Report

Ben Smith, Edward Shao, Wenrui Zhao

12/17/2022

Abstract

COVID-19 is a global pandemic with mortality rates comparable to those of other major diseases, calling for a need for effective prevention and treatment strategies. We used the dataset collected by Center for Disease Control (CDC) to study COVID-19, and several factors in this dataset such as states, age groups, or co-existing conditions were analyzed. To analyze this dataset, a time series plot was created for purposes of temporal and condition group differences and a spatial mapping of COVID-19 deaths over states was created. Given the non-linear temporal and spatial trends discovered in the initial analysis of the dataset, a generalized linear model was used to model the COVID-19 deaths from the factors in the dataset. The model found that all other condition groups had a significantly greater count of COVID-19 Deaths compared to the baseline condition group, Alzheimer's Disease. The largest increase from the baseline group was Respiratory Diseases ($\beta = 1.7996, p < 0.0001$), which corresponds to a multiplicative increase of 6.047 in the expected number of COVID-19 Deaths compared to baseline. Some other significant risk factors include Diabetes ($e^{\beta} = 5.173, p < 0.0001$) and Renal Failure ($e^{\beta} = 4.274, p < 0.0001$), indicating that individuals with these co-existing conditions are particularly vulnerable to COVID-19 Death.

Background

COVID-19, caused by the novel coronavirus SARS-CoV-2, has caused a global pandemic with significant morbidity and mortality. Understanding the factors contributing to COVID-19 deaths is crucial for developing effective prevention and treatment strategies.

Several factors have been identified as being associated with an increased risk of COVID-19 death. These include older age, male gender, certain underlying health such as obesity, hypertension, and diabetes, as well as certain racial and ethnic groups.

Other factors that may contribute to COVID-19 deaths include inadequate access to healthcare, lack of timely diagnosis and treatment, and inadequate infection control measures in healthcare settings. The effectiveness of public health interventions, such as mask-wearing and social distancing, may also play a role in the risk of COVID-19 death.

Additionally, the severity of the COVID-19 pandemic may vary based on the specific strain of SARS-CoV-2 and the presence of comorbidities and other health conditions. The impact of COVID-19 on individuals and populations may also be influenced by social determinants of health, such as poverty, overcrowding, and limited access to healthcare.

Overall, a combination of individual and societal factors contribute to the risk of COVID-19 death. Further research is needed to better understand these factors and develop effective strategies to reduce morbidity and mortality associated with COVID-19. In this study, we used data from the Center for Disease Control (CDC) to explore a few of the aforementioned factors contributing to COVID-19 death.

Data

The data were collected by the Center for Disease Control (CDC), and included a number of demographic variables. The full list of variables is included in the data dictionary included in the GitHub repository. The particular variables of interest are the Number of COVID-19 Deaths within the specified groups, as well as the Co-Existing Conditions within the patients. Other variables of interest are the State wherein the deaths occurred, the time (in months since Dec 2019) that the deaths occurred, and the Age Group of the individuals with COVID-19. The primary exposure of interest is considered to be the other conditions exhibited by the patients, with the other being considered primarily as variables that need to be controlled for, as they are additionally associated with the outcome variable. In regards to the primary outcome of interest, we additionally noted the significant amount of missing data. This was due to NCHS guidelines, which required death counts under 10 to be replaced with missing values, as values that were too low could be considered as identifying information. These values were imputed from a uniform distribution, which drew from integer values between 1 and 9, as these were the only possible values that would have been replaced with missing values.

Exploratory Analyses

Temporal and Condition Group Differences Given the particular interest in the differences in COVID-19 Mortality among the differing conditions, an exploratory analysis was conducted into the relationship between the COVID-19 Death counts and the coexisting conditions that were present in the patients. In order to visualize this relationship, a time series plot was created for the number of COVID-19 Deaths that occurred in each condition group across 35 months (January 2020 - November 2022). These values were aggregated over the Age Groups and the State in which the deaths occurred, as this exploratory analysis is primarily concerned with the differences among the condition groups. This time series plot can be found in the appendix (Figure 1).

The time series plot shows a clear difference between the condition groups, with two groupings of co-existing conditions having significantly higher counts of COVID-19 Deaths than the others. Particularly, co-existing conditions that fall into the categories “Respiratory Diseases” and “Circulatory Diseases” have significantly more COVID-19 deaths than any other category. This suggests that these different groupings of conditions have an impact on an individual’s likelihood of dying to COVID-19.

In addition, we can see a very clear temporal trend in the count of COVID-19 Deaths, corresponding to different waves of the COVID-19 pandemic, as well as different strains of the SARS-CoV-2 virus. Notably, we can see that the temporal trend is not linear, i.e. there are spikes at varying points in time, corresponding to a variety of events during the pandemic, including new strains of the virus and vaccine rollouts. This non-linearity in the temporal trend may suggest that using a linear spline - or some other non-linear model for time - may be beneficial in our model.

Spatial Analyses In addition to the temporal effects, an exploratory analysis was conducted into the spatial effects, namely how nearby states were related to each other. This was of particular interest due to the nature of infectious disease spread, and we considered it likely that nearby areas (in this case, States) would be correlated. As such, the COVID-19 deaths by each state, aggregated over all other explanatory variables, was plotted over time. In addition, the response variable was considered as a rate (COVID-19 Deaths per 100000) rather than a count for this exploratory analysis, in order to avoid the more populous states (New York, California, etc.) from dominating the visualization. A few of the months (April 2020 - October 2022, increments of 6 months) are given in the appendix (Figure 2).

Regression Modeling

Given the exploratory analyses, a few things were clear - first, there is indeed an association between the primary exposure variable, Condition Group, and the count of COVID-19 deaths experienced. Furthermore, there are indeed both spatial and temporal trends exhibited in the count of COVID-19 deaths, both of which will need to be accounted for in our regression modeling. Given the notable non-linearity in the temporal trend, a linear spline with knots at 4, 9, 13, 18, 21, 23, and 25 months was used in order to capture this non-linearity. The spatial trend was done simply by considering the State as a categorical factor in the linear regression model. This does not capture the entire spatial trend, i.e. the correlation exhibited between nearby states, but including a full spatial model while also including the other effects proved to be beyond the scope of this project.

Thus, we have identified our linear regression model as using COVID-19 Deaths as the sole response variable, with time, Condition Group, Age Group, and State being considered as predictor variables. Additionally, there is the consideration of a link function. In this case, the natural choice is that of a Poisson link function, or a log link, for the COVID-19 Deaths, as the deaths due to COVID-19 is a count variable. The model was adjusted to use this log link, as opposed to no link function (i.e. a Normally distributed outcome variable), in order to more accurately predict the outcome. The Poisson model was then tested for overdispersion, and it was found that the outcome variable could not be assumed to be Poisson ($p < 0.0001$). To correct this, a Negative Binomial link was instead assumed.

After fitting our final model, there are a number of things that we wish to note. Firstly, we note that the parameter estimate for the linear spline of time changes quite a bit between each knot, suggesting that there is indeed a large difference in the slope between these given time points - indeed, we knew this to be the case from the exploratory analyses. Furthermore, using Alzheimer's Disease as the reference group, we did confirm our suspicion from the exploratory analysis that Respiratory Diseases were associated with the greatest increase in risk of COVID-19 death ($\beta = 1.7996$), which corresponds to a multiplicative increase of 6.047 in the count of COVID-19 Deaths when compared to the group with Alzheimer's disease, holding all other factors constant. Interestingly, Diabetes was the group with the second largest increase in COVID-19 Death count ($\beta = 1.6435$), which corresponds to a multiplicative increase of 5.173 in COVID-19 Death count over the group with Alzheimer's disease, holding all else constant. This is surprising because Diabetes was well beneath both Respiratory Diseases and Circulatory diseases in our exploratory analysis. A table including all of the effect sizes for the Condition Groups is included in the Appendix (Table 1).

As for the other covariates of interest, the results were largely unsurprising. States with large population sizes (New York, California, etc.) were largely associated with an increased number of COVID-19 Deaths. This is primarily because the state population is so large, and as such we of course expect the raw counts of COVID-19 Deaths to be larger for these states. Ideally, rates would have been worked with instead (i.e. COVID-19 Deaths per 100000) in order to address this issue. However, while it is trivial to get population sizes by State, it is much more difficult to get group sizes for the other categories considered, and as such creating such a rate variable proved too difficult.

The Age Group variable was similarly unsurprising - lower age ranges were associated with lower expected counts of COVID-19 Deaths. Given the age group 0-24 as a reference, the count of COVID-19 Deaths increased with each age group, up to a an increase in the log expected death count of $\beta = 3.510$ at Age 85+, which corresponds to a multiplicative increase in COVID-19 Deaths of 33.448 in the Age Group 85+ when compared to that of the Ages 24 and under.

Conclusion

The time series plot for COVID-19 deaths over each condition group and spatial mapping of COVID-19 states were created. It was discovered that there was a likely significant differences between each condition groups for COVID-19 deaths with "Respiratory Diseases" and "Circulatory Diseases" both having more COVID-19 deaths than other groups. Additionally, the time series plot showed that there was non-linear temporal trend in the count of COVID-19 deaths. On the other hand, the spatial mapping showed that some of the states had more COVID-19 deaths than those of other states. It could be due to population density or laws regarding the COVID-19. Unforunately, we do not have enough data to investigate this. However, the spatial mapping confirms that there is spatial trend in the count of COVID-19 deaths.

Given the non-linear temporal and spatial trends confirmed in the count of COVID-19 deaths, a generalized linear model with a linear spline for time at 4, 9, 13, 18, 21, 23, and 25 months, as well as a Negative Binomial distribution for the outcome with a log link function was fitted after testing few other models.

After fitting the final model, the paramater estimate of linear spline of time changes very between each knot, which suggested that there is large difference in the slope between these given time points. Repspratory disease was confirmed to be associated with greatest risk in COVID-19 death with comparision to the reference group, which is Alzhemier Disease.

Additionally, states with large population and older age groups were also accounted large increase in COVID-19 deaths. This could because large population meant more people to spread COVID-19 to and older population were more vulnerable.

Hence, given the results of this project, we should aim to focus on preventing respiratory diseases and to encourages the uses of wearing masks. Furthermore, more efforts should be spent on states with large populations such as California or New York, and the older population.

Appendix

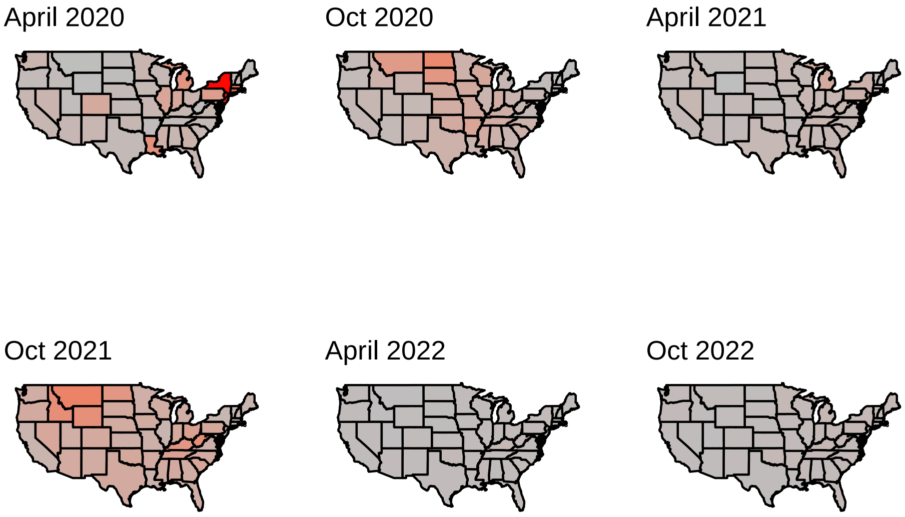

Figure 1: COVID-19 Deaths per 100,000 people by state at 6 month intervals, beginning in April 2020.

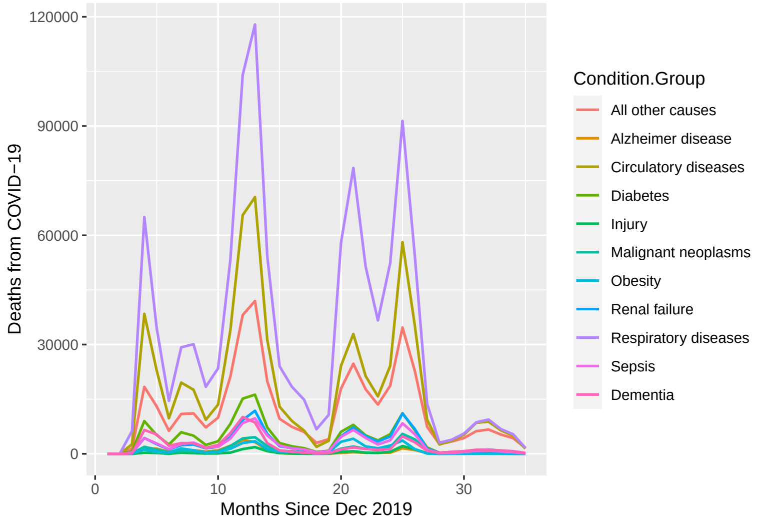

Figure 2: Time series analysis of COVID-19 death counts, separated by Condition Groups.

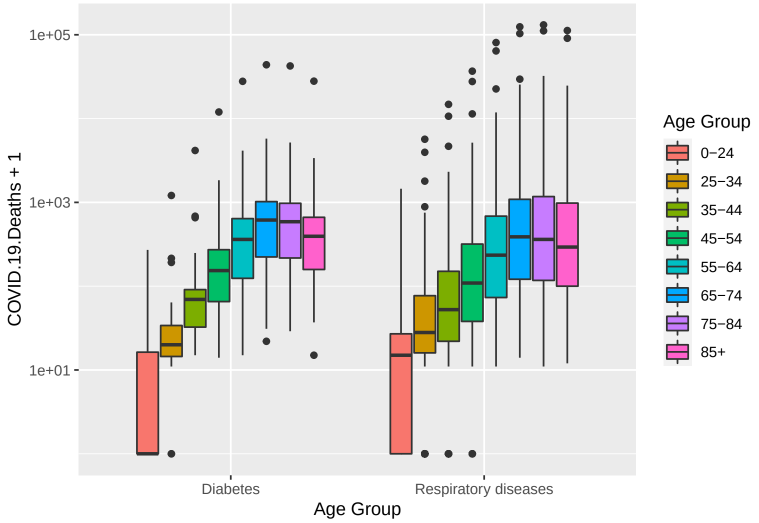

Figure 3: Boxplot of COVID-19 Deaths by Age Group and the top 2 Condition Groups (Diabetes and Respiratory Diseases)

	Effect Size	Mult. Change in COVID Deaths	95% CI Lower	95% CI Upper
Circulatory Diseases	1.262	3.534	3.426	3.644
Diabetes	1.626	5.082	4.891	5.280
Injury	0.782	2.185	2.101	2.273
Malignant Neoplasm	1.002	2.725	2.620	2.833
Obesity	1.348	3.849	3.704	4.001
Renal Failure	1.458	4.298	4.136	4.466
Respiratory Diseases	1.793	6.006	5.823	6.195
Sepsis	1.425	4.159	4.002	4.322
Dementia	0.691	1.995	1.917	2.076

Table 1: Effect sizes with exponentiation and CIs of Condition Groups from the final model fit. Alzheimer's Disease is baseline category.