Modeling and Forecasting COVID-19 Outcomes

Connor Tou, Lily Shang, and William Hou

***Abstract*—With the overwhelming number of deaths caused by SARS-CoV-2 and its increasing spread, computational methods to model COVID-19 outcomes are proving crucial to understand susceptibility factors as well as to inform when certain regulations should be implemented or lifted. In this work, we look to explore both of these two modalities. First, we implement a model which predicts outcomes based purely on state infrastructure and demographics data pre-COVID-19 in order to pinpoint features that are indicative of these outcomes - which can be targeted in preparation for future waves of this or other biological threats. Second, we utilized historical data to forecast deaths and cases via linear regression and ARIMA models. Ultimately, we were able to make decent predictions for both ends of this project; however, many computational improvements or alternatives, coupled with further data collection and biological/epidemiological understanding of this virus, are needed to increase the accuracy, usability, and impact of our models.**

I. Introduction

SARS-COV-2 has caused a global pandemic, with over 4-million cases worldwide and over 80-thousand deaths in the US alone [1]. This virus is characterized by its ability to cause upper and lower respiratory tract infections and is thought to be transmitted through contact of respiratory droplets from infected individuals. As such, risk groups as identified by Dr. Auwaerter from John Hopkins University include individuals with underlying health conditions, and adults over the age of 65 who are more susceptible to severe symptoms and death [2]. Many states in the US have begun to enforce mandatory lock down of public spaces; however, even with the self-isolation in place, confirmed cases are still rapidly rising as of May 2020 [3].

Due to the novel nature of SARS-CoV-2, much research needs to be done to understand its epidemiology - specifically transmission patterns and seasonality. While this data is not yet known, geographical, political, and demographic factors are certainly at play. Additionally, many sources, such as the Center for Disease Control (CDC), have been tracking historical data, which may be utilized to forecast future outcomes. We sought to explore both these modalities. In our project, we modeled COVID-19 outcomes using data from counties and states, including their confirmed cases and deaths counts, and in doing so, explored which features are important in predicting the number of COVID-19 related outcomes. Next, we utilized historical data in order to forecast future total deaths and total cases. Though, it should be noted that our lack of epidemiological understanding coupled with the lack of a confirmed cases plateau limits the accuracy of current models - as well as how we approached our model. Overall, with this work, we hope to better understand the critical factors which make a certain region more susceptible to COVID-19 and can be altered to minimize the number of new cases and deaths as well as investigate potential future outcomes based on current and historical data which may inform future regulatory and personal decision-making.

II. Mapping COVID-19 related deaths, cases, and mortality rate by state

The large range in total deaths, total cases, and mortality rates among different states are results of differences in regional characteristics as well as COVID-19 related regulations. In order to visualize these outcomes based on these differences, we geographically mapped them using geopandas and geoplotly packages at state and county-level granularity (Fig. 1). We looked to visualize which counties and states had higher and lower raw numbers in order to start investigating what state and/or county-specific factors contributed to these results. To ease our analysis, we only took into account the 50 U.S. states and excluded U.S. territories and D.C. We hypothesized from these maps, literature, and our intuition that features such as population, population density, medicare enrollment, racial makeup, population over 65, among others, would be important in predicting the number of deaths and cases. In the next section, we delve deeper into these and other features in developing our first model of this project.

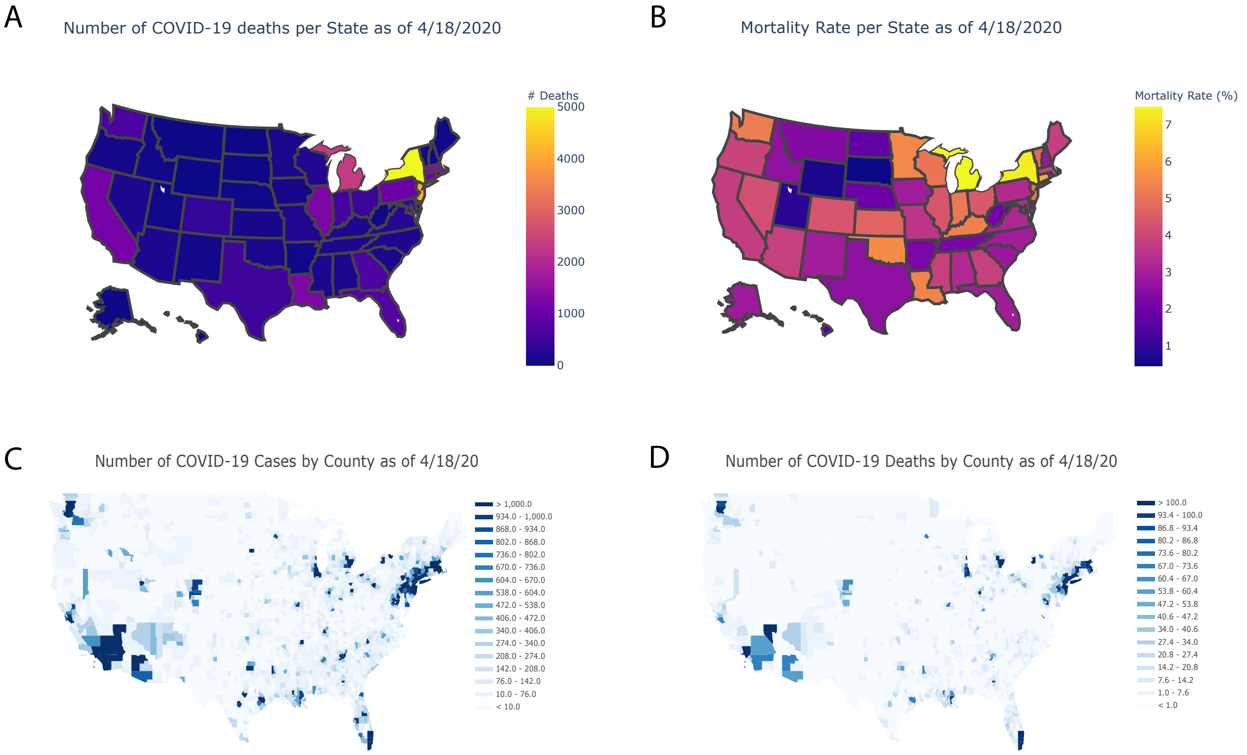


Fig. 1. **COVID-19 related deaths, cases, and mortality rate by state.** (A) Number of deaths per state (B) Mortality rate per state (C) Number of confirmed cases per county (D) Number of deaths per county. All numbers are as of 4/18/2020 (the reporting of this data)

III. Modeling the number of confirmed cases or deaths using critical state characteristics

A. *Data cleaning and preparation*

In order to investigate our hypotheses from Section 1, we looked to develop a model to predict how many deaths or cases a state would have as of 4/18/2020 (the reporting of this data) based only on a state's infrastructure and demographics pre-COVID-19 and potentially state guidelines. Importantly, since our target variables are at state-level granularity in the provided states dataset, and in order to ease our analysis, we first converted all of our features from county-level to state-level features. Therefore, our initial transformations involve these steps.

First, we created a dataframe containing the Gregorian dates corresponding to when a state implemented certain guidelines for each county. We converted these dates to a datetime object and calculated the number of days from January 1st to the date that the policy was implemented. Because each county in a given state had guidelines implemented at the same time, the first entry after grouping by state, was taken. The resulting dataframe contained integers for each state and for each set of specified guidelines. NaN values converted to our standardized integer system resulted in a value of -737424, which we filled with zeroes to represent that we lacked the initial date. Though this would correspond to 1/1/2020, and the state for that guideline would therefore be given the most conservative date, we preferred to deal with these values in this way instead of dropping any state or guideline measures.

As people of older age or those with preexisting health conditions generally have weaker immune systems, it is unsurprising that these attributes could be indicative of COVID-19-related outcomes [2].

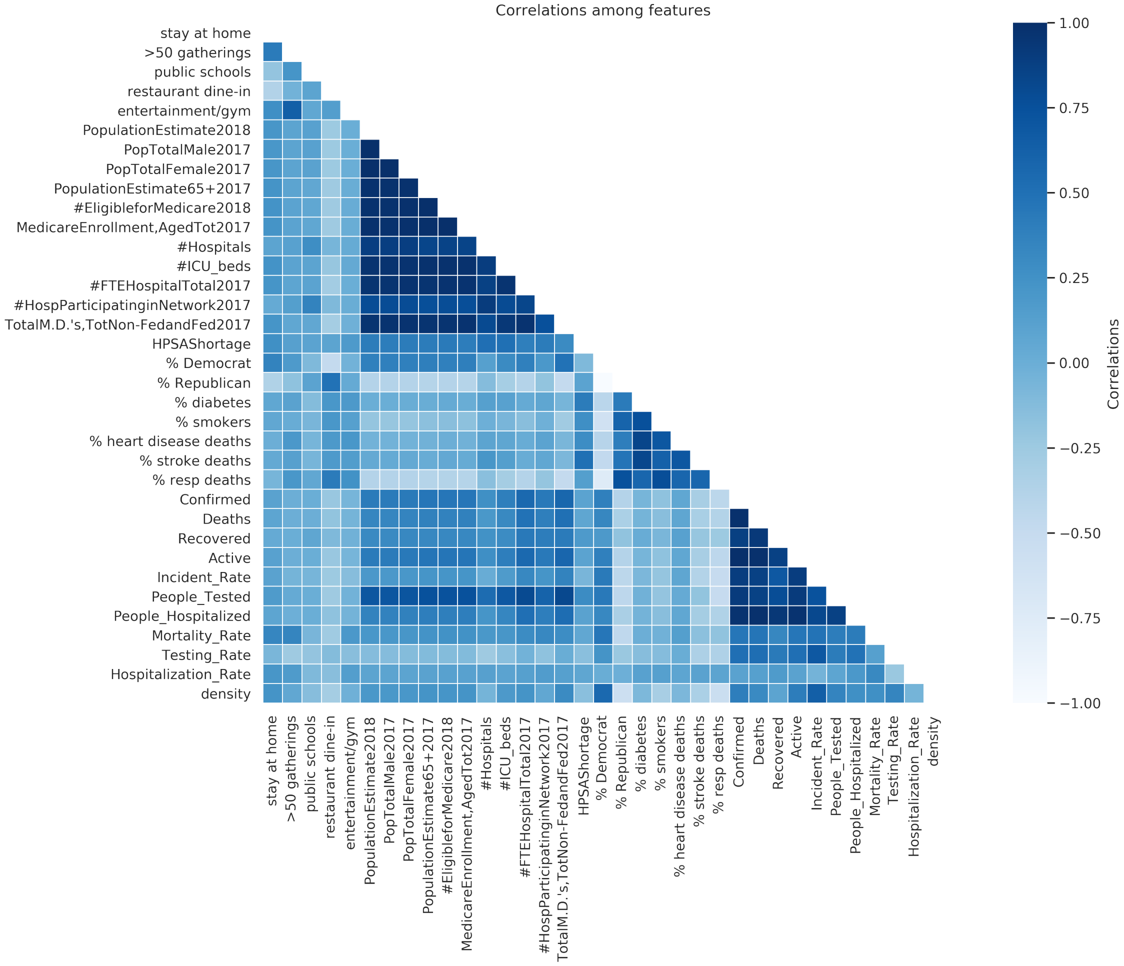


Fig. 2. **Correlations among features from merged state and modified county dataframes.** Strength of correlations between a given feature and other features/target variables ('Confirmed' & 'Deaths') are shown.

Moreover, we hypothesized that the partisan makeup of a state could influence both when state guidelines were implemented and the adherence to these guidelines. To convert our county-level data to state-level data, we calculated the total number of smokers, people with diabetes, Democrats, Republicans and the total number of deaths from strokes, respiratory disease, and heart disease for each county using respective percentages and mortality rates multiplied by a given county's population. Counties were grouped by state, and all values, including other features included in the original counties data deemed to be potentially useful (e.g. age demographics, #eligible for medicare, #hospitals, #ICU beds, etc), were summed. This underlies a large assumption that all counties in a state are present in this dataset. Percentages for each of the specified features were then back-calculated using each totaled value and the total population. Therefore, the resulting dataframe contained the % smokers, % diabetes, % Democrat, % Republican, % deaths from strokes, % deaths from respiratory disease, and % deaths from heart disease per state in addition to raw values for other features for each state. This dataframe and the dates dataframe described in the previous paragraph were then merged on state name.

Lastly, we looked to the provided states dataset containing the number of confirmed cases and the number of deaths per state - our target variables we are trying to predict. We merged this dataframe on state name with our previously created merged dataframe, filled any NaN values with zero, and only

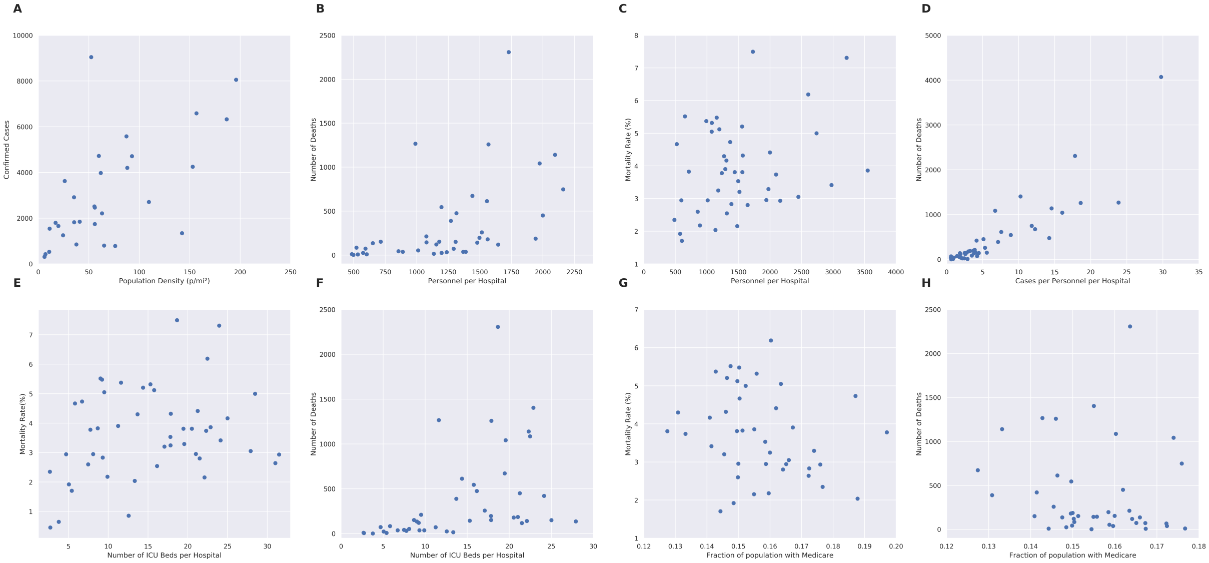


Fig. 3. **Feature visualizations/EDA.** Number of confirmed cases, number of deaths, or mortality rate were plotted against various engineered features for us to evaluate whether the particular features would benefit our model.

kept our target columns and additional features that we thought could be informative.

B. Feature selection/engineering and modeling

With our dataframe now cleaned, we proceeded to seek the critical features for building the confirmed cases and deaths prediction models.

First, we created a heatmap using Seaborn package to evaluate the correlations among the features themselves (Fig. 2). Based on the correlation heatmap, we picked features that looked highly correlated with our target variables of ”Deaths” and ”Confirmed”: state population density, #ICU beds, #Hospitals, #FTEHospitalTotal2017, TotalM.D.’s, TotNon-FedandFed2017, and MedicareEnrollment, AgedTot2017.

Next, we hypothesized from intuition and from reports by the Population Reference Bureau (PRB) that population density would be a key factor in COVID-19 spread and therefore the number of cases [4]. Hence, we plotted the state population density against the number of confirmed cases and observed an expected positive trend between the two variables (Fig. 3A). We also hypothesized that a larger number of medical personnel per hospital would be indicative of both a lower mortality rate and a lower number of deaths of deaths overall. We added up the number of healthcare staffs and the number of doctors per state, divided this sum by the number of hospitals per state, then plotted this ratio against these two outcomes (Fig. 3C/3D). From these plots, we observed an initially perplexing positive correlation for mortality rate, and past a threshold of around 1000 personnel per hospital, the same trend for the number of deaths. Upon further thinking, we realized that states with larger hospitals, and therefore likely with more personnel per hospital, will also see many more COVID-19 related (and other health) cases. Larger, more equipped, hospitals will tend to be located in areas with more people overall and a lower density of hospitals. Therefore, due to the overall larger number of patients these types of hospitals see, the number of deaths would increase in number despite having more personnel - which was what we observe. Moreover, the mortality rate would increase as the number of patients at these hospitals is exponentially more than the relatively higher number of personnel. Therefore, for predicting the number of deaths or the mortality rate, we concluded that we wanted to use the ratio of the number of cases in a state to the number of personnel per hospital. Indeed, when we plotted this ratio against the number of deaths, we observed a clear positive correlation (Fig. 3D). However, because the goal of our model is to predict the number of deaths or cases based purely on state infrastructure and demographic data, we did not use this feature in our model (described later). Instead, we concluded that our original feature (Personnel per Hospital) would be a valuable metric. Next, we wanted to visualize the relationship between mortality rate or number of deaths and the number of ICU beds per hospital. Again, we found a confuddling occurrence in that the correlation was positive, though weakly, overall (Fig. 3E/3F). Using the same logic as in the previous example, hospitals with more ICU beds likely served more people. There are also perhaps less of these larger hospitals in a state. Those with more ICU beds served many more people; therefore, the number of deaths would logically increase. Additionally, as before, the mortality rate would increase with increase ICU beds per hospital since the number of patients is much larger in proportion to the relatively higher number of beds in the hospital (i.e. the number of beds designated to a given hospital is not linearly and equally proportional to how many people it serves). However, we decided that this feature would still be useful. Next, We reasoned that the fraction of people enrolled in Medicare would affected a stateâs COVID-19 related outcomes. We assumed that the patients with healthcare plans would be more likely to and able to seek medical treatment early, and therefore, they would be less likely to die. First, we plotted this fraction against mortality rate and observed a weak negative trend as expected (Fig. 3G). We plotted this ratio against the number of deaths, and we did not observe a clear correlation instead of an expected negative correlation (Fig. 3H). However, overall state population may account for this - two states may have the same ratio, but the number of deaths will be different based on total population, density, and other factors discussed.

With these new features, we plotted and observed an updated correlation heatmap (Fig. S1) and began defining our linear regression model. We were motivated to see if we could predict the number of deaths or cases based purely on a state's characteristics from pre-COVID-19. We thus did not want to factor in COVID-19 related data except, of course, the response variable we were trying to predict. Therefore, the goal of this model was to only use data corresponding to state infrastructure or demographics in order to predict these outcomes as of 4/18/2020 (the reporting of this dataset) per state. We imported a linear regression model and split our data into training and test datasets according to an 80:20 ratio. Notably, our X-train and X-test were standardized since many of our features were on different scales. We then constructed a linear model for confirmed cases and a similar model for number of deaths using the same set of features that we deemed to be crucial. After fitting our model with our training data and calculating and R^2 value using *model.score* on our training data, we calculated the root mean squared error (RMSE) for both our training and testing datasets using the predictions from each dataset and our Y-train and Y-test data. Our scores and error values for our two models are shown in Table S1, and our final features used for both models are contained within our attached notebook. We believe the models reflect relatively strong predictions based only on state's characteristics data.

Notably, throughout this process, we also discovered features that we thought were going to be useful but turned out to be less predictive of our target variable and correlated with other features in unanticipated ways. One example was the partisanship landscape of a given state and the date a state instructed a stay at home order. Originally, we had hypothesized that, in general, states with a republican majority would have instructed stay at home orders later. Looking at our correlation heatmap, however, we observe that in general, as the % Democrat increases, the number of days the state instructed stay at home orders increases as well. This was the opposite as we look at increasing % Republican - which shows a negative correlation. Additionally, we hypothesized that people in increasingly red states might be less adherent to state guidelines overall, which may influence the number of cases and therefore the number of deaths. Notably, we found that including this demographics data and/or the integer dates corresponding to the number of days after January 1st state guidelines were implemented only increased our model fit by 0.01 and only slightly decreased our RMSE values. However, this slight increase in fit and decrease in error motivated us to keep this factor. Importantly, since % Democrat, % Republican, and day of stay at home orders were correlated with each other, we only chose one of these variables to avoid colinearity.

With all of these considerations in mind, our final model includes key features such as state population density, number of ICU beds and number of health professionals per hospital, and medicare enrollment status. Overall, we concluded that we could reasonably model how many deaths or cases a state will have incurred based purely on a given state's infrastructure and demographic characteristics recorded far before SARS-CoV-2 reached the U.S and uncovered certain influential characteristics that state's may want to address when preparing for future wide-spread health threats.

IV. Using historical data to forecast future trends in COVID-19 related deaths and confirmed cases

A. Development and implementation of linear regression models

Historical data provides an important scaffold to build a predictive model partly because the features which contribute or do not contribute to certain outcomes are generally inbuilt. Therefore, we sought to create a model based on the provided time-series datasets in order to predict the number of cases and deaths per state over time. While new state guidelines or alterations in a stateâs characteristics will also change how new death relates to past data, for the sake of this goal, we are making the assumption that these regulations and characteristics remain unchanged at the time of this data‘s download.

In processing the data for this task, we did not consider any county that lacked a county FIPS since we were not confident in any data recorded in a non-existent, invalid, or potentially FIPS reassigned county. As before, we only considered counties in the 50 U.S. states. Counties were grouped by state and the total number of deaths or cases were summed. The resulting two dataframes - one for confirmed cases over time per state and another for deaths over time per state - were used to plot the total number of deaths and the total number of cases over time. From these plots, we observed that the cases and deaths largely rose linearly or exponentially from the first death or case up until the reporting of this data. Therefore, we recognized that we could apply simple log and log-log transformations to linearize applicable data and fit a linear regression model to this data for each state.

Next, we looked to see if this approach would be efficacious in predicting outcomes at future dates. First, we split our cleaned dataframes into training and test sets according to an 80:20 ratio, pretending that the test data occurs in the future. Three linear regression models were fitted using data for each state, one with data that was untransformed, one with log transformed data, and one with log-log transformed data. Training R^2 values were compared and the best transformation-fitted model was utilized. Training and test RMSE values were calculated based on predictions compared to the training values and predictions compared to the test values respectively. The training accuracies were relatively high across all states for the model utilized, meaning the fit to the data was relatively good, and the training and test rmse values were generally neither far apart nor large in magnitude (Fig. S2). Therefore, our approach to modeling can produce decent predictions for dates that are not that far out from the current date.

The data used to validate our model was already known. Next, we wanted to predict the total number of deaths and the total number of cases that will have occurred at some date in the future. For the purposes of our model, we predicted 30 days into the future, but this can be modified in our code. To do this, we trained models for each state on the entire dataset by following the same process previously described and provided future dates in order to predict deaths per state (Fig. 4) or cases per state (Fig. 5) for those dates. Importantly, the axis labels differ based on which transformed dataset worked best for fitting a linear model. The training score values are reported in our notebook under the names 'modeled_deaths_df' and 'modeled_cases_df' (most are above 0.9). For both outcomes, the predictions look promising for states that have very linear or linear-looking data at later dates in the known data (e.g. Arizona and Louisiana for deaths data and Mississippi, Minnesota, and Wisconsin for cases data). However, many states' data is not as linear as expected when transformed or kept as its original values. For such cases, our model is likely overpredicting - sometimes egregiously, especially at dates farther from the current date (e.g. New Jersey, Indiana). The overprediction is compounded for log transformed or log-log transformed data points. Thus, linear regression works well for some cases and not as well for others since a lack of actual linearity in original or transformed data or

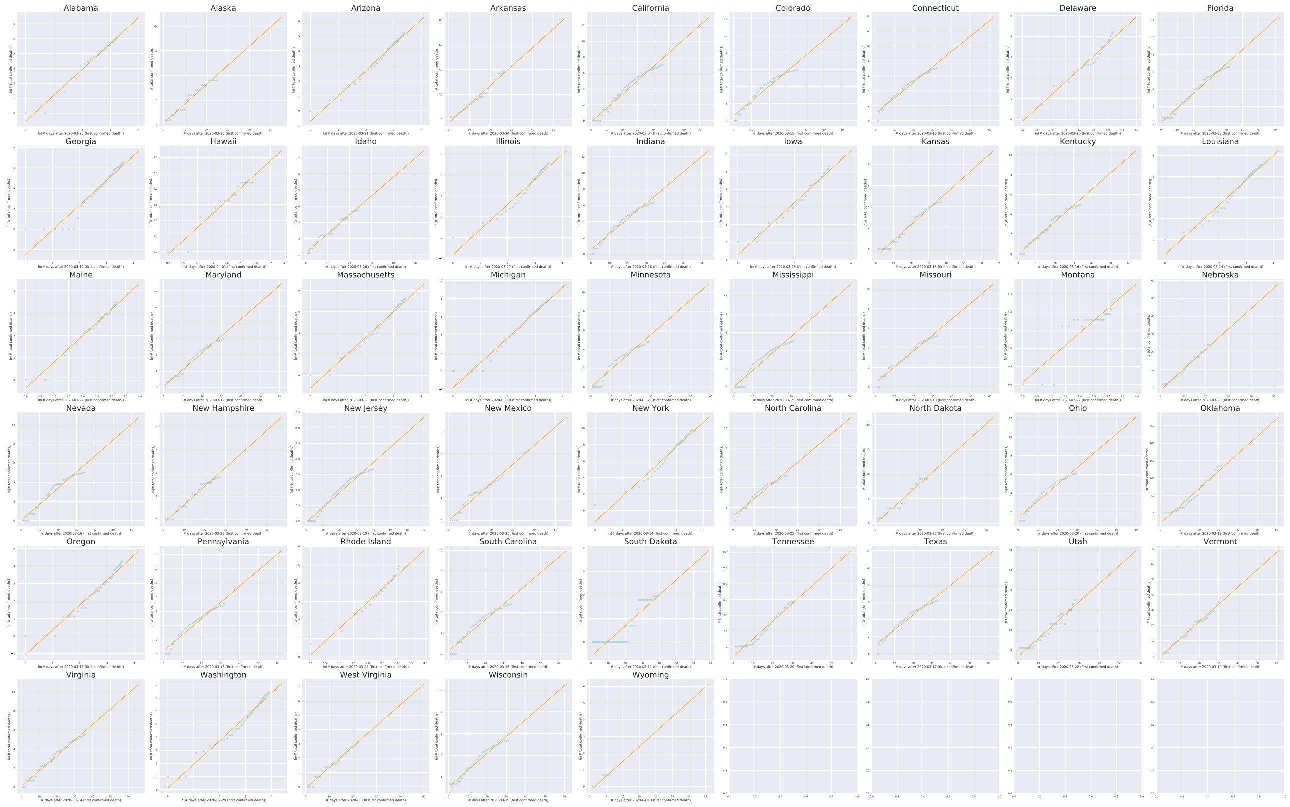


Fig. 4. **Linear regression model to predict the number of deaths per state.** The blue dots represent the actual data while the orange line represents our model extended out to 30 days into the future (past the last recorded date of 4/18/2020). Note that the x and y axes vary based on whether untransformed, log-transformed, or log-log transformed data led to the best fitted model.

a change in the linearity of the data significantly affects the error. If the data does not continue in this trend, as it might for very sensitive data like COVID-19 data, our linear regression model becomes more erroneous, especially predicting at dates farther out. Relatedly, we should note that our model becomes increasingly inaccurate once a state's local or global plateau occurs - which is unknown when these will occur - since our model continues to predict upwards. At the end of this work, we attempt to correct for some of this overprediction by instead utilizing mortality rate and predicted cases.

To address some of these concerns, we attempted to use an auto-ARIMA model, which as its name implies, uses an auto-regressive moving average (mathematically known as the Box-Jenkins method) to better forecast values. First, the cleaned data for deaths and cases was split into training and test sets according to an 80:20 ratio, the same split utilized for our linear model. We visualized both the prediction and actual values on the same graphs as a line plot for each state (Fig. 6, Fig. S3). While the forecast curve seems visually quite close to the actual curve in some cases, in others, it is relatively far off. This is due to several shortcomings. First, the seasonal and biological patterns for SARS-CoV-2 are much less understood than those for other viruses in the coronaviridae family, largely due to the relatively small amount of time this new strain has been affecting the human population. Despite the lack of knowledge for SARS-CoV-2, future generations of this model should seek to incorporate seasonality time series data for SARS-CoV, MERS-CoV, or for other well-documented respiratory viruses [5]. In this case, a SARIMA (seasonal ARIMA) model can be tested and further optimized. Second, as with our linear model, our ARIMA model is not as strong when outside factors, such as new state guidelines, availability and efficacy of potential treatments, quarantine adherence, testing availability/efficacy, and others, play a strong role

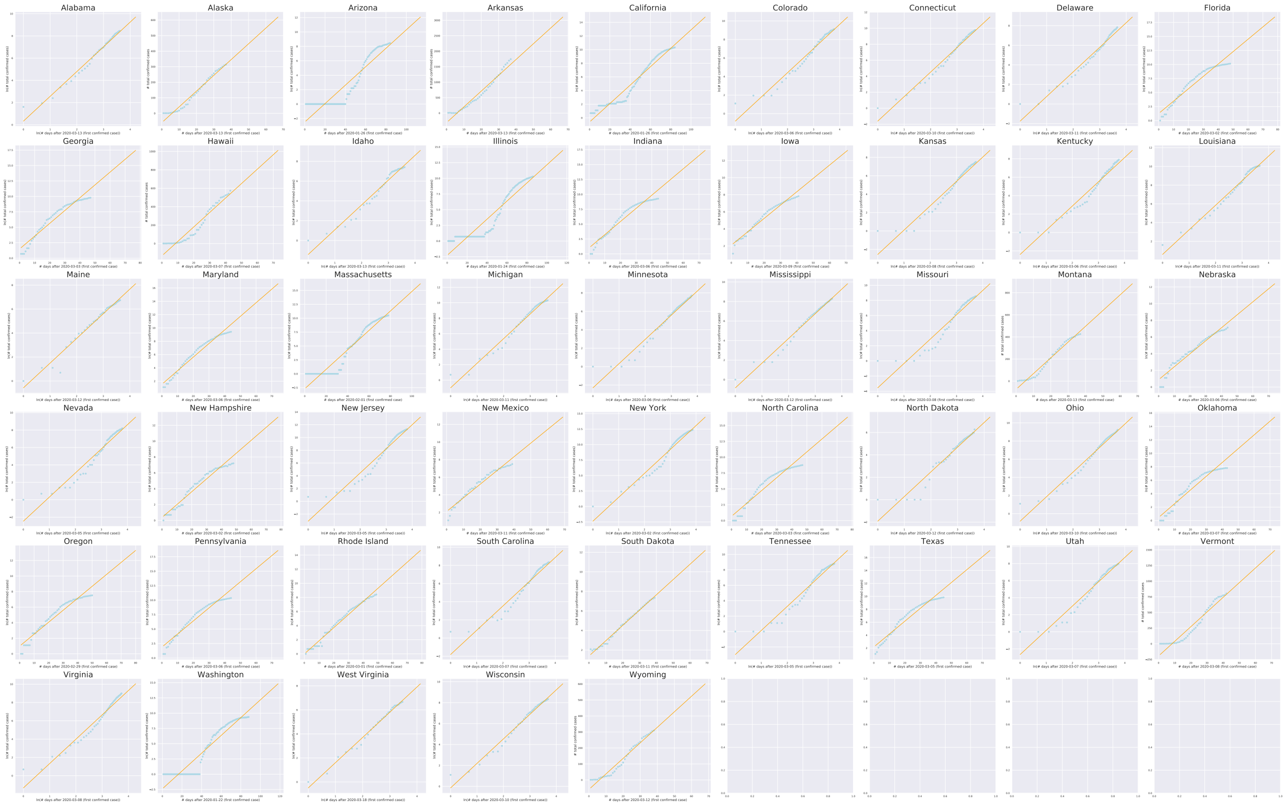


Fig. 5. **Linear regression model to predict the number of cases per state.** The blue dots represents the actual data while the orange line represents our model extended out to 30 days into the future (past the last recorded date of 4/18/2020). Note that the x and y axes vary based on whether untransformed, log-transformed, or log-log transformed data led to the best fitted model.

because our forecasts are dependent on past and current data, which assume the same outside factors are held constant or change in the same way as in past seasons.

To evaluate the strength of our linear regression model compared to our ARIMA model in predicting the number of deaths or the number of cases, we compared their test RMSE values per state for each outcome (Fig. S4). Notably, because we divided our training and test sets by the same ratio for each model, our RMSE values are calculated from the same number of data points corresponding to the same dates. In most cases, our linear regression model had lower RMSE values than our ARIMA model. Therefore, we used the predictions for each state generated from our linear regression model for the following work.

B. Geographical plotting predictions

Using our predictions from our linear regression model, we next looked to plot these values onto a map for a specified date, including future dates. We first merged our cleaned time-series data with the predictions data taken from our model and used geo-plotly to map out the number of deaths and the number of cases 1 week after the reporting of the given data (i.e. 4/25/2020) (Fig. 7). As expected for reasons previously described, when compared to actual reported values for this date (and testing and comparing other dates) our model overpredicted, sometimes egregiously, cases and deaths - particularly for the latter. Also as expected, predictions for deaths and cases at farther dates from 4/18/2020 are much farther off than those closer to this date. We sought to somewhat correct these values by instead multiplying the mortality rate for each state by our predicted confirmed cases to achieve numbers more in the range of the actual values. We recognize that this calculation is dependent on the predictions for our confirmed cases data, which contains the same limitations as our deaths data, and the assumption

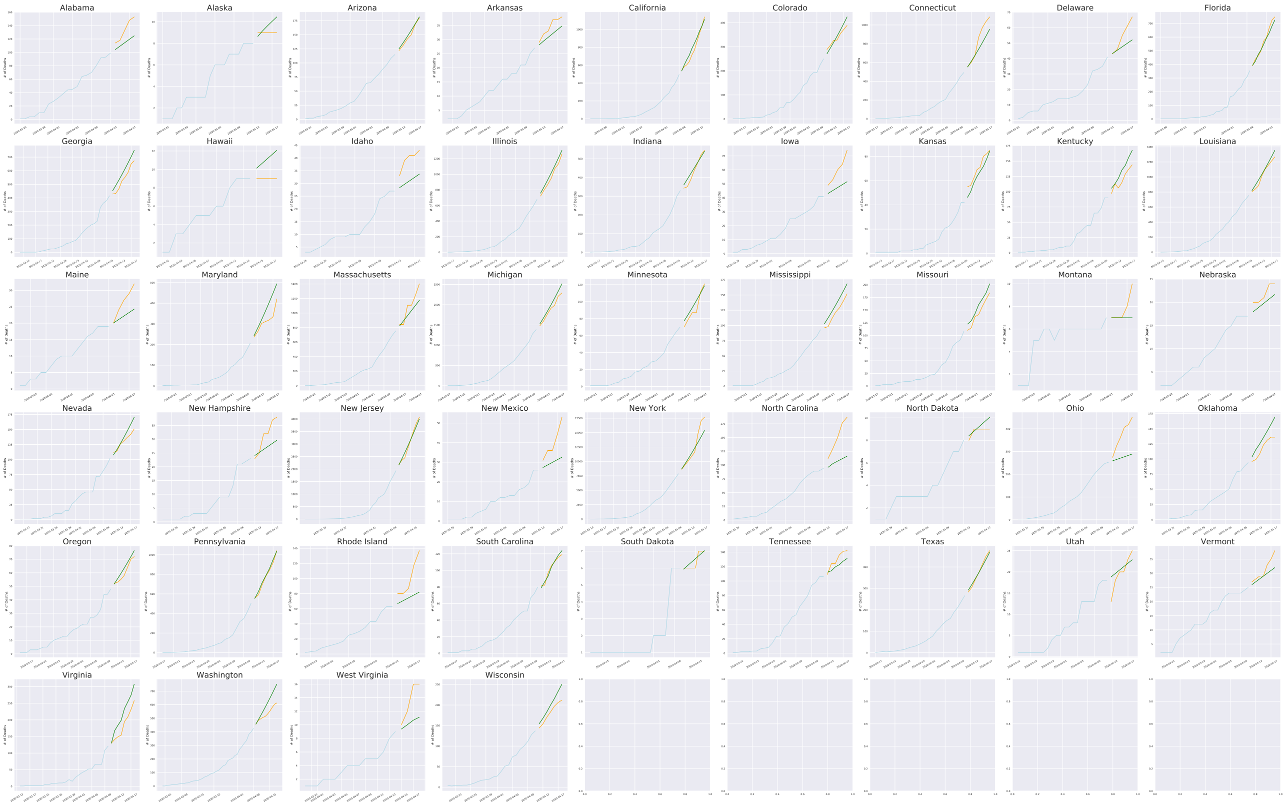


Fig. 6. **Using an Auto Regressive Integrated Moving Average (ARIMA) model to forecast number of deaths.** The actual data is shown as an orange line while our forecast is shown as a green line. Wyoming is not present because too few points were provided in the dataset, resulting in an invalid forecast. Note that the ARIMA model for cases is shown in FigS3.

that the mortality rate will remain relatively constant. However, this method did, in fact, successfully bring our numbers closer to the reported ranges.

V. Conclusion

SARS-CoV-2 cases are still rapidly increasing. Through our project, we were able to model the number of deaths and cases based only on a state's infrastructure and demographics as well as use historical data to forecast future outcomes. During this process, we were challenged with several ethical dilemmas, first notably how we were cleaning our data (elimination of certain data, filling NaN values with zeros, etc.) and the stated assumptions we were making. Additionally, other issues that we were aware of were merging datasets and pulling from features that were amalgamated in different years which could lead to a disparity in actual numbers.

Furthermore, consistency and accuracy of data reporting was an issue in many of the datasets for several potential reasons (accuracy or availability of testing, misrecording of data, response bias, broad approximations, etc.). Therefore, using these data points with knowledge of these inaccuracies and inconsistencies, especially in a global and health-related application, posed another ethical challenge. Overall, we recognize that these challenges are ethically ridden as such models (if used in any official setting on any scale) would inform regulatory or personal decisions that impact peopleâs lives and livelihoods.

If we had more time moving forward, we would want to include information about ethnicity. A recent study showed that more Black people were hospitalized due to the virus than any other ethnicity in the US [1]. We also hypothesize that differences in resources which contribute to race-related gaps would be a potentially strong indicator on a county or a state's COVID-19 related outcomes. Other studies include exploring features related to a state's restriction rules and examining the impact

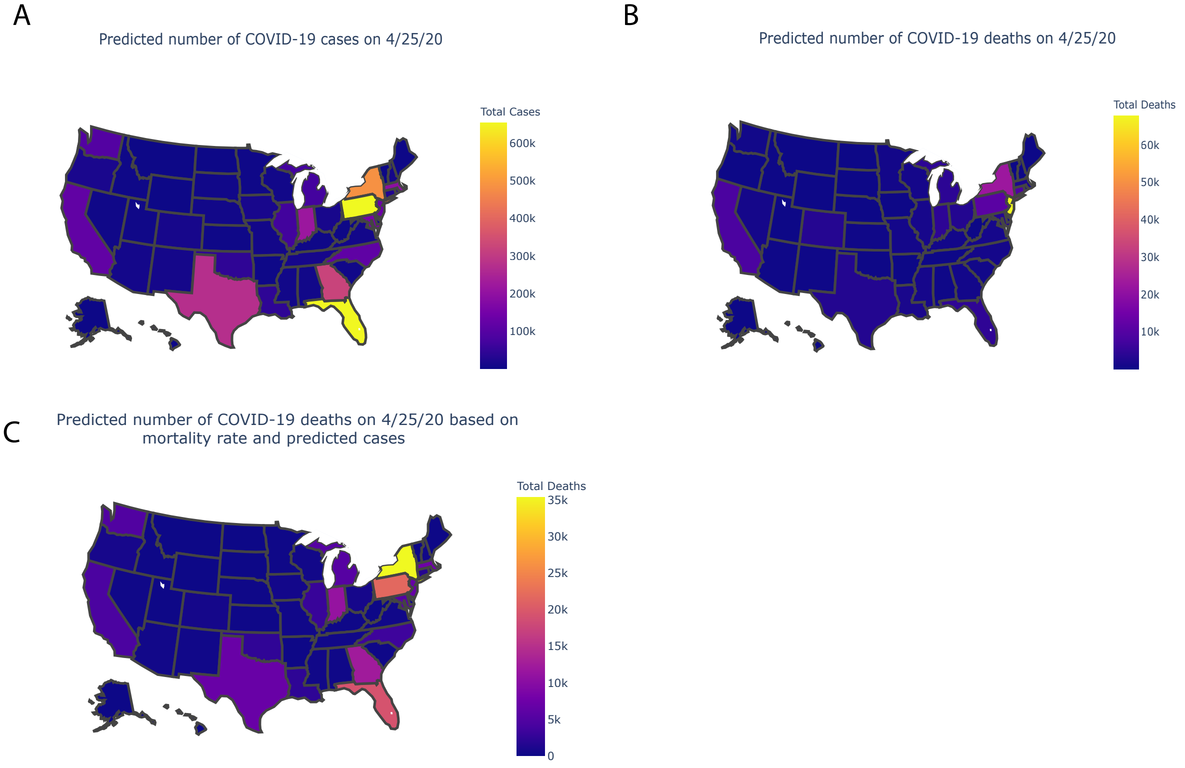

A
Predicted number of COVID-19 cases on 4/25/20
Total Cases
600k
500k
400k
300k
200k
100k
B
Predicted number of COVID-19 deaths on 4/25/20
Total Deaths
60k
50k
40k
30k
20k
10k
C
Predicted number of COVID-19 deaths on 4/25/20 based on mortality rate and predicted cases
Total Deaths
35k
30k
25k
20k
15k
10k
5k
0

Fig. 7. **Predicted number of COVID-19 related deaths and cases 1-week into the future based on a linear regression model.** (A) Predicted number of cases (B) Predicted number of deaths (C) Predicted number of deaths using predicted cases (as in A) and mortality rate in order to correct for some of our overprediction from our linear regression model in shown in B

of features such as stay at home, >50 gatherings, restaurant dine-in, etc. more in depth and how each of these relate to our predictions. As stated prior, we would also explore further avenues to more accurately forecast outcomes based on historical data. Although a lack of plateau makes prediction at this time difficult, routes to improve our model could include considering alternative types of models (e.g. SIR), factoring in seasonality data for well-characterized respiratory viruses (or other viruses of the coronaviridae family), and real-time updating and factoring in state regulations.

Ultimately, we hope that our work will help reveal the critical factors in predicting COVID-19 related outcomes that could be targeted when preparing for future waves of this and other biological threats. Lastly, building and implementing forecasting models have given us a mode to explore a highly relevant topic with some accuracy, especially as forecasting future deaths and cases is proving critical to how each nation adapts and responds to politically, financially, and medically precarious times.

References

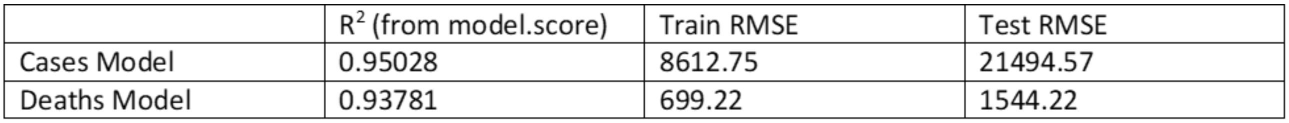
[1] Cases in the U.S. (2020, May 7). Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

[2] Auwaerter, P. G. (n.d.). Coronavirus COVID-19 (SARS-CoV-2): Johns Hopkins ABX Guide.

[3] Mervosh, S., Lu, D., Swales, V. (2020, March 24). See Which States and Cities Have Told Residents to Stay at Home.

[4] How Demographic Changes Make Us More Vulnerable to Pandemics Like the Coronavirus. (2020, March 26).

[5] Moriyama, M., et al. Seasonality of Respiratory Viral Infections. Annu. Rev. Virol. 2020. 7:2.12.19



	R^2 (from model.score)	Train RMSE	Test RMSE
Cases Model	0.95028	8612.75	21494.57
Deaths Model	0.93781	699.22	1544.22

TABLE S1

FINAL R^2 AND RMSE VALUES FOR A LINEAR REGRESSION MODEL PREDICTING NUMBER OF CASES OR NUMBER OF DEATHS BASED ONLY ON STATE INFRASTRUCTURE AND DEMOGRAPHICS DATA USING ORIGINAL AND ENGINEERED FEATURES

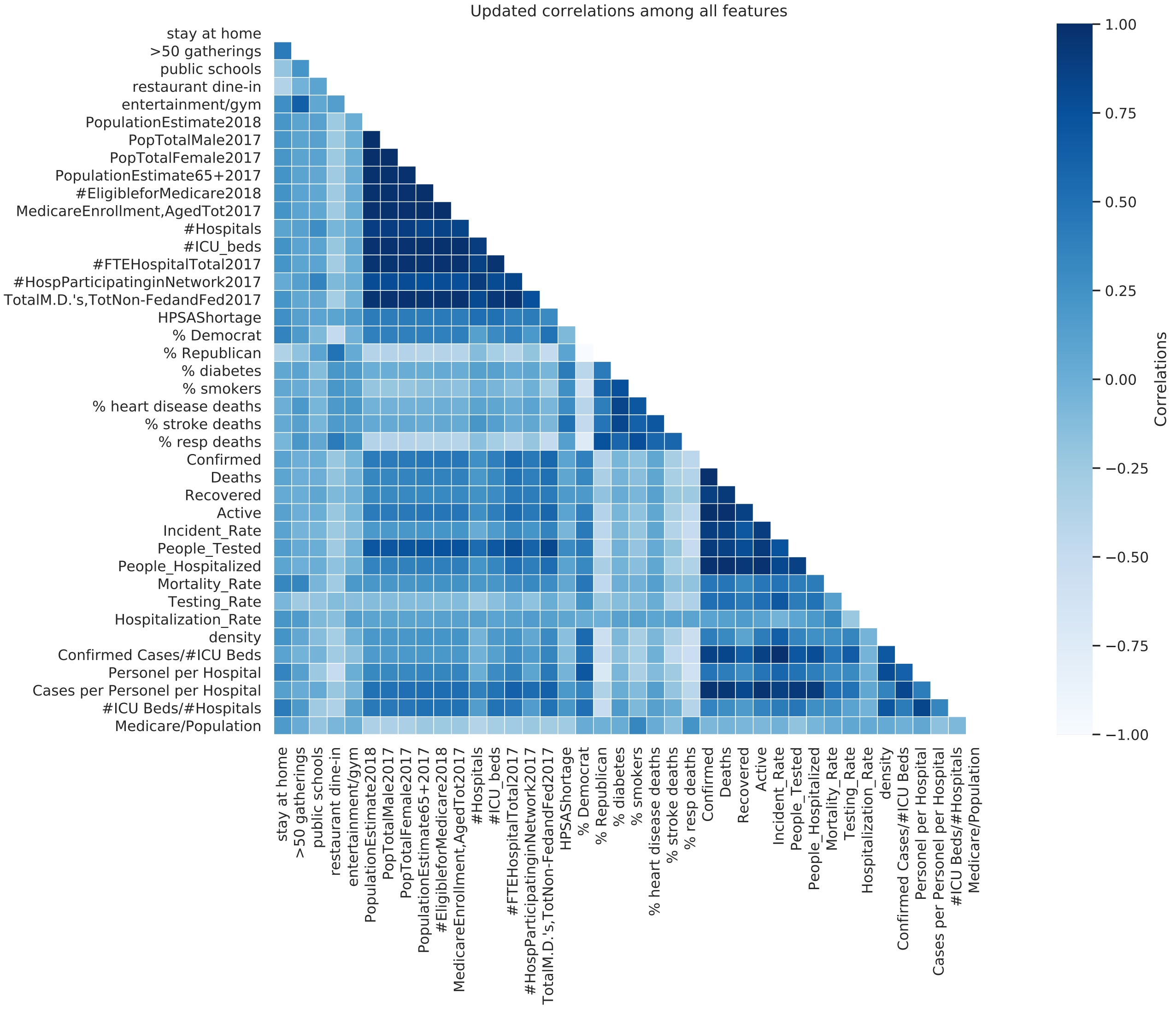


Fig. S1. **Updated heatmap showing the correlations after customized features were added.**

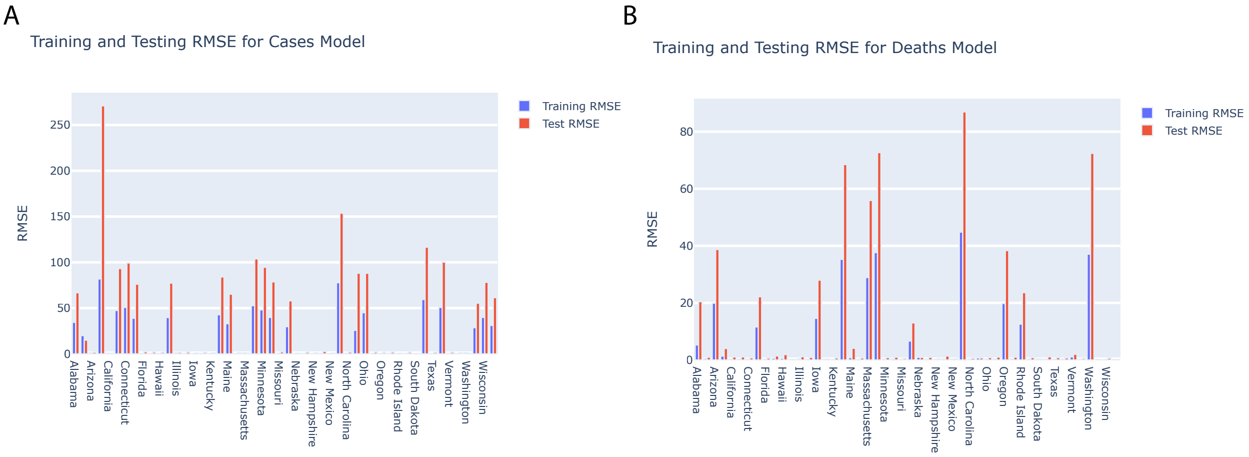


Fig. S2. **Training and test RMSE comparison by state for linear regression models to forecast future number of cases or deaths.** Note that untransformed, log-transformed, or log-log transformed data was used based on which transformation resulted in the best model fit to the training data.

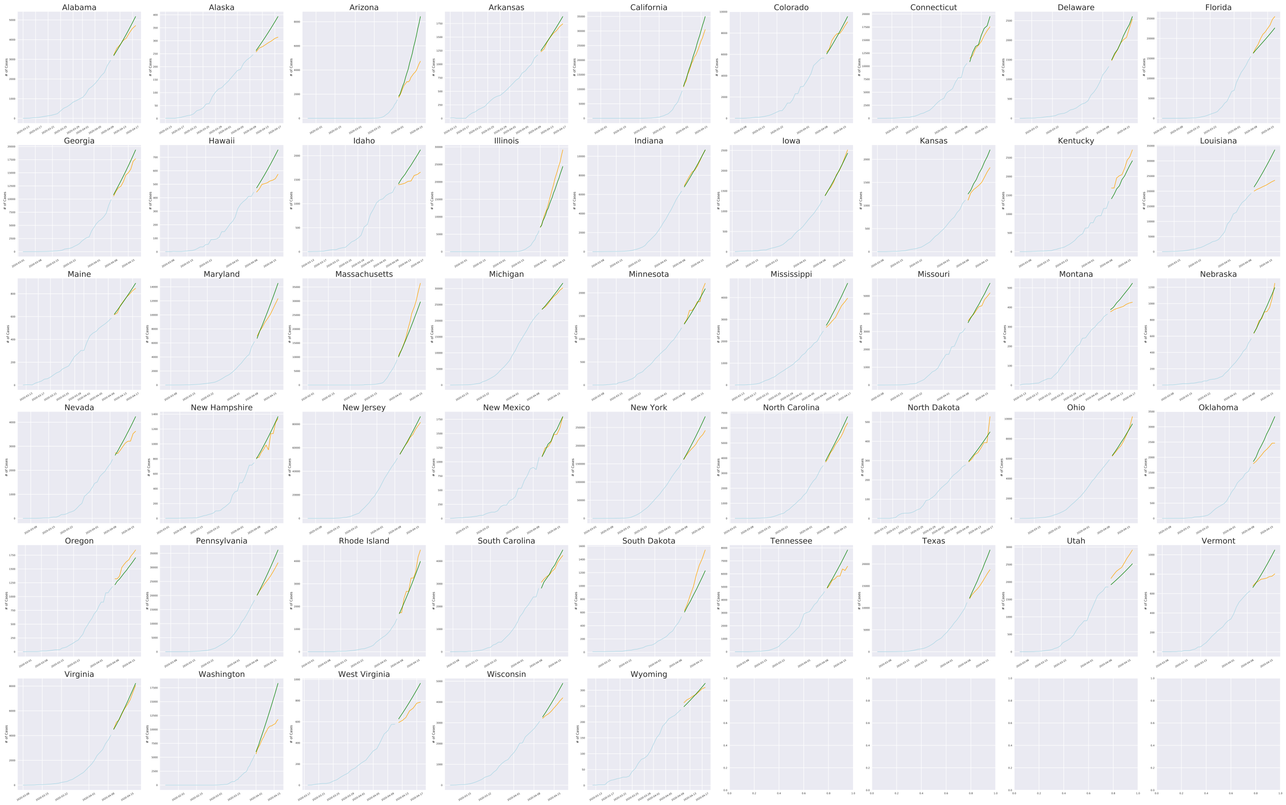


Fig. S3. **Using an Auto Regressive Integrated Moving Average (ARIMA) model to forecast number of cases.** The actual data is shown as an orange line while our forecast is shown as a green line.

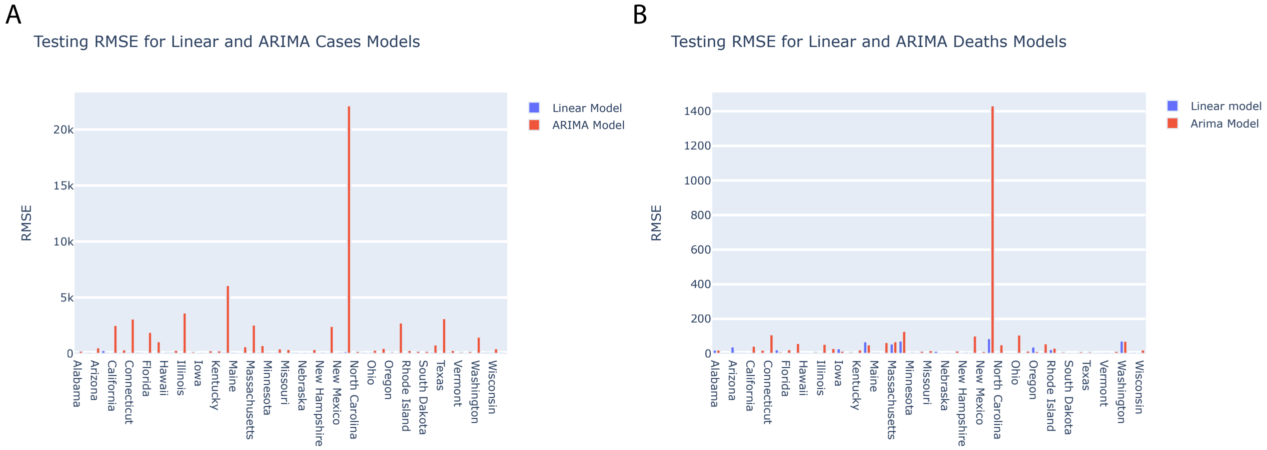


Fig. S4. **Comparison of test RMSE values for Linear and ARIMA models.** Notably, in almost all cases, the linear model had lower RMSE values. Both models forecasted the same future dates