Analyzing the Statistical Effectiveness of Diagnostic Tests for COVID-19 Using the ROC Curve

Student Summer Research

Authors: Mengle Hu, Simeng Li, Ruiyi Liu

Research Advisor: Dr. Paul Kvam

Department of Mathematics & Statistics
University of Richmond
05/2022 - 07/2022

1 Background

Diagnostic tests that use markers to determine whether a patient is diseased or healthy are standard tools in medical screening. For the diagnosis of many modern diseases, the difference in marker measurements used to screen healthy patients from diseased patients can be subtle, and statistical researchers work to develop the most effective tool to discern this difference. Misclassification costs are often asymmetric; that is, the cost of misclassifying a healthy patient into the diseased group (a false-positive result) is often less than the cost of misclassifying a diseased patient into the healthy group. We will focus on a tool that has been especially useful in recent decades called the receiver operating characteristic (ROC) curve.

2 Dataset Description

This section is separated into two subsections: introduction of data and the process of cleaning data.

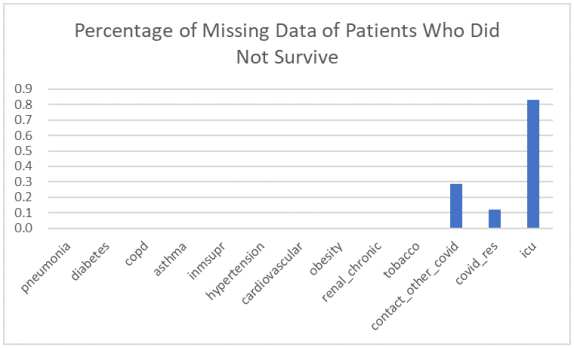
a. Introduction of Data

The data was released by the Mexican government. On group collected diagnostic test data worked to Covid 19 survived outcomes over the past two years, one goal is to learn how advanced statistical diagnostics can be used to predict Covid 19 survived and which risk factors are most useful in making those predictions. The dataset contains information on over 500,000 patients who were treated for Covid 19 infection. Along with survived status, these are 23 variables describing different patient risk factors, which are id, sex, patient type, entry date, date symptoms, date died, intubed, pneumonia, diabetes, copd, asthma, inmsupr, hypertension, cardiovascular, obesity, renal chronic, tobacco, contact other Covid, Covid res, and icu.

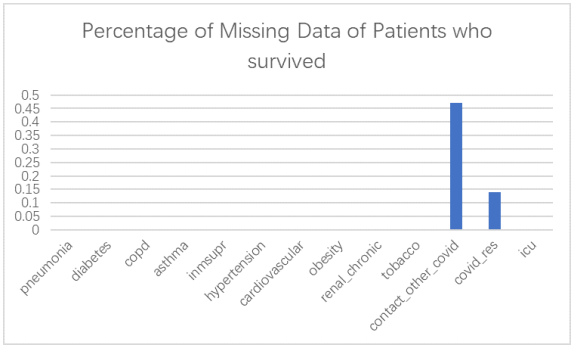
b. Process of Cleaning Data

Many observations were played by missing values. We first considered deleting this missing information. However, if we delete all missing values and unusable columns such as 'Pregnancy','other disease',and 'patient type', there are only around 40,000 patients left, and the logistic model we build present a 20 percent death rate, which is potentially biased and misleading.

Thus, we decided to see if there are some columns with an expected high number of missing values, and we considered deleting only these columns. By analyzing the data of survived and not survived patients, the columns of 'Contact Other Patients', 'Covid_res', and 'icu' have unexpected high number of missing values for both data.



(a) Percentage of missing data of Patients Who Did Not Survive



(b) Percentage of Missing Data of Patients Who Survived

We first deleted these three variables alone with other that would not be useful in this study, then we delete patients with missing information. The remaining set of data included 563286 patients with death rate of 6.4%.

Finally, by using the columns of 'date of entry' and 'date of sympton', we calculated the number of days before a patient went to the hospital for medical treatment. Then we created another column called 'number of risk factors', which calculated the sum of columns' values of each patients for further analysis that if there is a relationship between number of diseases and the death possibility of a patient. The range of this variable is from 0 to 11, and we suspect with higher number of risk factors, the possibility of death from the Covid is higher. Now we have a cleaned dataset that is prepared for data analysis.

In addition, there are two patient responses variables: intubation (patient had to be intubated to assist breathing) and death. Since majority of intubation data has missed (78 percent), we decided to focus only on patient survived or not as a diagnostic response.

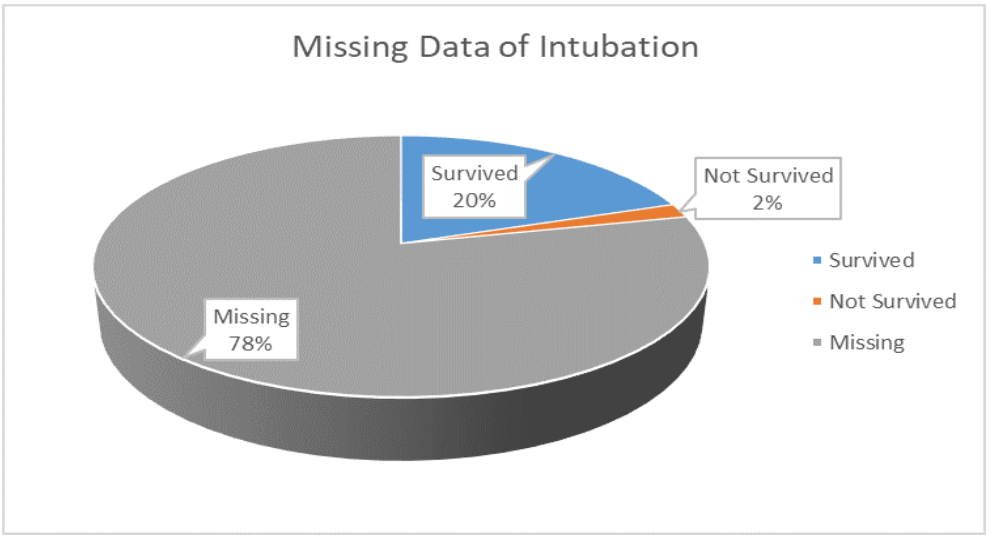


Figure 2: Percentage of Missing Data of Intubation

3 Logistic Regression

In this research, we learned about logistic regression and used it to make predictive analysis.

Logistic regression, a general linear model (GLM), is a way to predict a binary response with numerical predictors. It is used to describe data and explains the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. Usually, a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).

For binary response, we need to think of modeling not the zero/one outcome, but the "probability" of a success. In our research, "one" indicates the probability of death. In that way, the outcome is two stage: the explanatory variables help predict the probability of success p, and the response is thought of as the outcome of a coin-flip where the coin has probability p of flipping heads. So with every observation, we have recorded $(y_i, x_{1i}, x_{2i}, \cdots, x_{ki})$ but also the unrecorded $0 < p_i < 1$. So we model the unobserved p:

$$p_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon.$$

Because the relationship between p_i and the explanatory variables is only rarely a linear one, we need to consider a transformation of p the way we chose a transformation for the regression response y in a previous lab. This transformation is called the *link function*. Because $0 < p_i < 1$, we have special functions to draw out its range (so it is not compacted between zero and one), and the most popular transformation is called the logit transformation. We assigned this logit transformation to ρ:

$$\rho = logit(\hat{p}_i) = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon.$$

Therefore, given any input values $x_1, \cdots, x_k$, the model will generate a prediction $logit(\hat{p})$, which we can map back to get p. The predicted response p can be used to make a categorical prediction; for example, we might predict whether age, number of risk factors, and hospitalization days affect the probability that an admitted patient survives under COVID-19. Then, we took any two of the three predictor variables and group them in pairs for the logistic regression. Therefore, we fitted a logistic regression model using age and risk factors as the predictor variables and survival results(survive or not) as the response variable.

The following output shows the results of the logistic regression model:

```
Coefficients:
             Estimate Std. Error z value          Pr(>|z|)
(Intercept) -6.148415   0.022448    -274 <0.0000000000000002 ***
age          0.053665   0.000389     138 <0.0000000000000002 ***
nriskfactors 0.550312   0.004425     124 <0.0000000000000002 ***
```

Figure 3: logistic regression

This table means a 0.053665 increase in age or a 0.550312 increase in the number of risk factors will result in an increase in $logit(\hat{p}_i)$ if all other variables remain fixed. Now, if $\log(\frac{\hat{p}_i}{1-\hat{p}_i})$ increases by 0.053665, that means that $\frac{\hat{p}_i}{1-\hat{p}_i}$ will increase by $exp(0.053665) = 1.055$. This is a 5.5% increase in the odds of death (assuming that the variable number of risk factors remains fixed). Similarly, if $\log(\frac{\hat{p}_i}{1-\hat{p}_i})$ increases by 0.550312, that means that $\frac{\hat{p}_i}{1-\hat{p}_i}$ will increase by $exp(0.550312) = 1.734$. This is a 73.4% increase in the odds of death (assuming that the variable age remains fixed).

This table could help the logistic regression draws a (straight) line across the plot for classification, indicating observations on one side of the line would be classified as 0 (survive) and the other side as 1 (not survive). The line would be solved as:

$$\log(\frac{\hat{p}_i}{1-\hat{p}_i}) = -6.148415 + 0.053665x_a + 0.550312x_r.$$

After that, we can calculate the intercept and slope of the line, and thus got the following graph:

Figure 4: logistic regression as a classifier

In Figure 4, it shows that the left side of the line is predicted to be the patient who survives, while the right side is predicted to be the patient who does not survive.

Because of the density of observations, this figure fails to effectively show how the two groups differ in age and the number of risk factors. For this reason, we considered a contour plot. Contour mapping is a graphic technique that represents a three-dimensional surface in a two-dimensional format. Compared to the logistic regression plot, this contour plot more effectively shows that admitted patients in the bottom left corner have higher survival rates, while those in the upper right corner tend not to survive.

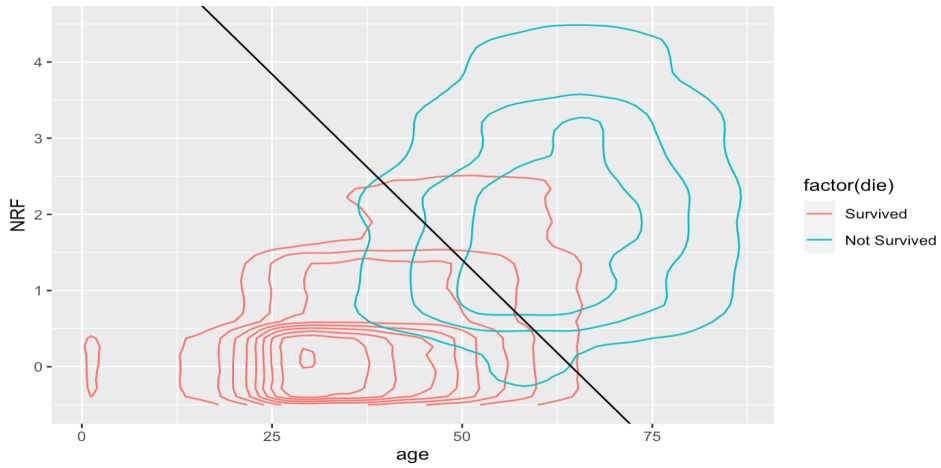


Figure 5: Contour Plot

Based on these two graphs, we can say both age and number of risk factors are positively correlated with death. However, logistic regression has a high rate of misclassification (proportion of people who are predicted to be alive but not survived) when used as a classifier. There are no obvious clusters of zeros or ones in the scatterplot. The logistic regression found the best line that separates the two groups, but misclassification is bound to happen. For example, the left side of the line should be all red, but there are several blue spots.

In this case, we are trying to figure out the error rate of the model. Specifically, if we use the threshold of $\hat{p} = 0.0633$ to predict whether an admitted patient would be survive based on their age and number of risk factors, we can test our classifier as a method of model checking.

```r
A01 <- covid$die
phat <- fitted.values(covid.logit2)
C0.1 <- A01[phat<0.5] # Proportion of people who are predicted to be alive but survived (misclassification)
C1.1 <- A01[phat>0.5] # Proportion of people who died in the whole population
c(mean(C0.1),mean(C1.1))
```

```
[1] 0.060022 0.363695
```

```r
A01 <- covid$die
phat <- fitted.values(covid.logit2)
C0.2 <- A01[phat<mean(covid$die)] # Proportion of people who are predicted to be alive but survived (misclassification)
C1.2 <- A01[phat>mean(covid$die)] # Proportion of people who died in the whole population
c(mean(C0.2),mean(C1.2))
```

```
[1] 0.017013 0.203147
```

Figure 6: Error Rate

As the graph shows, we contrast the actual death results with the model prediction inherent with $\hat{p}_1$, where 1 indicates the admitted patient not survived. We designed two cases to compare the error rates using death rate and with 0.5 to see how they differ. We found error rate is lower if we use death rate.

4 Receiver Operating Characteristic Curve (ROC curve)

In this research, we learned about the ROC curve and its applications in bio-statistics.

a. Introduction of ROC Curve and TPR/FPR

A receiver operating characteristic curve, or a ROC curve, is a graphical plot that shows the ability of diagnosis of a binary classifier system with a varied discrimination threshold. The ROC curve is made by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings.

The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test. For example, if we have a predictive model predicting that people at age 30 years old survived if they had covid, then the proportion of people of age 30 who did not survive will be the FPR.

b. ROC Curve using Age as Discrimination Threshold

For the first part of our ROC analysis, we use patients' age(c) as the discrimination threshold, and we divided our patients into two subgroups based on if they survived(Y = 0) or not(Y = 1). The TPR (the proportion of survived patients above this age) is $P(X \geq c|Y = 1)$ and the FPR (the proportion of patients who did not survive above this age) is $P(X \geq c|Y = 0)$. Then, we plotted the ROC using R:

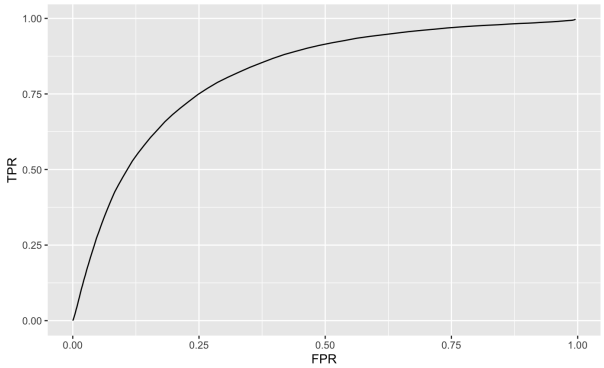


Figure 7: ROC Curve Using Age

An overall ROC curve is most useful in the early stages of evaluation of a new diagnostic test. Since the curve shows probabilities, we also calculate the Area Under the Curve (AUC) to see if the model performs ideally. The closer the AUC is to 1, the more accurate the model is. On the other hand, since the diagonal between points (0, 0) and (1, 1) shows the useless tests, if AUC is close to 0.5, the test is basically useless.

The AUC we calculated for this ROC curve is 0.81, which indicates a reasonable test.

The point on the curve that is farthest above the 45-degree line represents an optimal test, and is at around FPR=0.25, TPR=0.75, which corresponds with using Age=55 years old at the threshold value and might be the best threshold to choose for a decision rule.

Once the diagnostic ability of a test is established, only a portion of the ROC curve is usually of interest, for example, only regions with high specificity and not the average specificity over all sensitivity values. For our ROC curve, the most important part is at the upper left, where we have the high TPR (above 0.5) and low FPR (below 0.25). In this case, the partial ROC may give the best result and we found that the age range is from 51(highest TPR and lowest FPR) to 61 years old. However, the area of partial AUC is alomost 0 for TPR above 0.75 and FPR below 0.25.

c. ROC Curve using Predicted Values as Discrimination Threshold

For the next part of our ROC analysis, we did a logistic regression on three variables: age, number of risk factors, and days before patients go to hospital after they got covid-19. The regression will give us a series of values we will label $\hat{p}$ - the predicted values of each patient's possibility to be not survived - We extracted all the predicted phat values, and their values range from 0 to 1. Notice that the closer the phat values to 1, the less likely the patient will have survived. Then we use these values as the discrimination threshold for ROC analysis.

In this case, the TPR will be the proportion of survived patients above a certain predicted value, and FPR will be the proportion of patients who did not survive above a certain phat value.

The AUC we calculated for this ROC curve is 0.86, which indicates a reasonable test. This value is greater than the ROC using age as discrimination threshold, and it indicates an improvement from the previous one.

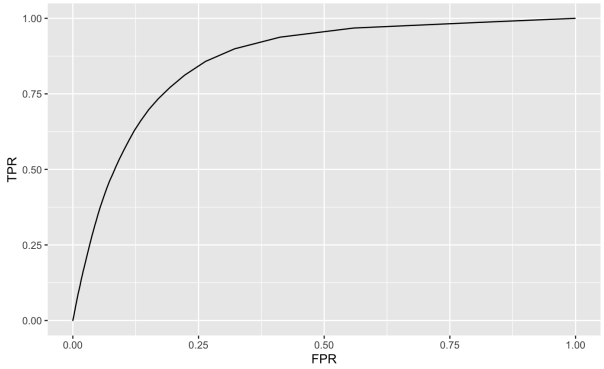


Figure 8: ROC Curve Using Predicted Values

Similarly, we found the point on the curve that is farthest above the 45-degree line is at around TPR=0.77, FPR=0.19, and $\hat{p}$=0.07, which might be the best threshold to choose for a decision rule.

For the Partial AUC, this ROC has a larger PAUC where TPR is above 0.5 and FPR is below 0.25 as we can just see from the plot. Also, this ROC has a part of curve where TPR is above 0.75 and FPR is below 0.25 which will be worth looking at later.

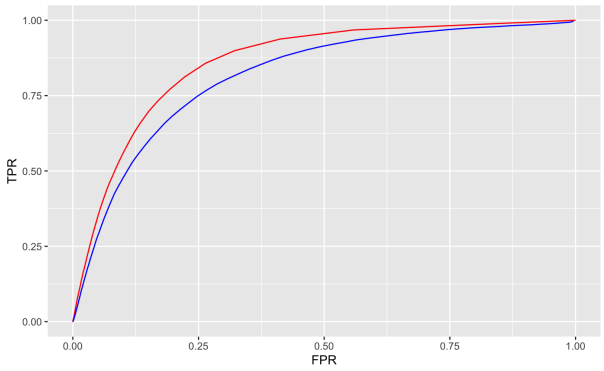


Figure 9: ROC Curves on the Same Plot

Figure 9 puts two ROC curves together on the same plot. The red one is the ROC curve using predicted values and the blue one is the ROC curve using age as discrimination threshold. We can clearly see that the AUC for the red ROC curve is larger and might be a better model.

d. Error Trade-off and ROC

In our research, we divided all patients into two subgroups for further analysis: Survived/Not Survived.

Figure 9 plots I,II, and III put two subgroups together, using age 55, 65, and 75 as the thresholds, where the left curve indicates survived patients and the right curve refers to the patients who did not survive. We can see a trend that patients survived are generally younger as expected.

The colored parts show an error trade-off between the two subgroups, where the orange part shows the patients who are predicted to be not survived but actually survived, while the green part shows those patients who are predicted to survive but actually not.

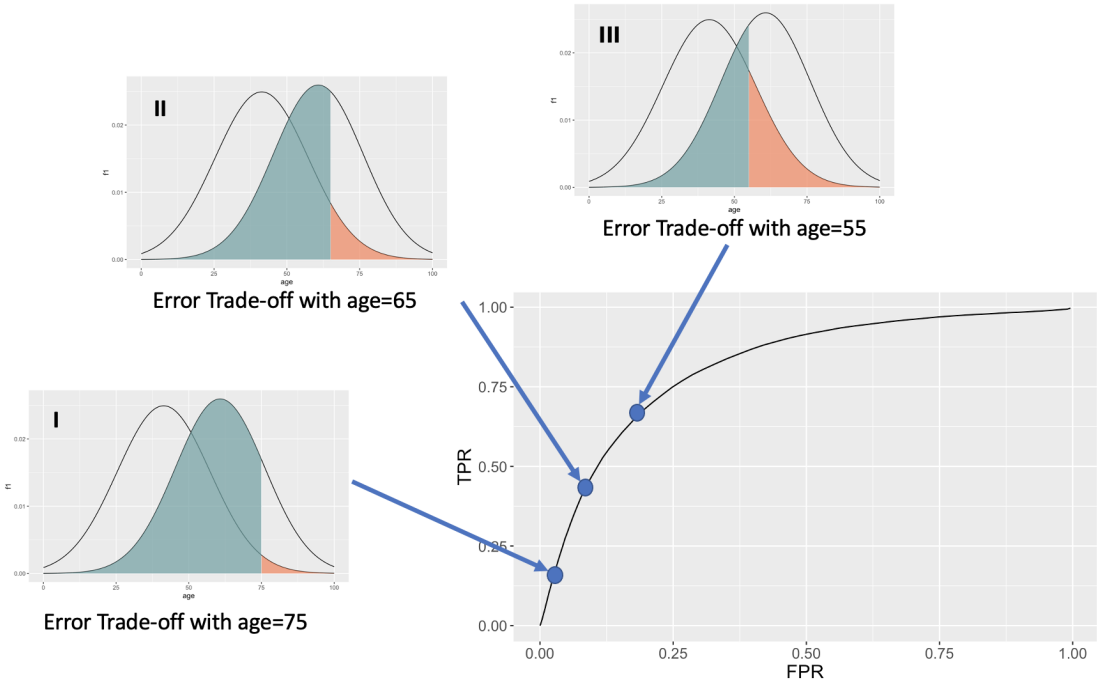


Figure 10: ROC Curve Using Age and Error Trade-off

Plot I: Error trade-off where Age=75, False Possibility Rate(orange)=0.028, False Negativity Rate(green)=1-TPR=0.831.

Plot II: Error trade-off where Age=65, False Possibility Rate(orange)=0.076, False Negativity Rate(green)=1-TPR=0.602.

Plot III: Error trade-off where Age=55, False Possibility Rate(orange)=0.182, False Negativity Rate(green)=1-TPR=0.342.

In plot I, for example, we see the orange part is larger than in plot II and plot III. This makes sense because the threshold we used in plot1 is a younger age such that more people older than 55 will be predicted as not survived but actually survived, and this increases the error rate of prediction(orange). Similarly, the area of the blue part increases as age increases, since more people will be predicted as survived but actually not, thus increasing the error rate.

To see the decision rule, we show the where exactly the error trade-offs are on a ROC curve. We put the error trade-off plots together with this ROC curve as Figure 10 shown above.

From Figure 10 we can see that the error trade-off with age=55 not only has the most even distribution of error trade-off, but it also falls on the point on ROC which is farthest above the 45-degree line. Thus, age=55 might be the best decision point for our model.