A statistical analysis of COVID-19 severity and pre-existing medical conditions

Nihaal Subhash

DAT_SCI 421 Final Project

Coordinator: Prof. Michael Schmitt

December 3, 2021

1. Introduction

Pre-existing medical conditions or comorbidities tend to impact the severity of covid symptoms[1]. Examples of these include - obesity, diabetes, pneumonia, etc. Age and sex are also prime factors in impacting covid severity. This study aims to corroborate this preexisting evidence and find new correlations amongst symptoms and covid severity. We shall also focus on a combination of factors – how patients having multiple comorbidities (such as both diabetes and pneumonia) can impact covid severity. We shall also build different classifiers for predicting covid severity and compare their relative performance.

2. Data

The dataset[2] is prepared by the Mexican government that contains the following individual information about the patients:

Column	Description	Range of Values
id	unique identification number for each patient	-
sex	sex of patient	1 Female 2 Male
patient_type	hospitalized status of the patient	1 for not hospitalized 2 for hospitalized
entry_date	date patient went to the hospital	-
date_symptoms	date patient first showed symptoms	-
date_died	date patient died	9999-99-99 if the patient did not die else a regular date
intubed	whether or not patient needed to be put on a ventilator	1 yes 2 no 97/98/99 not specified
pneumonia	whether patient has pneumonia	1 yes 2 no 97/98/99 not specified
age	age of patient	-
pregnancy	whether or not patient is pregnant	1 yes 2 no 97/98/99 not specified
diabetes	whether or not patient has diabetes	1 yes 2 no

		97/98/99 not specified
copd	whether or not patient has chronic obstructive pulmonary disease	1 yes 2 no 97/98/99 not specified
asthma	whether or not patient has asthma	1 yes 2 no 97/98/99 not specified
inmsupr	whether or not patient is immunosuppressed	1 yes 2 no 97/98/99 not specified
hypertension	whether or not patient has hypertension	1 yes 2 no 97/98/99 not specified
cardiovascular	whether or not patient has any heart related disease	1 yes 2 no 97/98/99 not specified
obesity	whether or not patient is obese	1 yes 2 no 97/98/99 not specified
renal_chronic	whether or not patient has chronic renal disease	1 yes 2 no 97/98/99 not specified
tobacco	whether or not patient consumes tobacco	1 yes 2 no 97/98/99 not specified
other_disease	whether or not patient has any other disease	1 yes 2 no 97/98/99 not specified
contact_other_covid	whether or not patient has been in contact with any other covid patient	1 yes 2 no 97/98/99 not specified
icu	whether or not patient has been admitted to an intensive care unit	1 yes 2 no 97/98/99 not specified
covid_res	whether or not patients result has come covid positive or not	1 yes 2 no 3 awaiting results

- The dataset contains information about 566,602 patients.
- Please note that while conducting statistical data analysis, this dataset has been slightly modified – for example, the label 2 for ‘no’ has been replaced with 0 and the labels 97/98/99 for ‘not specified’ have been replaced with a NaN value.
- The age column of the dataset has been normalized to range between 0 and 1.
- Also, while calculating co-relation coefficients, the rows with NaN values in that particular column have been ignored.

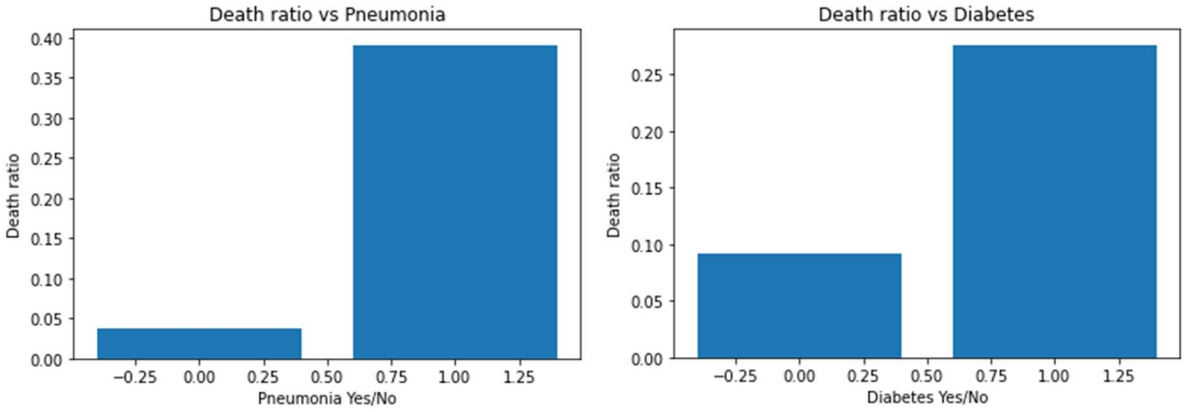
3. Methods

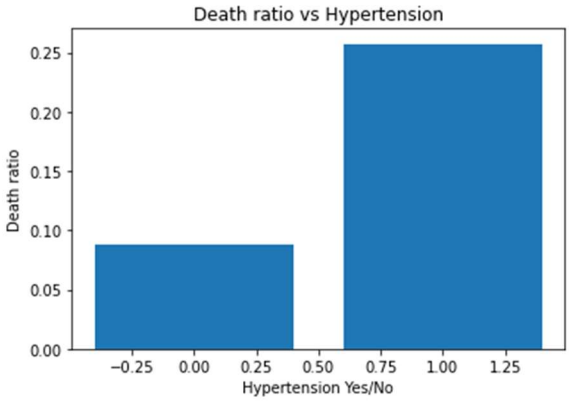
3.1 Calculating co-relation coefficients for different medical preconditions (and other factors) and the chances of death

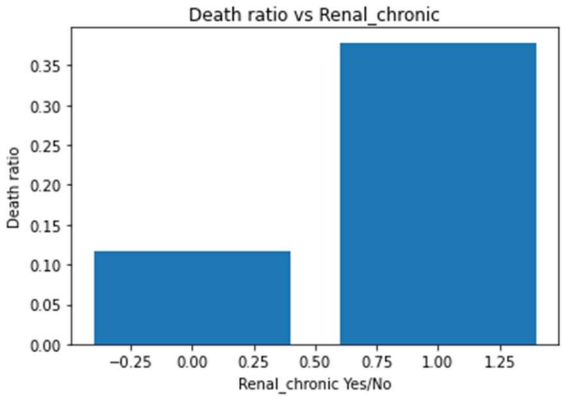
Medical precondition	Correlation coefficient with death
pneumonia	0.4591
age	0.3513
intubed	0.2187
diabetes	0.2081
hypertension	0.2063
renal_chronic	0.1165
icu	0.1003
copd	0.0904
sex	0.0844

As demonstrated by the co-relation coefficients, the patients who have pneumonia, are older, are men, or have other medical conditions like diabetes, hypertension and chronic renal disease seem to be more prone to dying.

The following bar graphs indicate these trends. Here the first bar shows the ratio of patients who died who explicitly stated that they did not have said pre-existing medical conditions. The second bar shows the ratio of dead patients who explicitly stated that they had the pre-existing medical condition. Patients who did not give information about that medical condition have been ignored in this graph.



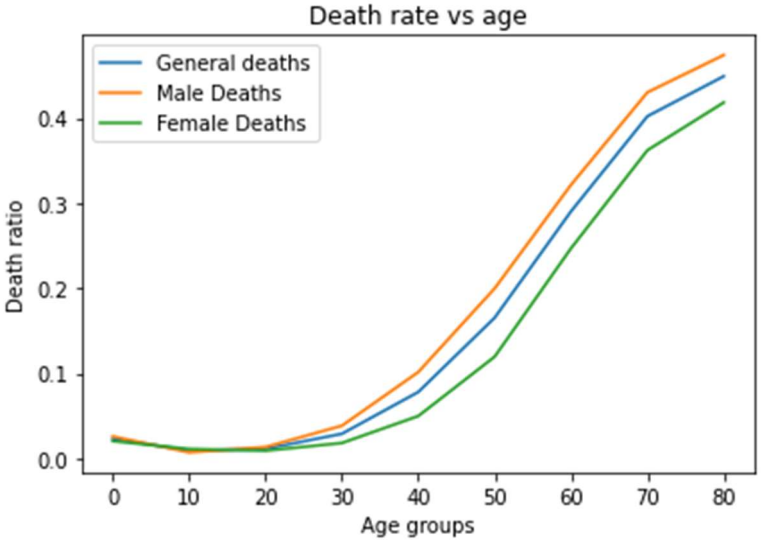




Here, patients who had to be put on ventilators or admitted to ICUs also show co-relation with dying but this is a trivial connection because obviously the patients who are in critical conditions are the ones who are put on ventilators in the first place.

Age is a very critical factor in determining the chances of death, older patients are more likely to die than younger patients. Sex is another important factor. Men die more frequently to covid than women do. Calculating the fraction of patients in each age group and gender who died, we get:

Age group	General Death ratio	Female Death Ratio	Male Death Ratio
0 to 10	0.0233	0.0205	0.0259
10 to 20	0.0092	0.0112	0.0072
20 to 30	0.0113	0.0092	0.0135
30 to 40	0.0290	0.0182	0.0385
40 to 50	0.0781	0.0499	0.1015
50 to 60	0.1654	0.1196	0.1999
60 to 70	0.2904	0.2470	0.3216
70 to 80	0.4021	0.3622	0.4302
80 to 90	0.4491	0.4183	0.4740



3.2 Calculating co-relation coefficients for multiple medical preconditions (and other factors) at a time and chance of death

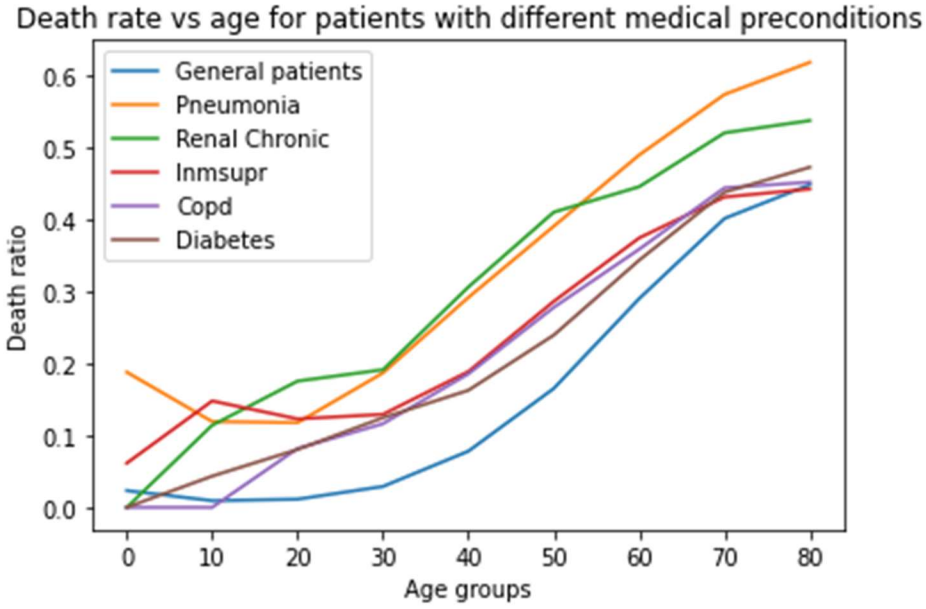
According to the CDC[3], patients who have multiple medical preconditions have a higher chance of facing covid severity. In this section, we shall try to corroborate this claim.

Here we have appended new columns to the dataset where the value in a particular column is the mean of the values in two other columns. Then correlation coefficient has been calculated w.r.t death and the new column.

While this is not the most accurate way of finding combinations of dangerous preconditions, it gives us a baseline. After finding the combination of medical preconditions which had the highest correlation with death, we calculated the ratio of the patients who had both medical preconditions and died.

Medical preconditions	Correlation coefficient with death	Death rate
Pneumonia & age	0.4980	(explored ahead)
Pneumonia & renal_chronic	0.4592	0.5533
Pneumonia & copd	0.4552	0.5280
Pneumonia & inmsupr	0.4496	0.4701
Pneumonia & cardiovascular	0.4473	0.4121
pneumonia & other_disease	0.4408	0.4773
Pneumonia & diabetes	0.4376	0.4689
Pneumonia & hypertension	0.4351	0.4790
Pneumonia & pregnancy	0.4304	0.1343
Pneumonia & asthma	0.4249	0.3564
Pneumonia & tobacco	0.3928	0.4052
Pneumonia & obesity	0.3580	0.4027
Pneumonia & sex	0.3476	0.4098
Age & renal_chronic	0.3107	-
Age & inmsupr	0.2931	-
Age & copd	0.2929	-
Age & diabetes	0.2875	-
Age & hypertension	0.2767	-
Age & cardiovascular	0.2717	-
Age & other_disease	0.2557	-
Age & pregnancy	0.2535	-
Diabetes & hypertension	0.2505	0.3187
Diabetes & renal_chronic	0.2246	0.4600
Pregnancy & hypertension	0.2225	0.0754
Hypertension & renal_chronic	0.2215	0.4322
Diabetes & copd	0.2201	0.3836

Age and other existing preconditions at the same time is also a deadly combination. For example, if we calculate the death rate of different age groups with different medical preconditions and covid, we get:



Age group with covid	General death ratio	Death ratio with pneumonia	Death ratio with chronic renal	Death ratio with immune suppressed	Death ratio with chronic obstructive pulmonary disease	Death ratio with diabetes
0 to 10	0.0233	0.1882	0.0000	0.0615	0.0000	0.0000
10 to 20	0.0092	0.1194	0.1142	0.1481	0.0000	0.0434
20 to 30	0.0113	0.1179	0.1757	0.1231	0.0819	0.0805
30 to 40	0.0290	0.1871	0.1916	0.1294	0.1164	0.1247
40 to 50	0.0781	0.2921	0.3066	0.1890	0.1857	0.1629
50 to 60	0.1654	0.3913	0.4109	0.2867	0.2783	0.2397
60 to 70	0.2904	0.4904	0.4463	0.3750	0.3593	0.3440
70 to 80	0.4021	0.5748	0.5213	0.4320	0.4448	0.4386
80 to 90	0.4491	0.6194	0.5384	0.4431	0.4527	0.4735

With some exceptions, these numbers are significantly higher than the numbers we observed in general for different age groups (without taking any medical conditions into account).

3.3 Training different classifiers for predicting whether the patient will die, be admitted to ICU, and/or be put on a ventilator

In this section, we shall attempt to train classifiers to predict whether the patient will die, need to be admitted to the ICU and/or need to be put on a ventilator. The input features to each classifier are – sex, pneumonia, pregnancy, diabetes, copd, asthma, inmsupr, hypertension, other_disease. cardiovascular, obesity, renal_chronic, tobacco, contact_other_covid, age.

For the input features, 0 for sex signifies female and 1 for sex signifies male. For age, the ages of patients have been normalized to between 0 and 1. For the remaining features, 0 means no and 1 means yes.

The classifiers will attempt to predict 3 values – 0 or 1 on whether the patient is probably going to die, whether patients will need to be admitted to the ICU, and whether patients will need ventilator support.

The dataset has been split in an 80-20 ratio for training and testing purposes. This means there are 453,281 samples in the training set and 11,3321 samples in the testing set.

We will consider the following classifiers here:

- Decision Tree
  - Here we are using entropy (information gain) as the criteria for the split. We aim to decrease entropy (maximizing information gain)
  - Max depth has not been set; tree stops splitting when information gain is less than some given threshold
- Random Forest
  - Same hyperparameters as decision tree
- K – nearest neighbor
  - Here we have taken k=3 because after some experimentation we found that it gives consistently good results for predicting ICU requirements, ventilator requirements as well as chances of death.
- Weighted K-nearest neighbor (weighted according to the distance between points)
  - Here again k=3
  - Here the points are weighted by the inverse of their distance from the input point i.e., more closer samples in the training data (to the input data) have a greater influence than the points further away.
- Naïve Bayes Classifier
  - This classifier is based on the Bayes theorem. It is called naïve because it 'naively' assumes that the input features are conditionally independent of each other. This is not always true in real life applications but as seen in our experiment it performs decently as well.
- Neural Network (with hidden layer sizes (5,1))
  - Activations: relu
  - Solver/Optimization algorithm: lbfgs
  - Learning rate: 0.00001
  - Maximum allowed iterations = 200
- Neural Network (with no hidden layer – just a single neuron)

Model	Death Accuracy	ICU Accuracy	Ventilator Accuracy
Decision Tree	0.9332	0.8478	0.8327
Random Forest	0.9362	0.8600	0.8415
K – nearest neighbor	0.9316	0.7919	0.8415
Weighted K-nearest neighbor	0.9350	0.8562	0.8377
Naïve Bayes Classifier	0.8812	0.8495	0.7944
Neural Network	0.9424	0.8818	0.8705

Single Node Neural Network	0.8748	0.8818	0.8587

The neural network outperforms all the other models. However, it is also the slowest to train.

The Naïve Bayes and KNN classifiers have essentially no training time.

The KNN classifier on the other hand is the slowest during runtime since all the computation is done at runtime.

3.4 The Accuracy-Recall Story

While most of our classifiers are able to achieve impressive performance in terms of accuracy, they are actually doing so while maintaining low levels of recall. In medical applications, we ideally want a higher level of recall while sacrificing some amount of precision, because in medical applications more false negatives can be dangerous.

For example, while predicting the requirement of ICU beds, predicting less than the required number of beds could be extremely harmful. However, predicting a little over the required number is not as harmful.

Therefore, we will attempt to tune the model to maximize recall while maintaining a decent level of accuracy.

For this example, consider the decision tree. Its accuracy is 93% but the recall is low here: only 21.4%. This means that only 21.4% of the patients who need ICU beds are classified correctly. We can tune this decision tree by giving weights to the classes. We give a weight of 0.98 to the 'need ICU' label and 0.02 to the 'don't need ICU' label. These weights are then used to calculate entropy during splitting and building the decision tree. Using this method, we get an accuracy of 88% (which is only 5% lower than our standard model) but an increased recall of 69.1% compared to the standard model's 21.4%.

These kinds of changes can also be applied to the other models. For example, in neural networks, we can use a weighted loss function where the model is penalized more heavily for a false negative and thus it tries to minimize the false negatives and the recall score improves.

4. Conclusion

- The worst preexisting medical conditions to have are – pneumonia, diabetes and hypertension.
- Surprisingly, pregnancy as a preexisting medical condition does not seem to have extreme adverse effects on death rates.
- Age and sex are also important factors. Older patients have higher death rates than younger patients. Male patients have higher death rates than female patients.
- Multiple medical preconditions at the same time can further exacerbate covid severity.

- Decision Trees, Random Forest, K – nearest neighbor, Weighted K-nearest neighbor, Naïve Bayes, and Neural Network Classifiers can be used for predicting death, ventilator requirements and ICU requirements based on age, sex, and medical preconditions.
- Recall scores can be improved for the classifiers by weighting the labels to minimize false negatives.
- A fully fledged neural network consistently outperforms other classifiers, although the difference in accuracy is not that large, and it is also the slowest classifier to train.

5. Future work

- The current study has been limited to data gathered from the country of Mexico only. It can be expanded to include other countries. This will help build classifiers that generalize better.
- The modification to the classifiers (for increasing recall) can be applied to not just the decision trees but other models like the neural networks as well.
- If these classifiers (like need for ventilator classifiers) are used in conjunction with other models (like image classification on lung ultrasounds to predict covid severity), the classification accuracy can be improved by a significant margin and robust models can be built.

6. References

1. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/underlyingconditions.html#:~:text=Certain%20underlying%20medical%20conditions%20increased,condition%20increased%20with%20age.
2. https://www.gob.mx/salud/documentos/datos-abiertos-152127
3. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html