Understanding COVID-19

Brian Chin, Christine Nguyen, Oscar Hu

***Abstract* —** *As the outbreak of COVID-19, or more colloquially known as coronavirus, continues to spread rapidly across the United States and the globe, many questions arise surrounding this unprecedented pandemic. However, living in the age of technology and data, the answers to these questions and further insights are much more accessible than in previous pandemics. This paper summarizes our findings from the COVID-19 datasets and discusses the methods we used for exploratory data analysis, feature selection, and modeling.*

I. INTRODUCTION

With the uncertainty surrounding the spread and effects of COVID-19, many questions arise: Does heavy government interference improve the wellbeing of the nation/state during a pandemic? Why are certain areas experiencing more severe repercussions than other areas? What population demographics are most vulnerable during a pandemic? Is it possible to predict the number of cases/deaths with the current data we have and what we know about past pandemics?

The aim of this paper is to (1) draw insightful patterns and trends from the data, (2) find what county attributes are more influential in predicting the number of confirmed cases, and (3) predict the number of confirmed cases and deaths over the next week. First we will introduce the datasets and data frames that were used, then we will provide a general summary of the data including some visualizations and observations that we made. Then we will discuss our two models: linear regression model to predict the number of confirmed cases in a county, and the time series model to predict the future number of cases and deaths.

II. THE DATA

We used datasets containing different information about COVID-19 [1][2]. The states data frame includes general statistics from nine unique countries/regions such as the number of cases, deaths, and recovered, and several different rates. Although we did not use the states dataframe for modeling purposes, we used it to extract several general trends across countries and across states in the U.S. that will be discussed in more detail later in the paper. The counties data frame consists of over 80 observations for each county in the U.S., containing information about general population demographics and the dates in which the different shelter in place laws were enforced. The confirmed and deaths data frames are time series from January 22 to April containing the number of confirmed cases and deaths respectively in each state in the U.S. The infections data frame is a time series of the confirmed cases on the county level in the U.S.

III. SUMMARY OF THE DATA

Initially, we compared the data on the country level. In Fig. 1 we can see the percentages of active cases, recovered, and deaths by country on April 18th. The percentages were calculated by taking the column of interest and dividing it by the number confirmed, and this is assuming that Active + Deaths + Recovered = Confirmed. The percentages for the U.S. adds up to over 100% so there might have been some error in the data collection process, as the total is greater than the number of confirmed cases.

Some interesting things to note is that a high percentage of infected people in the Netherlands die. This could be due to a number of reasons such as the country could have an older population, people who are more susceptible to passing away due to illness; however, this is just speculation, as we do not have the data to prove this. Surprisingly, the U.S. has the lowest percentage of confirmed cases who do not recover from the coronavirus. It would be interesting to also compare the testing rates of these countries but we only had the data for the testing rate of the U.S.

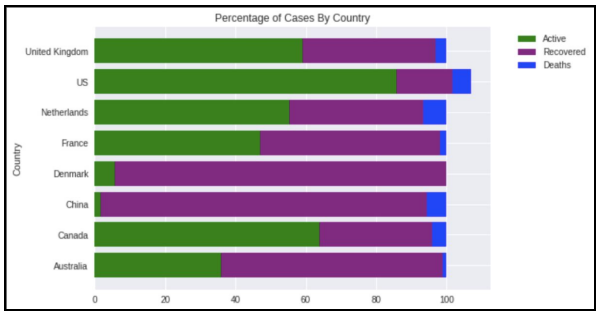

Fig. 1 Percentage of Cases By Country

Moving forward and looking at the data on the state level, we wanted to see how the mortality rate, number of cases, number of deaths, and testing rate compare across each state and if there is any correlation statewise. Unfortunately, hospitalization rate data was not available in some states.

With Fig. 2.1.3 and Fig. 2.1.4, it is evident that New York's number of confirmed cases and recorded deaths are significantly higher than the rest of the states, showing that New York is at the epicenter of the pandemic in the U.S. It is also clear that the states surrounding New York also have high numbers of infected people because of proximity to New York. States further away from New York and at the border of the U.S., such as California, Texas, Washington, and Florida also have high numbers of confirmed cases. Much of the inland US, especially the Northwest around Montana and Idaho are relatively unaffected. In Figure 2.1.2, the testing rate, a percentage of total people tested per 100,000 persons, shows that the state of New York has the highest testing rate probably due to having such a high number of confirmed cases and because of the exceedingly high concentration of people in New York. Other states with a high testing rate include Louisiana, and states surrounding New York, possibly in an attempt to curb the spread of Covid-19 out from New York. In Figure 2.1.1, which shows the mortality rate, or the rate of deaths over confirmed cases, it is observed that states in the Great Lakes region have a relatively higher mortality rate compared to the other states in the US, with exception to New York, Oklahoma, Washington, and Louisiana.

State Level Heatmaps:

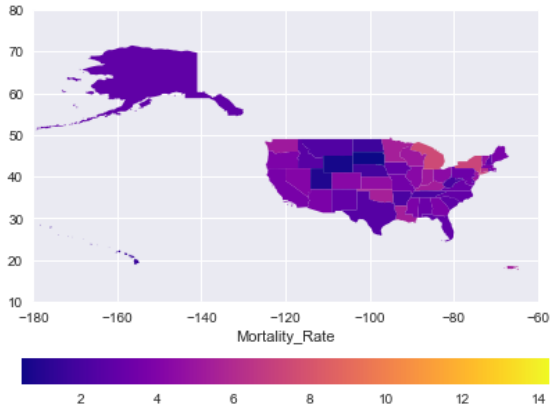

Fig. 2.1.1: Mortality Rate Heatmap by State

*(Number recorded deaths * 100 / Number confirmed cases)*

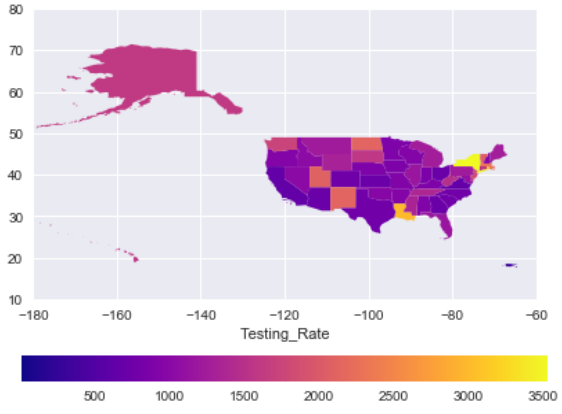

Fig. 2.1.2: Testing Rate Heatmap by State
(People tested per 100,000 persons)

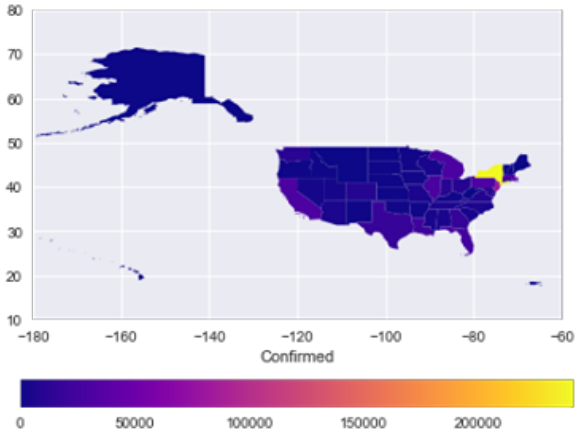

Fig. 2.1.3: Confirmed Cases Heatmap by State

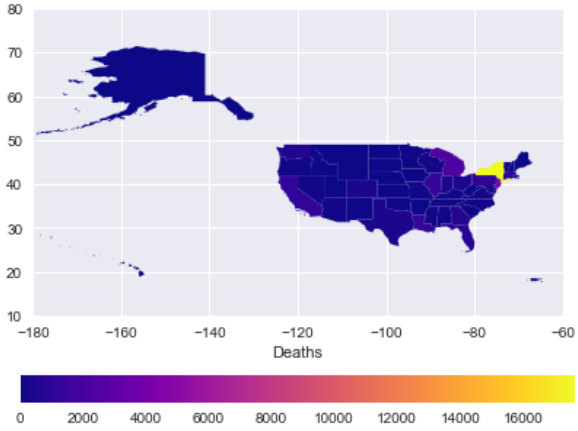

Fig. 2.1.4: Deaths Heatmap by State

In an attempt to further trace the epicenter of the outbreak in the state of New York, we examined the county level of the US, all 3142 of them. When taking the same heatmap on the confirmed cases and deaths in the US, but at a county level, almost all counties had a confirmed cases less than five thousand. Besides a sparse number of exceptions, most of the counties that have a higher number of confirmed cases lie in New York, as shown in Figure 2.2.1. As of April 18th, 2020, the highest number of cases in any county in the US is Queen's County, New York, at 40216 confirmed cases. Counties around this county also have extremely high cases. Long Island's county, Suffolk County, has 26143 cases, while King's County has 35763 cases. It is observed that the areas around this epicenter are certainly influenced by the epicenter, as there is a radius of higher than average cases around Queen's County. The recorded deaths on the county level has a similar perspective, where the counties that have the highest number of recorded deaths are also in New York, while the rest of the counties in the US have less than 500 deaths. The neighboring county of Queen's County, King's County, which aforementioned had the second highest number of confirmed cases, has the highest number of recorded deaths as of April 18th at 3612 deaths. Queen's County has the second highest at 3466 deaths. Similarly to the heightened radius of confirmed cases around these two counties, there is a radius of a higher number of deaths than average.

County Level Heatmaps (of New York):

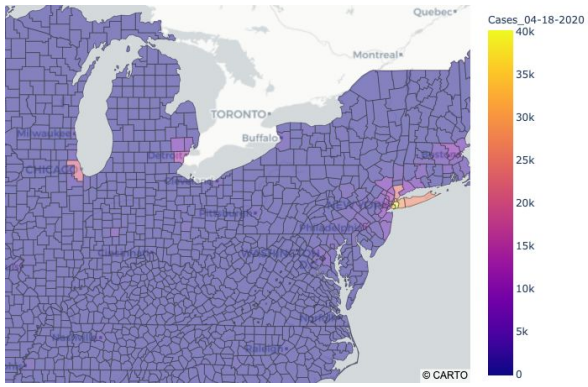

Fig. 2.2.1: Confirmed Cases Heatmap of New York

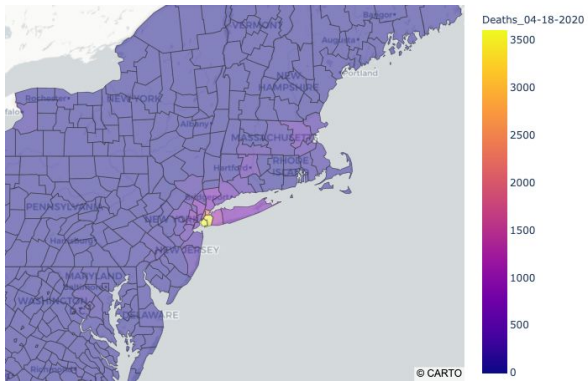

Fig 2.2.2: Deaths Heatmap of New York

We also thought it would be interesting to visualize when different states enacted the *shelter in place* laws in order to help flatten the curve. California was the first state to enact the shelter in place starting on March 19th, and South Carolina was the last state starting on April 7th. In Fig. 3 many states began the shelter in place around March 23-24.

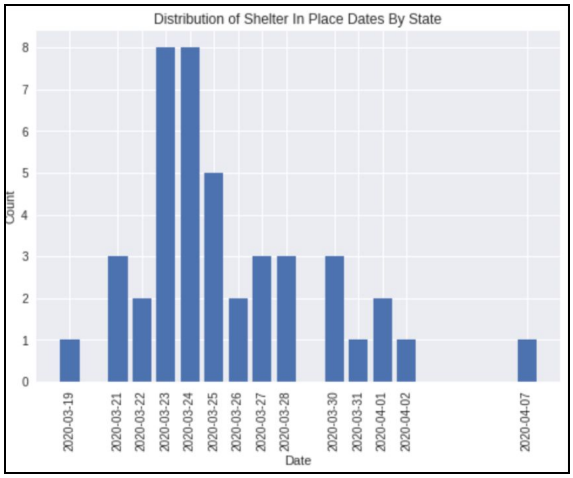

Fig. 3. Distribution of Shelter in Place Dates By State

IV. MODELS AND RESULTS

A. *Linear Regression*

In order to decide which features in the counties data frame would be most influential in predicting the number of cases in a county, we decided to use a linear regression model.

Cleaning the Data

Since we were doing the predictions by the county model, we had to merge the infected and counties data frames in order to get the features of the counties and the cumulative number of confirmed cases the county had on April 18th. After merging the infected and counties data frames, we dropped the columns with ordinal values. When checking for null values, we saw that there were some features that had thousands of null values, such as the columns "3-YrMortality" for each age group. We decided to drop the features that had over 1000 null values, as those would not be helpful in our model. South Dakota also had no data so we dropped that state from our data frame.

Feature Engineering

The columns with categorical data such as CensusRegionName, CensusDivisionName, and State had to be one hot encoded. We also made a heatmap of the correlation of the features in order to determine which features were dependent. There were many columns regarding population size and this created multicollinearity in our data. We decided to only keep the column PopulationEstimate2018. The columns with dates such as "stay at home," "public schools," etc. were converted to the number of days since the first date in its column. This is because when running the linear regression model, we can not have dates.

Results

First we ran the model with all the covariates and to no surprise, it performed horribly with a Test RMSE of 2,530 and a Train RMSE of 923. The model is clearly overfitted and is predicting the test very poorly. After running KFold Cross Validation, we received a RMSE of 1138 but the Test RMSE was still high at 2455. The final model included 13 features and had a Train RMSE of 1338 and a Test 1384. Fig 4. Shows the residuals from the final model and shows a relatively straight line along the x-axis.

Some features that were not helpful in the model were the number of hospitals, percentage of people with diabetes, percentage of people who smoke, and respiratory mortality rate from 2014. Possible reasons as to why these features were not helpful could be that the number of hospitals was too closely related to the population of the county. We originally expected smoker percentage to be a useful feature as well, but upon further analysis we realized that it had little to no correlation with the number of cases. This may be

because smokers do not necessarily have a higher chance of spreading the disease. While it is true that both COVID-19 and smoking create a number of respiratory issues, the act of smoking may not increase physical contact with others and thus no strong correlation with new cases.

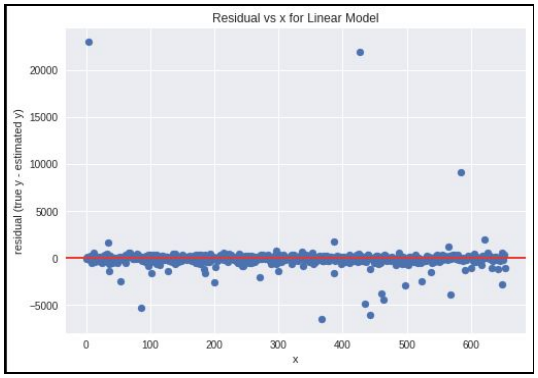

Fig.4 Residual vs X for Linear Model

B. Time Series Regression

One convoluted question about the outbreak is when the curve of cases and deaths flattening. We decided to try implementing a Time Series Regression model based on the previous four months, starting our data collection from January 22nd and ending on May 10th. We used the usafacts_infections.csv dataset that contained county-level data on the cumulative number of cases and number of deaths per county per day. The first row of this dataset was not a valid county, so it was dropped.

Building the Pipeline

The Time Series Model uses a lag function and a forward chaining k-fold validation function. Forward chaining k-fold validation is basically chaining an initial training set of the first *n* days and a validation set of a consecutive subset of *m* days. This chained set would be the new training set and another consecutive subset of *m* days would be used as the new validation set. Typical k-fold validation that involves shuffling the data and splitting it won't work for Time Series as each day influences the next day. The lag function used calculates:

1. The prior day's cases/deaths
2. The difference between the current day's and prior day's cases/deaths.

The lag function creates these features to be used with the current day's cases/deaths and then inputted into the model as training data. For the Time Series model, we opted for a Random Forest Regressor customized to be used on Time Series. In retrospect, a more Time Series-like regressor, such as autoregression or moving-average regression would be more suited for this dataset. We also decided to use RMSLE (root mean squared log error) for determining error in our cases model and RMSE (root mean squared error) for determining error in our deaths model as they counteracted the large outliers in our dataset.

Evaluation of Model

For the Cases Model, we used just the cases as the initial feature set and yielded a .268 RMSLE error when predicting the latest in our data collection, May 10th. Figure 5 shows a visualization of this error per county, but no apparent observations can be taken from it, except perhaps that Minnesota holistically was predicted poorly.

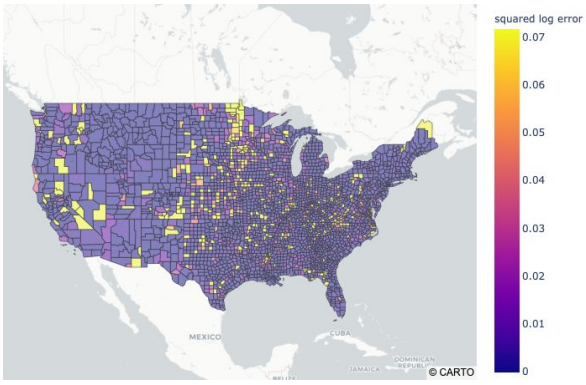

Fig. 5: Squared Log Error of Predicted Cases by County

Instead, we looked at the maximum RMSLE yielded from our model, and tried to see where our model is faulty. The county with the highest error was Dawes County in Nebraska. In this county, there was only 1 case in the last two days where our data collection ended. May 9th and May 10th. The predicted abrupt rise in cases indicates a trend that the model did not, or could not account for with the features given, which were the time series of just the cases and the lack of more data collection. This is why this county has the highest error when predicted by the model.

Similar cases to the Dawes County are apparent during the Covid-19 Outbreak, showing that these outbreaks are hard, with the features given, to even ballpark the predicted number of cases in the next few days, much less weeks. This is why there is so much shrouding the expectancy of stay at home martial laws and when the curve will flatten. Since each county holds the same weight in this time series model, there is too much variance between the trends of each county to develop a proper and accurate model with just the features given in usafacts_infections.csv. More modeling, such as using autoregressive models or moving average models as well as having a feature set that is not as fluctuating as the current one, will result in much more accurate predictions. Random Forest Regression is just not up to par with this dataset.

As for the deaths model, we used the same model but instead used the deaths per county per day as our initial feature set. Using RMSE as our loss function, we yielded a 3.78 RMSE error when predicting May 10th deaths in each county. Figure 6 delivers the visualization of each county's error. Here, compared to the cases error visualization, we see a clear centralization of high error along the coast near New York and Pennsylvania.

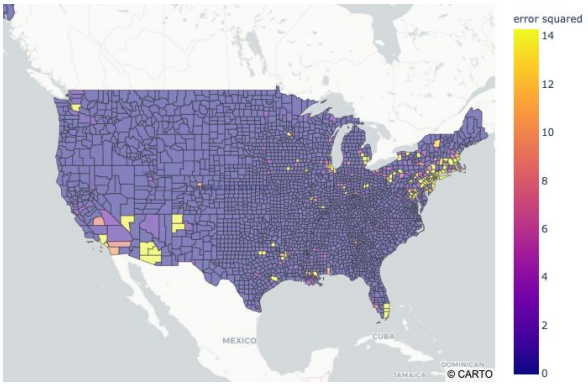

Fig. 6: Squared Error of Predicted Deaths by County

We likewise found the county of the maximum error in our death model to see how our model was faulty. The county with the highest error was Cook County, Illinois. Here we see the complementary case of the maximum error in our case model. This county increased in the last two weeks but started stagnating very recently. The model tries to compensate for these two ends of the spectrum by averaging the trends out.

Prediction and Results

Finally, we predicted the week of 5/11 - 5/17 for both cases and deaths in each county by stacking predictions. That is, predicting one day then using it as a datapoint for the next day's prediction. Figure 7 shows the predictions with the actual data for the number of cases per county, respectively. Here is an alarming reason behind why the model flattens on some cases but skyrockets on others. Since the majority of counties have very little cases that jump but then stagnate, counties that have a large amount of cases tend to stagnate in the prediction, while counties that have very little cases have drastically increased predicted number of cases within the 7 day prediction window. In the graph

above, only 9 counties have cases above 20000.

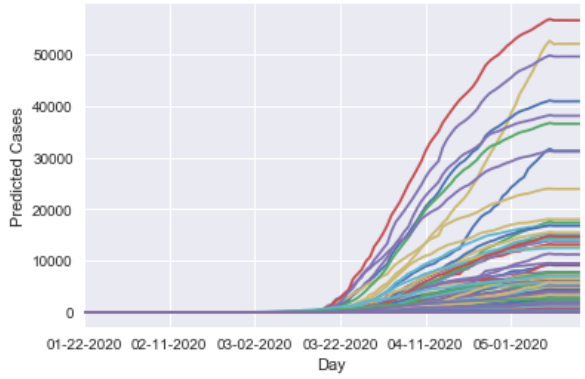

Fig. 7: Squared Log Error Heatmap by County

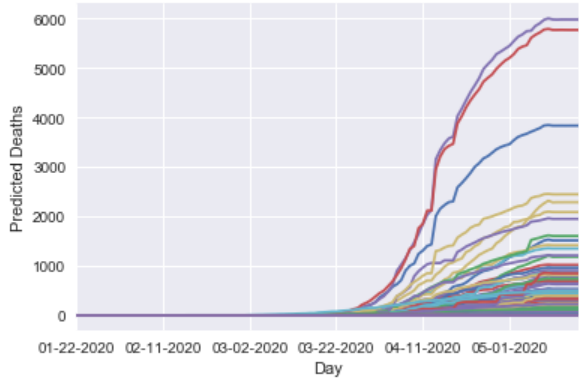

Fig. 8: Squared Log Error Heatmap by County

Figure 8 shows the predicted deaths alongside the actual data similarly to Figure 7. Similar to the cases model, but much less apparent, there are a few counties that jump very quickly, but since all counties have equal weight, the overall trend the model predicts flatlines at a high death rate and jumps at a low death rate before flatlining again.

V. DISCUSSION

With the dataset used in Time Series Regression, just merely the number of cases and deaths per county is not enough to develop an accurate model. In truth, there are a lot of other factors and features that are in play when determining the direction of an outbreak. Certain features, such as social activity in each county were definitely contemplated but for clarity's sake were left out of the final feature set. More time and more resources at our disposal would allow us to explore more complicated feature sets than just the elementary feature set used in our time series model. Also, as seen in our results, county-level data is heavily fluctuating, as a majority of counties are minimal in cases and deaths, while a few counties hold a high amount of both. Likewise, the trends of these counties varies heavily, based on features that are not implemented in our feature set. Certain time series regressors such as ARIMA alongside more complex feature sets would be better equipped to give an accurate prediction. The infection dataset also had a window of 60 days where the highest number of cases per county were around 2 to 3 and with a majority of them at 0. This is clearly shown for both cases and deaths in Figure 7 and 8, where there is no fluctuation in the counties until around March 22nd, two months after the beginning of data collection. We assume that these data points, although seemingly redundant, should be factored into the model, but are not entirely sure if it is necessary or it may even be detrimental to the final result.

For the linear regression model, we found it interesting how the ratio of democrats to republicans in a county was effective in predicting the number of cases. We expected the respiratory mortality feature to be impactful as COVID-19 is a respiratory infection that attacks the lungs. However, this turned out to be false after further graphing and visualizing the feature. One possible explanation is that this feature data is from 2014, which could very likely be outdated in a span of 6 years. Similarly, the percentage of smokers per county was not effective despite also targeting the respiratory system. This may be due to the fact that the act of smoking does not imply increased physical contact, which is needed for the spread of COVID-19.

Feature selection was probably the hardest obstacle to overcome. We found it difficult to narrow down and identify the features

that had a high impact on predicting cases. The limitation of the analysis was that the linear regression and time series regression only included features from individual datasets. If we had more county-level features that encompassed other potential aspects of the outbreak such as a wide hospital-level dataset, there would be a more complex feature set to choose from. We saw that the democratic to republican ratio was significant in the linear regression model. However, this part of the data possibly only accounts for people who had answered their political affiliations.

Perhaps unethically, we dropped numerous rows with missing data when attempting to work with the data. As a result, the data may be biased towards the data that was present. For our time series dataset, there were no missing values, but we did drop a row that did not correspond to any county, since its countyFIPS was invalid. A significant part of the data was the population estimate per county. However, the data collected for this does not include the undocumented people living in the U.S. who are still being affected by COVID-19.

Given more conclusive and descriptive data regarding government interference (laws, acts, etc.), we can more accurately predict which features were effective in predicting the number of cases. Furthermore, additional data on specific races and ethnic percentages would allow us to determine the effect on any particular race. The repercussions of COVID-19 are not equalized among different races due to a variety of social factors. As a result, it would be beneficial to identify where these issues occur and what specifically causes these racial discrepancies.

With access to more data, a concern that arises is the privacy of those whose information is being used. Gathering data on COVID-19 victims and family is an ethical concern that should be abided by and respected.

REFERENCES

[1] “JHU CSSE COVID_19 Dataset,” Available: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

[2] Yu Group, University of California, Berkeley Available: https://github.com/Yu-Group/covid19-severity-prediction/tree/master/data