Hannah Wang
Claire Luo STA 210

THE IMPACT OF PATIENT ATTRIBUTES ON COVID-19 DEATHS

Introduction

Over the past years, the COVID-19 pandemic has had a significant impact globally, resulting in significant mortality and health challenges. Our research aims to discover the patterns of COVID-19-related mortality, focusing on how demographic variables like age and gender correlate with mortality rates. This focus is driven by historical observations from past pandemics where there have been distinct mortality patterns, such as during the 1918 Spanish Flu, according to the National Library of Medicine. Notably, while typically higher mortality rates are observed in older adults due to weaker immune systems, the Spanish Flu interestingly exhibited a peculiar W-shaped mortality curve with unexpectedly high death rates among young adults. The highest mortality rates were surprisingly held by the young adult demographic around 1%, doubling the older generation. Furthermore, the National Library of Medicine mentioned that sex is another attribute often studied as many viruses appear to interact with different sexs' bodies differently. Inspired by these historical trends, our research delves into how potential various factors can play a significant role with COVID-19 mortality.

Literature Review

In the research paper "Role of Sex and Age in Fatal Outcomes of COVID-19: Women and Older Centenarians Are More Resilient," Caruso et al. investigate the influence of age and gender on COVID-19 fatality rates, exploring three specific key areas: (1) the resilience of women to the virus compared to men, (2) the mortality rates of centenarians versus other elderly groups, and (3) the resistance of centenarians born before 1919 to SARS-CoV-2(COVID-19 Virus) relative to their younger counterparts. Their findings suggest that women generally exhibit greater resilience to COVID-19 than men, with men displaying twice the mortality rate of women. The study also notes that centenarians do not necessarily show lower mortality rates than other older age groups. Moreover, centenarians born before 1919 appear to possess increased resilience compared to younger centenarians. Building on this foundation, our research seeks to delve deeper into the variability of mortality rates across different sexes, ages, and previous underlying health conditions. By examining both demographics and existing health conditions, our study aims to provide a more nuanced understanding of how physiological and demographic attributes can affect COVID-19 mortality rates, thus expanding on the insights offered by Caruso et al. (2023).

In response to this, we have chosen to introduce a more distinct approach into understanding COVID-19 mortality by separately analyzing how various physiological and demographic attributes influence the mortality rates through separate models. Our research will primarily dissect the roles of demographic conditions and attributes that could potentially influence health risks, such as underlying diseases, sex, and age. Recognizing these disparities is crucial, as it can help identify attributes that are particularly vulnerable to severe outcomes from COVID-19. The motivation for our research is to be able to adequately analyze and predict COVID-19 patient data, specifically the health conditions of those patients who lived and those who died. This research topic has significance to the researchers because family members from China, including elderly relatives, have suffered severe health consequences from COVID-19. Over 1.1 million people have died from COVID-19 in the United States, and 6.86 million people worldwide. The novelty of our research lies in the isolated examination of these two feature subgroups, enhancing the clarity of our findings. Being able to accurately explore and predict this data can allow for tailored and improved treatment, care, and sufficient supervisions of current and future COVID-19 patients in hospitals. For example, demographic insights found could inform public health strategies, while physiological insights may help guide clinical treatment protocols. The significance in this research lies specifically in the mortality prediction given the specific attributes of each patient. Moreover, we want to be able to identify health conditions that may signal higher risk of mortality.

Data

The dataset we utilized for our research was from Kaggle, provided by the Mexican government and contains anonymized patient-related information including pre-existing conditions and demographic

information. The raw dataset consists of 21 unique attributes and 1,048,576 observations, with each observation representing a single patient. Most of the attributes contain binary data, with zeros representing "no" and ones representing "yes." Null values or missing values are denoted with a 97 and 99. Below is a detailed description provided by Kaggle on each attribute of the dataset and the meaning of the values.

Sex: 1 for female and 2 for male.

Age: of the patient.

Pneumonia, Diabetes, Asthma, Hypertension, Cardiovascular, Chronic Obstructive Pulmonary Disease, immunosuppression, Cardiovascular disease, renal chronic, Diabetes, Other disease:

1 for 'Has Condition' and 2 for 'Does Not Have'

Date died: If the patient died, indicate the date of death, and 9999-99-99 otherwise.

We have formulated our research questions to identify which demographics are more susceptible to developing severe conditions and passing away from the virus. Our first research question asks: How do preexisting health ailments or conditions influence the likelihood of someone dying from COVID-19? To broaden our research, we thought it would be interesting to analyze patient demographics compared to their associated mortality rates. Our second research question addresses: How do various demographic factors like age and sex influence COVID-19 mortality rates?

Exploratory Data Analysis

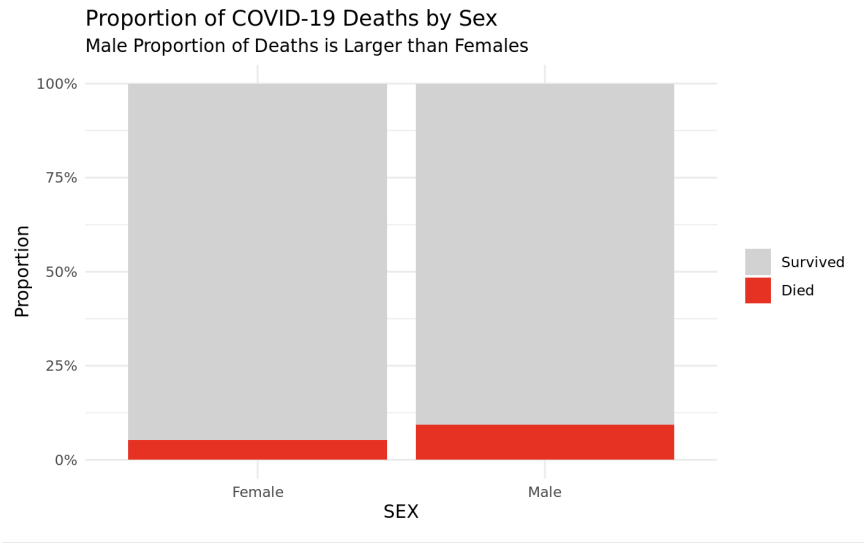

Figure 1

From Figure 1, we can visualize that the proportion of men who died in our COVID-19 dataset is larger than that of females. This aligns with Caruso et al.'s study indicating that men may be generally at a higher risk of severe outcomes from COVID-19 compared to women, possibly due to biological and behavioral factors. The difference in mortality rates between the sexes highlights the importance of considering sex as a significant predictor in our COVID-19 mortality model, thus potentially tailoring public health interventions accordingly.

Hannah Wang
Claire Luo

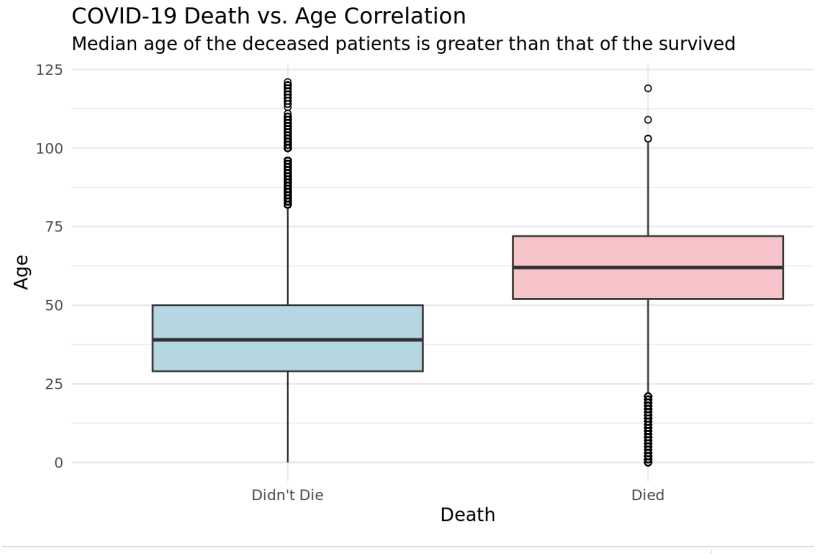

Figure 2

Figure 2 reveals that the median age of patients who died is notably higher than that of patients who survived. Furthermore, the group that did not die displays a lower age range, starting close to 0 and extending into the upper middle age, with fewer outliers that reach into the higher age range. This may suggest that for our dataset, the younger individuals tend to survive more frequently. The group that died spans a wider age range, beginning in middle age and stretching well into the elderly years. The plot strongly suggests that age is a significant factor in COVID-19 mortality, with older patients having a higher risk of death. This aligns with existing research indicating that older age is associated with higher mortality risks in COVID-19 due to factors such as diminished immune response and the presence of comorbidities.

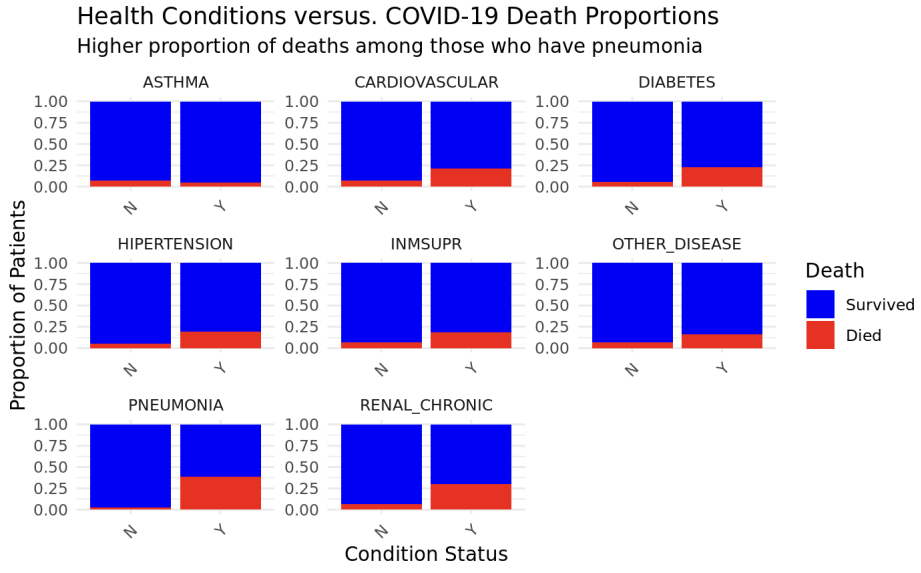

Figure 3

Figure 3 compares the mortality rates of COVID-19 patients across various health conditions, with a specific focus on the distinction between those who have the conditions (labeled as "Y") and those who do not ("N"). The plot for pneumonia stands out significantly, displaying a substantial proportion of deaths among patients with pneumonia compared to those without. This suggests that pneumonia is a critical risk factor for mortality in COVID-19 patients, most likely due to the additional burden on respiratory functions already compromised by the virus.

Methodology:

For our research, our methodology consisted of data cleaning and preparation, model selection, and our model assumptions.

Data Cleaning

For our research, we employed a structured approach for data cleaning and for our exploratory data analysis and modeling, focusing specifically on the binary outcome of patient survival(Died or Not). Since the majority of our features were categorical, initial data cleaning involved transforming multiclass disease status variables (PNEUMONIA, DIABETES, etc.) into a binary format, where 2 represents 'No Condition' and 1 indicates 'Has Condition.' From the dataset information, there were specific numerical values representing missing or null data (97, 99) were converted to NA, and rows containing these NAs were subsequently dropped in order to maintain data integrity. Additionally, the DATE_DIED variable was transformed into a binary 'died' variable, where the non-existent date ‘'9999-99-99' was coded as 0 (survived), and any other date was coded as 1 (died). This step was crucial for aligning the dataset with the binary nature of logistic regression analysis.

Model Selection

For our model, we decided to utilize logistic regression since our dependent variable ‘died,’ is a binary variable(Died or Not Died). Logistic regression was suitable since it allowed us to estimate the odds ratio for each predictor, which provided insights into the strength and association between the predictor (physiological and demographic features) and the patient death outcomes. We believed that developing distinct models for the different sets of predictors was a more focused approach to ensure the clarity in interpreting the separate effects of various health conditions on the COVID-19 deaths. This approach allowed each of our models to specifically address different aspects of research questions, facilitating a more focused analysis on physiological and demographic. Initially, we were deciding between incorporating interaction terms in our models, but we decided in order to maintain model interpretability and clarity, we would not include them since our primary goal is to understand the independent effect of each physiological and demographic. By isolating them in separate logistic regression models, we decided to provide clear, detailed insights into how each category of variables independently affects the risk of death from COVID-19. Furthermore, we decided to incorporate an AUC curve, as well as interpreting sensitivity, specificity, positive predicted values, and negative predicted values in order to evaluate the accuracy of both our logistic regression models. These measurements ensure that our findings are not only statistically significant but also clinically relevant in today’s world. The AUC curve is particularly valuable since it measures the model's performance across all possible classification thresholds, highlighting its ability to discriminate between patients who died from COVID-19 and those who survived. By integrating sensitivity and specificity, we were able to account for the model's accuracy in identifying true positives and true negatives, respectively, thus reflecting its effectiveness in detecting actual cases of mortality and survival. These incorporations allowed us to ensure that the model not only predicts mortality accurately but does so in a way that is practical and implementable in clinical real-world settings, allowing healthcare providers to make more informed decisions that could potentially save lives.

Model Assumptions

For our logistic models, we had to check for the assumption of linearity, in order words, if the logit transformation of the probability of the dependent variable, mortality, is a linear function of the continuous

predictor. This means that the relationship between age and the log odds of the outcome should be linear. We incorporated the empirical logit plot to assess whether the logit of the probability of dying is linearly related to age. Furthermore, the second assumption for our logistic model is independence. We can assume that the assumption of independence is satisfied since each observation in our dataset corresponds to a unique patient, with data collected independently from that of any other patient. Gathering information about one observation/patient would not inform us about another observation/patient. With our assumptions, we were able to appropriately incorporate our models.

Results

First, let's look at the output from our logistic regression model using the pre-existing health conditions. Since all the predictors for this model were strictly categorical, we did not have to check for the linearity assumption satisfaction.

```
Call:
glm(formula = died ~ as.factor(PNEUMONIA) + as.factor(DIABETES) +
    as.factor(ASTHMA) + as.factor(INMSUPR) + as.factor(HIPERTENSION) +
    as.factor(CARDIOVASCULAR) + as.factor(RENAL_CHRONIC), family =
"binomial",
    data = covid_health)

                          Estimate  Std. Error z value  Pr(>|z|)
(Intercept)               -3.940151   0.007526 -523.53   <2e-16 ***
as.factor(PNEUMONIA)1      2.939022   0.009073  323.93   <2e-16 ***
as.factor(DIABETES)1       0.716696   0.010671   67.16   <2e-16 ***
as.factor(ASTHMA)1        -0.513693   0.030762  -16.70   <2e-16 ***
as.factor(INMSUPR)1        0.285193   0.027739   10.28   <2e-16 ***
as.factor(HIPERTENSION)1   0.749181   0.010299   72.75   <2e-16 ***
as.factor(CARDIOVASCULAR)1 0.230301   0.022394   10.28   <2e-16 ***
as.factor(RENAL_CHRONIC)1  0.555161   0.021603   25.70   <2e-16 ***
```

Raising *e* to the power of each of the coefficient estimates from our model output above produces the odds of a patient with the health condition dying from COVID-19 in relation to the baseline level of each predictor variable (the baseline is the absence of the disease). The resulting values are shown in Table 1 below.

```
Results of exp(coef(logit_mod_health))
```

Intercept	PNEUMONIA	DIABETES	ASTHMA	INMSUPR	HYPERTENSION	CARDIOVASCULAR	RENAL_CHRONIC
0.0194453	18.897359	2.047656	0.598282	1.330018	2.115267	1.258979	1.74222116

Table 1

We notice that the presence of pneumonia, hypertension, and diabetes in a patient are the most influential in

affecting the likelihood of that patient dying from COVID-19. However, the most notable is pneumonia, with approximately 18.887 times the odds. The presence of asthma for a patient is the least influential in affecting the likelihood of that patient dying from COVID-19.

A patient with pneumonia is predicted to have approximately 18.887 times the odds of dying from COVID-19 compared to a patient without pneumonia, while adjusting for the presence of diabetes, asthma, immunosuppression, hypertension, cardiovascular disease, and chronic renal disease.

A patient with diabetes is predicted to have approximately 2.046 times the odds of dying from COVID-19 compared to a patient without diabetes, while adjusting for the presence of pneumonia, asthma, immunosuppression, hypertension, cardiovascular disease, and chronic renal disease.

A patient who has hypertension is predicted to have approximately 2.116 times the odds of dying from COVID-19 compared to a patient who does not have hypertension, while adjusting for the presence of pneumonia, diabetes, asthma, immunosuppression, cardiovascular disease, and chronic renal disease.

A patient with asthma is predicted to have approximately 0.598 times the odds of dying from COVID-19 compared to a patient without asthma, while adjusting for the presence of pneumonia, diabetes, immunosuppression, hypertension, cardiovascular disease, and chronic renal disease.

Table 2 evaluates the performance of the health conditions model by comparing the model's predictions of whether a patient died from COVID-19 to the actual mortality result. Model predictions are on the left hand column (either "did not die" or "died") and actual values are on the top row (0 = did not die, 1 = died).

	0	1
Did not die	937687	61126
Died	13203	13706

Table 2

Sensitivity: TP / (TP+FN) = 13706 / (13706 + 61126) = 0.183

In the context of our model, sensitivity measures the proportion of actual deaths)correctly identified. Here, the model correctly predicts 18.3% of all actual deaths. This low sensitivity suggests the model is not very effective at predicting all patients who will die.

Specificity: TN / (TN + FP) = 937687 / (937687 + 13203) = 0.986

Specificity measures the proportion of patients that survived that are correctly identified. The model has a high specificity of 98.6%, meaning that it is very effective at identifying patients who will not die.

Positive Predicted Value: TP/(TP + FP) = 13706 / (13706 + 13203) = 0.509

The PPV indicates the likelihood that patients predicted by the model to die actually end up dying. In this context, when the model predicts death, there is a 50.9% chance that the patient actually died. This suggests moderate reliability of the model's positive prediction.

Negative Predicted Value: TN/(TN + FN) = 937687 / (937687 + 61126) = 0.939

The NPV tells us the probability that patients predicted not to die actually did not die. There is a 93.9% chance it is correct, which is quite high.

The model performs well at predicting patients who will survive COVID-19, as reflected by its high specificity and NPV. However, it struggles significantly with correctly identifying all patients who will die, as shown by the low sensitivity. The moderate PPV indicates that about half of the death predictions made by the model are

correct, suggesting room for improvement in predicting actual mortality. The high specificity and NPV are valuable in a clinical setting where it is important to accurately identify patients who are not at risk of dying, possibly helping to prioritize resources for those at higher risk.

Finally, we will analyze the AUC number and the ROC curve to evaluate the predictive capability of our logistic regression model.

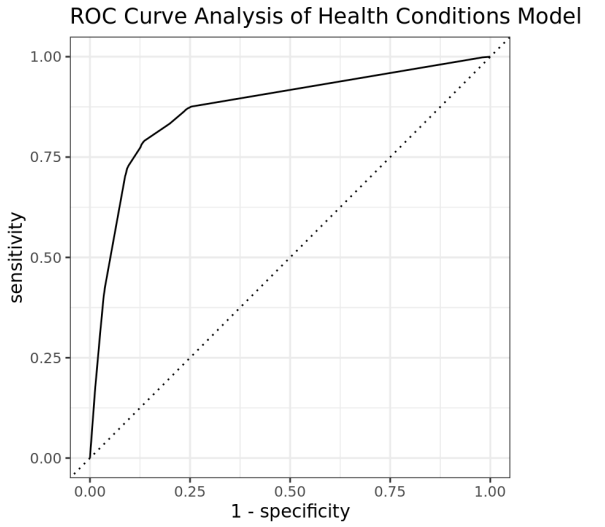

Figure 4

The AUC we achieved by physiological and health condition model is 0.8718065. This value indicates that the model has an 87.18% probability of correctly distinguishing between a patient who died from COVID-19 and one who did not, based on the model's predicted probabilities. As discussed before, the AUC is particularly important because it demonstrates the model's utility in clinical settings. Our AUC score is relatively good, and healthcare providers could use the model's predictions to prioritize interventions or monitor patients more closely if they are predicted to have a higher risk of severe outcomes.

Now, let's look at the output from our second logistic regression model, this time based on demographic factors like gender and age. First, we have to ensure that we satisfy our assumption of linearity for our continuous variable, age.

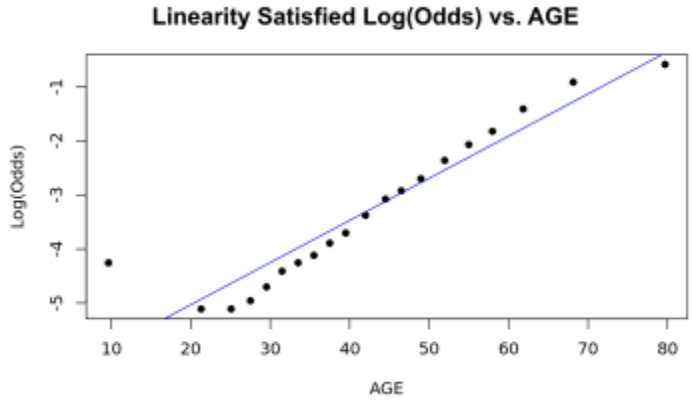

Figure 5

From Figure 5, we see that it displays a linear relationship between the log odds of dying and age. The graph displays points that are relatively linearly uniform and follow the trend of the line. From the graph, we

do not see that it contains any distinct trend that would challenge our assumption of linearity. We have discussed satisfying the assumption of independence, and since the assumption of linearity and independence are satisfied, we can move forward with our logistic regression model using age as a predictor for the likelihood of dying from COVID-19.

```
Call:
glm(formula = DIED ~ as.factor(SEX) + AGE, family = "binomial",
    data = covid_health)


Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.7507458  0.0168321 -401.06   <2e-16 ***
as.factor(SEX)Male  0.6375300  0.0085209   74.82   <2e-16 ***
AGE                 0.0760520  0.0002632  288.91   <2e-16 ***
```

From our model output, we can make the following conclusions:

Holding age constant, we predict the odds of a male patient in the COVID-19 dataset passing away to be around e^0.63753001(1.8918) times that of a female patient.

Holding sex constant, we predict that for each additional year in age of the patient, the odds of passing away are multiplied by e^0.07605202(1.0790).

Again, we evaluate the demographics model's predictions by comparing it with the actual mortality result. In the Table 3 below, the predictions are along the left column and the actual results are along the top row.

	0	1
Did not die	943940	70364
Died	6950	4468

Table 3

Sensitivity: TP / (TP+FN) = 4468 / (4468 + 70364) = 0.0597

Specificity: TN / (TN + FP) = 943940 / (943940 + 6950) = 0.993

Positive Predicted Value: TP/(TP + FP) = 4468 / (4468 + 6950) = 0.391

Negative Predicted Value: TN/(TN + FN) = 943940 / (943940 + 70364) = 0.931

There is a similar pattern for this model with our health conditions model, where the specificity and the Negative predicted value are significantly higher than the sensitivity. This also suggests that this model is excellent at predicting patients who will survive COVID-19, as reflected by its high specificity and NPV, but it struggles significantly with correctly identifying all patients who will die, as shown by the low sensitivity.

While both of our models are reliable for ruling out death, it is less effective at confirming cases of death, signifying a need for further refinement to enhance its sensitivity and overall predictive accuracy. This could involve incorporating more features, adjusting the model's parameters, or exploring different modeling techniques.

Finally, we will analyze the AUC number and the ROC curve for this model as well.

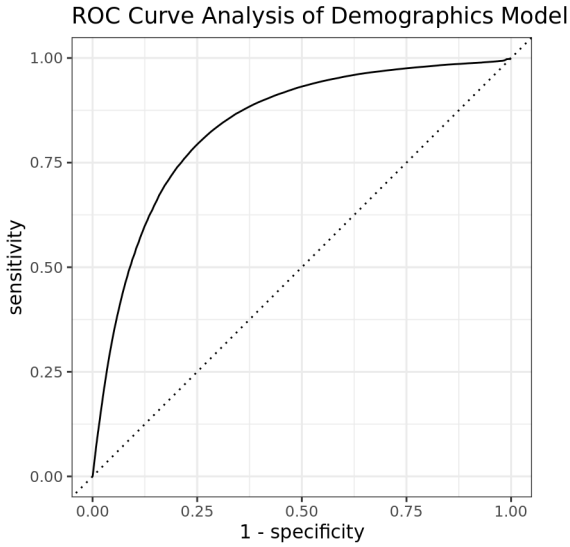

Figure 6

The AUC we achieved by the model is 0.8395. This value indicates that the model has an 83.95% probability of correctly distinguishing between a patient who died from COVID-19 and one who did not, based on the model's predicted probabilities. This is a decently high score, suggesting that this model does a fairly good job of distinguishing the patients that did pass away and did not. This means that our model is fairly reliable, but still has room for improvement. Low sensitivity and low PPV scores for both models may pose potential risks, such as missed early interventions since it was not predicted, and potential resource allocation. We will discuss the implications of this in our discussion.

By identifying the most susceptible groups, we intend to provide valuable insights that could guide public health decisions and interventions aimed at reducing mortality among the most affected populations. For example, individuals with pneumonia, diabetes, and hypertension have an increased likelihood to die from COVID-19 compared to individuals who don't have these health conditions, suggesting that we can implement preventative measures and policies particularly protecting these groups. Knowing that male patients and older patients are more likely to die from COVID-19 in comparison to women and younger patients respectively (while holding the other variable constant) can be useful information to know in making treatment decisions.

Furthermore, to comprehensively assess the possible improvements in model performance, we made a final model incorporating all predictors. After checking for satisfaction of our assumptions, this comprehensive model was developed without including interaction terms, specifically to observe any changes in the sensitivity and PPV scores. Our objective was to determine whether a broader model would enhance our ability to accurately predict COVID-19 mortality outcomes. Below lies our confusion matrix and AUC for the comprehensive model.

	0	1
Did not die	931992	50028
Died	18898	24804

Table 4

Sensitivity: TP / (TP+FN) = 24804 / (24804 + 18898) = 0.5676

Specificity: TN / (TN + FP) = 931992 / (931992 + 50028) = 0.949

Positive Predicted Value: TP/(TP + FP) = 24804 / (24804 + 50028) = 0.3315

Negative Predicted Value: TN/(TN + FN) = 931992 / (931992 + 18898) = 0.9801

We notice that the sensitivity has increased by a decent amount compared to our demographic and health condition model, suggesting that incorporating a broader array of predictors, both the demographic and physiological features, enhanced the model's ability to correctly identify patients death from COVID-19. This improvement in sensitivity is particularly significant because it indicates a reduction in the number of false negatives, ensuring that fewer cases at high risk are overlooked. Consequently, this can lead to more timely and targeted interventions, potentially improving patient outcomes. On the other hand, we did not notice a notable increase in the PPV.

Discussion

Our results conclude that out of pre-existing health conditions, pneumonia is the most influential in increasing the likelihood of a patient dying from COVID-19, as gathered from our logistic regression models. Between sex and age, sex is more influential in increasing the likelihood of death from COVID-19. However, both models had low sensitivity and positive predicted value scores. When we fit an overall logistic regression model including all the predictors (health conditions and demographics), a higher sensitivity score was achieved. However, although the model performance was better with more predictors regarding sensitivity, no interaction terms were incorporated to encapsulate any combined effect between two. Since our research focused particularly on the two separate subgroups of physiological and demographic features, there may be more complex interactions that we did not dive into in our project that could capture more holistic interdependencies between the predictor variables. In the future, we'd like to experiment with various interaction terms and whether there is a relation between certain demographic and health-related factors (for example, old people are more likely to contract certain illnesses).

The dataset that we worked with presented several limitations. There were a considerable amount of null and missing values, labeled as the values 97 and 99, that we needed to filter out before fitting the logistic regression models to the data. In doing so, we lost valuable information. Additionally, the only numerical variable we worked with was age, so we seek to incorporate more numerical variables into our statistical analysis in the future. The reliability and validity of our findings are contingent on the accuracy and representativeness of the dataset and the appropriateness of the logistic regression analysis. Thus, for the future, we would like to incorporate data from a wide variety of areas, as the current dataset is from the Mexican government, not completely representative of the United States as a whole. Doing so would allow us to apply our results to a larger population. Future work could include incorporating more variables, more advanced modeling techniques, and utilizing a more comprehensive and representative dataset. By expanding the scope and depth of our analysis, we can improve the accuracy of mortality predictions and contribute to more targeted and effective public health responses. Our study not only sheds more light on how COVID-19 affects various groups but also lays the groundwork for subsequent research targeted at reducing the mortality rates.

Hannah Wang
Claire Luo STA 210

Works Cited

National Center for Biotechnology Information. "A comparison of causes of death in China and India, 2001-2003." Bulletin of the World Health Organization, vol. 90, no. 7, 2012, pp. 515-521. PubMed Central, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3291398/#:~:text=In%20contrast%2C%20age%2Dspecific%20death,20%E2%80%9340%20years%20of%20age.

MDPI. "Evaluating Deep Learning Techniques for RNA Editing Detection." International Journal of Molecular Sciences, vol. 24, no. 3, 2023, p. 2638. MDPI, https://www.mdpi.com/1422-0067/24/3/2638.