Jake Yang and Blake Mills

Wayne T. Lee

Applied Data Mining

19 April 2021

Covid and Individual Risk

Introduction

The current COVID-19 has had a major impact on the lives of many, particularly in the effect it has had on people's health. With a death count currently estimated to be above three million, it is clear that a virus of this magnitude and obscurity has wreaked havoc on the global healthcare system. In order to assist with possible mitigation of further deaths, we set out to build a model to identify risk at an individual level for having severe complications of COVID-19. We defined severe complications as being hospitalized or dying of the virus. The intended goal of this model is to identify patients most at risk of having severe complications in order to provide and prioritize them with access to healthcare in an attempt to mitigate the effects.

Goals and Audience

The goal for our model and analysis is to assist in developing scales for risk of COVID-19. Specifically, we would like to examine who is at risk for being hospitalized or dying from COVID-19. Our model's intended audience would be for medical professionals, healthcare workers, and public health employees. The intended use of the model would be to locate and identify people who are more likely to be hospitalized and die from Covid in order to proactively mitigate health issues before they become more intense. During the early stages of the pandemic, one of the major issues in large cities like New York was the overwhelming of the healthcare system and the struggle to prioritize who needed medical treatment most. The goal of the model is to identify those most at risk for having severe complications of Covid in order to prioritize their healthcare need to prevent further detrimental effects.

Data Sets and Collection

To conduct our analysis and construct our model, we gathered data sets from the Two Rivers Public Health Department in central Nebraska. The jurisdiction covered the counties of Buffalo, Dawson, Phelps, Kearney, Gosper, Harlan, and Franklin. These counties combined have a total population of approximately 100,000. All datasets from the department are collected, entered and managed by department staff. They are not publicly available, but can be obtained through a FOIA (Freedom of Information Act) request, as the department is publicly funded.

We utilized two data sets from the department. First, was the Investigations data set. This is an excel generated output that provides information from case investigations of COVID-19. The

data set contains approximately 12,000 patients with over 300 variables. While most are demographic variables (such as name, address, healthcare provider, etc. . .), many variables relate directly to the patient's experience with COVID-19 (symptoms, date of test, exposures). Due to the way case investigations are conducted, it was noted that some patients were not marked as "Confirmed" Covid cases, but rather "Not a Case" or "Suspect Case." It was discovered that these were people in the early stages of the pandemic that were being monitored due to COVID-like symptoms or an exposure. Due to the lack of a confirmed test, these people were removed from our dataset, bringing the total observations down to approximately 10,000.

To simplify the dataset, only variables relating to a patient's symptoms, pre-existing conditions, occupational information, and basic demographics (age, gender, city and county of residence) were kept. This brought the total number of variables down to approximately 30. After this, we removed any cases where the patient was not able to be contacted. While data for these unreachable patients would include information about age, gender, race, and ethnicity, we would not be able to ascertain their symptoms or preexisting conditions, thus building any regressions with these variables would not be possible. After unreachable cases were removed, the final dataset was left with 7,340 patients.

We also implemented the use of another data set provided by the health department to complement our analysis. The set used is referred to as the COVID-19 Vertical, and it contains all testing data. While the Investigations provides information about the patient who got tested, the Verticals data allows us to access all positive tests, allowing us to monitor how many people tested positive and when they tested positive. While we know the people in the investigation were positive, we are unable to see how many tests they took, or if they took multiple tests apart and remained positive. Thus, this additional set allows us to monitor how many people tested positive in a given area within a specific time span.

While the original dataset contained over 100,000 observations, we chose to filter out all negative COVID-19 tests, as our analysis was primarily concerned with monitoring risk, thus negative tests are less influential than positives. While this dataset is also monitored by the health department, all tests are uploaded automatically, thus all variables are presented in a standardized format. Demographic and personal information is included in the set, as well as specifics regarding what facilities the test was conducted at and what laboratory performed the test; however, for our analysis, the primary variables we were interested in is the town in which the patient resided and on what date was their specimen collected for testing.

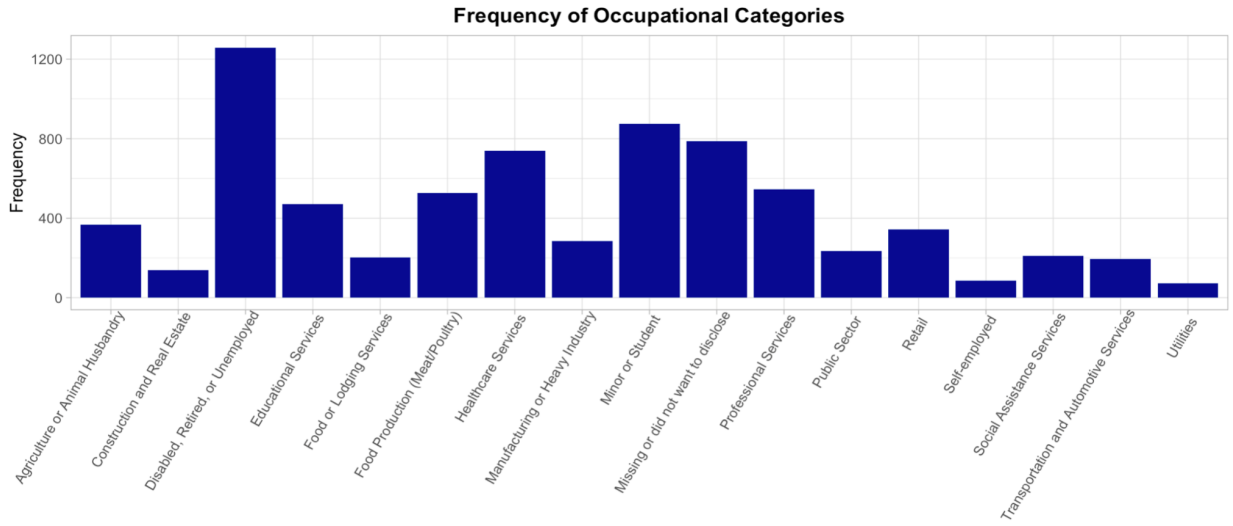
The last data we utilized was the Census data for city level population of all towns in the health department's jurisdiction. This was pulled from the Census website and utilizes the 2019 population estimates for each town.

Data Processing, Features, and Summarization

From the variables available, we were interested in creating a feature to categorize people into occupational categories in order to examine whether certain occupations were more prone to severe outcomes of illnesses than others. This notion is anecdotally supported by the reports of healthcare workers and workers in the meatpacking industry being more prone to catching and spreading the virus. The investigation data set we used has variables for "occupation," "employer," and "industry." However, all of these variables were manually entered by health department employees, and lacked a clear standardized set of rules. Because of this, none of the variables, as they stood, were able to place people into a standardized industry category. To overcome this issue, we utilized a random forest categorization model to do this sorting for us. Before running the forest, we placed all text in the "occupation," "employer," and "industry" variables into one string, removed all capitalization, punctuation, and numbers. We then removed any sparse terms (words that appeared in less than 0.3% of observations). Finally, to account for the fact that most children would not have an occupation, but they should not be read as "unemployed", we attached the word "minor" to the string if the patient was under 18 years old. The reason for appending the word and not just instantly categorizing people under 18 as "minor" is due to the fact that some older minor may have jobs. Due to this, we allowed the regression to sort them on the best fit.

The model was built using occupational categories relevant to the occupations of people in the jurisdiction. In total, 17 occupation categories were used (as recommended by the epidemiologist at the department, Dr. Aravind Menon). The frequency of these categories are presented in Figure 1 below. Full descriptions of the occupational categories and example occupations are included in the appendix.

Figure 1



The largest proportion of our observations include people who are disabled, retired, or unemployed. This may illustrate a bias in the data, as case investigations are only opened when someone gets a positive COVID-19 test, and people in this category would have the time to get tested, as they would not be working. Nonetheless, we still feel there are enough observations in all categories to run a proper analysis.

We also did preprocess cleaning on all of the symptoms and pre-existing condition variables. Originally, there were 18 dummy symptoms variables present in the data set. One of the main issues with the data is that patients responded to the symptoms themselves and were not medically diagnosed, thus introducing the possibility of bias in responses. Additionally, some symptoms were very similar to each other (had a fever vs felt feverish). For these reasons, we chose to collapse all the symptoms into 5 categories. The only symptom that was not collapsed was the indicator of whether patients developed pneumonia, as this would have been medically diagnosed. The frequencies of these symptoms categories, as well as what symptoms are included, and pre-existing conditions can be found in Table 1 below.

Table 1

Symptoms	Frequency (%)	Pre-existing Condition	Frequency (%)
ENT: runny nose, loss of taste or smell, sore throat	4,477 (61.0)	Liver Condition	45 (0.6)
Gastrointestinal: abdominal pain, diarrhea	1,858 (25.3)	Lung Conditions	613 (8.4)
Respiratory: cough, dyspnea (difficulty breathing), wheezing	4,051 (55.2)	Renal (Kidney) Conditions	207 (2.8)
Systemic: fatigue, fever, felt feverish, headache, loss of appetite, nauseous	5,184 (70.6)	Cardiovascular Conditions	645 (8.8)
Musculoskeletal: chills, rigors, myalgia (muscle pains)	3,582 (48.8)	Immunocompromised	248 (3.4)
Pneumonia	179 (2.4)	Diabetes	611 (8.3)
		Neurological or Psychiatric Condition	494 (6.7)
		Smoker or Vaper	245 (3.3)
		Obesity	188 (2.6)
		Substance Abuse	50 (0.7)

We also included selected demographic and outcome (death and hospitalization indicators) variables in our analysis. The only continuous demographic variable we included was the patient's age. Of the 7,340 observations, the mean age was 43.6 years with standard deviation of

21.16 years. All other variables were categorical and the distribution of them is presented in Table 2 below.

Table 2

Variable		Freq (%)
Race		
	American Indian or Alaska Native	22 (0.3)
	Asian	49 (0.7)
	Black or African American	139 (1.9)
	Native Hawaiian or Pacific Islander	10 (0.1)
	White	6,507 (88.7)
	Other Race	166 (2.3)
	Unknown	447 (6.1)
Ethnicity		
	Hispanic or Latino	1,417 (19.3)
	Not Hispanic or Latino	5,516 (75.1)
	Unknown	407 (5.6)
Gender		
	Male	3,442 (46.9)
	Female	3,894 (53.0)
	Unknown	4 (0.1)
County of Residence		
	Buffalo	3,624 (49.4)
	Dawson	2,008 (27.4)
	Gosper	148 (2.0)
	Harlan	148 (2.0)
	Franklin	162 (2.2)
	Kearney	405 (5.5)
	Phelps	675 (9.2)
	Unknown	170 (2.3)
Hospitalized for COVID-19		424 (5.8)
Died of COVID-19		136 (1.9)

For our other data set, the Vertical, we preprocessed the data by filtering out any negative COVID-19 tests. As stated prior, the positives are more influential in determining risk of the virus than negative tests. This brought the total observation down from over 100,000 to 17,415. From there, we grouped all observations by the city the patient resided in, and calculated a 14 day rolling sum for each city, each day of the pandemic. Thus, this new feature was able to show us how many people in the last 14 had tested positive for the virus. We used this as a way to represent how many people in a town at a given time would be infected with COVID-19. The fourteen day threshold was chosen as it was the recommended quarantine time by the CDC in the beginning. While the CDC has since lowered this number to 10 day with 3 days being symptom free, we chose to keep fourteen days to create a more conservative model of the prevalence of the virus in a given area. Given that some town are much larger than other, we chose to standardize the rolling sum by dividing the count of people sick in the last 14 days by the population of the town from the Census data, thus giving us a number that represented the proportion of the town that had tested positive for the virus in the last 14 days. The logic behind creating this variable was to create a feature that would represent the danger of COVID-19 in the area. As seen with places like New York and Arizona, mortality rates of the virus can increase when an area is overwhelmed with cases, as access to quality healthcare becomes harder to attain as it is demanded by more. Thus, we use this new variable as a proxy to represent access to quality care and prevalence of the virus in a certain place and certain time.

After this was complete, we merged this set with the Investigations dataset based on the date tested and city. Thus, along with all the other variables of the Investigation set, we were able to add the new variable of what percentage of the city had tested positive for COVID-19 in the last 14 days when the patient tested positive themselves.

In our last step of cleaning, we removed any of the demographic variables that had “Unknown” levels for race, ethnicity, gender, or age. After this was completed, the final set used in the regression has 7,140 observations

Algorithm and Validation

In order to measure risk of Covid, we decided to build a logistic regression model on Death and a logistic regression model on Hospitalization. After data cleaning we had 7170 observations with records of them being actually dead or actually hospitalized. Both of the models are using similar methods.

First, we decide to do a stepwise selection by AIC to decrease the dimension preliminarily. The variables we threw into the initial logistic regression include All the symptoms, pre-existing

conditions, after effects, age, the industries they are working in, as well as percent14(which is the percentage of Covid positive population within the patient residential city on the day the patient gets tested). Moreover, we would like to make sure there is no collinearity between independent variables, thus we then check the VIF for each variable that has been selected through stepwise selection and keep those with VIF lower than 10 which most of them have a VIF around 1.

Until this step we had 13 variables left for the death model. For the death model, since we are assessing medical data, we need to warn people with higher risk of dying. The risk of not warning is much more costly than making a wrong warning. Thus, we would like the recall to be higher than 80% (which means 80% of positive are recognized). In this step, we used cross validation to select the best model with higher than 0.8 recall score and highest precision score. Each fold is using slightly different variables, and we picked the folds that generate the best recall and precision scores. On the other hand, for the hospitalization model we care both precision and recall equally thus, we used F1 score instead of 0.8 recall as metrics. We screen through all cutoff points to find the variables that generates the best model

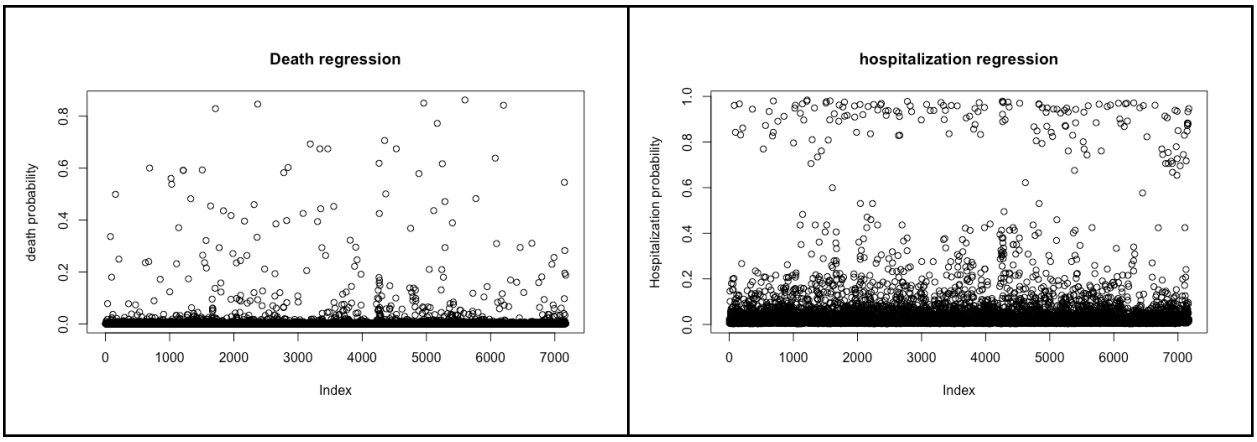
Death model cross validation results

```
[1] "cross validation fold 1"
[1] "variables used in model: RespSymp\nSysSymp\nMuscSymp\nRenalCond\nCVCond\nNeuroCond\nPneum\nAgeNum\nNewIndustry_Mi
ssing or did not want to disclose"
[1] "Probability higher than  0.091  is consider dead"
[1] "Recall : 0.827586206896552"
[1] "Precision : 0.452830188679245"
[1] "cross validation fold 2"
[1] "variables used in model: GISymp\nSysSymp\nMuscSymp\nRenalCond\nCVCond\nNeuroCond\nPneum\nAgeNum"
[1] "Probability higher than  0.041  is consider dead"
[1] "Recall : 0.814814814814815"
[1] "Precision : 0.366666666666667"
[1] "cross validation fold 3"
[1] "variables used in model: GISymp\nSysSymp\nRenalCond\nNeuroCond\nPneum\nAgeNum\nNewIndustry_Missing or did not wan
t to disclose"
[1] "Probability higher than  0.036  is consider dead"
[1] "Recall : 0.80952380952381"
[1] "Precision : 0.293103448275862"
[1] "cross validation fold 4"
[1] "variables used in model: GISymp\nSysSymp\nMuscSymp\nCVCond\nNeuroCond\nPneum\nAgeNum"
[1] "Probability higher than  0.017  is consider dead"
[1] "Recall : 0.848484848484849"
[1] "Precision : 0.325581395348837"
[1] "cross validation fold 5"
[1] "variables used in model: MuscSymp\nRenalCond\nNeuroCond\nObese\nPneum\nAgeNum"
[1] "Probability higher than  0.024  is consider dead"
[1] "Recall : 0.84"
[1] "Precision : 0.276315789473684"
```

Hospitalization model cross validation results

```
[1] "cross validation fold 1"
[1] "variables used in model: ENTSymp\nRenalCond\nCVCond\nPneum\nAgeNum"
[1] "Probability higher than  0.021  is consider hosipitalized"
[1] "F1 : 0.547619047619048"
[1] "Recall : 0.55421686746988"
[1] "Precision : 0.541176470588235"
[1] "cross validation fold 2"
[1] "variables used in model: GISymp\nENTSymp\nSysSymp\nRenalCond\nCVCond\nObese\nPneum\nAgeNum"
[1] "Probability higher than  0.016  is consider hosipitalized"
[1] "F1 : 0.475247524752475"
[1] "Recall : 0.571428571428571"
[1] "Precision : 0.406779661016949"
[1] "cross validation fold 3"
[1] "variables used in model: GISymp\nENTSymp\nSysSymp\nRenalCond\nCVCond\nObese\nPneum\nAgeNum"
[1] "Probability higher than  0.068  is consider hosipitalized"
[1] "F1 : 0.496240601503759"
[1] "Recall : 0.392857142857143"
[1] "Precision : 0.673469387755102"
[1] "cross validation fold 4"
[1] "variables used in model: GISymp\nENTSymp\nRenalCond\nCVCond\nObese\nPneum\nAgeNum"
[1] "Probability higher than  0.034  is consider hosipitalized"
[1] "F1 : 0.482269503546099"
[1] "Recall : 0.404761904761905"
[1] "Precision : 0.596491228070175"
[1] "cross validation fold 5"
[1] "variables used in model: GISymp\nENTSymp\nRenalCond\nSmoker\nObese\nPneum\nAgeNum"
[1] "Probability higher than  0.017  is consider hosipitalized"
[1] "F1 : 0.451282051282051"
[1] "Recall : 0.523809523809524"
[1] "Precision : 0.396396396396396"
```

Lastly, we use the entire dataset and the variables selected from best folds to perform a final logistic regression and predict the probability of dying or hospitalizing because of Covid. The probability scatter plot is looking good, for the death regression, the majority of the patients are not in risk of dying which was shown by the plot on the left hand side. On the right hand side, the plot shows a gap between patients who need hospitalization and those who don't. The plot is reasonable since most people cure on themself, but the severe case will need to be hospitalized.



After the regression models were built and validated, we generated predicted probabilities for each patient being hospitalized and dying of COVID-19. Afterwards, these two values were averaged. Patients were then ranked in order of the resulting value, with the lowest value being 1 and the highest being 7,140 in our set. From this, we are able that people with the highest ranks are most likely to have severe complications of COVID-19.

After discovering that our percent14 variable was not significant in our models, we chose to discretize the variable into five categories (Very Low, Low, Medium, High, and Very High) to represent community wide risk of the virus. We agreed that while this variable may not be statistically significant in the regression, it was important to factor into the analysis in one way or another, as measuring community risk will be relevant (maybe not strongly correlated) to COVID-19 as a whole. Therefore, we choose to break any ties in ranks with this ordered variable. Essentially, if someone had a tie in rank with a patient, the tie was broken with how prevalent (percent14) COVID-19 was in the community at a given time, with more severe prevalence making someone rank lower (higher rank value). Thus, we use both individual (hospitalization and death probability averages) and community (percent 14 category) level risk to rank and prioritize patients who have tested positive for COVID-19.

Usefulness and Limitations

To conduct one last check to analyze the relevance of our prioritization algorithm, we pull the 135 patients with the lowest (highest value) ranks. The choice of 135 observations stems from the total number of deaths in the set. Ideally, all 135 would be identified as death. When we pulled the cases based on those with the lowest ranks for *only* the death prediction, prioritized 73 patients that died of COVID-19 and 102 that were hospitalized. We compared this to the ranks of the average of the hospitalization and death prediction values. From this, the algorithm

prioritized 76 people that died of COVID-19 and 113 that were hospitalized. While the improvement is small, we feel that this displays that the inclusion of the hospitalization metric allows us to correctly identify more people that are more at risk of having severe complications of COVID-19. While not perfect, we feel that our algorithm allows healthcare workers to prioritize people who are most likely to develop severe complications of COVID-19 by recognizing their chances of being hospitalized and dying and providing them with access to care. We feel that our algorithm is especially useful at saving healthcare workers time by quickly sorting people by risk of developing COVID-19. While it would be difficult to keep track of several thousand people at once who are infected, giving healthcare workers the options to identify the most at risk will save time and allow resources to be properly allocated to those who need them most.

In terms of the percent14 variable, we are still unsure the entirety of its usefulness. Given that the population from the data we had access to never had a severe overwhelming of the healthcare system compared to other locations like New York and Arizona, it may not be entirely useful. However, since we do no include the variable in the regression, but rather paired it with the risk-rank metric, we feel that the extra information allows healthcare providers to use their own discretion to interpret its meaning. While it may not be relevant in a rural area when case numbers are usually low, it may be useful in larger, metropolitan areas where outbreaks can affect ease of access to quality care. Applying the model to a more diverse population would allow us to gain a better understanding of this features usefulness; however, as it does not affect the regression and provides healthcare workers with more clarity about the community level risk of COVID-19, we feel it is important to include.

There are a few limitations that come in our data. First, due to the fact that patient data has to be collected through an interview, there is a strong chance of non-response and self-reporting bias. In our data set, we were only able to use the information of people who contact tracers could reach to collect their data, there is no guarantee that the data that was reported to them was accurate of the patient, as it was never confirmed by a medical professional. Further, since our model was built using only complete cases, there is a possibility that the non-response patients would not fit the model. Beyond issues with data collection, our current model is only applicable to a Nebraska cohort. While this particular data collection method and questionnaire is standard across Nebraska, different states have different systems and different questions they ask. While we tried to use variables that should be common across states, the particular symptoms and pre-existing conditions asked may vary from state to state. Moreover, in our analysis, we found that a patient's occupation was not significant in measuring their risk of being hospitalized or dying from the virus. While this was the case for us, it is possible that industries prevalent in other states may be significant. As mentioned, the dataset comes from Nebraska, which has a very different set of prevalent industries compared to other states. For instance, coastal states may have a heavy fishing industry prevalence which would not exist in Nebraska due to its

distance from the ocean. While the occupational categories we picked were not found to be significant, it is possible there is still a relationship between industry and risk of severe complication of COVID-19, just not in industries prevalent in our region.

Data Dredging

We are confident that our results are a valid metric of measuring risk on an individual level and are not a result of data dredging. Through our k-fold cross validation and stepwise logistic regression, we were able to validate the significance of the variable in our regression analysis. Moreover, our objective of our model was not necessarily to correctly predict who will die or be hospitalized due to COVID-19, but rather to simply identify those most at risk for such severe outcomes in an attempt to intervene before they can occur. Thus, in our model, while we would like to reduce the number of false positives in order to narrow down the population which needs priority access to medical care, a false positive is much less detrimental than a false negative (an instance where we believe a person will not have a severe complication of the virus and then does). While there are some biases in our data set, as described in the limitations above, we feel that our analytic methods and object of our model prevent a data dredging outcome.

Conclusion

While the novel nature of the coronavirus makes building models to predict outcome somewhat difficult, we believe our algorithm is a decent metric for measuring risk of an individual's level of risk of developing COVID-19. While not perfect, our algorithm gives healthcare professionals the ability to quickly identify those most at risk of having severe complications in order to intervene with care and prevent detrimental outcomes from occurring. While more data from different regions of the country would allow us to fine tune our algorithm, we feel that it still meets the intended goal of quickly filtering a large population to identify a target population (those most at risk).