs plotting various features of a county against the county's case rate and death rate. We first calculated the case rate for each county by dividing the number of confirmed cases by total population, as well as the death rate by dividing the number of deaths by the number of confirmed cases. We also imported external data regarding the economic state of United States counties (unemployment and poverty rate). We then directed our focus onto features that could be used to reasonably predict case or death rates for a county. Additional data cleaning included merging the `counties`, `confirmed`, and `deaths` datasets on the primary key `countyFIPS`. We converted `countyFIPS` to numbers to be able to match that column to the `FIPS` column in the `confirmed` and `deaths` dataframes for future merges. We then removed any NaN values in the `FIPS` column, since all valid counties had a `FIPS` value.

The `FIPS` column serves as a primary key since it is a standardized key to identify counties, even in the datasets that we found externally. Then, we began the merging of different tables. We first inner merge merged with confirmed on the `FIPS` columns, specifically only looking at the `5/12/20` column since that contains the most updated number of cases per county. We then rename this to a more usable name so it will be easier to reference later. We then similarly inner merge the new merged with deaths on the `FIPS` columns, specifically only looking at the `5/12/20` column since that contains the most updated number of deaths per county. We then rename this to a more usable name so it will be easier to reference later. Finally, we merge the two USDA datasets in a similar fashion to what we did above. We convert the `FIPS` columns in `poverty` and `unemployment` to numeric data again and select the columns from each that we want to use in our analysis before performing an inner merge.

We then create a new copy of all the merged datasets. We want to first add a `case_rate` column, which is just the number of cases divided by the total population of a county. We also multiply this by 100 to make it easier to read. We want to then add a `death_rate` column, which is the number of deaths divided by the number of confirmed cases. Before we this, we remove the 272 counties that have 0 confirmed cases so we don't have a division by 0 error. Since this project is also aiming to look at effects of certain factors on the `case_rate` and `death_rate`, counties with no cases don't add anything to our data. We also multiply this by 100 to make it easier to read. We then added three more columns:

1. `old` is the percent of the population that is age 65+
2. `inMedicare` is the percent of Medicare eligible people who are actually enrolled
3. `medicare_rate` is the percent of the population that is eligible for Medicare

We also noticed in the dataset that counties with latitude and longitude equal to 0 or missing were other territories or not valid counties or lacking important information, so we remove all instances where that is true. We decided we should analyze which columns were missing more data than others, which ended up including include ["3-Yr Diabetes", all the "3-YrMortality", "stay at home" (396), "HPSA" (1013)]. We also checked on what states were included in the data after cleaning. The data now covered 48 states (excluding Alaska and Hawaii since they're not continental and including District of Columbia).

The features we looked at during our initial data exploration stage included political party, dates of stay-at-home orders, median income, and age. Our reasoning for looking into these features was that they either affected the case rate (e.g. not staying at home would result in more people getting exposed and infected) or death rate (e.g. age affects health, which affects how likely one is to recover).

For some of these features, we found it necessary to remove outliers that prevented a relationship in the data from being displayed. In analyzing the distribution of case and death rates among all counties, almost all data points were clustered into one bin (see top of Figure 2/3). Upon further exploration of the data, we noticed that the mean number of cases in a county was 480.75, but New York alone had a case count of 186,123. much higher than even the second-greatest number of cases: 55,000. Excluding New York as an outlier, the correlation in the box plot became significantly clearer (bottom of Figure 2/3).

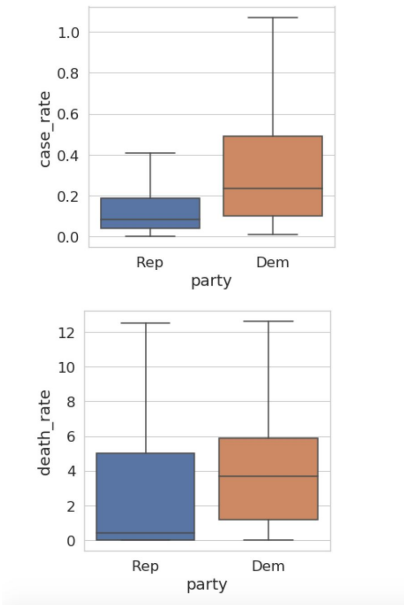


Figure 1: Political Party vs. Case/Death Rate

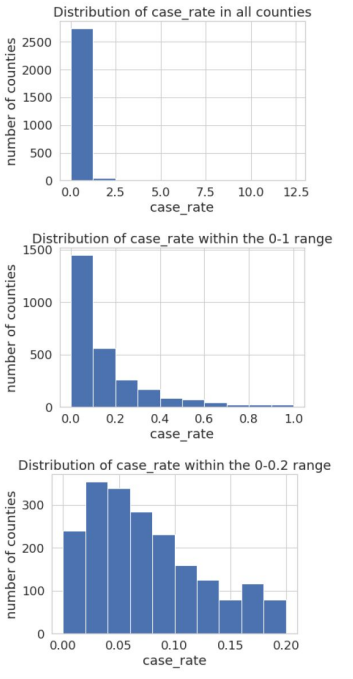


Figure 2: Distribution of case rate (1), excluding bad data (2) and outliers (3).

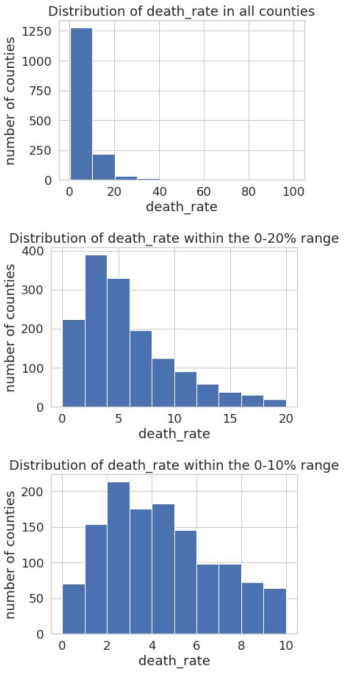


Figure 3: Distribution of death rate.

We used Figure 2 and Figure 3 to gain a better understanding of the distribution of the case rate and the death rate of different counties. We originally were going to include political affiliation in our models (Figure 1), but it did not fit nicely into the three categories we ultimately used to model and predict the data.

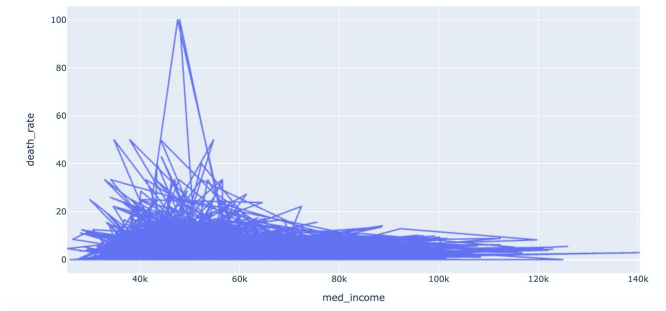


Figure 4: Line Plot ofMedian Income vs. Death Rate

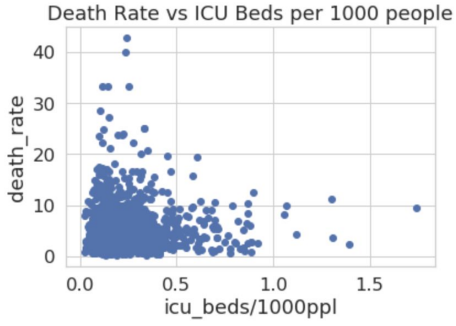


Figure 5: Scatterplot of #ICU beds vs. Death Rate

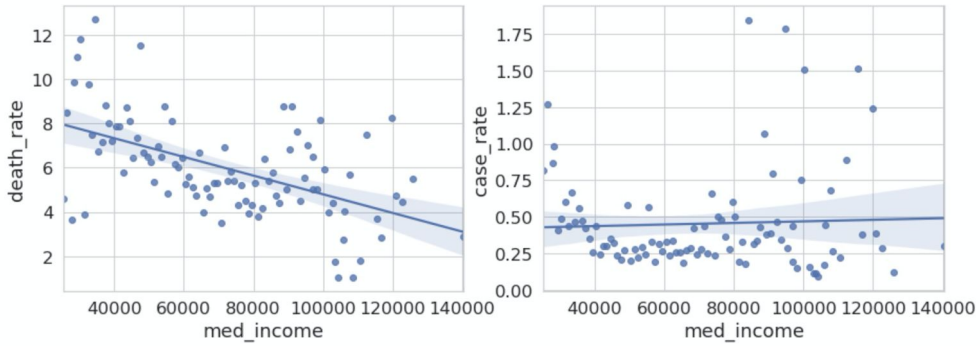


Figure 6: Scatterplot with Best Fit Line of Median Income vs. Death Rate (left) and Case Rate (right)

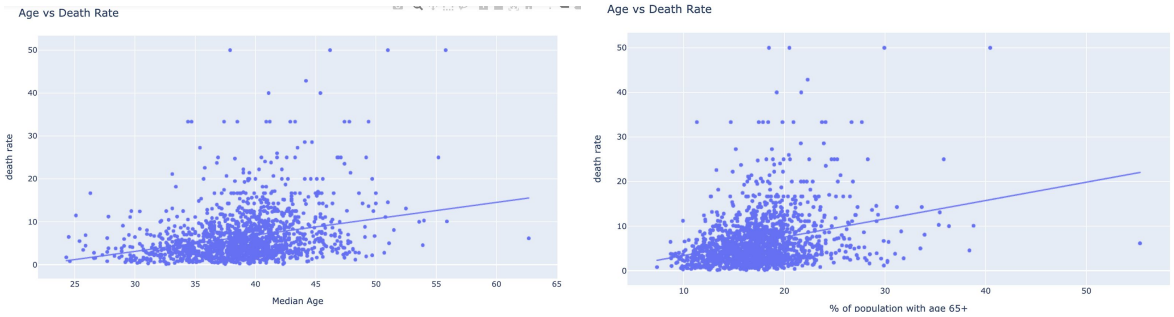


Figure 7: Median Age (left) and % of Population Over 65 (right) vs. Death Rate

We made a simple line plot (Figure 4) between the median income and the death rate, and saw a general negative correlation between the two: We wanted to get a closer look at that trend, so we did some more thorough investigating. We then grouped the income by the thousands (Figure 6) so we would have fewer data points and created a scatterplot with a best fit line to show the correlation even more. Coupled with the case rate graph, the death rate vs median income graph shows that when income is higher, the death rate of the counties is lower because the case rate stays generally constant throughout all income levels.

We had originally thought that counties that had moore ICU beds would have a lower death rate compared to other counties (Figure 5). However, with this graph, we saw that there really was no strong correlation between number of ICU beds in a county and the death rate of the county. On the other hand, with our scatterplots in Figure 7, we noticed a positive correlation between factors related to age (specifically, median age and percent of the population over 65 years old) and death rate. This aligns with what we know about the COVID-19 virus: its mortality rate is higher among people with weaker immune systems, particularly the elderly.

Methods

To answer our research question, we want to build a model that is able to predict the COVID-19 case/death rate for a county based on certain characteristics of the county: e.g. population size, number of hospitals, etc. Since case/death rate is continuous rather than binary, we used a linear regression model that learns weights to give to each given feature to predict a reasonable case/death rate. We split the data into a training set (85%) and test set (15%).

Our first model uses features related to the mortality of common health issues among the county's population to predict COVID-19 death rate, i.e. how likely it is for a confirmed case of COVID-19 in that county to result in a death (or if we're doing case rate: how likely a resident of that county is to be infected with COVID-19). The specific features we chose from the dataset to characterize the general health of a county's population were heart disease mortality rate, stroke mortality rate, respiratory illness mortality rate, the percentage of smokers, and the percentage of the population with diabetes.

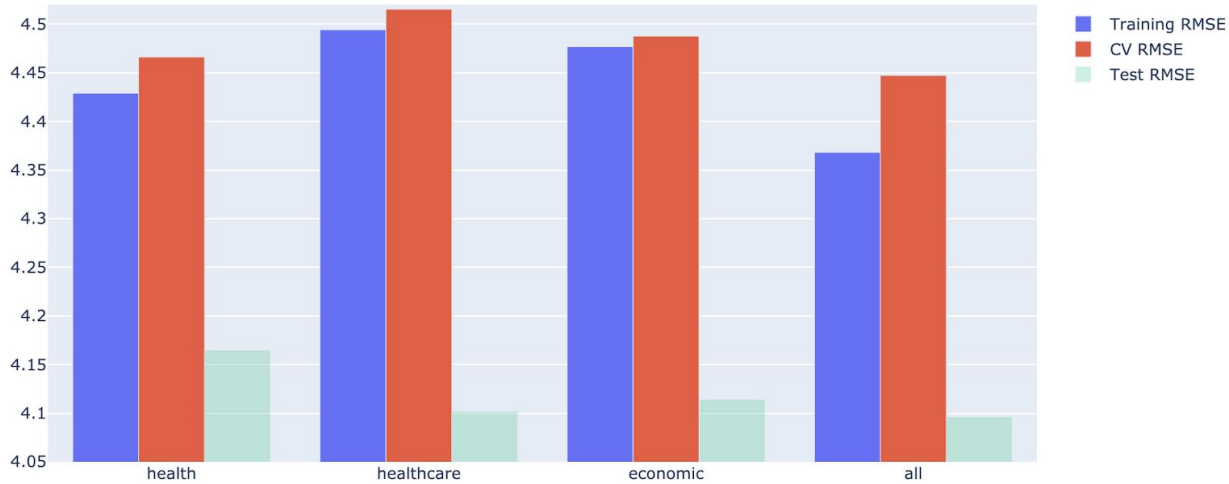
Our second model uses statistics relevant to the healthcare system of a county to predict the county's COVID-19 death rate. The specific features we chose from the dataset to characterize a county's healthcare system were the number of hospitals per 1,000 people in the county, the number of ICU beds per 1,000 people in the county, the percent of people enrolled in Medicare who are eligible for Medicare, the number of MDs in each county in 2017, and the number of full-time employees at hospitals in 2017 in the county.

Finally, our third model used data regarding the economic state of the population of each county. In this model, the aspects we took into account were the county's SVI percentile (the CDC's Social Vulnerability Index), the percent of the population eligible for Medicare in the county, the percentage of people in poverty in the county, the unemployment rate of the county, and the median income of the county.

To create and evaluate each model, we used a Linear Regression and found the RMSE values of the training data. To better evaluate the models, we added cross validation and took the mean of those values as the "CV RSME". We only used the test set at the end, and primarily used the training data to evaluate the quality of our models.

Results

The results (different RMSE values for each model) are shown below. The different models we created performed relatively similarly showing that all three groups contain features that can help predict the death rate for a county. As expected, when we combined the factors, we got the lowest RMSE values, and that was our best model.



Our second model, based on healthcare quality in a county, had a high training RMSE but the test RMSE was fairly similar to the model with all features. Throughout our study, we show that many features such as age and income level that are hypothesized to be correlated to higher risks for COVID-19 do indeed have a higher correlation with death rate. We also show that a combination of these features can serve as a good predictor for death rate, especially when considering counties that are growing in cases and deaths.

Our approach provided a useful categorization of the features which can show what groups have the biggest implications for death rate. We could have evaluated our model with different subsets of the features to see what specific individual features had the biggest impact and thus were the most consequential in trying to analyze death rate.

Discussion

The two most interesting features we came across for our question were how income level and population age statistics affect death rate. We found a strong negative correlation between income level and death rate, as we would have expected; people with more money are more likely to have better access to healthcare to keep them alive. Similarly, there were positive correlations between death rate against both the median age of a county and the percentage of the county's population over the age of 65. As immune systems grow weaker with old age, it makes sense for a county with an elderly population to have a higher death rate, as the virus could be more lethal for the county's residents.

A feature we thought would be useful in predicting death rate was the number of ICU beds per 1000 people; theoretically, the fewer ICU beds available, the less prepared the county is to handle a pandemic where many people need hospitalization, resulting in worsened care for infected patients and a higher death rate. When plotting the `#ICU_beds` column against death rate, however, the data showed no clear correlation and a best-fit line with a slope near 0. Thus, we found `#ICU_beds` relatively ineffective as a feature for our models, and had to combine it with other relevant features regarding the county's healthcare quality to improve our model's accuracy.

The largest challenge we faced with our data was looking for features with a noticeable correlation to case/death rate, as the issue we faced with the `#ICU_beds` column also occurred with other features we wanted to use. We created several graphs plotting various columns of the table, combinations of columns, and even the logs of columns against case rate and death rate, but many graphs showed no correlation at all or were crowded because of outliers.

A limitation of our analysis is that some of the data that we have is from the past (as in, 2014-2016, 2017, 2010 even). For example, the mortality rates we have access to are from 2014-2016, the median age is from 2010, and our information regarding MDs and hospital workers is from 2017. We assumed that this data would accurately represent the present day data, even if it may truly not. However, this is the only data we had access to in full, so we used it for our model -- our model may have turned out to be a bit different if we had up-to-date numbers on the mortality rates, ages, MDs, etc. for each county.

Additionally, in our study, we removed counties where the cases and deaths were zero. This could have emphasized other factors in our model that would have had less of an effect if we had taken into account the counties with zero cases and deaths. There could have been a reason those counties had no cases and no deaths that we had not considered when analyzing the data.

Additional data that would allow us to test other hypotheses would be racial demographic information about the deaths due to COVID-19. There have been many reports on the news that talk about

certain races have a higher chance of dying due to COVID-19, and it would be interesting to test that data out ourselves to see if that is the case, or if it would be because they are located in counties with reduced access to healthcare or more poverty.

Ethical concerns one may have with this data would be if this data were to be used in future elections (see Figure 1). Differing political parties could use the data regarding the case ratio and death ratio of the parties against the other (or for themselves). For example, Democratic counties have a higher death rate and case rate than Republican counties do, so the Republican party can used that to their advantage. On the other hand, the Democratic party can show that larger, more urban counties are more likely to be Democratic so their death and case ratios are only higher because the populations are higher, and therefore the Republican counties actually have a higher chance of getting or dying to COVID-19. Regardless, using this data for anything political would be extremely unethical. To address these concerns, we could incorporate other factors into the study to determine whether the political party affiliation of the population actually has an affect, or if it is really another confounding variable (like the rural vs urban density of the area) that is having the effect on the case and death ratio.

Future work to further expand our exploratory data analysis would be to incorporate more recent data, as previously mentioned, since our data and plots would be more accurate. Additionally, since we expected to see more relationships between certain features and case/death rate than we actually did, perhaps incorporating more accurate recent data would help clarify these cases. Our model also has much room for improvement, so finding features with a stronger linear relationship to case rate or death rate would be a good next step.