ABSTRACT

The COVID-19 pandemic has had a significant impact on global health and has placed enormous strain on healthcare systems worldwide. The ability to identify, isolate, and treat infected individuals is crucial for controlling the spread of the virus. Reverse transcription polymerase chain reaction (RT-PCR) tests remains the the gold standard for COVID-19 diagnosis. However, due to limited testing capacity, it is not always possible to test all admitted patients in hospital settings. Therefore, it would be helpful to find alternative methods for detecting the virus in patients who cannot be tested using RT-PCR. This study aims to identify laboratory tests that could be used to identify patients who are likely infected with COVID-19. We also aim to discover whether correlating demographic factors such as gender, race, and age have any statistical significance with infection. By analyzing laboratory tests and demographic information from 608 patients admitted to a US hospital, the study aimed to find correlations between specific laboratory tests and the likelihood of a patient having COVID-19. The results of this study could have significant implications for COVID-19 diagnosis in hospital settings, as it could help healthcare professionals to identify infected patients quickly and accurately, without having to rely solely on RT-PCR testing.

Our study utilized logistic regression to formalize the following results. We found that there exists significant statistical relationship between vaccinations and chance of being infected. Furthermore, there is evidence to support that Hemoglobin, Leukocytes, and Eosinophils are viable laboratory tests to predict the odds of being infected by Covid-19. More specifically, this study found that being vaccinated reduces the odds of being infected by 36.8%. And furthermore, every unit increase of Hemoglobin increases odds of increase by 100.4% whilst a decrease per unit of Leukocytes and Eosinophils decrease odds of being infected by 10.4% and 6.2% respectively.

STUDY DESIGN AND SETTING

This study will perform a logistic regression to deduce statistical significance between result of COVID-19 infection among admitted patients with vaccination, collected laboratory tests and patient demographics.

Assumptions Since we are using logistic regression, there are basic assumptions that must be met by the data. We will be using the canonical logit link which is depicted below:

$$logit(p) = \log \frac{p}{1-p}, \quad p = Prob(infection = positive) \tag{1}$$

And since the relationship between independent variables and the logit of the dependent should be linear, we get the following equation below for our relationship between response and predictor:

$$logit(p) = Intercept + Variable_1 * x \tag{2}$$

Another assumption we make is that the observations should be independent of each other and that the response variable only takes in binary values. In this case, the response is the `infection` variable which takes in only values of `negative` or `positive`.

Furthermore, independent variables should not be highly correlated and there shouldn't exist a dependent variable that perferctly calculates the result of the response, creating perfect separation.

THE DATA

The data is described by twenty-seven variables. Out of which, six pertains to patient demographics, vaccination, and infected status whilst the remaining twenty-five pertains to their laboratory result.

In total, there are 608 rows of observations with 151 rows of complete observations. Missing values only exist for laboratory test results. There exist 457 observations with at least one missing laboratory test results.

Below is a sample of the first five observations of the data that includes only the first two laboratory test results.

Table 1: First 5 observations of data

ID	infection	age	race	vaccinated	gender	Hematocrit	Hemoglobin
2	negative	43	white	yes	female	0.2250854	0.0061882
9	negative	48	black	no	male	-1.6013554	-0.7862806
16	negative	59	others	no	male	-0.6560539	-0.6506212
19	negative	35	black	yes	male	1.0249743	0.8866446
23	negative	71	white	no	female	0.1397312	-0.1014216

Note:
The above table only contains first two columns out of the twenty six lab test results

RESPONSE VARIABLE `infection` will be the response variable in our study. It denotes a binary result that confirms whether a particular patient tested positive or negative with COVID-19.

1. `infection`: positive or negative

PREDICTOR VARIABLES out the remaining 26 variables, this study will employ `age`, `race`, `vaccinated`, `gender` and all laboratory test variables as predictor variables. We exclude `ID` as it serves only as an identifier variable. Thus, in total, we will have 25 predictors for our response variable `infection`.

Demographic Variables Out the six demographic variables listed below, we will utilize `age`, `race`, `vaccinated`, and `gender` as predictors. Thus, we exclude `ID` from our model as it exist purely as a identifier variable, which does not serve to have any statistical relationship on the outcome of our response variable `infection`.

1. `ID`: patient ID
2. `age`: patient age in years
3. `race`: white, Hispanic, black, or others
4. `vaccinated`: yes or no
5. `gender`: male or female

Laboratory Tests Note all laboratory test results are standardized with values of normal mean of zero, and standard deviation of one.

The laboratory tests include:

`Hematocrit, Hemoglobin, Platelets, Mean platelet volume, Red blood Cells, Lymphocytes, Mean corpuscular hemoglobin concentration (MCHC), Leukocytes, Basophils, Mean corpuscular hemoglobin (MCH), Eosinophils, Mean corpuscular volume (MCV), Monocytes, Red blood cell distribution width (RDW), Serum Glucose, Neutrophils, Urea, Proteina C reativa, Creatinine, Potassium,` and `Sodium`.

Thus, in addition to our five predictor variables from above, an additional 20 laboratory test variables will be added to deduce our response variable.

EXPLORATORY ANALYSIS

In our findings, we found that there exists multiple

One of the key statistical relationship we wish to examine is whether or not vaccinations actually have significant impact on overall infection rate. The 2x2 contingency table below reveals counts of infection individual crossed with status of vaccination in our observed data.

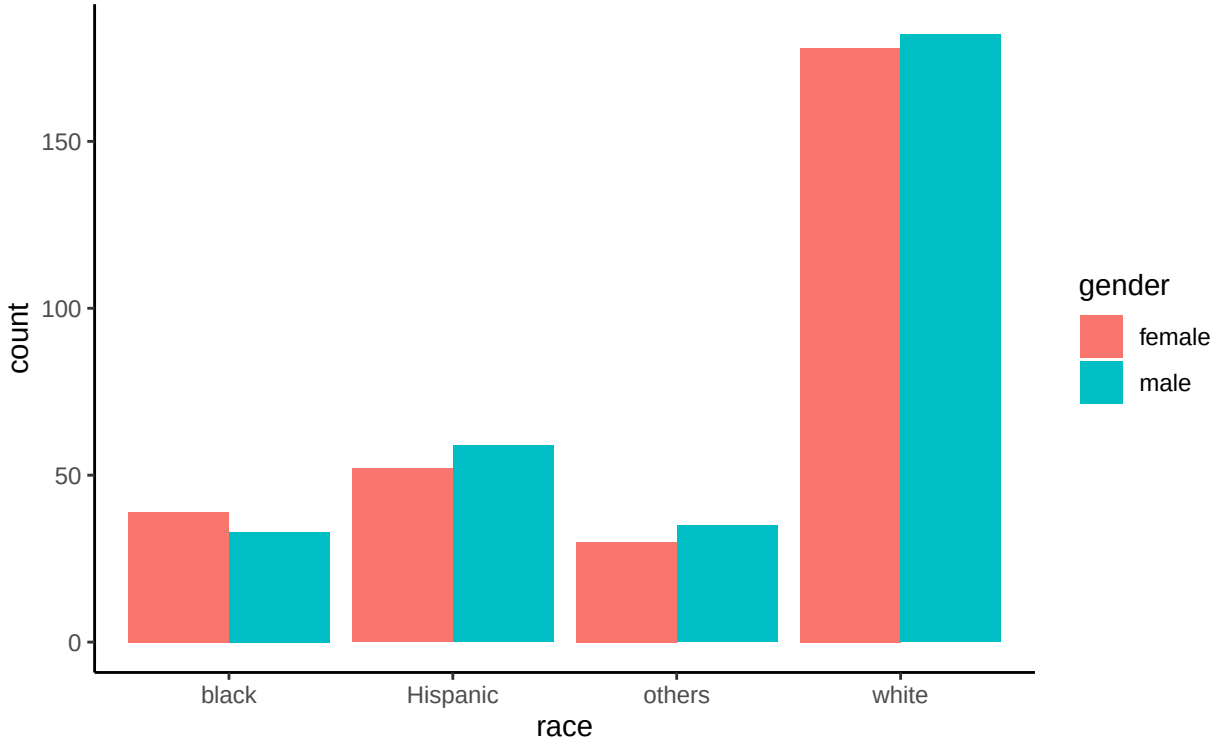
Table 2: 2x2 contingency table between infection and vaccination status

	no	yes
negative	250	274
positive	54	30

From the table above, it is clear that the data suffers from an imbalanced data set. There exists far greater amounts of non infected patients compared to infected patients.

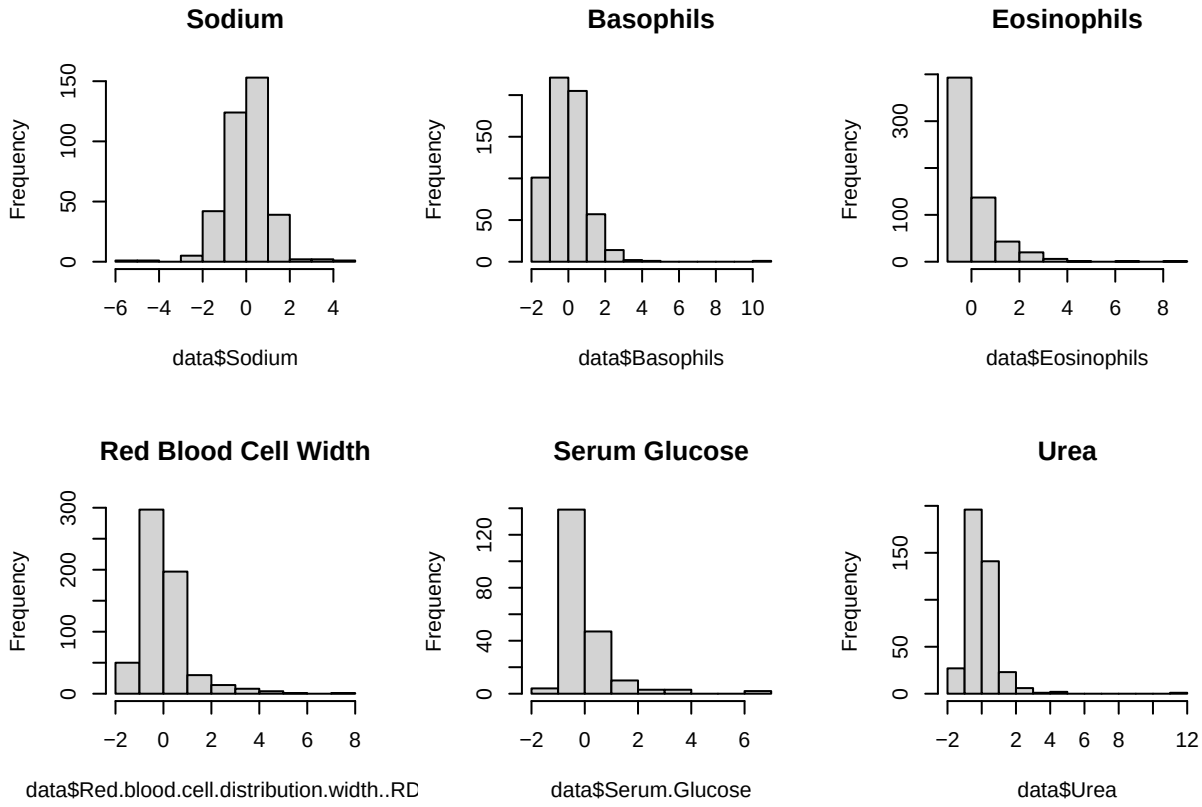
This is troubling as unbalanced data suggests that traditional measures such as accuracy are not viable as the model is prone to bias. Thus, later on, we will use Cohen's Kappa as a statistic to evaluate our model's actual performance in relation to the imbalance that exists in our dataset.

Figure 1: Bar graph depicting race and gender



Furthermore, it appears that there also exist a disproportionate amount of patients in terms of race. From the graph above, it is evident that white patients outnumber every other race 2:1. Thus, it is clearly evident that the data set is not balanced in terms of both the response and recorded predictor variables.

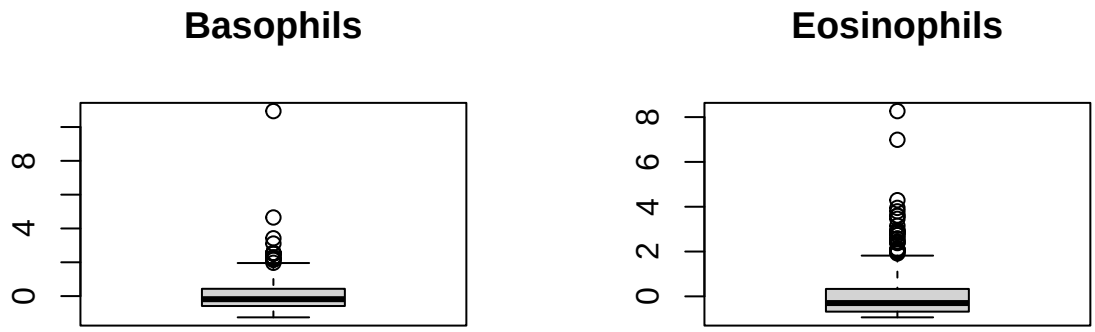
Figure 2: Histogram of skewed lab tests



Furthermore, it is revealed that several of the laboratory tests contained results that deviated far from the mean. Thus, this may be indicative of substantial amounts of outliers that may exist in our data.

Below are two boxplots that demonstrate this issue.

Figure 3: Boxplot of two laboratory tests with skewed distributions



Above are boxplots for Basophils and Eosinophils, two of which are both skewed distributions. We see that they both contain an abundant amount of outliers. Thus it is crucial that later fits of the model is checked against influential points to ensure that there is no violation of our basic assumptions for logistic regression.

MISSING DATA AND IMPUTATION

In total, there exists 457 missing observations, accounting for a total of 1553 missing observation points.

Below is a table that outlines the number of missing observation points that exist in each of the respective laboratory tests.

Table 3: Total missing values per each laboratory test

	Total Missing Values
Hematocrit	5
Hemoglobin	5
Platelets	6
Mean platelet volume	9
Red blood Cells	6
Lymphocytes	6
Mean corpuscular hemoglobin concentration (MCHC)	6
Leukocytes	6
Basophils	6
Mean corpuscular hemoglobin (MCH)	6
Eosinophils	6
Mean corpuscular volume (MCV)	6
Monocytes	7
Red blood cell distribution width (RDW)	6
Serum Glucose	400
Neutrophils	95
Urea	211
Proteina C reativa	102
Creatinine	184
Potassium	237
Sodium	238

Table 4: Frequency of missing observation per each observation

Missing Data	0	1	2	3	4	5	6	7	16	19	20
Frequency	151	133	76	41	24	129	47	1	1	1	4

As seen from `Table 3` above, the majority of missing observations do not share overlap on missing laboratory tests for a patient.

Due to heavy skew in many of the variables above, median values of each respective laboratory tests were imputed for missing values. Since the above tests differ biologically between males and females, the imputed values is further separated by the median between each group of male and females for laboratory tests.

MODEL SELECTION

In this report, we deduced three models that caters towards three scenarios. We considered the best models under each of the assumptions below:

1. `Scenario 1` - with `gender`, `race`, `vaccinated`, and `gender` as predictors
2. `Scenario 2` - with all predictors present and removal of missing observations
3. `Scenario 3` - with all predictors present and imputed data for missing observations

The best fitted models under each scenario above were fitted through a logistic regression model with canonical logit link using the the `GLM` function. The models were fit in full with under each condition and were trained on their respective appropriate data sets. Model selection was then performed to obtain the best fitted model using the `stepAIC` function from the **MASS** package. This selection process involved choosing the best model based on the Akaike Information Criterion, to find and measures significance of predictor variables in the best-fit model against the z-test statistic.

Model 1 Below is a summary output of `Model 1`, where *age* and *vaccinated* were identified to be significant predictors. Since there were no missing observations non laboratory variables, there were no missing observations in the data used to construct the model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.833	0.2054	-8.925	4.46e-19
gendermale	0.5708	0.2434	2.345	0.01903
vaccinatedyes	-0.7172	0.2457	-2.919	0.003507

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	488.4 on 607 degrees of freedom
Residual deviance:	474.7 on 605 degrees of freedom

As seen from above, *gender* and *vaccinated* were both identified to be significant predictors in relation to our response variable *infection*. Without the inclusion of laboratory tests, this is the model that best fit our model.

Recall that p equals to the probability that a patient is infected with Covid-19. Interpretation of the model above goes as follow:

$$logit(p) = -1.833 + Male(0.5708) - Vaccinated(0.7172) \tag{3}$$

Model 2 Below is a summary output of `Model 2`, where *vaccinated*, *Hemoglobin*, *Leukocytes*, and *Eosinophils* were identified to be significant predictors. This model was built on only complete observations in the data. Thus, this model was built on 151 observations out of the total 608 observations in the data.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.106	0.6965	-4.459	8.244e-06
vaccinatedyes	-1	0.5733	-1.745	0.08098
Hemoglobin	0.696	0.2887	2.411	0.01591
Leukocytes	-2.257	0.5583	-4.042	5.296e-05
Eosinophils	-2.773	0.8779	-3.158	0.001586

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	141.81 on 150 degrees of freedom
Residual deviance:	80.73 on 146 degrees of freedom

`Model 2` reveals that with more information, gender no longer remain as a significant predictor of our data. However, please note that this may be due to a loss of information because of a reduction in sample size.

Interpretation of the model above goes as follow:

$$\begin{aligned} logit(p) = &- 3.106 - Vaccinated + Hemoglobin(0.696) \\ &- Leukocytes(2.257) + Eosinophils(-2.773) \end{aligned} \tag{4}$$

Model 3 Below is a summary output of `Model 3`, where *vaccinated*, *Hemoglobin*, *Leukocytes*, and *Eosinophils* were identified to be significant predictors. This model was built on a data set with imputed median values for each laboratory tests. Thus, this model was obtained on 608 observations.

	Estimate	Std. Error	z value	Pr(>|z|)
intercept	-2.964	0.3357	-8.829	1.052e-18
vaccinatedyes	-0.7839	0.2985	-2.626	0.00863
gendermale	0.4682	0.299	1.566	0.1174
Hemoglobin	0.6202	0.1794	3.457	0.0005463
Platelets	-0.5638	0.2375	-2.373	0.01762
Lymphocytes	-0.6958	0.2992	-2.325	0.02006
MCHC*	-0.3665	0.1841	-1.991	0.04651
Leukocytes	-1.512	0.2987	-5.063	4.127e-07
Eosinophils	-1.345	0.3336	-4.031	5.555e-05
MCV*	-0.2578	0.1661	-1.552	0.1207
RDW*	-0.347	0.1953	-1.776	0.07572
Neutrophils	-0.8835	0.322	-2.744	0.006073
Proteina C Reativia	0.6296	0.1852	3.4	0.0006743

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	488.4 on 607 degrees of freedom
Residual deviance:	318.5 on 595 degrees of freedom

See footnotes 1,2,3 near the bottom of page for `MCHC`, `MCV`, and `RDW` laboratory tests [1] [2] [3]

From above summary results, we see that the best fit model for the imputed data suggests that `vaccinated`, `gender`, `Hemoglobin`, `Platelets`, `Lymphocytes`, `MCV`, `Leukocytes`, `Eosinophils`, `MCV`, `RDW`, `Neutrophils`, and `Proteina C Reativia` are all significant predictors in determining Covid-19 `infection`.

$$\begin{aligned} logit(p) = &- 2.964 - Vaccinated(0.7839) + Hemoglobin(0.6202) - Platelets(0.5638) - Lymphocytes(0.6958) \\ &- MCHC(0.3665) - Leukocytes(1.512) - Eosinophils(-1.345) - MCV(0.2578) \\ &- RDW(0.347) - Neutrophils(0.8835) + Proteina(0.6296) \end{aligned} \tag{5}$$

Cross Validation Results of Models Using the `train` function from the **caret** package, we are able to deduce the effectiveness of each model above. More specifically, the `train` function provides us with validation statistics of accuracy and Cohen's Kappa score of cross validation repeated 10 times using 10 folds.

[1] MCHC: Mean corpuscular hemoglobin concentration
[2] MCV: Mean corpuscular volume
[3] RDW: Red blood cell distribution width

As explored from our explanatory analysis, the response variable `infection` is heavily imbalanced. Thus the Cohen's Kappa is a valuable score that relays how much better the classifier does compared to just random guessing.

Below is the formula for Cohen's Kappa:

$$\begin{aligned} \kappa &= \frac{p_0 - p_e}{1 - p_e} \\ p_0 &= observed \quad agreement \\ p_e &= expected \quad agreement \end{aligned} \tag{6}$$

Cohen's Kappa always takes on values from 0 to 1 with a lower Cohen's Kappa relaying that the model is no different from guessing where as a higher score indicates that the model is much more effective than just random chance.

The training and testing data set is split 80:20 proportionally. Thus, in `model 1` and `model 3`, the resulting split hovers around 120 out of 608 observations for the testing data set. Whilst for `model 2`, the resulting split hovers around 30 out of 151 observations for the testing data set.

Table 11: Output of Model CV results

parameter	Accuracy	Kappa	AccuracySD	KappaSD
Model 1	0.8213890	0.0000000	0.0133369	0.0000000
Model 2	0.8636908	0.4926561	0.0511893	0.1969235
Model 3	0.8796217	0.3740538	0.0182900	0.0931294

Note:
Trained Cross Validation - 5 folds, repeated 10 times

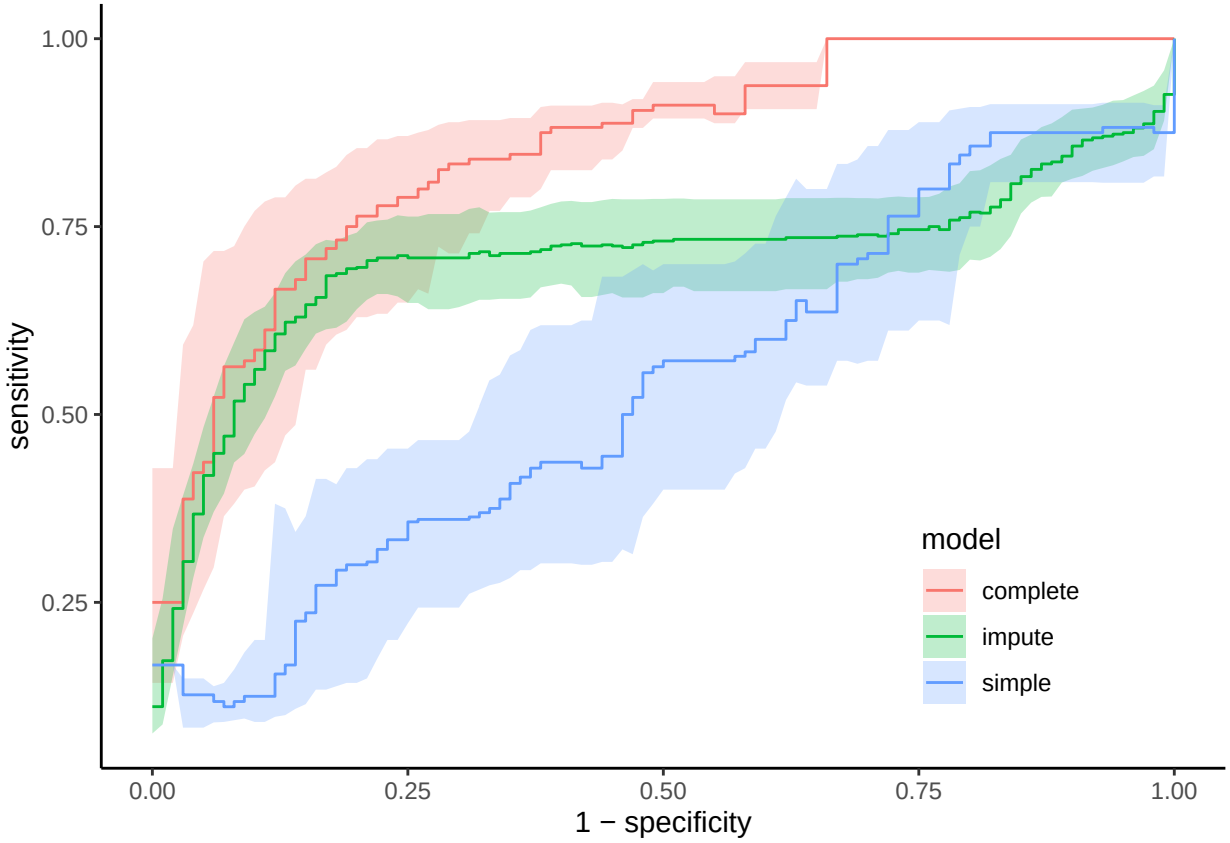
From the results above, we see that `Model 1` has virtually no difference compared to guessing whereas `Model 2` does an adequate job compared to just random guess. Thus, even though the accuracy is high for `Model 1`, it is invalid as it most likely just the consequence of our imbalanced response variable rather than actually being a strong predictive model.

`Model 3` does not perform as well as `Model 2` but still does much better than `Model 1`. It's Cohen Kappa score also suggests that its predictions are better than just guessing off random chance.

Cross Validation AUC graphs for each model Another way to validate the performance of the model above is to construct AUC graphs. Below is a graph that depicts pooled AUC graphs obtained from random sampling of 70% training and 30% testing dataset for all three models. `Model 1` was tested on the original dataset, `Model 2` was tested on the completed observations only dataset and `Model 3` was tested on the imputed dataset.

The confidence band is obtained through calculating the quantile for specificity results with probability of prediction set at 0.25 and 0.75 respectively. Each model was ran 100 times.

Figure 4: Pooled Cross Validated AUC Curves



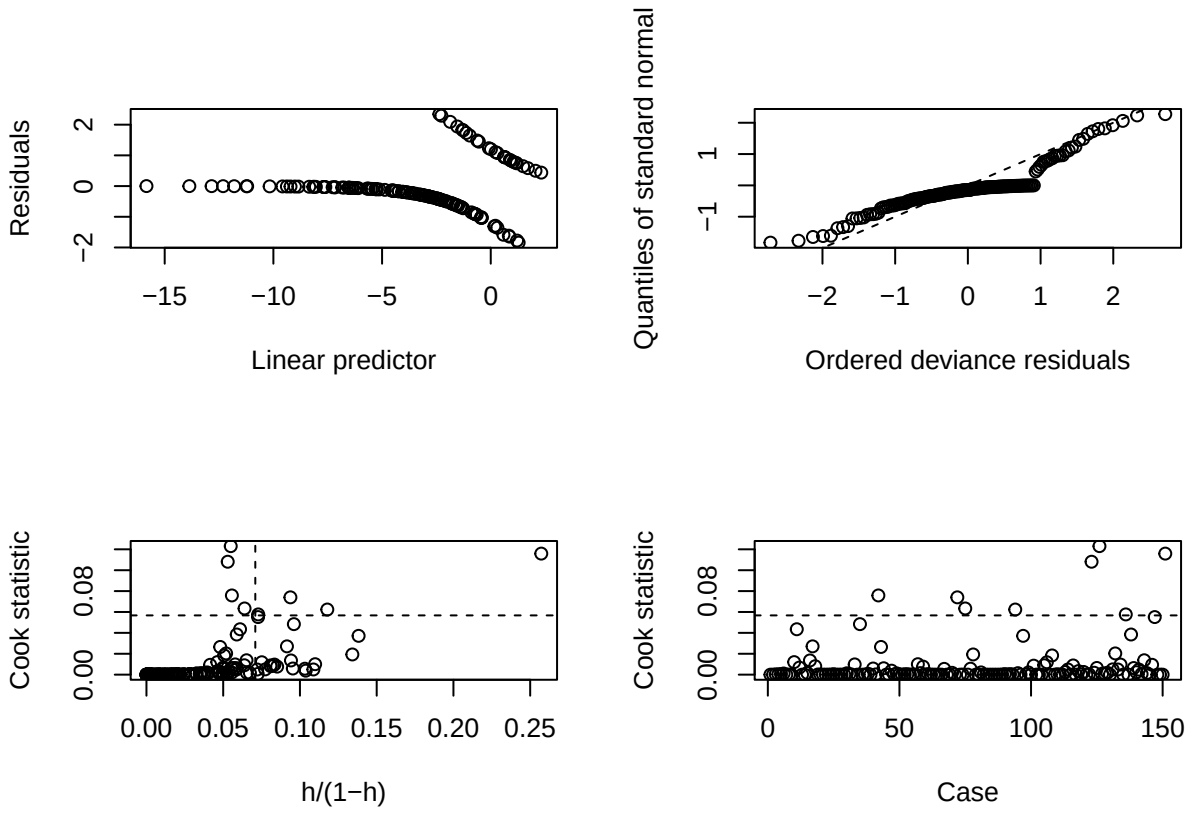
As seen from above results above, `Model 2` outperforms `Model 1` and `Model 3` by a large margin.

Thus, this AUC graph supports the result of our model summary and cross validation results above.

Influential Data Points In the assumptions above, it is stated that logistic regression is more sensitive to outliers as compared to that of other regression models. Thus, it is important to check if influential data points exist that may be skewing the results of our model or making our model obsolete.

Using `glm.diag.plots` from the **boot** package, we construct the following plots that reveal possible influential points that affects our model:

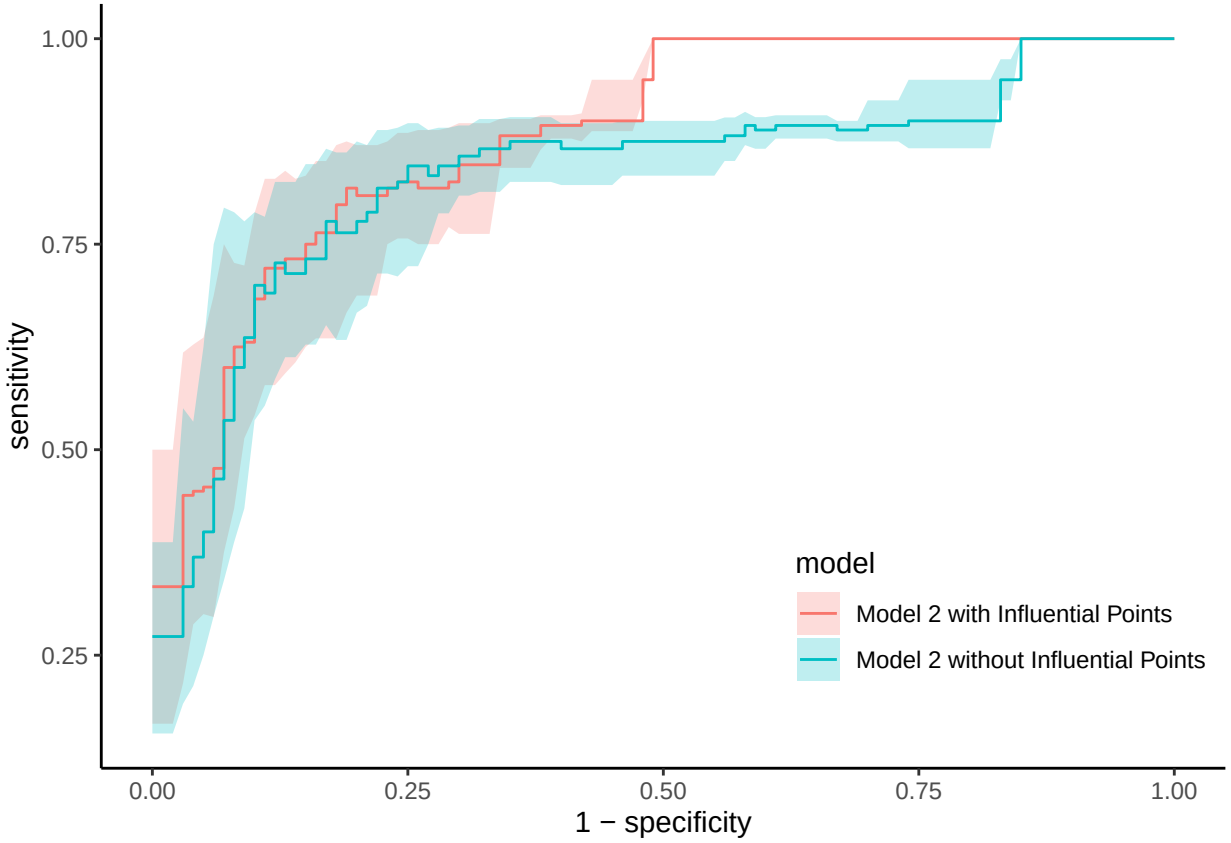
Figure 5: GLM diagonstics plots for Model 2



From the above graphs, we see that influential points exist beyond 0.06. To test if removing influential points greatly affect the accuracy of the model, we remove data points that are considered influential greater than 0.7. This new model will follow the exact same significant predictors as `Model 2`.

Below is an pooled cross validate AUC graph using the same procedure as above between the results of `Model 2` and `Model 2` with influential points removed.

Figure 6: Pooled Cross Validated AUC Curves for influential points



Limitations of the Data One of the issues with the current data lies in the amount of missing observations that is present in laboratory tests. This is especially prevalent in the laboratory test pertaining to `Serum Glucose`. With over 400 data points missing, it wipes out over two-thirds of available data needed for a complete and robust analysis.

The trouble with the final deduced model is that although it performs well on the current data, the loss of over 50% of data greatly diminishes statistical power and precision. Furthermore, this also makes it difficult to replicate findings.

Another one of the issues that currently plagues the current is the imbalanced data that is our response variable. With only 84 of our observations being positive cases of Covid-19, it creates a heavy bias for the model to predict negative results to achieve high accuracy. Furthermore, the data does a poor job of representing actual statistical relationship between patients who tested positive and the predictor variables as there just does not exist a substantial amount of information to make the relationship concrete and robust.

Essentially, the amount of missing data as well as poor balance of data creates a insufficient capture of information to properly represent the outcome of the binary class.

Conclusions From the data above, there is not enough strong evidence to support the claim that Covid-19 infection is related to Race, Gender, or Age. However, there is strong statistical evidence to suggest that vaccinations reduce log odds of being infected by 0.368. In other words, reduce the odds of being infected by 36.8%. Furthermore, Per every unit increase of Hemoglobin, the odds of infection increase by 100.4%. Furthermore, a decrease per unit of Leukocytes and Eosinophils correlates to a decrease in odds of being infected by 10.4% and 6.2% respectively.

To improve future research in this field, efforts should be made to obtain a more balanced dataset, ensuring

a sufficient representation of the minority class. This could involve targeted sampling techniques or data augmentation methods.

Furthermore, exploring alternative modeling techniques, such as ensemble methods or advanced resampling strategies, could potentially enhance the performance of the model and address the challenges posed by imbalanced data.

Despite the limitations, our findings provide insight into how Covid-19 infection rates might relate to a patient's demographic and laboratory tests, paving the way for further research and practical applications.