Using Healthcare Data to Study the Association Between Pre-existing Respiratory Disease and COVID Death Rate in Massachusetts

Abstract

The varying range of patients' responses to COVID, spanning from mild symptoms resembling a common cold to severe cases resulting in death, has inspired investigation of potential associations between pre-existing respiratory diseases and fatalities among COVID patients. A dataset collected by hospital staff throughout the United States starting in 2020 and published by the Centers for Disease Control and Prevention(CDC) was used to create a multiple linear regression model. The data analyses suggest that patients with influenza/pneumonia experience more COVID deaths than patients with pre-existing respiratory conditions of the other category. There is no significant difference in COVID deaths between patients with influenza/pneumonia and patients with respiratory failure. However, limitations from expansive data removal, a focus on only Massachusetts hospitals, and potentially uninvestigated confounding variables, suggest a need for additional study.

Background and Introduction

Although the height of the COVID pandemic has passed, COVID is still a prevalent disease that is especially dangerous for those with chronic health issues. Because COVID targets the respiratory system, there is research that focuses on the relation between different respiratory diseases and varying responses to COVID. For example, one study from 2022 found that chronic lung disease contributes significantly to COVID mortality (Kilic et al. 2022), and another study concluded that patients with interstitial lung disease, a specific respiratory disease, have increased odds of death by COVID (Esposito et al. 2022). Despite these findings, current literature investigating pre-existing respiratory diseases and COVID outcomes remains limited (Lohia et al.) Thus, the study's guiding question is to determine which specific respiratory disease is related to increased COVID deaths in Massachusetts.

As COVID continues to pose a significant threat for those with respiratory diseases, our study aims to provide valuable information to the healthcare industry. By identifying the specific respiratory diseases that are associated with high COVID mortality, our findings can help healthcare professionals improve patient outcomes. Furthermore, by contributing novel insights into patient prioritization protocols, our study can help enhance the overall quality of care for hospital patients. Thus, our research fills a critical knowledge gap in COVID patient prioritization and has significant implications for healthcare in Massachusetts.

Data and Exploratory Analysis

The study utilized a dataset of COVID deaths from patients with pre-existing health conditions, which was collected by hospitals throughout the United States and published by the CDC ("Centers for Disease Control and Prevention").[1] After filtering out extraneous data, each row represents one case, or a one-month period, in Massachusetts between January 1, 2020 and March 25, 2023. However, it is important to note that there may be some selection bias as only the most severe cases are hospitalized. The response variable is the quantitative number of deaths attributed to COVID with a range of 10 to 646. Based on the histogram of Number of Mentions from Previous Month (Figure 1) and the histogram of COVID Deaths in a Month (Figure 2), it is notable that both are right-skewed with multiple outliers.

The predictor variables the data subset initially included are age groups, pre-existing respiratory conditions, and the number of mentions from the previous month. Age groups 0-24, 25-34, and 35-44 were deleted due to the scarce data for the younger generations, and thus the study only focused on age groups 45-64, 65-74, 75-84, and 85+. The side-by-side boxplots comparing Age Group and Number of COVID Deaths show that all groups have a distribution that is skewed right with multiple high outliers and roughly the same median (Figure 8). The variation appears to increase as age increases.

Three condition groups were formed, which are influenza/pneumonia, respiratory failure, and other respiratory diseases, which includes adult respiratory distress syndrome, chronic lower respiratory diseases, and respiratory arrest. The barplot of Frequency of Respiratory Conditions has about the same number of samples in each category (Figure 3), and the barplot of Frequency of Patient Age Groups suggests that as age increases, the number of samples in each category increases as well (Figure 4). Additionally, the side-by-side boxplots comparing Condition and Number of COVID Deaths reveal that all groups have a distribution that is skewed right with multiple high outliers (Figure 9). It seems that respiratory failure and influenza/pneumonia have roughly the same median and similar variability, while the group of other diseases has a slightly lower median of COVID deaths and a smaller variability, as indicated by the shorter length of the boxplot (Table 1, Figure 9).

Lastly, the study included a quantitative predictor variable, the number of reported COVID patients from the previous month with a range of 10 to 653. The scatterplot between

[1] Please refer to the appendix for the detailed figures and tables mentioned in the paper.

COVID Deaths and Number of Mentions indicates a weak, positive linear relationship (Figure 5). There appears to be two trends: one that is vertical and one that is positive and linear (Figure 5). The color-coded scatter plots based on group for condition and age suggest that there is no apparent relationship between groups and the number of mentions from the previous month to predict COVID deaths (Figure 6, 7). Therefore, the different groups do not explain the two trends in the scatterplot.

Model and Results

A multiple linear regression model was used to analyze the data. To improve the fit of the data to the multiple linear regression model, the boxcox function suggested transforming the response variable, COVID deaths, to the negative square root. After applying the transformation, the plots to diagnose multiple linear regression assumptions had improved. To be specific, the Residual versus Date plot has a random scatter, indicating independence (Figure 10). The Residual versus Predicted plot with transformed COVID deaths improved mean zero with a greater balance of points above and below the residual=0 line. Also, the transformation improved constant variance, indicated by points more spread out along the residual=0 line, and linearity, indicated by more randomness in the points. The points in the QQ plot follow the normality line and the histogram of residuals has a unimodal bell-shaped distribution, which both suggest normality. In summary, the transformation was used since it showed evidence of largely improving our model**.**

To find the best subset of variables for the model, the forward, backward, and both-direction selection process was performed, which all suggested the same model. The selection processes indicated a removal of all interactions and age groups from the full model to produce the final model. For influenza/pneumonia, the least squares regression line (LSRL) is $\widehat{(COVID\ Deaths)}^{-½}$ = 0.214 -0.00042(Mentions). For respiratory failure, the LSRL is $\widehat{(COVID\ Deaths)}^{-½}$ = 0.2127 -0.00042(Mentions). For other respiratory diseases, the LSRL is $\widehat{(COVID\ Deaths)}^{-½}$ = 0.241 -0.00042(Mentions).

Table 1. Summary of Model Coefficients for $\widehat{(COVID\ Deaths)}^{-½}$

Coefficient	Estimate	P-value (2-sided)	P-value (1-sided)	Confidence Interval
$\hat{\beta}_0$ (Intercept)	0.214	$< 2 \times 10^{-16}$	$< 1 \times 10^{-16}$	0.2018 to 0.2268
$\hat{\beta}_{\text{Respiratory Failure}}$	-0.0013	0.87098	0.43549	-0.017 to 0.015
$\hat{\beta}_{\text{Other Respiratory Diseases}}$	0.027	0.00127	0.000635	0.011 to 0.044
$\hat{\beta}_{\text{Mentions}}$ (Slope)	-0.00042	$< 2 \times 10^{-16}$	$< 1 \times 10^{-16}$	-0.0005 to -0.0003

$\hat{\beta}_0$ is 0.214, which represents the average value of $\widehat{(COVID\ Deaths)}^{-½}$ when the number of COVID cases in the previous month is zero and the patient has influenza/pneumonia. We are 95% confident that when there are no COVID cases in the previous month and holding other variables constant, we expect the true value of $\widehat{(COVID\ Deaths)}^{-½}$ to be between 0.2018 to 0.2268 transformed deaths.

$\hat{\beta}_{\text{Mentions}}$ is about -0.00042, which suggests that every increase by one mention of COVID cases of the last month is associated with about 0.00042 decrease in $\widehat{(COVID\ Deaths)}^{-½}$, holding all other variables constant. With a very small p-value, the number of mentions has a significant, strong negative linear relationship with $\widehat{(COVID\ Deaths)}^{-½}$ (Table 2), given the number of mentions from the previous month is held constant. We are 95% confident that

holding all other variables constant, we expect that each additional increase in the number of COVID cases from the previous month by one, is associated with a decrease between 0.0003 to 0.0005 $(\widehat{COVID\ Deaths})^{-½}$.

$\widehat{\beta}_{\text{Respiratory Failure}}$ is -0.0013, which means that when holding the number of COVID cases from the previous month constant, the average number of $(\widehat{COVID\ Deaths})^{-½}$ for patients with respiratory failure are about 0.0013 smaller than for patients with influenza/pneumonia. However, the large p-value is about 0.43549, and thus there is no significant difference in the average number of transformed COVID deaths between respiratory failure and influenza/pneumonia, given the number of mentions from the previous month is held constant in the model.

$\widehat{\beta}_{\text{Other Respiratory}}$ is 0.027, which indicates that holding the number of COVID cases from the previous month constant, the average number of $(\widehat{COVID\ Deaths})^{-½}$ for patients with other respiratory diseases is about 0.027 larger than for patients with influenza/pneumonia. There is a significant difference in the average number of transformed COVID deaths between other diseases and influenza/pneumonia, given the number of COVID cases from the previous month is held constant (Table 2).

Lastly, this model is appropriate for the study because it has a reasonable R^2_{adj} of 0.2547, which is the highest of the models we found from forward, backward, and two-sided variable selection subsetting, and has the lowest AIC value of -2020.9.

Discussion/Conclusions

The study's guiding question is to determine which specific respiratory disease patients experience more COVID deaths in Massachusetts. One key finding is that as the number of mentions from the previous month increases, the number of COVID deaths increases as well, given the condition group is held constant. This may be due to the nature of COVID spreading in waves; an increased amount of infected people may result in more deaths in the following weeks as their health declines. Also, the data suggests that patients with other respiratory diseases experience less COVID deaths than influenza/pneumonia patients. This may be because patients who are coinfected with influenza/pneumonia and COVID may experience more severe disease effects. Lastly, there is no difference in COVID deaths between influenza/pneumonia and respiratory failure patients. This is expected as both conditions are more prevalent in older individuals. However, these conditions differ in their underlying causes, with influenza/pneumonia being caused by infections and respiratory failure being a result of the lungs' inability to provide enough oxygen. This represents a point of potential further study.

Limitations in the study include removing a significant portion of the dataset, which may have distorted the statistical analyses above. Furthermore, the results may not be generalized to states other than Massachusetts, which limits the applicability of the study's findings. Additionally, the study raises the question of a potential confounding factor of socioeconomic class, since many low-income Americans do not have the same access to hospital care as others. In the future, including research on other states to compare the impact of pre-existing respiratory conditions on COVID outcomes may be insightful to see if conclusions of this study are applicable to other regions. Also, an investigation of potential confounding variables such as socioeconomic class that could have influenced this study's results could be beneficial to evaluate our model effectiveness. Lastly, the scope of this study can be expanded to examine the impact of additional respiratory diseases and emerging variants of COVID on patient outcomes to provide further insights into patient prioritization protocols.

References

Centers for Disease Control and Prevention. (2023). Conditions Contributing to COVID Deaths, by State and Age, Provisional 2020-2023[Data file]. Retrieved from https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-COVID-by-age-group

Esposito, Anthony J., et al. "Increased Odds of Death for Patients with Interstitial Lung Disease and COVID: A Case–Control Study." American Journal of Respiratory and Critical Care Medicine, American Thoracic Society, 2021, https://doi.org/10.1164/rccm.202006-2441LE. PubMed ID: 32897754.

Kilic, Hatice et al. “Effect of chronic lung diseases on mortality of prevariant COVID pneumonia patients.” *Frontiers in medicine* vol. 9 957598. 13 Oct. 2022, doi:10.3389/fmed.2022.957598

Lohia, P., Sreeram, K., Nguyen, P. *et al.* Preexisting respiratory diseases and clinical outcomes in COVID: a multihospital cohort study on predominantly African American population. *Respir Res* 22, 37 (2021). https://doi.org/10.1186/s12931-021-01647-6

Appendix

EDA

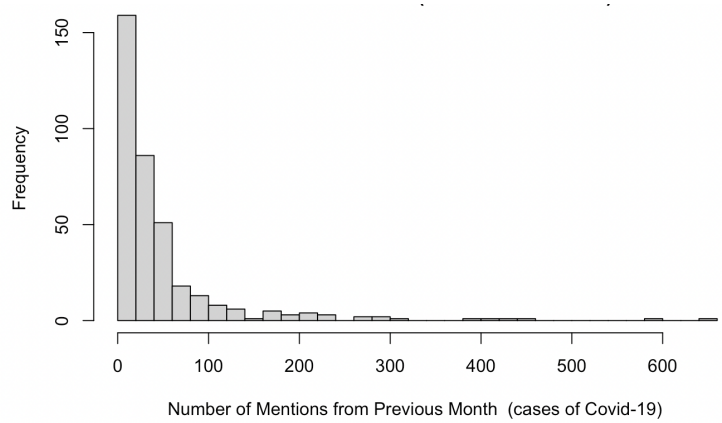

Figure 1. Histogram of the Number of Mentions from Previous Month (Cases of COVID)

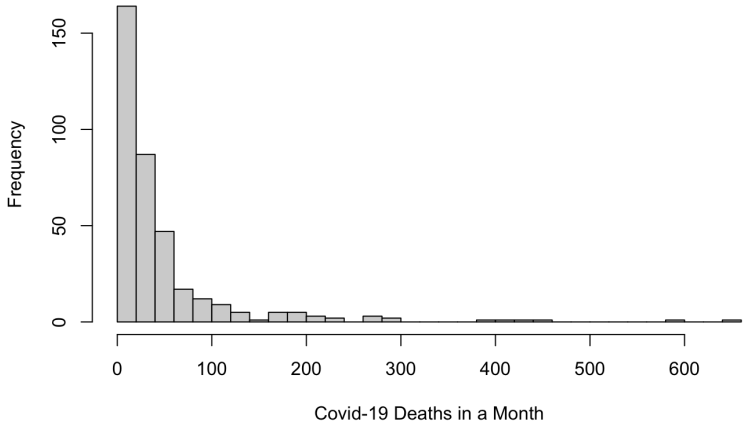

Figure 2. Histogram of the Frequency of COVID Deaths from Previous Month (Cases of COVID)

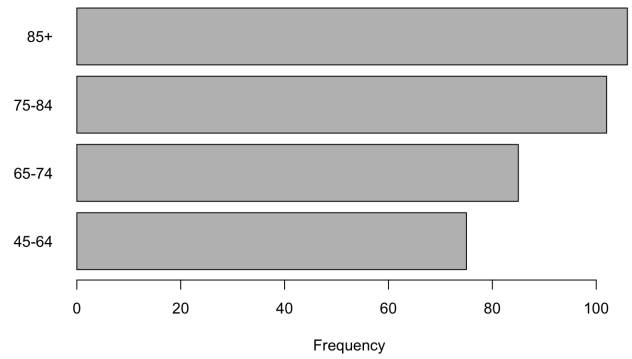

Figure 3. Barplot of the Frequency of Patient Age Groups

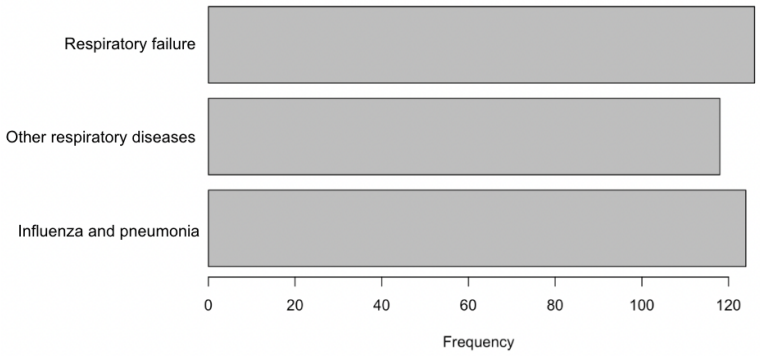

Figure 4. Barplot of the Frequency of Pre-Existing Respiratory Conditions

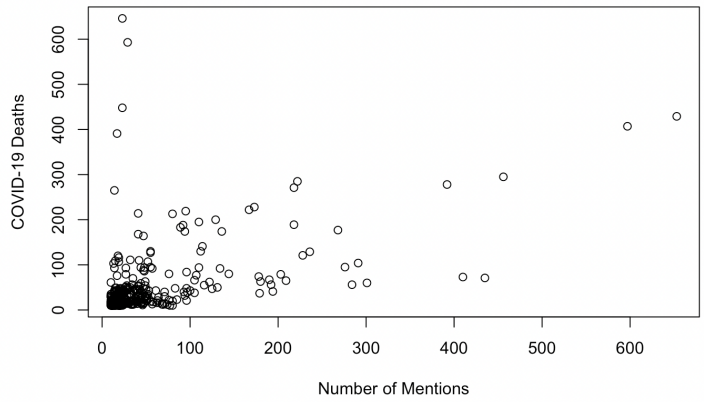

Figure 5. Scatterplot of Number of Mentions and COVID Deaths

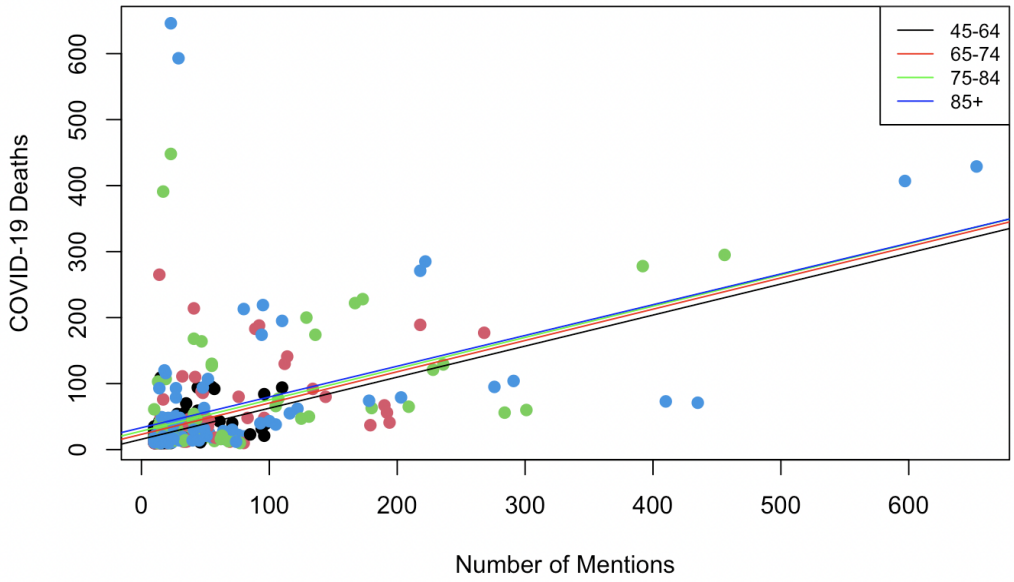

Figure 6. Color-Coded Scatterplot of Number of Mentions and COVID Deaths by Age Group

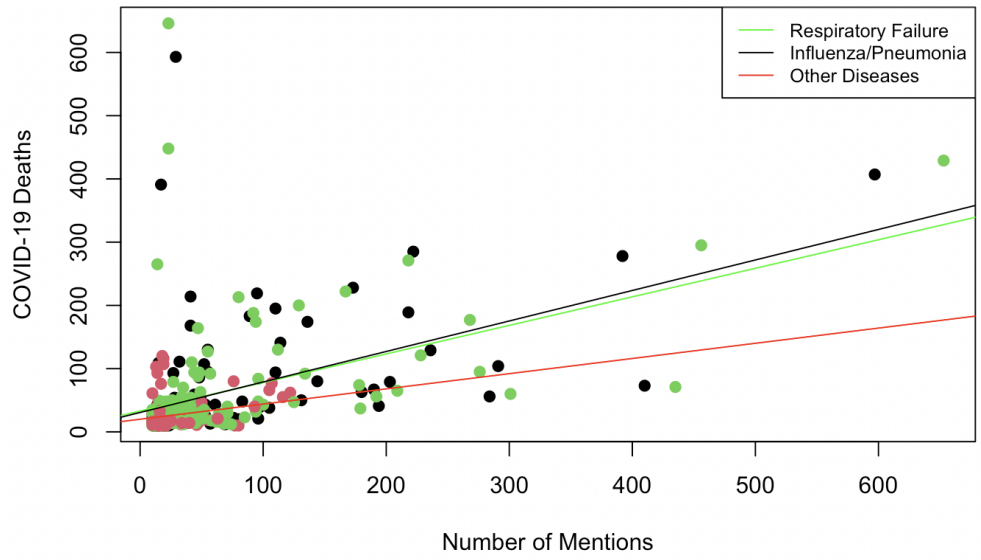

Figure 7. Color-Coded Scatterplot of Number of Mentions and COVID Deaths by Condition

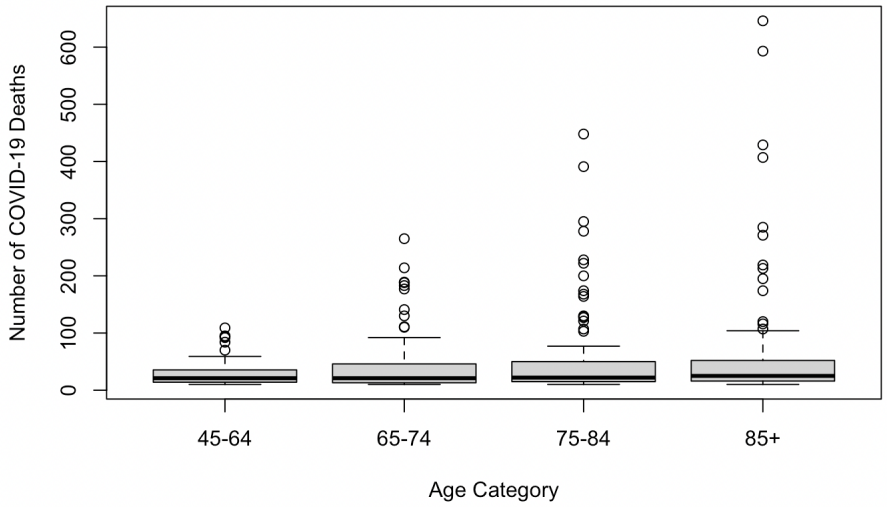

Figure 8. Side-by-side Boxplot of Age Category and COVID Deaths

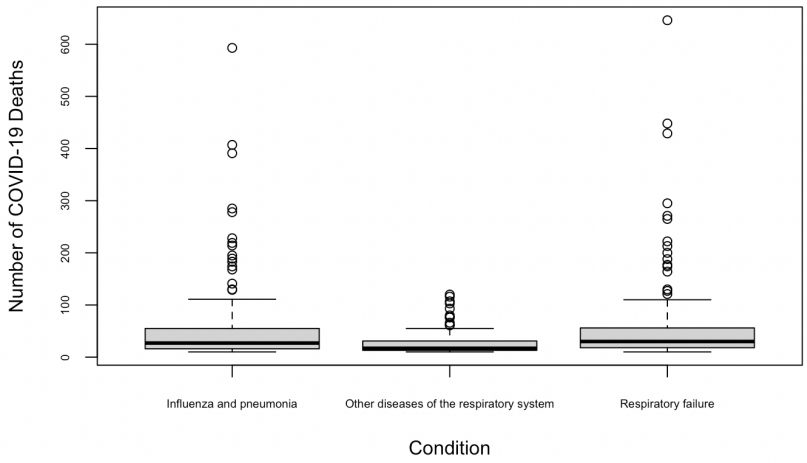

Figure 9. Side-by-side Boxplot of Respiratory Condition and COVID Deaths

Model and Results

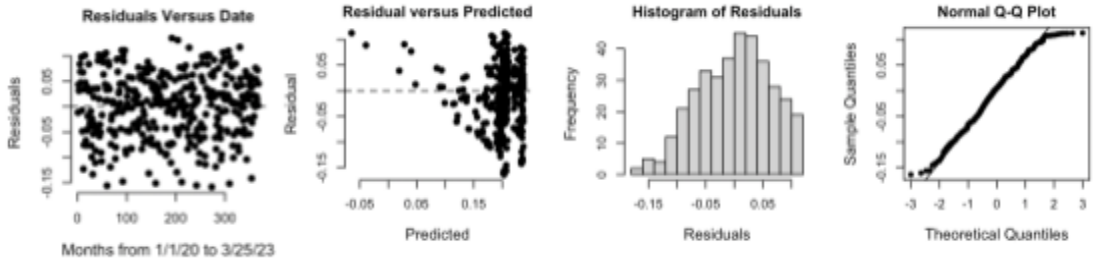

Figure 10. Residual Plots of Multiple Linear Regression Model for $(\text{COVID Deaths})^{-½}$

Table 2. Five-Number Summary of $(\text{COVID Deaths})^{-½}$ by Respiratory Condition

Respiratory Condition	Minimum	First Quartile	Median	Third Quartile	Maximum
Influenza/ Pneumonia	10	16	27	55	111
Other Respiratory Diseases	10	13	17	31	55
Respiratory Failure	10	18	30	56	110