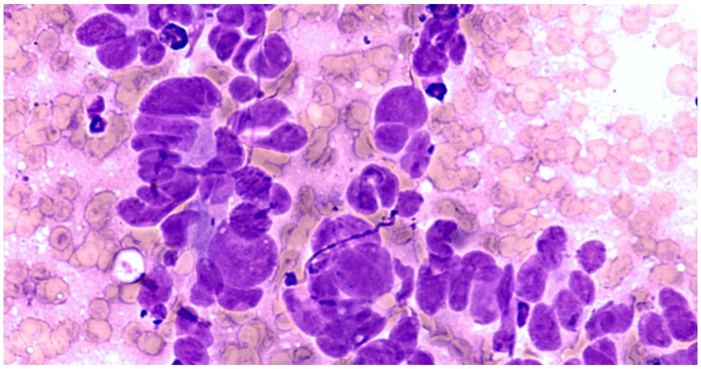
# A Non-Invasive Method of Detecting Breast Cancer: Determining its Effectiveness

Iffah Batrisyia Asmawi, Student ID: 31209289



## 1. Abstract

Tumours are big indicators of cancer, and checking to see if they present as malignant often needs invasive surgical procedure. Fine needle aspiration is a new, non-invasive method of diagnosing tumour status. In a data set from the R package, 'Faraway', we have ratings of abnormality of the tissue samples extracted. Using logistic regression models, we want to fit the data set and determine its effectiveness in diagnosis. We aim to study whether this newfound process is effective and if they can be used in medical practices.

Not only that, we also discuss which characteristics of a cell are important indicators that would lead us to detecting cancer within the breast tissue. Not all cell features are significant for diagnosis. Once we determine the variables to include in our model, the final model will be tested to see how accurate it will be in allocating the samples into the right class.

## 2. Dataset

### 2.1 Biostatistics

By definition, biostatistics is a branch of statistics that works with biomedical data or data from living organisms so that we are able to find out any patterns or processes (Berger and Matthews 2006). This branch deals heavily with research topics, all in relation to biology, medicine and public health by using statistical methods that can prove if new findings are true. For example, we can prove if a new type of medicine is effective against diseases, or survival rates.

The discipline of biostatistics provides tools and techniques for collecting data and then summarising, analysing, and interpreting it. We can analyse samples taken from random groups of people in order to make inferences about the whole population. Rigorous analysis allows us to see whether the correlations within the dataset are causal or coincidental.

In this project, in which we try to find out whether fine needle aspiration is an effective tool for breast cancer diagnosis, it is categorised as biostatistics.

### 2.2 Breast Cancer and Diagnosis

Breast cancer is said to be the world's most prevalent cancer, with 7.8 million women alive who were diagnosed with breast cancer in 5 years by the end of 2020 (Breast cancer n.d.). Survival improvements started in the 1980s in countries with early detection programmes and improvements in healthcare, particularly more research leading to more treatments.

Common presentations of breast cancer is the formation of lumps in the breast tissues or thickening of the tissues. Not all tumours formed in the breast are cancerous, however it is still important for them to be checked. Early detection and treatment of malignant tumours has proven successful in lowering the mortality rate of people diagnosed.

Diagnosis of breast cancer has multiple stages, one of which is extraction of breast tissues, also known as a biopsy. However, biopsy procedures are often difficult, involving invasive surgical methods. It is also costly and ends with too many complications for the patient, namely post-surgery care, even if open-surgery biopsies are highly accurate (Bruening et al. 2009). Fine needle aspiration is a new, non-invasive method of extracting small samples of breast tissue and studying them in order to determine tumour status.

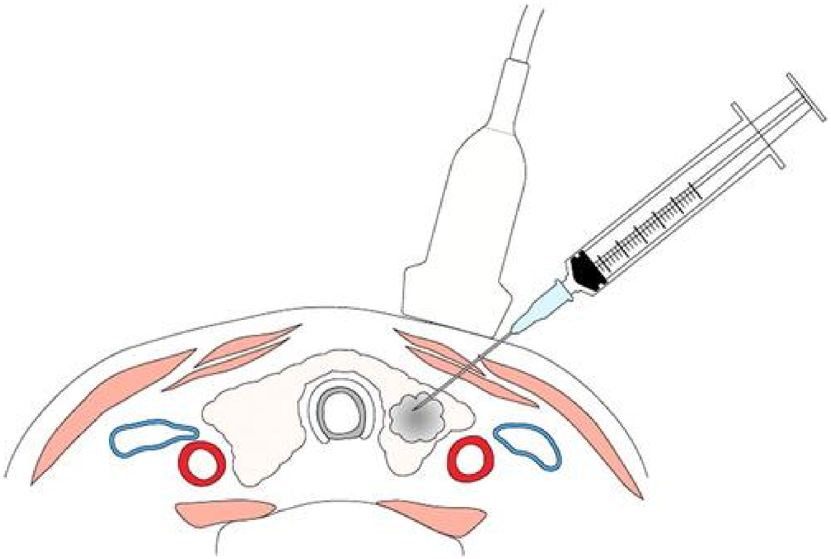


*Fig 1. Biopsy Collection by Needle Aspiration. Source: https://pubs.rsna.org/doi/10.1148/rg.287085033*

With the new method, it would be much easier for people to get tested for breast cancer.

In order for us to prove that fine needle aspiration is a good method of biopsy, we will want to observe the cells within the extracted tissue and figure out if the sample extracted is a good sample to predict tumour status.

### 2.2.1 Determining Malignancy

Tumours are made up of abnormal or damaged cells that do not break down normally and continually multiply. In simple terms, tumours that appear in tissues are either malignant or benign.

Benign tumours are normally of no real concern, but may cause disruptions in body functions such as becoming large and compressing other structures in the site (Patel 2020). They are removable by surgery and do not often grow back.

Malignant tumours, or synonymously known as cancerous tumours, are also lumps of cells that are unable to break down normally. The key difference from benign tumours is that they are able to invade and spread to other parts of the body and form more tumours. This process is known as metastasis. This kind of tumours that spread can cause severe damage to how the body functions, and most people who die of cancer, die of metastatic disease (What Is Cancer? - National Cancer Institute n.d.).

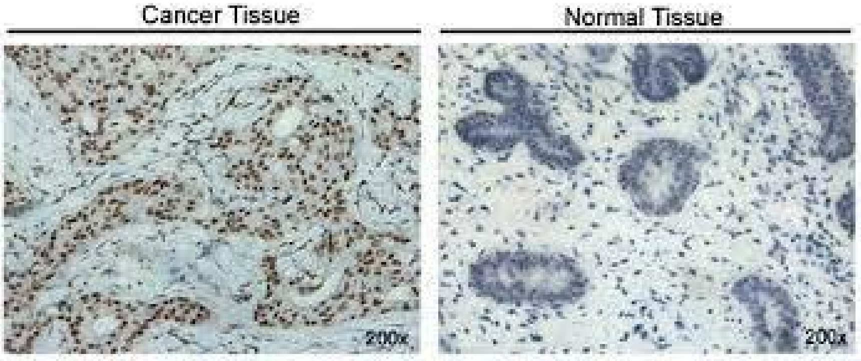


*Fig 2. Comparison of breast tissue with and without cancer. Source: https://www.spandidos-publications.com/10.3892/or.2013.2798*

Although the two classes of tumours share many traits, the appearance of the cells differ. Cell structures between the two types are different, such as variation in shape and size of the cell and nucleus (Kumar, Srivastava, and Srivastava 2015). However, we are not overly concerned about the detailed reports on how to tell apart the tumour cells. The data set that we will be working with already addresses these differences, which allows us to do analysis.

## 2.3 Understanding the Data Set

In the study, we observe the cells of the breast tissue and observe the appearance of the cells. The data set comes from a study of breast cancer in Wisconsin. There are 681 samples and observations of potentially malignant tumours, where 238 of them are actually cancerous. There are 10 variables within the data set.

Here, there are 9 characteristics of a cell that we are observing. Namely, the marginal adhesion, the bare nuclei, bland chromatin, epithelial cell size, mitoses, normal nucleoli, clump thickness, cell shape uniformity and cell size uniformity. A doctor was called in to review 681 samples and provided observations for each sample and the characteristics of interest.

In our data set, each characteristic is given predictor values of each cell feature and are rated from a scale of 1 (normal) to 10 (abnormal). The data set also includes the classification of the tumour (0 if malignant, 1 if benign).

In this report, we want to find the relationship between the tumour status, and the 9 characteristics. We will want to determine which predictors are important in identifying malignant cells. We are investigating which features of a cell, with a higher rank of abnormality, would lead us to identify cancerous cells.

We will attempt to apply a logistic regression model onto the data. Using methods like AIC and BIC to determine the right parameters. After determining the appropriate models, we want to then observe the specificity and sensitivity of the models.

## 3. Model Fitting

### 3.1 Logistic Regression Model

Logistic regression is a type of regression analysis. It is a statistical analysis method that tries to find the relationship between dependent and independent variables, and uses collected data and observations to make predictions on future data. The more data that comes in, the more accurate our predictions become. Logistic regression works similarly to linear regression, but with a binomial response variable (Sperandei 2014).

This type of regression can also help us find probabilities and odds of an event occurring. Logistic regression can link multiple predictors that may influence the odds of an event and the final outcome (Tolles and Meurer, 2016).

It can be used to predict a binary outcome of the response (dependent variable) from the predictors (independent variables). Typically, response would place a single data to be put into one of two categories. For example, ‘yes’ or ‘no’.

As for the predictors, there are 2 main categories that logistic regression model can utilise:

- Continuous:  It is data that is categorised as either interval data, or ratio data.
- Discrete: Ordinal discrete data is data which can be placed on a scale. Nominal discrete data is data which fits into named groups.

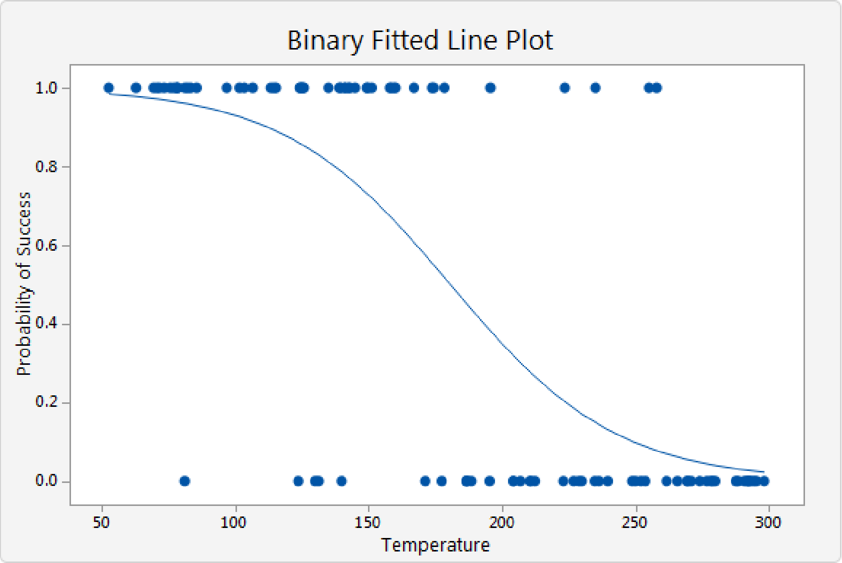


*Fig 3. Simple logistic regression graph. Source: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/simple-binary-logistic-regression/interpret-the-results/all-statistics-and-graphs/graphs/*

Figure 3 shows what a simple logistic regression graph would look like, with one variable. It can show the spread of the data and the frequency of the independent variable being classified at each class depending on their value. Typically, the steeper the curve, the more receptive the model would be in classifying the data points.

### 3.1.1 How it is beneficial for this data set

With our data set, using the logistic regression model is appropriate. We are investigating the relationship of the response (abnormality ranking of cell characteristics) and the predictors (malignant or benign). The predictors are categorical, which is best suited with using logistic regression.

There are also many predictors within our data, but conveniently, using the logistic model makes it easier to handle more than two explanatory variables simultaneously. We also know that it works for both discrete and continuous predictors. As we will discuss in section 3.2.1, our data set explanatory variables are continuous.

## 3.2 Model Fitting

Before getting into the model, we want to define what it is and how we interpret the values. In logistic regression, we use what is also known as the logit function. We define the logit function as the natural log of the odds (Belyadi and Haghighat 2021). The odds are defined as the probability of the event occurring divided by the probability of the event not occurring. It is important to note that odds and probability are not the same.

$\rho$ indicates the probability of an event.

$$Odds = (\frac{\rho}{1-\rho}) \Leftrightarrow \rho = \frac{Odds}{1 + Odds}$$

$$logit(\rho) = log(\frac{\rho}{1-\rho})$$

In logistic regression, we mostly calculate the odds. Unlike probability, where the value of it lies between 0 and 1, the range of odds is between 0 to positive infinity.

$$log(\frac{\rho}{1-\rho}) = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \cdots + \beta_m\chi_m$$

$\beta_i$ are the regression coefficients with $i = 0 \cdots m$ associated with the reference group, and the $\chi_i$ explanatory variables (Sperandei 2014).

In our model, $m = 9$ and will end up looking as below:

$$log(\frac{\rho}{1-\rho}) = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3 + \beta_4\chi_4 + \beta_5\chi_5 + \beta_6\chi_6 + \beta_7\chi_7 + \beta_8\chi_8 + \beta_9\chi_9$$

The variables are in this order: *Adhes, BNucl, Chrom, Epith, Mitos, NNucl, Thick, UShape* and *USize*.

$\beta_0$ represents the intercept term. We also call it the reference group, where it groups all the samples with reference levels of each variable $\chi_i$ . The exponential of the term will give us the mean odds of tumours being at class 0. As defined earlier in section 2.3, class 0 is if the tumours are malignant. The $\beta_0$ term is used as a baseline, so when we work out $exp(\beta_i)$, for the other coefficients, it will give us a value that is read as: how much bigger the chance of the tumour being benign times the value of $exp(\beta_0)$ (Sperandei 2014).

That is how we can interpret the coefficients of the model. So now the question is, how can the model tell us if the outcome is one or the other?

We will know by looking at the logit value. In the most basic setting, if logit value $< 0$, then the outcome is that the particular data falls into class 0, and if logit value $\geq 0$, then the data falls into class 1. Within the context of our dataset, if the logit is < 0, then tumour is classed as malignant and vice versa for benign classification. In section 4.4.1, we will discuss this implication in more detail.

### 3.2.1 Predictors

At face value, we see that each predictor in our dataset has 10 levels of abnormality, which looks like discrete data. However, in a regular logistic regression model fit, the independent variables have two levels (for example, ‘presence’ and ‘non-presence’). In other words, it is often binomial.

In the case we continue to view the predictors as discrete, then we can describe the predictors as being multinomial and there will be $n - 1$ binary variables for each predictor. An example of this is from (Sperandei 2014), with satisfaction levels.

Satisfaction has three levels: low, medium and high. Then, we will obtain this logistic regression model

$$log(\frac{\rho}{1-\rho}) = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2$$

There are 2 binary variables $\chi_1$ and $\chi_2$. There are 3 outcomes of this logit model when $\chi_1$ and $\chi_2$ are constrained to 0 and 1 depending on the satisfaction level (note that they cannot both have the value of 1).

Keep in mind that this is only for a single predictor with 3 levels. In comparison, our model has 9 predictors with 10 levels each. Evidently, this would lead to our model looking very long and complex, but the model would be less ambiguous as our predictors continue to be categorical.

Another way to do analysis on this data set is by treating the predictors as continuous data, but it wouldn't be as precise, but it would lead to a more parsimonious model. In this report, I will be using the latter method, by treating predictors as continuous data.

### 3.2.2 Fitting Model in R

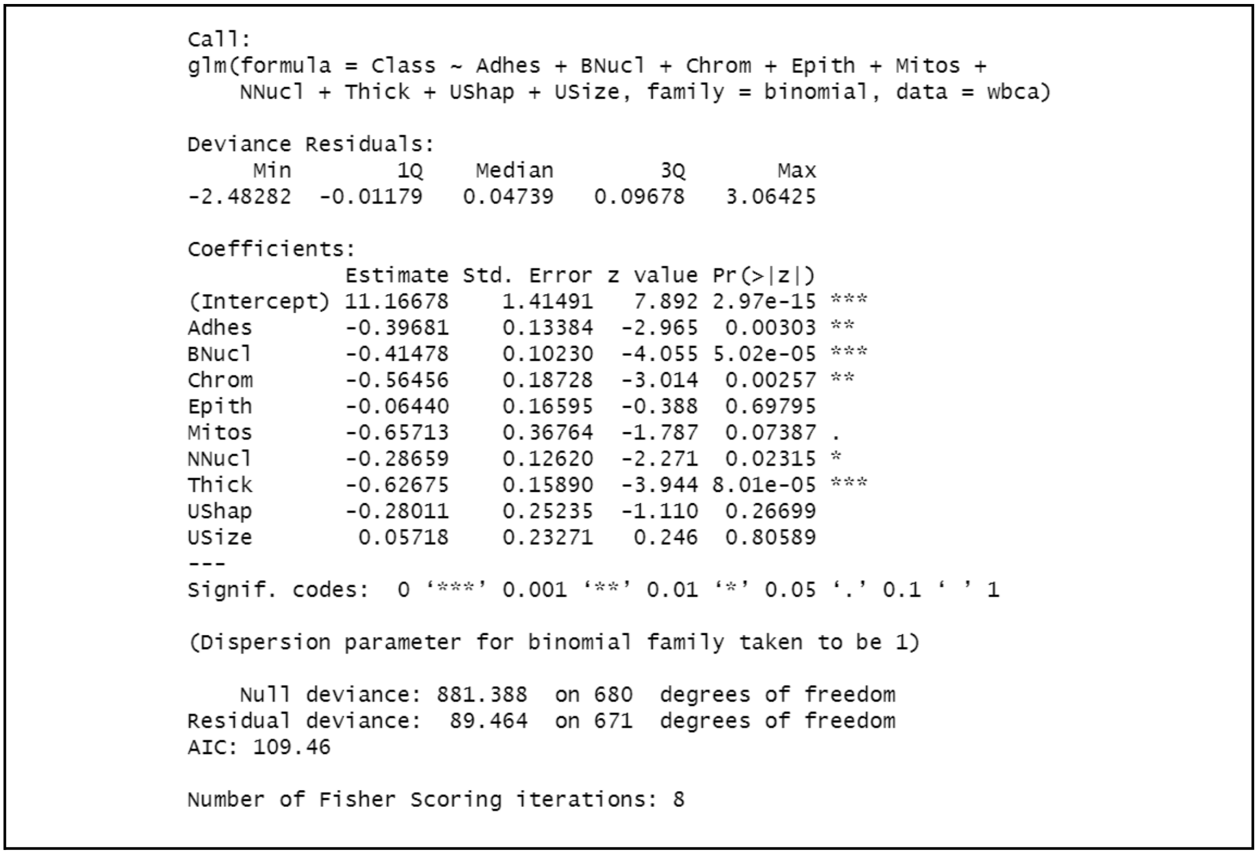
First, we load the data into R. The data is extracted from the R Package Faraway.

```
> library("faraway")
> data(wbca, package="faraway")
```

We want to fit the model into a generalised linear model.

```
> wbca.logit <- glm(Class ~ ., family=binomial, data=wbca)
> summary(wbca.logit)
```

Output:



```
Call:
glm(formula = Class ~ Adhes + BNucl + Chrom + Epith + Mitos +
    NNucl + Thick + UShap + USize, family = binomial, data = wbca)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.48282  -0.01179   0.04739   0.09678   3.06425

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.16678    1.41491   7.892 2.97e-15 ***
Adhes       -0.39681    0.13384  -2.965  0.00303 **
BNucl       -0.41478    0.10230  -4.055 5.02e-05 ***
Chrom       -0.56456    0.18728  -3.014  0.00257 **
Epith       -0.06440    0.16595  -0.388  0.69795
Mitos       -0.65713    0.36764  -1.787  0.07387 .
NNucl       -0.28659    0.12620  -2.271  0.02315 *
Thick       -0.62675    0.15890  -3.944 8.01e-05 ***
UShap       -0.28011    0.25235  -1.110  0.26699
USize        0.05718    0.23271   0.246  0.80589
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 881.388  on 680  degrees of freedom
Residual deviance:  89.464  on 671  degrees of freedom
AIC: 109.46

Number of Fisher Scoring iterations: 8
```

With this output, we are able to obtain the regression coefficients for each predictor.

$$log(\frac{\rho}{1-\rho}) = 11.167 - 0.397\chi_1 - 0.415\chi_2 - 0.565\chi_3 - 0.064\chi_4 - 0.657\chi_5 - 0.287\chi_6 - 0.627\chi_7 - 0.28\chi_8 + 0.057\chi_9$$

The coefficient values are the maximum likelihood estimates of each variable, and they are important in order for us to determine the impact of each variable on the outcome. For example, for every unit of *Adhes* that is decidedly more abnormal, the log odds of it being benign increases by -0.397 times. We may also interpret it as, for every unit it is more abnormal, the log odds of it being malignant increases by 0.397.

## 4. Variable Selection

### 4.1 Why the need for variable selection?

Before we look at methods of variable selection, a notable question comes up: why do we need to select variables and why do we not simply include all of it in our final model? When looking at statistical models, a goodness of fit is important and there is usually a better fit when more variables are included.

This is due to the principle of parsimony that is also important when deciding models, where the less variables, the better.

Furthermore, including too many variables can lead to cases of overadjustment where some of the variables that have an apparent effect are actually due to coincidences (Schneider, Hommel, and Blettner 2010). In further analysis, we will want to decide the best balance between fit and complexity.

### 4.2 Selection based on test statistics and variance

After fitting the dataset into the logistic regression model, we will want to analyse which of the 9 variables are significantly relevant for the model. From the summary of the generalised linear model from section 3.2.2, we see that the variables *USize*, *UShap* and *Epith* are not statistically significant. These variables correspond to the sample's cell size uniformity, cell shape uniformity and epithelial cell size respectively.

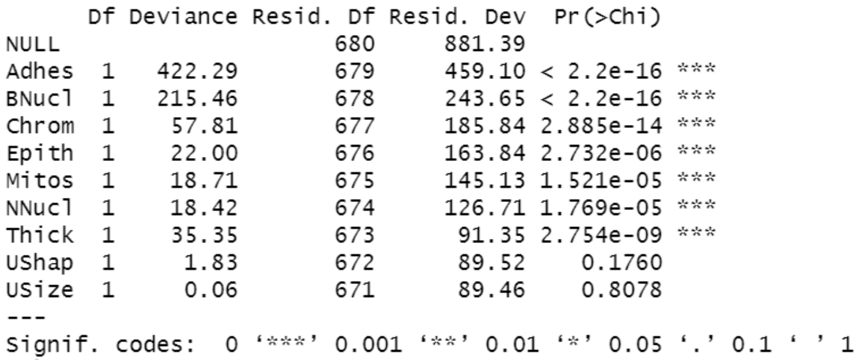
Essentially, within the summary output for z-value, it is the test statistic and we are performing hypothesis tests for all the variables. The null hypothesis is set for each corresponding variable equals to 0, and we reject it when the variable is shown to be significant at a certain confidence level. For example, the variable *BNucl* is significant at a 99.9% level, and has the lowest p-value among all the variables. What this tells us is that the abnormality of the bare nuclei has a strong association with the probability that the tumour sample is malignant.

On the other hand, the *Mitos* variable is significant at 90% level. For most statistical tests, we would usually aim for at least a 95% significance level. For now, the chosen model based on test statistic and p-values has all the variables, excluding the three that are considered statistically insignificant.

Using an ANOVA function in R, we can also conduct a log-likelihood ratio test of whether each explanatory variable is needed in the model.

```
> anova(wbca.logit, test= "LRT")
```

Output:



```
      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                    680     881.39              
Adhes  1   422.29       679     459.10 < 2.2e-16 ***
BNucl  1   215.46       678     243.65 < 2.2e-16 ***
Chrom  1    57.81       677     185.84 2.885e-14 ***
Epith  1    22.00       676     163.84 2.732e-06 ***
Mitos  1    18.71       675     145.13 1.521e-05 ***
NNucl  1    18.42       674     126.71 1.769e-05 ***
Thick  1    35.35       673      91.35 2.754e-09 ***
UShap  1     1.83       672      89.52    0.1760    
USize  1     0.06       671      89.46    0.8078    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the ANOVA table, in the column of p-value based on chi-squared test, there are only two variables that are not significant, which is UShap and USize. Clearly, these two variables should not be included in our final chosen model.

## 4.3 AIC, BIC and Stepwise Selection

There are other methods which are more effective than selection based on basic test statistics. Namely, by using the Akaike information criterion (AIC) and Bayesian information criterion (BIC).

What exactly is AIC and BIC, and what do they calculate? AIC is a technique that utilises in-sample fit to estimate the likelihood of a model to estimate the future values, and BIC is another technique for model selection that calculates the trade-off between model fit and complexity of the model (Mohammed, Naugler, and Far 2015).

The formulas for calculating the values are given as:

$$AIC = -2ln(L) + 2k$$

$$BIC = -2ln(L) + 2ln(N)k$$

L is the likelihood, N is the number of recorded measurements, and k is the number of variables.

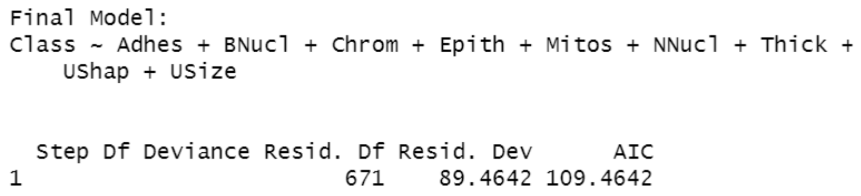
In essence, we formulate multiple models with varying combinations of variables, and then the model with the lowest value of AIC and BIC is the ideal model. However, doing this by hand will be tedious, even if we remove the three variables above that we found are statistically insignificant. Hence, we will compute this in R and utilise stepwise selection.

For definition, stepwise selection is a mixture of forward and backwards selection. Forward selection is first having a model which starts with no predictors, then iteratively adding the most contributive predictors, and stops when the improvement is no longer statistically significant. Backwards selection is similar, but instead of adding predictors, it removes them iteratively (Stepwise Regression Essentials in R - Articles - STHDA n.d.).

Sometimes, all three selections give the same model, but that is not always the case, and as we will see, it is not the case for our data set either. We will be using the MASS R package in order to calculate the AIC.

```
> library(MASS)
> fwd.wbca <- stepAIC(wbca.logit, direction = "forward", trace = FALSE)
> fwd.wbca$anova
```

Output:



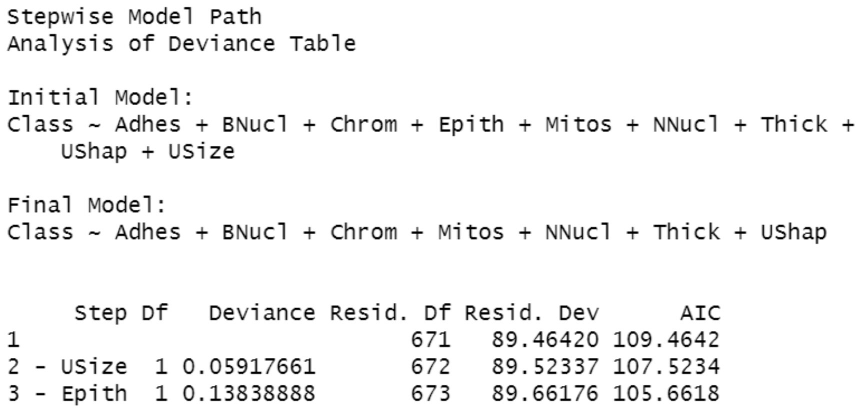
```
Final Model:
Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
    UShap + USize


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                        671    89.4642 109.4642
```

As we see, the iteration stops at the first model where all variables are included. With reference to our p-value selection from earlier, we know that this is a case of overadjustment and not all of the variables are meaningful to us. This is where stepwise selection helps, as forward selection stops its iterations too early for us to make use of the results.

```
> step.wbca <- stepAIC(wbca.logit, direction = "both",
              trace = FALSE)
> step.wbca$anova
```

Output:



```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
    UShap + USize

Final Model:
Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap


     Step Df   Deviance Resid. Df Resid. Dev      AIC
1                             671   89.46420 109.4642
2 - USize  1 0.05917661       672   89.52337 107.5234
3 - Epith  1 0.13838888       673   89.66176 105.6618
```

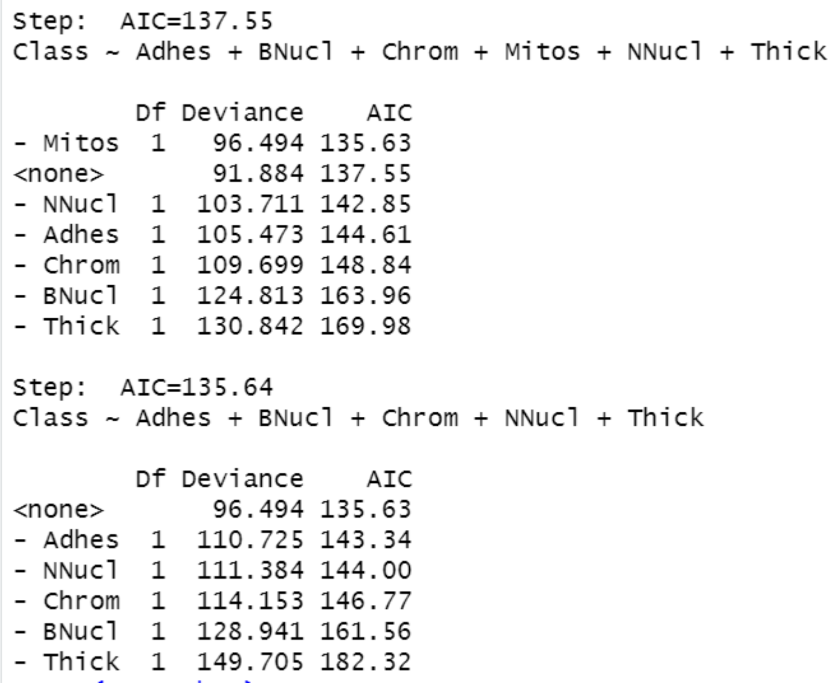
The final model chosen by the stepwise selection, it only removed 2 variables, namely *USize* and *Epith*. Earlier, we have also shown that these 2 variables are statistically insignificant using p-value selection. The AIC value is 105.6618.

Why is this the case within the context of using AIC selection? Within the AIC formula, there is a term in which it penalises larger models. We aim to balance the lack of fit and the model complexity with models having smaller AIC values indicating a better balance between these two important aspects (Narisetty 2020).

Now, we want to look at the models selected by BIC. In order to do that, we will make use of the stepAIC() function that we've been using by assigning different criteria. The default argument k in the function is set to 2, which is for AIC (Zhang 2016).

```
> BIC <- stepAIC(wbca.logit, k=log(nrow(wbca)))
```

Output:



```
Step:  AIC=137.55
Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick

        Df Deviance    AIC
- Mitos  1   96.494 135.63
<none>       91.884 137.55
- NNucl  1  103.711 142.85
- Adhes  1  105.473 144.61
- Chrom  1  109.699 148.84
- BNucl  1  124.813 163.96
- Thick  1  130.842 169.98

Step:  AIC=135.64
Class ~ Adhes + BNucl + Chrom + NNucl + Thick

        Df Deviance    AIC
<none>       96.494 135.63
- Adhes  1  110.725 143.34
- NNucl  1  111.384 144.00
- Chrom  1  114.153 146.77
- BNucl  1  128.941 161.56
- Thick  1  149.705 182.32
```

Although the output above states AIC, we have configured the stepAIC() function in order for it to calculate the BIC in its place. The model with the lowest BIC value removes even more variables, as BIC is known for penalising complex models more than the AIC. Compared to the model selected from AIC, the variables *Mitos* and *UShap* have been removed. This makes sense as BIC usually results in a more parsimonious model than the AIC so it favours a much less complex model (Zhang 2016).

### 4.3.1 Limitations of Stepwise Selection

In regression model building and epidemiological analysis, stepwise selection is a very popular method among researchers to figure which variables are significant. However, there has been much criticism for the heavy and somewhat overreliance on this selection method. One of the main issues is that the standard statistical test would assume a single test from a pre-specified model. Choosing the variables using stepwise involves multiple steps. With the previous assumption applied, instead of acknowledging the multi-step process, it would lead to the issue of bias, overfitting and exaggerated p-values (Smith 2018).

However, stepwise selection is integrated in many statistical programs and tools already and it would still give us good insight on best selected variables. What most experts agree is that using stepwise selection is still viable, but should be coupled with other methods of variable selection.

### 4.4 Choosing the Final Model

Now we have gone through many tests, we are now left to choose which variables and model we will be our final model. The table below is the compiled models we have defined from section 4.2 and 4.3, with the variables excluded, method of variable selection used, and the BIC and AIC counts.

*Table 1. Generalised linear models with calculated AIC and BIC*

| Model | Selection | Variables Excluded | AIC | BIC |
|---|---|---|---|---|
| wbca.logit | Full Model | - | 109.46 | 154.70 |
| an.wbca | ANOVA (chi-squared test) | UShap, USize | 107.35 | 143.54 |
| step.wbca | Stepwise (AIC) | Epith, USize | 105.66 | 141.85 |
| p.wbca | Summary | Epith, UShap, USize | 105.88 | 137.55 |
| bic.wbca | BIC | Mitos, Epith, UShap, USize | 108.49 | 135.64 |

Taking into account the test statistics from section 4.2 as well as table 4,1, we can now discuss on how we will choose the appropriate model.

Firstly, it is easy to exclude the variables *Epith* and *USize* from our final model. From the table, we see that removing *Epith* and *USize* shows a low number of BIC and AIC, and is the model chosen from the stepwise selection method. It would also be best to remove *UShap* from the model, as the variable is shown to not be significant from both test statistics we have performed in section 4.2. Furthermore, the difference between is seen if we compare model p.wbca and step.wbca, with the former having a much lower BIC number. step.wbca may have a smaller AIC, but the difference is pretty much negligible (with only a difference of 0.22).

Should we also exclude *Mitos*, as this variable was left out from our BIC selection? No, it would be best to include it in our final model. Aside from BIC selection, this variable is shown to be significant in all our other tests.

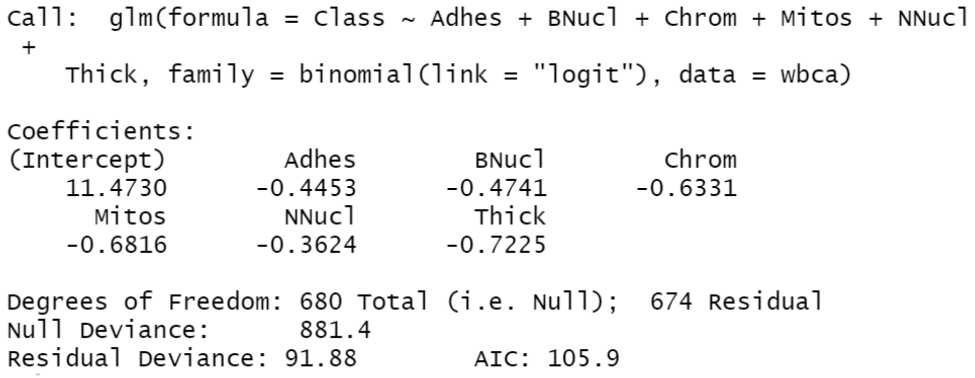
To bring the variable back into the context of cancerous cells, mitosis of the cells play a large role in defining whether a cell is cancerous. As a reminder, cancerous cells divide infinitely and never really die (Mercadante and Kasi 2021). Hence, leaving this variable out of the model would not be a wise idea, despite it showing lower significance compared to the other chosen variables (significant at a 90% level, as we normally aim for 95% confidence level).

The variables that we choose to include into our model are *Adhes, BNucl, Chrom, Mitos, NNucl,* and *Thick.* These 6 variables have shown to be significant in all the tests that we have performed.

In R, we have already defined this logistic regression model with

```
> p.wbca
```

Output:



```
Call:  glm(formula = Class ~ Adhes + BNucl + Chrom + Mitos + NNucl
 +
   Thick, family = binomial(link = "logit"), data = wbca)

Coefficients:
(Intercept)        Adhes        BNucl        Chrom
    11.4730      -0.4453      -0.4741      -0.6331
      Mitos        NNucl        Thick
    -0.6816      -0.3624      -0.7225

Degrees of Freedom: 680 Total (i.e. Null);  674 Residual
Null Deviance:      881.4
Residual Deviance: 91.88        AIC: 105.9
```

Our chosen model, with its coefficients is as below:

$$log(\frac{\rho}{1-\rho}) = 11.473 - 0.445\chi_1 - 0.474\chi_2 - 0.633\chi_3 - 0.682\chi_4 - 0.362\chi_5 \\ - 0.362\chi_6 - 0.723\chi_7$$

The variables are in this order: *Adhes, BNucl, Chrom, Mitos, NNucl,* and *Thick*.

It is important to note that the result is not to say that the test statistic with z-value is the best method of selecting variables. The variables we decide to exclude are based on a number of evidence and tests that we have performed in this report.

### 4.4.1 Making Predictions with the Model

Now that we have our final model, it can be used to predict whether a new sample with its corresponding predictor values of abnormality, is cancerous or not.

As mentioned previously in section 3.2, the logit value will tell us the outcome of a sample. When all the $\chi_i$ values are inputted into the equation, the logit will be equal to any value within the range of $[-\infty, \infty]$. If the value is negative, it corresponds to a probability of less than 0.5 for the positive outcome, and if the logit value is positive, the probability for a positive outcome is more than 0.5. If the probability is less than 0.5, the model would normally allocate the sample into class 0.

Of course, we are able to use R to help us compute predictions for large incoming datasets, as we have also used the program to build the model and all associated calculations. We do this by using the ‘predict’ function built into R, linking the model and the new dataset.

However, this automatic setting of 0.5 being the probability threshold of our model may not represent an optimal interpretation of the predicted probabilities for our dataset. In section 5, we will discuss the sensitivity and specificity of the model, where we determine the best probability threshold for our dataset.

The plus side of removing a few variables from the original model, is that the doctors who are examining the tissue samples don't have to account for the abnormality of unhelpful features of the cells. When there are multiple samples to go through, this can help speed up the process, albeit the impact may not be so noticeable.

## 5. Sensitivity and Specificity

### 5.1 Definition

Ensuring that our results and findings are accurate is important especially when it comes to determining diseases, and in healthcare, these sort of diagnostic tools need to be as accurate as possible in order to receive the best treatment needed. In the case of breast cancer, it is time sensitive and correct results are crucial for the livelihood of the patients. Fine needle aspiration is what we are referring to when talking about diagnostic tests in this report, and the model that we created from the previous section would need to undergo testing in order to prove that this diagnostic test is accurate.

Sensitivity and specificity is calculated and used in order to determine the accuracy of the given dataset and model, and are essential indicators of it. This is how we are able to test whether our diagnostic method is a good tool in determining breast cancer. The two terms have often been used to quantitatively assess the validity of a diagnostic test (Coughlin et al. 1992).

There are other values that we are interested in calculating when it comes to figuring out the accuracy of a model aside from just sensitivity and specificity. We also want to find values of positive predictive values (PPV) and negative predictive values (NPV).

We would want to compare our screening test to the gold standard test. The gold standard test is considered a benchmark test in order to evaluate the diagnostic tool's ability to clearly tell apart patients with breast cancer and patients without (Coughlin et al. 1992). The ideal gold standard test would have both sensitivity and specificity at 100%, but trying to reach that standard is often unrealistic, but instead, most researchers would try to get as close as possible to that value (Umemneku Chikere et al. 2019).

Let us first define all these terms. In order to make sense of the definition, we refer to the table given below. Afflicted patients are defined as the patients that have breast cancer and unafflicted patients are defined as the patients that do not have breast cancer. The values TP, FP, FN and TN are all numerical values corresponding to the number of patients in each category.

*Table 2. Classification of breast cancer patients for gold standard test*

| | Afflicted Patients | Unafflicted Patients |
|---|---|---|

| | | |
|---|---|---|
| Tests positive for breast cancer | True Positive (TP) | False Positive (FP) |
| Tests negative for breast cancer | False Negative (FN) | True Negative (TN) |

Sensitivity is the proportion of the number of patients who correctly test positive for breast cancer to the total number of afflicted patients, including the ones that were incorrectly diagnosed.

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Specificity is the opposite of sensitivity, where it is the proportion of the number of patients who correctly test negative for breast cancer to the total number of unafflicted patients, including the ones that are incorrectly diagnosed.

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

There is a trade-off between the two. If we want a more sensitive model or result, we would often sacrifice specificity. If we want less false positives, then specificity would decrease, leading to more false negatives, and vice versa. In the context of our dataset, if we want more afflicted patients to test positive, then there will be more unafflicted patients to test positive for breast cancer as well. This issue arises due to the fact that some patients will show indications of breast cancer that are right on the line between testing positive or negative (Cardoso et al. 2014).

For PPV, it is the ratio of true positives to the total tested as positive, and for NPV, it is the ratio of true negatives to the total tested as negative. We can interpret the value of the PPV as the probability that someone who tests positive actually has breast cancer, and similarly for NPV. The higher the value, the better (Bartol 2015).

$$PPV = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$NPV = \frac{True\ Negatives}{True\ Negatives + False\ Negatives}$$

## 5.2 Calculating in R

Calculations for the sensitivity and specificity are fairly straightforward. From (203.4.2 Calculating Sensitivity and Specificity in R | Statinfer n.d.), we have an R code that simplifies the calculations for us.

```
> threshold=0.5
> pred.values<-ifelse(predict(p.wbca,type="response")>threshold,1,0)
> actual.values<-p.wbca$y
> conf.matrix<-table(pred.values,actual.values)
> conf.matrix
```

*Table 3. Calculated number of patients under gold standard test, (probability threshold = 0.5)*

| | Afflicted Patients | Unafflicted Patients |
|---|---|---|
| Tests positive for breast cancer | 228 | 9 |
| Tests negative for breast cancer | 10 | 434 |

```
> library(caret)
> sensitivity(conf.matrix)
> specificity(conf.matrix)
> negPredValue(conf.matrix)
> posPredValue(conf.matrix)
```

*Table 4. Sensitivity, specificity, NPV and PPV of our dataset, (probability threshold = 0.5)*

| Sensitivity | Specificity | NPV | PPV |
|---|---|---|---|
| 0.95798 | 0.97968 | 0.97748 | 0.96203 |

As mentioned in section 4.4.1, setting the probability threshold to be 0.5 may not be the optimal interpretation of the predicted probabilities for our dataset and model.

In the code, we set the threshold at the midpoint, where specificity and sensitivity are balanced. As discussed in the previous section, there is a trade-off between the two and we will have to decide which we would want to prioritise. We can adjust which side we want to maximise by adjusting the threshold.

```
> threshold2=0.75
> pred.values2<-ifelse(predict(p.wbca,type="response")>threshold2,1,0)
> conf.matrix2<-table(pred.values2, actual.values)
> conf.matrix2
```

*Table 5. Calculated number of patients under gold standard test, (probability threshold = 0.75)*

| | Afflicted Patients | Unafflicted Patients |
|---|---|---|
| Tests positive for breast cancer | 234 | 14 |

| | | |
|---|---|---|
| Tests negative for breast cancer | 4 | 429 |

```
> sensitivity(conf.matrix2)
> specificity(conf.matrix2)
> negPredValue(conf.matrix2)
> posPredValue(conf.matrix2)
```

*Table 6. Sensitivity, specificity, NPV and PPV of our dataset, (probability threshold = 0.75)*

| Sensitivity | Specificity | NPV | PPV |
|---|---|---|---|
| 0.98319 | 0.96840 | 0.99076 | 0.94355 |

Even after adjusting the threshold, we see that sensitivity and specificity still has a high value of above 0.95. This means that this adjustment is viable. We would want to increase the threshold, as we want to maximise sensitivity. It is far more important to reduce the number of FN than FP.

What exactly are the real life consequences of this adjustment? More afflicted patients would receive the treatment they need for breast cancer, but at the same time more unafflicted patients would also receive treatment that they don’t actually need. This would be costly for the patient and healthcare providers.

Another way to determine the best test cut-off value is by plotting and studying the receiver operating characteristic (ROC) Curve.

### 5.2.1 ROC Curves

ROC curves allow us to visualise all possible cut off values for the probability threshold and represent the trade off between FN and FP. An ROC plot is obtained from plotting sensitivity against (1-specificity) of every observed data (Altman and Bland 1994).

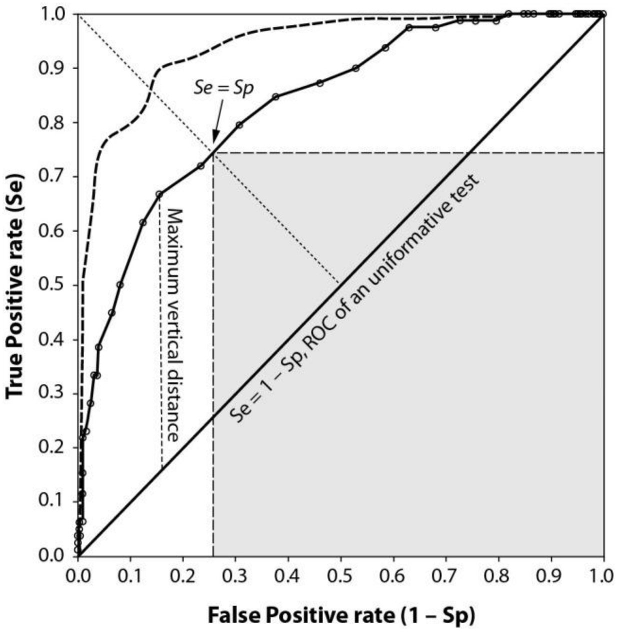


*Fig 4. General structure of an ROC Curve. Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5082211/*

Diagnostic accuracy of the test is also the value of the area under the curve (AUC). A perfect test is one where AUC equals to 1, but normally, a good test would have the AUC anywhere between a range of 0.5 and 1. At an AUC of 0.5, however, there is no distributional difference between sensitivity and specificity of our diagnostic test (Unal 2017).

Now, the point in the graph that we wish to find, corresponds to the probability threshold that we want to apply to our test. There are many different criterions which we can choose from, with each having its own benefits over the other. We will discuss 2 methods.

One of the more frequently used criterions is any point on the curve where sensitivity and specificity are equal. Over the curve, there is a range, when sensitivity increases, specificity decreases so there will be a point where the two values are equal. In figure 4, this point on the graph is where both sensitivity and specificity are maximum, and subsequently, the AUC is also maximum, creating a square from the shaded region (Habibzadeh, Habibzadeh, and Yadollahie 2016). This may not be the criterion that we choose to use with our model tests, as we deliberately want to prioritise one over the other.

Another method is known as the Youden's index (J). The threshold is chosen where calculated Youden's index is maximum and this helps us evaluate the effectiveness of our diagnostic test. The formula of Youden's index is presented as a function. J is a function of Se(c) (sensitivity) and Sp(c) (specificity), such that

$$J(c) = [Se(c) + Sp(c) - 1] = [Se(c) - (1 - Sp(c))]$$

c represents all cut off points from the ROC curve, and the value of c that we end up choosing is one where J is maximum. It is also known as the optimal cut off point. (Unal 2017).

These two are the most common methods of gauging out the best probability threshold for us to test our diagnostic test.

### 5.4 Improvements

The model that we fitted in this report utilises the whole dataset, i.e we obtained the best fitted model. However, as we do not have additional samples to test our best fitted model with, it is often difficult to gauge out how accurate the model is. Specifically, whether the model has any overestimation when it comes to calculating values such as sensitivity and specificity.

A remedy to this problem would be to split the data, ideally with a ratio of 4:1, where 80% of the data will be dedicated to model fitting and the other 20% will be relegated to be used as test samples. This solution will help us see the extent of how well the model ends up performing as the test samples are not fitted into the model building. If a high proportion of the test samples are classified correctly, then, we say that the model is a good fit. Sensitivity and specificity have reduced chances of overestimations.

Not only that, predicted values of the test samples can be used for data visualisation, when constructing graphs and wanting to see comparisons.

As far as improvements for the dataset itself, replications may help. In which different doctors may come in, review the same samples and give their own scores of abnormality. This would help smooth out any biases that may come from only one doctor reviewing every sample. On the other hand, this would make calculations and model building more complex, but we would end up with a far more accurate model.

## 6. Conclusion

Prior methods of diagnosing breast cancer is by doing biopsies and inspecting the tissues for abnormality, and they are often heavy ordeals that can involve surgery. What we aimed to show is that fine needle aspiration, a safer and less risky method of biopsy, is able to extract a good enough sample to allow medical experts to diagnose breast cancer.

Throughout this report, we have shown that the dataset given and the model built from the dataset, are good, with significant variables and high values of sensitivity and specificity. This ultimately leads us to conclude that fine needle aspiration is, in fact, a good method of biopsy for diagnosis of breast cancer. How did we reach this conclusion? If we conclude that the dataset and the model built are good, this means that tissues extracted using fine needle aspiration were not damaged or conformed, which allows the medical expert to accurately score abnormalities of cell structures. Finally then, we conclude that fine needle aspiration can be used by medical experts in the process of diagnosing breast cancer.

## Bibliography

"203.4.2 Calculating Sensitivity and Specificity in R | Statinfer." https://statinfer.com/203-4-2-calculating-sensitivity-and-specificity-in-r/ (April 13, 2022).

Altman, D. G., and J. M. Bland. 1994. "Diagnostic Tests 3: Receiver Operating Characteristic Plots." *BMJ : British Medical Journal* 309(6948): 188. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2540706/ (May 7, 2022).

Bartol, Tom. 2015. "Thoughtful Use of Diagnostic Testing: Making Practical Sense of Sensitivity, Specifi City, and Predictive Value." *Nurse Practitioner* 40(8): 10–12. https://journals.lww.com/tnpj/Fulltext/2015/08000/Thoughtful_use_of_diagnostic_testing__Making.2.aspx (April 12, 2022).

Belyadi, Hoss, and Alireza Haghighat. 2021. "Supervised Learning." *Machine Learning Guide for Oil and Gas Using Python*: 169–295.

Berger, Vance W., and J. Rosser Matthews. 2006. "What Does Biostatistics Mean To Us." *Mens Sana Monographs* 4(1): 89. /pmc/articles/PMC3190464/ (April 10, 2022).

"Breast Cancer." https://www.who.int/news-room/fact-sheets/detail/breast-cancer (April 10, 2022).

Bruening, Wendy et al. 2009. "Comparative Effectiveness of Core-Needle and Open Surgical Biopsy for the Diagnosis of Breast Lesions." *Comparative Effectiveness of Core-Needle and Open Surgical Biopsy for the Diagnosis of Breast Lesions*. https://www.ncbi.nlm.nih.gov/books/NBK45220/ (April 10, 2022).

Cardoso, Jefferson Rosa, Ligia Maxwell Pereira, Maura Daly Iversen, and Adilson Luiz Ramos. 2014. "What Is Gold Standard and What Is Ground Truth?" *Dental Press Journal of Orthodontics* 19(5): 27. /pmc/articles/PMC4296658/ (May 7, 2022).

Coughlin, Steven S. et al. 1992. "The Logistic Modeling of Sensitivity, Specificity, and Predictive Value of a Diagnostic Test." *Journal of Clinical Epidemiology* 45(1): 1–7.

Habibzadeh, Farrokh, Parham Habibzadeh, and Mahboobeh Yadollahie. 2016. "On Determining the Most Appropriate Test Cut-off Value: The Case of Tests with Continuous Results." *Biochemia Medica* 26(3): 297. /pmc/articles/PMC5082211/ (May 9, 2022).

Kumar, Rajesh, Rajeev Srivastava, and Subodh Srivastava. 2015. "Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features." *Journal of Medical Engineering* 2015: 1–14.

Mercadante, Anthony A., and Anup Kasi. 2021. "Genetics, Cancer Cell Cycle Phases." *StatPearls*. https://www.ncbi.nlm.nih.gov/books/NBK563158/ (April 10, 2022).

Mohammed, Emad A., Christopher Naugler, and Behrouz H. Far. 2015. "Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics." *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology: Algorithms and Software Tools*: 577–602.

Narisetty, Naveen Naidu. 2020. "Bayesian Model Selection for High-Dimensional Data." *Handbook of Statistics* 43: 207–48.

Patel, Aisha. 2020. "Benign vs Malignant Tumors." *JAMA Oncology* 6(9): 1488–1488. https://jamanetwork.com/journals/jamaoncology/fullarticle/2768634 (April 10, 2022).

Schneider, Astrid, Gerhard Hommel, and Maria Blettner. 2010. "Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications." *Deutsches Ärzteblatt International* 107(44): 776. /pmc/articles/PMC2992018/ (April 10, 2022).

Shreffler, Jacob, and Martin R. Huecker. 2021. "Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios." *StatPearls*. https://www.ncbi.nlm.nih.gov/books/NBK557491/ (April 12, 2022).

Smith, Gary. 2018. "Step Away from Stepwise." *Journal of Big Data* 5(1): 1–12. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6 (May 5, 2022).

Sperandei, Sandro. 2014. "Understanding Logistic Regression Analysis." *Biochemia Medica* 24(1): 12–18. https://www.biochemia-medica.com/en/journal/24/10.11613/BM.2014.003 (April 10, 2022).

"Stepwise Regression Essentials in R - Articles - STHDA." http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/ (April 10, 2022).

Tolles, Juliana, and William J Meurer. 2016. "Logistic Regression Relating Patient Characteristics to Outcomes JAMA Guide to Statistics and Methods." *JAMA August* 2: 533. http://jama.jamanetwork.com/ (April 10, 2022).

Umemneku Chikere, Chinyereugo M et al. 2019. "Diagnostic Test Evaluation Methodology: A Systematic Review of Methods Employed to Evaluate Diagnostic Tests in the Absence of Gold Standard – An Update." https://doi.org/10.1371/journal.pone.0223832.g001 (April 13, 2022).

Unal, Ilker. 2017. "Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach." *Computational and Mathematical Methods in Medicine* 2017. /pmc/articles/PMC5470053/ (May 10, 2022).

"What Is Cancer? - National Cancer Institute." https://www.cancer.gov/about-cancer/understanding/what-is-cancer (April 10, 2022).

Zhang, Zhongheng. 2016. "Variable Selection with Stepwise and Best Subset Approaches." *Annals of Translational Medicine* 4(7). /pmc/articles/PMC4842399/ (April 10, 2022).