Contingency table vs Bayesian Inference vs Random Forest for multi-class outcome in the context of medical statistics

Minh Chau Van Nguyen *[1] and Hung Dung Van M.D, Ph.D[2]

[1]*School of Mathematics and Statistics, Victoria University of Wellington*
[2]*The Heart Institute of Ho Chi Minh city, Pham Ngoc Thach University of Medicine*

Abstract

In this study, we experimented three statistical methodologies to achieve the aim of finding a potential association between annuloplasty Band and the side effect of Mitral Regurgitation in patients with mitral valve disease, where the data application is a multi-class problem. The first method was the analysis of contingency table, examining odds and probabilities. The second method was Random Forest classification in terms of confusion matrix, ROC-AUC curve and feature importance. Finally, the report looked at Bayesian approach to multinomial logistic regression modelling in the context of likelihood, prior and posterior distributions. DIC model selection, MCMC diagnostics and posterior predictive checks were performed to determine the best model, confirm estimates' convergence to the target distribution and assess goodness-of-fit of the sampling model, respectively. The results derived from these methods showed there is a positive relationship between annuloplasty Band and Mitral Regurgitation.

Keywords— *Bayesian, Random Forest, DIC, multiple multinomial logistic regression, confusion matrix, contingency table, ROC-AUC curves, rjags, posterior predictive checks, MCMC diagnostics*

Background

A follow-up study included 119 patients age ranged from 0 to 18 years with mitral valve disease was developed to study the long-term side effects of posterior Band annuloplasty and artificial classic Carpentier-Edwards (C-E) ring annuloplasty in mitral valve repair surgery, and their association to the re-operation rate caused by Mitral Regurgitation (MR) and Mitral Stenosis (MS). The mean follow up is 134.5±15.4 months. Preliminary results showed that small C-E ring (<26mm) can lead to mitral stenosis and the rate of recurrent mitral regurgitation seem to be higher in patients with posterior bands annuloplasty.

Research question : Does an association exist between the side effects of patients who underwent mitral vale repair using annuloplasty Band and the clinical outcome, Mitral Regurgitation?

Data obtained from the follow-up study are summarised in the below 3x3 contingency table. We can learn from this table that the categorical variables of interest are **Type of Conditions** (outcome variable) and **Type of Implants** (explanatory variable). The Type of Condition feature con-

	Type of Condition		
Type of Implant	MR	MS	None
Band	6	1	27
Ring	2	15	61
None	0	0	7

Table 1. *3x3 contingency table of type of annuloplasty devices implant by long term clinical condition*

[*] Corresponding author
Email address: minh.chau@outlook.co.nz
Last modified on September 13, 2020

sists of three levels, “MR" short for ***Mitral Regurgitation***, is a condition in which the heart’s mitral valve doesn’t close tightly, allowing blood to flow backward in your heart; “MS" short for ***Mitral Stenosis***, is a narrowing of the mitral valve opening that obstructs blood flow from the left atrium to the left ventricle; and “None" simply means the patient has neither of the above heart conditions. The Type of Implants feature also has three categories, “Band" short for ***posterior Band annuloplasty***, “Ring" short for classic ***Carpentier-Edwards (C-E)*** annuloplasty, “None" means the patient has neither of those implants. Annuloplasty is a common procedure in mitral valve repair which involves the implanation of annuloplasty devices such as bands and CE rings. The purpose of annuloplasty is to help maintaining the natural shape, motion, and flexibility of the annulus of the heart [13].

1 Introduction

For under 18 years old patients with mitral valve related diseases, the rate of re-operation caused by recurrent mitral regurgitation (MR) in the long run seems to be higher in those having annuloplasty Band done in mitral valve repair surgery, as shown in the recent follow-up study. The present study introduce and implement three different statistical methods to analyse the multi-class data problem and interpret the results with the goal is to answer to the research question, whether there exists a significant relationship between posterior Band annuloplasty procedure and its potential side effect in the long term, Mitral regurgitation. The methods and their results presented in this report are Contingency table analysis, Bayesian probability inference and Bayesian Multinomial Regression, and Machine Learning Random Forest clasisfication.

The role of contingency table is very common in clinical research as many useful statistics can be derived from the information presented, such as detecting the presence of a significant relationship between two categorical variables. For each variable, we can summarise the data by counting the number of observations in each category or level [2]. In this study, we outline the analysis of contingency table and interpret the degree of association necessary to understand the relationship between Type of Implant and Type of Condition variables, using Odds, Odds ratio and Relative Risks.

In a series of papers and technical reports, it has been demonstrated that substantial gains in statistical classification and regression accuracy can be achieved by using forests (or ensembles) of trees, where each tree in the forest is grown in accordance with a random parameter (feature randomness). By using the algorithm’s regressor, final predictions of the data can be obtained by aggregating over each ensemble [20]. As the base constituents of the ensemble are tree-structured predictors, and since each of these trees is constructed using an injection of randomness, these procedures are called “random forests” [4]. Beside feature randomness, the classification also allows each individual tree to randomly sample from the data set with replacement known as bagging, resulting in different trees and reducing correlation. As a result, the trees protect each other from their individual errors. Using a (3x3) confusion matrix table and ROC-AUC curves [19], we aim to interpret and evaluate the reliability of the model’s performance and predictive ability with precision, post fitting the random forest classification in a multi-class setting. Lastly, we apply the feature importance method using Mean Decrease Accuracy and Mean Decrease Gini to draw conclusions about what features contribute most to the decision making in the model, and help the user to comprehend the drivers behind the model.

Bayesian inference is the process of using data to update prior probabilistic beliefs about parameters of the data generation process [14]. In the context of probability inference, the Frequentist method sees probability as frequency or proportion of the data and then compares the actual data and how often it did happen to what were expected (predictions). The Bayesian method is subjective (results influenced by personal opinions) and uses a priori beliefs to define a prior probability distribution on the possible values of the unknown parameters [15]. Bayesian approach have many benefits, including better small sample results and greater flexibility in model selection [10]. This report focuses on the Bayesian data application to single-level regression and estimation of the unknown model parameters. The latter can be done using a Markov chain Monte Carlo (MCMC) procedure which provides a practical and tractable estimation method. To demonstrate the use of prior belief for deriving the posterior estimates for the given data, we construct a multinomial logistic regression model from a Bayesian perspective using the `rjags` package [12][5] and use this model to estimate the unknown parameters. The choices of prior are chosen to be compatible with the sampling model of the data outcome [11]. Using posterior predictive checks [6], we test the sampling model ability to fit the data well by plotting simulated data and observed data and visually compare them. We then compare amongst candidate models to select the model that is the closest fit to our data using Deviance Information Criterion (DIC) [18] [7]. Additionally, MCMCM diagnostics [16] are implemented to determine whether the parameters from the chosen model converge to the target distribution.

This paper is organised as follows. **Section 2** describes the analysis of a 3x3 contingency table and the concept of odds, odds ratio and relative risks in clinical research. **Section 3** reviews the basics of Random forest classification, its key results based on a 3x3 confusion matrix and shows the ROC-AUC curves for imbalanced and balanced data. **Section 4** introduces Bayesian probability inference, and summarises the stages of building and evaluating a multinomial logistic regression in `rjags`. **Section 5** then combines the results derived from the different methods to identify any association between Band implant and MR condition. Finally, **Section 6** presents our conclusions.

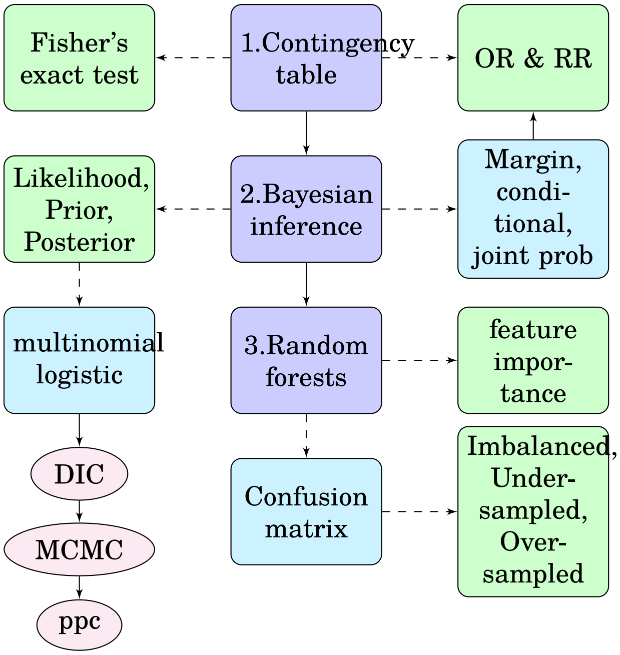


Figure 1. *Report keywords & writing system diagram (OR = Odds Ratio; RR = Relative Risk; DIC = deviance information criterion; MCMC = Markov Chain Monte diagnostics; ppc = posterior predictive checks*

2 Contingency table analysis

2.1 Test of independence

Independence tests are used to determine if there is a significant relationship between two categorical variables. There exists two different types of independence test:

- Chi-square test (χ^2)
- Fisher's exact test

Generally, Fisher's exact test is preferable to the chi-squared test because it is an exact test whereas χ^2 is an approximate test. The chi-squared test should be particularly avoided if there are few observations (e.g. less than 5) for individual cells. Since Fisher's exact test is more compatible with small sample sizes, it is a more suitable test of independence for this data example.

The hypotheses of the Fisher's exact test are the same than for the Chi-square test, that is:

Hypothesis Testing

H_0: the variables are independent, there is no relationship between the two categorical variables. Knowing the value of one variable does not help to predict the value of the other variable.
H_1: the variables are dependent, there is a relationship between the two categorical variables. Knowing the value of one variable helps to predict the value of the other variable.

The Fisher's exact test in R can be computed using the count data provided in **Table 1**. This yields a `p-value` < 0.05, hence we can reject the null hypothesis for `p-value` is less than the significance level at 5%. Rejecting the null hypothesis for the Fisher's exact test of independence means that there is a significant relationship between Type of Implant and Type of condition, i.e. they are not independent.

2.2 Odds, Odds Ratios and Relative Risks

In this section, we introduce parameters that describe the association and present inferential methods for those parameters with the purpose of comparing proportions amongst cells in 3x3 contingency tables.

Notation

Denote **X** and **Y** as Type of Implant (explanatory) and Type of Condition (response), respectively and **n** as the total sample size. The following notations are used to described the three types of probabilities, *joint*, *marginal* and *conditional*.

Abbreviation	Description
n_{ij}	cell counts
n_{i+}	row totals
n_{+j}	column totals
$\pi_{ij} = n_{ij}/n$	joint distribution of X & Y
$\pi_{i+} = \sum_j \pi_{ij}$	marginal distribution of X
$\pi_{+j} = \sum_i \pi_{ij}$	marginal distribution of Y
$\pi_{j\|i} = \pi_{ij}/\pi_{i+}$	conditional distribution of Y given X
$\pi_{i\|j} = \pi_{ij}/\pi_{+j}$	conditional distribution of X given Y

Table 2. *Notation for cell counts and cell probabilities*

Table 4, **Table 6** and **Table 8** summarize the marginal distributions, joint probabilities and conditional probabilities, respectively. Using these tables, we can compute odds, odds ratio and relative risks.

Odds Ratio (OR)

For a probability of success π, the odds of success are defined to be

$$\text{odds} = \pi/(1-\pi)$$

The odds are non-negative, with value greater than 1.0 when a success is more likely than a failure. For example, when odds = 4.0, a success is four times as likely as a failure. In the following example, we are going to look at the odds of patients having MR condition (odds of success) versus the odds of patients having MS condition and the odds of patients having neither condition (both are odds of failure) in the Band (row 1) and Ring (row 2) groups. Using the conditional probabilities in **Table 8**, the odds ratio can be calculated as follows:

Within row 1,

$$\text{odds}_1 = \frac{\pi_{1|1}}{1-\pi_{1|1}} = \frac{0.1765}{1-0.1765} = 0.2143291$$

Within row 2,

$$\text{odds}_1 = \frac{\pi_{1|2}}{1-\pi_{2|2}} = \frac{0.026}{1-0.026} = 0.02669404517$$

Odds ratio,

$$\text{OR} = \frac{\text{odds}_1}{\text{odds}_2} = \frac{0.214}{0.027} = 8.029098$$

Interpretation: Odds of having MR (success) for patients with Band implants are estimated to be 8 times the odds of having MR for patients with Ring implants.

Sample Odds Ratio

Alternatively, the odds ratio can be calculated based on the cell frequencies alone. Since we are comparing the patients with Band implants group to the rest of the The sample odds ratio for the red sub table from the 3x2 table (see **Table 4**) is (6x76)/(28x2) = 8.1, implying that the odds of having MR condition for patients with Band implant are estimated to be 8.1 times higher than the odds of having MR condition for patients with a Ring implant.

95% confidence interval for sample OR

The 95% confidence interval for the above sample odds ratio can be calculated as follows:

$$\text{Upper } 95\% \text{ C.I} = \exp\big(\ln(\text{OR}) + 1.96\text{SE}\big)$$
$$\text{Upper } 95\% \text{ C.I} = \exp\big(\ln(\text{OR}) - 1.96\text{SE}\big)$$

$\rightarrow$ 95% C.I: (1.114688, 3.079594).

where SE = $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$.

Interpretation: Since the 95% C.I of 1.11 to 3.08 does not include 1.0, the increased odds (OR = 8.1) of having MR condition is statistical significant between Band group and MS plus None group.

Relative Risk (RR)

Whereas the odds ratio is the ratio of two odds, the relative risk (RR) is a ratio of two probabilities and RR can be any non-negative real number.

$$\text{RR} = \frac{\pi_{1|1}}{\pi_{1|2}} = \frac{0.1765}{0.026} = 6.788462$$

Interpretation: Patients with Band implants group has 6.8 times the risk of getting MR in the long term against patients with Ring implants group.

Y X	MR	MS	None	Total
Band	$n_{11} = 6$	$n_{12} = 1$	$n_{13} = 27$	$n_{1+} = 34$
Ring	$n_{21} = 2$	$n_{22} = 15$	$n_{23} = 61$	$n_{2+} = 78$
None	$n_{31} = 0$	$n_{32} = 0$	$n_{33} = 7$	$n_{3+} = 7$
Total	$n_{+1} = 8$	$n_{+2} = 16$	$n_{+3} = 95$	n = 119

Table 3. *Table 1 in terms of cell count notations*

Y X	MR	MS & None	Total
Band	$n_{11} = 6$	$n_{12} = 28$	$n_{1+} = 34$
Ring	$n_{21} = 2$	$n_{22} = 76$	$n_{2+} = 78$
None	$n_{31} = 0$	$n_{32} = 7$	$n_{3+} = 7$
Total	$n_{+1} = 8$	$n_{+2} = 101$	n = 119

Table 4. *MS and None columns combined*

Y X	MR	MS	None	Total
Band	π_{11}	π_{12}	π_{13}	π_{1+}
Ring	π_{21}	π_{22}	π_{23}	π_{2+}
None	π_{31}	π_{32}	π_{33}	π_{3+}
Total	π_{+1}	π_{+2}	π_{+3}	1.0

Table 5. *Joint distribution of X and Y*

Y X	MR	MS	None	Total
Band	0.050	0.008	0.227	0.285
Ring	0.017	0.126	0.513	0.656
None	0	0	0.059	0.059
Total	0.067	0.134	0.799	1.0

Table 6. *Joint distribution of X and Y*

Y X	MR	MS	None	Total
Band	$\pi_{1\|1}$	$\pi_{2\|1}$	$\pi_{3\|1}$	1.0
Ring	$\pi_{1\|2}$	$\pi_{2\|2}$	$\pi_{3\|2}$	1.0
None	$\pi_{1\|3}$	$\pi_{2\|3}$	$\pi_{3\|3}$	1.0

Table 7. *Conditional distribution of Y given X*

Y X	MR	MS	None	Total
Band	0.1765	0.0294	0.7941	1.0
Ring	0.026	0.192	0.782	1.0
None	0	0	1.0	1.0

Table 8. *Conditional distribution of Y given X*

3 Random Forest Multi-class classification

Random forest is a supervised learning algorithm that can be used for both classification and prediction. A random forest consists of many decision trees and merges these together to get a more accurate and stable prediction. It applies the bagging method (also known as bootstrap aggregation) when building each individual tree to produce uncorrelated forests of trees, whose prediction by “forests" is more precise than that of any individual tree due to the fact that an ensemble of trees usually increases the overall result. Random forest also adds feature randomness to the model, searching for the best feature among a random subset of features instead of searching for the most important feature while splitting a node. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.

This section shows how to manually compute performance metrics using a 3x3 confusion matrix, evaluates the AUC-ROC curves for imbalanced and balanced data and performs feature importance. The data introduces a new explanatory variable, *Age group*, consisting of four categories (0-4 years, 5-9 years, 10-14 years & 15-18 years). The procedure for analysing the data using the random forest classification is as follows,

- Split the data into training (80%) and test (20%) sets
- Fit the model using the training data
- Obtain the confusion matrix created from predicting new data (test data)
- Evaluate the model performance based on *Accuracy*, *F1-score*, *ROC curves*, and *AUC*
- Examine the importance of the predictors in the model

3.1 Confusion Matrix Analysis

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualisation of the performance of an algorithm and easy identification of confusion between classes e.g. one class is commonly mislabelled as the other. This section shows how to calculate the performance measures (**Table 9**) using a 3x3 confusion matrix, done on imbalanced and balanced data.

Metric	Formula
Accuracy	$\frac{TP}{\text{All cells}}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
NPV	$\frac{TN}{TN+FN}$
F1-score	$2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

Table 9. *Performance evaluation of Random Forest*

3.1.1 Imbalanced data

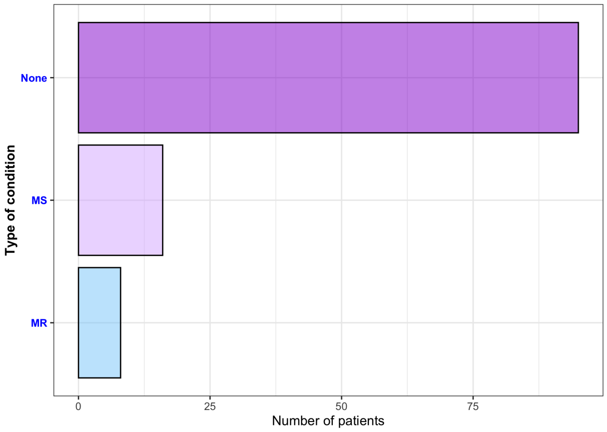


Figure 2. *Distribution of the Type of condition outcome variable. The prevalence for MR, MS & None classes are 8%, 15% & 77%, respectively*

Figure 2 shows the proportion of Type of condition variable where the majority of the data belongs to “None" class, implying a class imbalance problem to our data structure. If we use this data for prediction model, it will be dominated not only by “None" category, similarly when accuracy calculated using such data will be highly influence with the larger group of the data [9].

This results in the following 3x3 confusion matrix where the data is also highly imbalanced, where there are only 8 observations for **A** (MR), 14 observations for **B** (MS) and 74 observations for **C** (None).

		Actual data ⟶		
		A	*B*	*C*
Prediction	*A*	0	0	0
↓	*B*	0	0	0
	C	8	14	74

Table 10. *Type of condition confusion matrix where A is MR, B is MS and C is None*

⋆Note: It is very important to pay attention to whether Prediction are rows or columns as the performance measures calculation are different depending on how the predictions are presented. In this case, the predictions are shown as rows and the actual data are shown as columns. The following matrices represent the location of *True Positives*, *True Negative*, *False Positive* and *False Negative* for A, B and C. Using these matrices, the key performance measure of the Random Forest model such as *Precision*, *Recall*, *Accuracy* and *F1-score* can be calculated manually or obtained from the model output in R.

1. Compute TP, TN, FP and FN

True Positive

From the True Positive matrix, we can clearly see that no observation out of 8 was correctly identified as A and no observation was correctly identified as B out of 14 while 74 observations out of 74 were correctly identified as C.

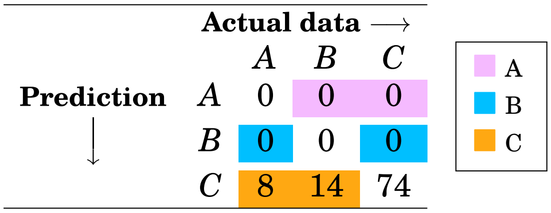
		Actual data ⟶		
		A	*B*	*C*
Prediction	*A*	0	0	0
↓	*B*	0	0	0
	C	8	14	74

False Positive

Measures	Definition
True Positive	**data items are correctly included in the positive class**
True Negative	data items are correctly included in the negative class
False Positive	**data items are incorrectly included in the positive class**
False Negative	data items are incorrectly included in the negative class
Prevalence	**number of cases in a defined population at a single point in time and is expressed as a decimal or a percentage.**
Specificity	measure of the correctly identified negative cases from all the actual negative cases/the percentage of true negatives
Recall or Sensitivity	**measure of the correctly identified positive cases from all the actual positive cases/the percentage of true positive**
Precision or PPV	measure of the correctly identified positive cases from all the predicted positive cases
NPV	**measure of the correctly identified negative cases from all the predicted negative cases**
Accuracy	measure of all the correctly identified positive cases
F1-score	**the harmonic mean of Precision and Recall**

Table 11. *Terminologies of characteristics of a diagnostic test*

From the False Positive matrix, no observation was incorrectly classified as A, no observation was incorrectly classified as B and 22 observations were incorrectly classified as C.

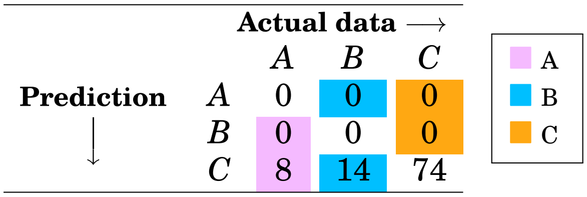


		Actual data ⟶		
		A	*B*	*C*
Prediction	*A*	0	0	0
↓	*B*	0	0	0
	C	8	14	74

🟪 A
🟦 B
🟧 C

False Negative

From the False Negative matrix, 8 observations were incorrectly identified as not A, 14 observations were incorrectly identified as not B and 74 observations were incorrectly identified as not C.

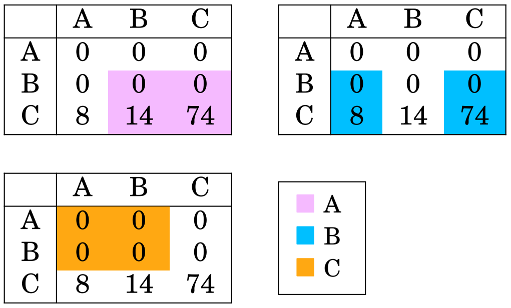


		Actual data ⟶		
		A	*B*	*C*
Prediction	*A*	0	0	0
↓	*B*	0	0	0
	C	8	14	74

🟪 A
🟦 B
🟧 C

True Negative

From the True Negative matrix, 88 observations were correctly identified as not A, 82 observations were correctly identified as not B and no observations were correctly identified as not C.



	A	B	C
A	0	0	0
B	0	0	0
C	8	14	74

	A	B	C
A	0	0	0
B	0	0	0
C	8	14	74

	A	B	C
A	0	0	0
B	0	0	0
C	8	14	74

🟪 A
🟦 B
🟧 C

2. Model evaluation for imbalanced data

Table 12 summarises the performance metrics computed for each of the three classes for the imbalanced data. We can average over three classes to compute the overall Precision and Recall metrics using either micro-average or marcro-average.

Metric \ Class	A	B	C
TP	0	0	74
TN	88	82	0
FP	0	0	22
FN	8	14	0
Precision/PPV	0	0	0.77
Recall/Sensitivity	0	0	1
NPV	0.92	0.85	0

Table 12. *Single metrics results for imbalanced data*

By definition, micro- and macro-averages compute slightly different things, and thus their interpretation differs. A macro-average will compute the metric independently for each class and then take the average (hence treating all classes equally), whereas a micro-average will aggregate the contributions of all classes to compute the average metric. For example, to calculate Recall $= \frac{\text{TP}}{\text{TP}+\text{FN}}$ using **Table 12**,

Macro-average:

$$\text{R}_{macro} = \frac{0+0+1}{3} = 0.33$$

Micro-average:

$$\text{R}_{micro} = \frac{0+0+74}{8+14+74} = 0.77$$

In a multi-class classification setup, micro-average is preferable if class imbalance is suspected. Micro-averaged and macro-averaged precision and recall can then be used to calculate F1-score. In general, accuracy works well on balanced data while F1-score is a better metric for interpreting imbalanced data. The results showed that the accuracy for random forest is relatively high while the macro-averaged F1-score of the classifier is much smaller due to small values of precision and recall for the imbalanced data. Micro-averaged F1-score of the model for the imbalanced data however was reported to be relatively high (77.08%) as expected.

3.1.2 Balanced data

Machine learning algorithms including random forests struggle with accuracy because they assume that the data set has a balanced class distribution when the dependent variable has an unequal distribution as seen in the data example. This causes the performance of existing classifiers to get biased towards majority class (the class with the most observations, i.e. “None"). In this section, we are going to compare the performances of *Undersampling* and *Oversampling* methods when applied to the imbalanced data.

1. Difference between undersampling and oversampling

The main difference between undersampling and oversampling methods is that undersampling works with majority class and oversampling works with minority class. More specifically, undersampling method reduces the number of observations from majority class to make the data set balanced and the latter replicates the observations from minority class to balance the data.

2. Results

	Undersampling	Oversampling
MR (A)	6	78
MS (B)	6	78
None (C)	6	78

Table 13. *Class distribution of undersampling and over sampling*

Table 13 summarises the equal distribution of the outcome variable using undersampling and oversampling methods. This results in the following matrices, where the data is still imbalanced although not as significant as the previous data. For example, undersampling data has 10 observations for A, 5 observations for B and 17 observations for C; and oversampling data has 10 observations for A, 19 observations for B and 17 observations for C.

Actual data ⟶

Prediction ↓

	A	B	C
A	2	0	2
B	0	5	10
C	0	0	5

	A	B	C
A	2	0	2
B	0	5	8
C	0	0	7

Table 14. *Confusion matrices for undersampling data (left) and oversampling data (right)*

Under	A	B	C
TP	2	5	5
TN	20	15	7
FP	2	15	0
FN	2	0	12

Over	A	B	C
TP	2	5	7
TN	20	15	7
FP	2	8	0
FN	0	0	10

Table 15. *Confusion matrices for undersampling data (left) and oversampling data (right)*

From **Table 16**, it is clear that the performance metrics for both Undersampling and Oversampling data have improved significantly in comparison to that for the imbalanced data, where F1-score is now higher than Accuracy as expected. In general, oversampling method performs better then undersampling method, with higher Accuracy and F1-score. This makes sense since undersampling works best when the data set is huge but we have quite a small data (N = 119 observations). It is however worth noting that a disadvantage of the oversampling method is overfitting problem, where the training accuracy of such data set will be relatively high, but the accuracy on unseen data will be worse (only 58.33%).

3.2 ROC curves and AUC

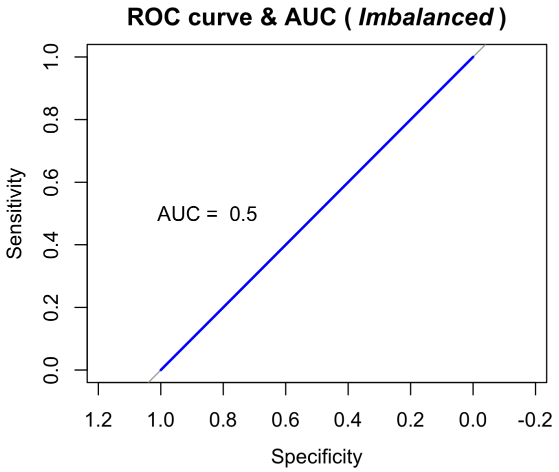


Figure 3. *ROC curve and AUC for imbalanced data*

⋆Note: the x axis goes decreasing from 1 to 0, which is exactly the same as plotting 1 - specificity on an x axis increasing from 0 to 1.

1. ROC curves

Receiver Operating Characteristic or ROC curves are commonly used to plot the True Positive rate against False Positive Rate, depicting sensitivity/specificity trade offs for a binary or a multi-class classifier [8].

$$\begin{aligned}\text{False Positive Rate}(FPR) &= \frac{FP}{FP+TN}\\ &= 1-\text{Specificity}\end{aligned}$$

$$\begin{aligned}\text{True Positive Rate}(TPR) &= \frac{TP}{TP+FN}\\ &= \text{Sensitivity or Recall}\end{aligned}$$

Most machine learning algorithm classifiers produce positive scores that correspond with the strength of the prediction that a given case truly belongs to their class. These scores can be classified into yes or no predictions which requires setting a threshold such that observations with scores above the threshold are classified as positive, and observations with scores below the threshold are predicted to be negative. Different threshold values give different levels of sensitivity and specificity. A high threshold (>0.5) means it is less likely to produce false positive results but also more likely to ignore cases that are in fact positive (lower rate of true positives). In contrast, a low threshold means positive labels are more likely to be produced, results in more false positives but also more true positives. For our data example, we will leave it to the random forest model to find the optimal threshold desired.

For multi-class model, we can **plot N number of ROC Curves for N classes using One vs All methodology**. In our example, we have three classes labelled MR, MS and None, then there will be a ROC for MR classified against MS and None, another ROC for MS classified against MR and None, and a third one of None classified against MR and MS.

2. AUC

		Macro-average			**Micro-average**		
Data	Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score
Imbalanced	77.08%	25.77%	33.33%	29.07%	77.08%	77.08%	77.08%
Undersampled	50.00%	60.00%	76.47%	67.24%	41.38%	46.15%	43.63%
Oversampled	58.33%	62.82%	80.59%	70.53%	58.33%	58.33%	58.33%

Table 16. *Marco-average and micro-average of the performance metrics for imbalanced, undersampled and oversampled data*

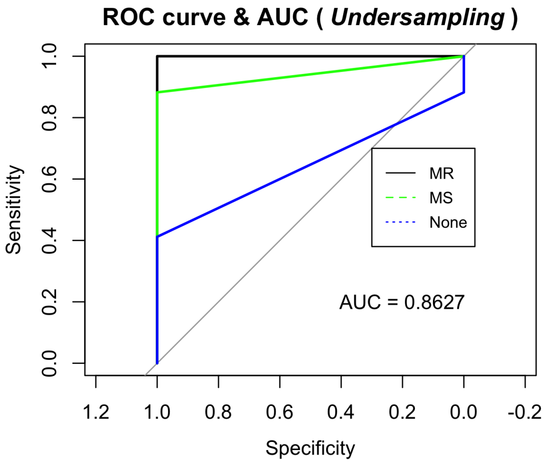


Figure 4. *ROC curve and AUC for undersampling data*

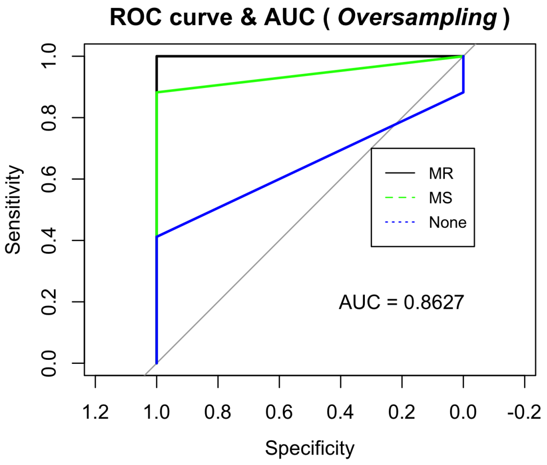


Figure 5. *ROC curve and AUC for oversampling data*

AUC (Area Under the Curve) is another performance metric for accuracy of a classification algorithm that is associated with the ROC curve. A model with greater accuracy should have a value closer to 1 for AUC. Therefore, when comparing the accuracy of multiple models using the AUC, the one with the highest AUC can be considered the best performing model.

The **diagonal line** in a ROC curve represents perfect chance. In other words, a test that follows the diagonal has AUC = 0.5 (half the area). Looking at **Figure 3**, it is clear that the Random Forest model when applied to the imbalanced data has no discrimination capacity to distinguish between classes as AUC for all classes is approximately 0.5, and that the ROC curves for all classes overlap with the diagonal line. On the other hand, the classifier used to fit the data post undersampling/oversampling have a much higher average AUC of the three classes as shown in **Figure 4** and **Figure 5**, meaning the model has a good measure of separability.

To summarise, the undersampling method and oversampling method have the same AUC but because the latter has a higher accuracy and F1-score, the optimal model is then the oversampling model.

3.3 Feature Importance

After choosing the optimal model, it is natural to ask which variables have the most predictive power. Variables with high importance are drivers of the outcome and their values have a significant impact on the outcome values. By contrast, variables with low importance might be omitted from a model, making it simpler and faster to fit and predict. This section focuses on the importance degree of the independent variables in terms of predicting the dependent variable.

There are two measures of importance given for each variable in the random forest. The first measure is based on how much the prediction performance of the model decreases when the variable is excluded (permutation/mean decrease in accuracy importance). This is further broken down by outcome class. The second measure is based on the mean decrease in impurity, also known as Gini importance. Assessing the decrease in GINI when a variable is omitted leads to an understanding of

	MR	**MS**	**None**	**MeanDecreaseAccuracy**	**MeanDecreaseGini**
Type of implant	6.92410	14.806164	-5.860717	11.653713	2.734225
Age group	-1.284382	8.282171	10.943925	9.275776	3.912649

Table 17. *OOB estimate of error rate: 33.33% (class error rate: MR (33.33%), MS (16.67%), None (50%)*

how important that variable is to split the data correctly (pure single class nodes). A higher Mean Decrease in Gini and a higher Mean Decrease in Accuracy both indicate higher variable importance. Gini importance however is overall inferior to permutation based importance as it is relatively more biased and unstable.

Table 17 shows the importance of *Type of implant* and *Age* when predicting Type of condition outcome with three classes. For example, Type of Implant is important for predicting that a patient with MR as well as predicting a patient with MS conditions, but not important for predicting a patient with either condition. The mean decrease in accuracy for Type of condition variable is higher than that Age, indicating the model performance accuracy is lower when Type of condition is excluded in comparison to excluding Age.

⋆`Note`: In **Table 17**, **Out-of-bag (OOB)** error was also mentioned briefly. It is a method of measuring the prediction error of random forests, here OOB error rate is about 33.33% which is high and the class prediction error rate error rate for MR, MS and None were 33.33%, 16.67% and 50%, respectively.

4 Bayesian Method

Bayesian Approach, which is concerned with the probability of having a certain outcome given the data, is a method of statistical inference in which Bayes' theorem is used to update the probability for a belief as more evidence or information becomes available. The differences between Frequentist and Bayesian lie between their interpretation on probability and in terms of what is fixed. The frequentist perspective is that data are repeatable random measures, it is always possible to run another experiment the same way since the population stays fixed. Meanwhile with the Bayesian approach, the data observed from a realised sample and the parameters of the population are unknown, but can be described probabilistically hence they're not sort of fixed values. However, the data are fixed. As previously mentioned, The Bayesian approach focuses on the application of Bayes' rule, relating the conditional probabilities with the prior beliefs. This allows taking your belief about the probabilities of certain events happening and update them when you more data is collected [1].

4.1 Marginal, Conditional, and Joint Probabilities

Margin probability, P(A)

• The margin probability of an event A, or P(A), is the probability of A without conditioning on the occurrence or non-occurrence of any other events.

Conditional probability, P(B|A)

• For any two events A and B, the conditional of B given A, or P(B | A), is the probability that event B will occur, given that we know that event A has occurred.

Joint probability, P(A, B)

• The joint probability of events A and B is the probability of both events occurring and is given by P(A,B).

Data example: Number of Patients who has Band or Ring Implant that are exposed to MR condition

Using **Table 3**, denote the following:

- **A** - the patient has a Band Implant
- **B** - the patient has MR condition
- **C** - the patient has a Ring implant

Table 9 summarise the margin, joint and conditional probabilities for events A, B and C. The conditional probability for B given A and the conditional probability for C given A can be calculated by two approaches, restricting the sample space or applying Bayes' rule. For example, by restricting the sample space to only the first row of the table (i.e. the outcome in A), the probability of drawing a patient diagnosed with Band implant that has MR condition is,

$$P(B|A) = \frac{6}{34} = 0.176471$$

We can also obtain the conditional probability P(B|A) using **Bayes' theorem**:

$$\boxed{P(B|A) = \frac{P(A|B)P(B)}{P(A)}}$$

This is also known as the *prior probability* that a randomly encountered patient (assuming we don't have the information whether they have a Band

Probability	Definition	Value
P(A)	The probability of drawing a patient with Band implant	34/119
P(B)	The probability of drawing a patient with MR condition	8/119
P(C)	The probability of drawing a patient with Ring implant	78/119
P(A,B)	The probability of drawing a patient with Band implant that has MR condition	6/119
P(B|A)	The probability of drawing a patient diagnosed with MR condition, given that this patient has a Band implant	
P(C,B)	The probability of drawing a patient with Ring implant that has MR condition	2/119
P(B|C)	The probability of drawing a patient diagnosed with MR condition, given that this patient has a Ring implant	

Table 18. *Probabilities explained in a Bayesian sense for the patients with heart disease data*

implant, a Ring implant or neither) will be diagnosed with MR condition.

$$\begin{aligned} P(B|A) &= \frac{P(A,B)}{P(A)} \\ &= \frac{6/119}{34/119} = 0.17641 \end{aligned}$$

Since $P(B|A) = 0.176471$, the chance is 0.176 that the patient is diagnosed with MR condition if they have a Band implant. Note that this prior conditional probability is the same (after rounding) as the conditional probability in **Table 8**.

Repeat this step to find the probability that a patient has MR, given that this patient has a Ring implant.

$$\begin{aligned} P(B|C) &= \frac{P(C,B)}{P(C)} \\ &= \frac{2/119}{78/119} = 0.02564103 \end{aligned}$$

If a patient has a Ring implant, the chance is 0.02564103 that this patient also has MR, which is much lower than that for those that has a Band implant.

4.2 Bayesian Multinomial Regression Modelling

A major difference between frequentist and Bayesian methods is that only the latter can incorporate background knowledge (or lack thereof) into the analyses by means of the prior distribution.

Data example

Using the same data set as the Random Forest methodology, we fit a regression model using a Bayesian approach. The aim is to investigate whether the occurrence of mitral regurgitation (MR) are associated with any of the independent variables:

(i) Type of Implant, categorical variables (3 levels)
   - None (baseline)
   - Band (X_1)
   - Ring (X_2)

(ii) Age groups, categorical variables (4 levels)
   - 0-4 years (baseline)
   - 5-9 years (X_3)
   - 10-14 years (X_4)
   - 15-18 years (X_5)

For the response variable, MR is a categorical variable with three levels, the following full multinomial regression model can be used to analyse the data,

$$\begin{aligned} log\left(\frac{Y=j}{Y=k}\right) =&\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3 + \\ &\beta_{j4}X_4 + \beta_{j5}X_5 \end{aligned} \quad \textbf{(1)}$$

for $j = 2, \ldots, k$ levels of the response variable where k is the baseline.

Listing 1 . *R output for the full model with Normal prior*

```
    model{
for (i in 1:N){

  ## sampling model for the outcome variable
  o1[i] ~ dbern(pi3[i])  # None
  o2[i] ~ dbern(pi2[i])  # MR
  o3[i] ~ dbern(pi3[i])  # MS

  ## predictors
  # r = Type of Implant, r1 is the baseline (None); a = Age group, a1 is the baseline (0-4 years)
  logit(pi1[i]) <- beta1[1] + beta1[2]*r2[i] + beta1[3]*r3[i] + beta1[4]*a2[i] + beta1[5]*a3[i] + beta1[6]*a4[i]
  logit(pi2[i]) <- beta2[1] + beta2[2]*r2[i] + beta2[3]*r3[i] + beta2[4]*a2[i] + beta2[5]*a3[i] + beta2[6]*a4[i]
  logit(pi3[i]) <- beta3[1] + beta3[2]*r2[i] + beta3[3]*r3[i] + beta3[4]*a2[i] + beta3[5]*a3[i] + beta3[6]*a4[i]
}
  ## priors
  for(j in 1:6){
    beta1[j] ~ dnorm(0, 1e-01)
  }
  for(j in 1:6){
    beta2[j] ~ dnorm(0, 1e-01)
  }
  for(j in 1:6){
    beta3[j] ~ dnorm(0, 1e-01)
  }

  ## derived quantity
  OR1 <- exp(beta2[2])
  OR2 <- exp(beta2[3])
}
```

Important: the link between logit, logistic and odds [21]

The logit is a link function / a transformation of a parameter. It is the logarithm of the odds. In terms of π, it is defined as follows:

$$logit(\pi_j) = log\left(\frac{\pi_j}{1-\pi_j}\right)$$

The logistic function is the inverse of the logit. If we have a value x, the logistic is:

$$logistic(x) = \frac{e^x}{1+e^x}$$

Thus, logit regression is:

$$log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3 + \beta_{j4}X_4 + \beta_{j5}X_5$$

and logistic regression is:

$$\pi_j = \frac{e^{(\beta_{j0}+\beta_{j1}X_1+\beta_{j2}X_2+\beta_{j3}X_3+\beta_{j4}X_4+\beta_{j5}X_5)}}{1+e^{(\beta_{j0}+\beta_{j1}X_1+\beta_{j2}X_2+\beta_{j3}X_3+\beta_{j4}X_4+\beta_{j5}X_5)}}$$

Exponentiating the logit will give the odds of an event occurring. Likewise, the odds can be obatined by taking the logistic and dividing it by 1 minus the logistic. That is:

$$odds = exp(logit(\pi_j)) = \frac{logistic(x)}{1-logistic(x)}$$

4.2.1 Likelihood, Prior and Posterior distributions

The posterior distribution is a compromise between the prior distribution and the sampling model. Inference we make under the Bayesian framework is sensitive to:

- the choice of prior distribution
- the strength of confidence we express in our prior information
- the sample size

We also split the data into training (80%) and test (20%) sets prior to modelling and use the same seed as random forests in order to produce the same sets. The model procedure for fitting a single-level regression model for a multi-class outcome in this data example is as follows,

Procedure

1. Fit the model in `rjags` using training data
2. Model selection using DIC
3. MCMC diagnostics of the unknown parameters
4. Posterior predictive check for lack of fit
5. ROC-AUC computation for the chosen model when fited to new (test) data

1. Fit the model in `rjags`

LIKELIHOOD, $\boldsymbol{p(y|\theta)}$

By treating the outcome variable as dummies, each class of the outcome variable then follows a Bernoulli distribution, i.e.

$$y_j \sim Bernoulli(\pi_j) \quad \text{for j = MR, MS, None levels}$$

where π_j in this case is equivalent to θ. This is known as the sampling model or likelihood.

PRIOR, $\boldsymbol{p(\theta)}$

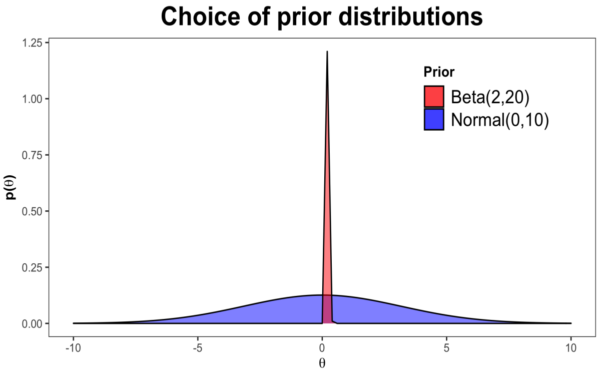


Figure 6. *Arbitrary choice of **prior** density plots*

Using the model from **Eq. (1)** and suppose that our prior belief is that β parameters are probably between -10 and 10 where the choice of prior is completely arbitrary. In this example, **Normal(0, $\boldsymbol{\sigma^2}$)** and **Beta(a, b)** were chosen as weak (non-informative) priors, and specified to have their intervals fall between the desired lower and upper limits for β. For the Normal prior to be weak, the most common choice for μ is zero, and σ^2 is usually chosen to be large enough to be considered as non-informative, common choices being in the range from σ^2 = 10 to σ^2 = 100. The prior sample size, denoted by $w = a + b$ corresponding to a weak Beta(a,b) is the strength of our confidence belief. If the prior confidence is much larger than the sample size, then the prior has a marked influence on the posterior. For Beta distribution, a and b parameters were specified as 2 and 20 to reflect on a small prior sample size for a weak prior.

$\boldsymbol{\beta_{jm}}$	$m = 1, \ldots, 6$ intercept + predictors coefficients
Likelihood	Bernoulli distribution
Prior	$\beta_{jm} \sim$ Normal(0, 10) $\beta_{jm} \sim$ Beta(2, 20)
iterations	200,000
burn-in	10,000
thinning	100

Table 19. *Likelihood, choices of prior & MCMC practical considerations*

⋆ `Note`: In `rjags`, the `dnorm` function is used to fit a prior Normal distribution, where the arguments are mean and precision. The **precision** is equivalent to an inverted variance, $\boldsymbol{\tau = 1/\sigma^2}$, implying that the smaller the variance, the greater the precision.

POSTERIOR, $\boldsymbol{p(\theta|y)}$

→ Beta distribution is a conjugate prior for the Bernoulli distribution. **Proof**:

Prior, $\boldsymbol{\theta \sim Beta(a, b)}$

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\Gamma(a,b)}, \quad 0 \leq \theta \leq 1$$

Likelihood, $\boldsymbol{y_i \sim Bernoulli(n, \theta)}$

$$p(y|\theta) = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}, \quad 0 \leq \theta \leq 1$$

Posterior,

$$\begin{aligned}\boldsymbol{p(\theta|y)} &= \frac{\boldsymbol{p(y|\theta)p(\theta)}}{\boldsymbol{p(y)}} \\ &= \frac{1}{p(y)\Gamma(a,b)}\theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n} \times \\ &\quad \theta^{a-1}(1-\theta)^{b-1} \\ &= \frac{1}{p(y)\Gamma(a,b)}\theta^{\sum_{i=1}^n y_i+a-1}(1-\theta)^{n-\sum_{i=1}^n y_i+b-1} \\ &\propto \underbrace{\theta^{(a+\sum y_i)-1}(1-\theta)^{(n-\sum y_i+b)-1}}_{Beta(a+\sum y_i, n+b-\sum y_i) \quad \text{kernel}}\end{aligned}$$

→ The posterior distribution via Bayes Theorem constructed from a Bernoulli likelihood and a Normal prior on the other hand does not have a closed form, as seen in **Eq. (2)** for $i = 1, \ldots, n$ patients; $j = 2, 3$ levels of outcome (MR and MS, respectively); $p = k, \ldots, 5$ independent variables.

2. Model selection using DIC

The best model with the ability to predict new data can be selected amongst the candidate models using the deviance information criterion (DIC),

$$p(\theta|y) = \prod_{i=1}^{n}\left(\frac{e^{\beta_{j0}+\beta_{j1}X_{i1}+\dots+\beta_{jp}X_{ip}}}{1+e^{\beta_{j0}+\beta_{j1}X_{i1}+\dots+\beta_{jp}X_{ip}}}\right)^{y_i}\left(1-\frac{e^{\beta_{j0}+\beta_{j1}X_{i1}+\dots+\beta_{jp}X_{ip}}}{1+e^{\beta_{j0}+\beta_{j1}X_{i1}+\dots+\beta_{jp}X_{ip}}}\right)^{1-y_i}$$
$$\times \prod_{i=1}^{p}\frac{1}{\sqrt{2\pi}\sigma_k} exp\left(-\frac{1}{2}\left(\frac{\beta_k-\mu_k}{\sigma_k}\right)^2\right) \qquad \textbf{(2)}$$

based on the Markov Chain Monte Carlo (MCMC) chain output. Comparison can be of nested models, or models with completely different predictor variables, generalised linear models with different link functions or hierarchical models with different numbers of levels. A key restriction is that the response variable must be of the same form in all models being compared. For example, DIC cannot be used to compare models in which the response variable is log-transformed, while in the other the response variable is not log-transformed [17].

DIC is similar to AIC used for model comparison in frequentist settings where it is also based on the trade-off between:

- **lack of fit** - increased number of parameters in the model almost always improves the goodness of the fit
- **lack of parsimony** - increased complexity is related to increased prediction error for new data

The DIC protects against over-fitting by penalising models with larger count of the number of unknown parameters in the model, known as the *number of effective parameters*. Smaller DIC values indicate better predictive ability for new data (of the same kind as that observed) predictions.

Computation of DIC

Let $D(\boldsymbol{y},\phi) = -2\text{log}p(\boldsymbol{y}|\phi)$, where ϕ is the vector of all the unknown parameters. Then the effective number of parameters an DIC are

$$\begin{aligned} p_D &= E_{\phi|\boldsymbol{y}}(D) - D(E_{\phi|\boldsymbol{y}}(\phi)) \\ &\approx \bar{D} - D(\bar{\phi}) \end{aligned}$$
$$\boldsymbol{DIC{=}\bar{D} + p_D = D(\bar{\phi}) + 2p_D}$$

If the absolute difference between the models is greater than 2, then the model with a smaller DIC value is a better model. In R, the `dic.sample` function can be used to find the DIC value for each model.

Unexpectedly, **Table 20** shows that the Intercept only or Null model with a Normal(0, 10) prior is the best model with the smallest DIC value. The second best model is the full model with the same prior as the best model, including both Type of Implant and Age group predictors. DIC of the second best model increases significantly (difference > 2) when dropping the Type of Implant variable, and

	Model	Prior	DIC
1	$\boldsymbol{\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3 + \beta_{j4}X_4 + \beta_{j5}X_5}$	**Normal(0,10)**	**383.2**
2	$\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2$	Normal(0,10)	383.9
3	$\beta_{j0} + \beta_{j3}X_3 + \beta_{j4}X_4 + \beta_{j5}X_5$	Normal(0,10)	404.3
4	$\boldsymbol{\beta_{j0}}$ **(Intercept only)**	**Normal(0,10)**	**275.2**
5	$\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3 + \beta_{j4}X_4 + \beta_{j5}X_5$	Beta(2,20)	510
6	$\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2$	Beta(2,20)	504.7
7	$\beta_{j0} + \beta_{j3}X_3 + \beta_{j4}X_4 + \beta_{j5}X_5$	Beta(2,20)	504.1
8	β_{j0} (Intercept only)	Beta(2,20)	457.7

Table 20. *DIC values of candidate models for j = 2, 3 levels (i.e. j = MR, MS levels).* 🟦 *best model ;* 🟥 *second best model*

decreases non-significantly (difference < 2) when dropping the Age group variable. This indicates that only Type of Implant significantly influence the prediction of the outcome, Type of condition based on the second best model and Model 2 is potentially a good enough model, without the inclusion of Age group.

Although DIC has determined that the best model for this data example is the null model, we will refer to the second best model for the sake of interpreting the relationship between Type of Implant and Type of condition variables in the next few sections.

3. MCMC diagnostics

Markov Chain Monte Carlo diagnostics (MCMC) for the β_{jm} parameters from the chosen model were performed to demonstrate that the estimates are based on a chain that have converged for all unknown parameters.

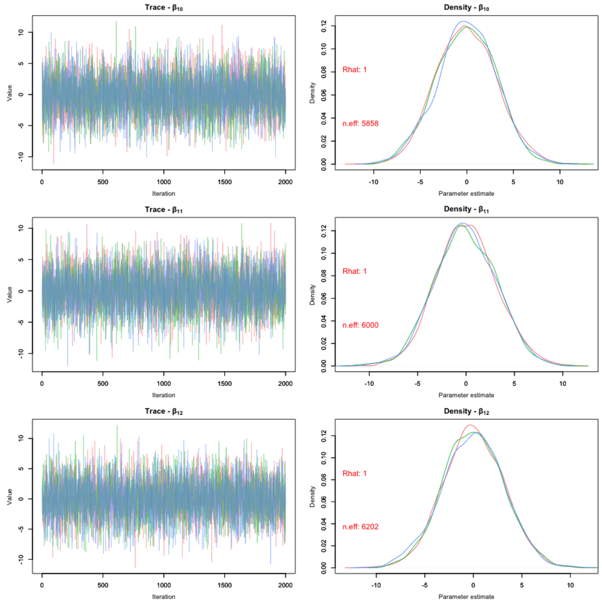


Figure 7. Trace plots and density plots for β_{10}, β_{12}, β_{13} posterior estimates of None category

There are a number of practical challenges in assessing MCMC reliability:

- ○ **Starting values** (optional) - A number of parallel chains with different starting values can be used to assess convergence informally. If chains with different starting values all converge to the same distribution, this maybe an indication that the chain does converge. Choice of starting value can affect the speed of convergence to the target posterior distribution and length of chain (number of iterations) required to achieve desired precision.
- ○ **Burn-in** - In practice, it is common to discard a certian number of the initial draws, known as burn-in. In doing so, the retained draws will become closer to the target distribution and less dependent on the stating points.
- ○ **Thinning** - In order to reduce autocorrelation caused by dependence between draws in the Markov chain, it it recommended to keep only every d^{th} draw of the chain after discarding burn-ins.

In this study, the `rjags` model was performed without specifying the starting values as this is optional. **Table 19** summarises the required parameters for the model and other MCMC practical considerations such as burn-in and thinning.

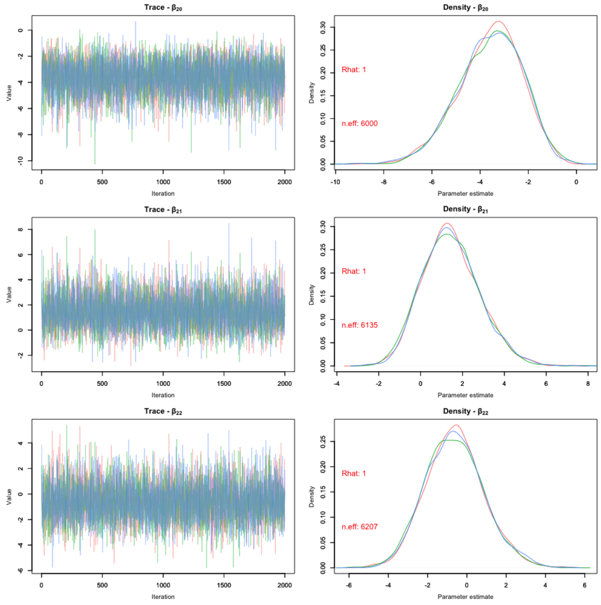


Figure 8. Trace plots and density plots for β_{20}, β_{21}, β_{22} posterior estimates of MR category

Figure 7, **Figure 8** and **Figure 9** show the trace plots on the right and the density plots on the left on each figure for β_{j0}, β_{j1} and β_{j2}. The trace and density plots for the rest of the β_{jm} parameters can be found in the appendices. The trace plots is only showing 2000 iterations which means these plots include thinning but exclude burn-ins. They look like a hairy caterpillars which means the chains explore the sample space freely. The density plots show each line (representing a chain) where these lines overlap to form one line, indicating the convergence to the target posterior distribution is reached. $\hat{R}$ for all parameters is lower than 1.2 so we can conclude that convergence to the target distribution is achieved (i.e. all values in the chains come from the same distribution).

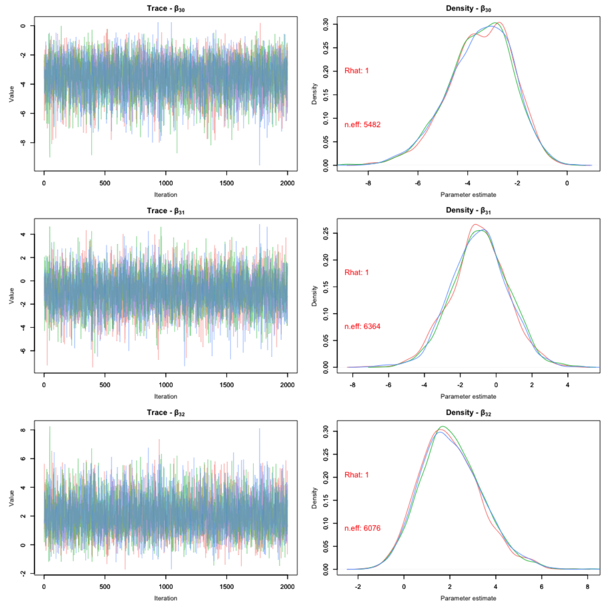


Figure 9. *Trace plots and density plots for β_{30}, β_{31}, β_{32} posterior estimates of MS category*

The effective sample sizes S_{eff} as shown in the above figures is the number of independent Monte Carlo (MC) samples necessary to observe the same precision as the MCMC samples. For example, **Table 21** shows $S_{eff} = 6000 < S = 200,000$ for β_{20}, implying high auto-correlation between draws.

4. Posterior predictive checking for lack of fit of the sampling model

If an assumed model fits, then simulated data using the model should look similar to the observed data and the observed data should look plausible under the posterior predictive distribution. We can check this by drawing replicated data values from the posterior predictive distribution and compare replicates to observe data using a discrepancy measure (denoted as T). In this section, we demonstrate the use of posterior predictive checks for evaluating how well the sampling model fits the data. The Bernoulli distribution has a Mean = θ and Variance = $\theta^2(1-\theta)$. The discrepancy measure is then

$$T = \frac{\theta}{\theta^2(1-\theta)} = \frac{1}{\theta}, \quad 0 \le \theta \le 1$$

$T > 1$ is expected if Bernoulli is true. In this example, the replicated data is then generated from the posterior predictive distribution constructed from multiplying the prior distribution and posterior distribution together.

Posterior $\propto$ Likelihood $\times$ Prior

$$\propto Beta\left(a + \sum_{i=1,j=1}^{n} y_{ij}, n + b - \sum_{i=1,j=1}^{n} y_{ij}\right)$$

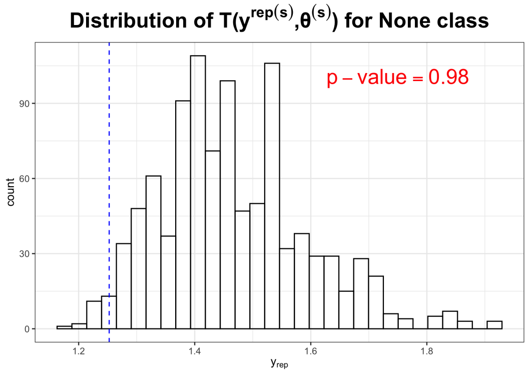


Figure 10. *Posterior predictive check for None class model using MCMC method (S = 1000 replicates)*

A *posterior predictive p-value* is the tail posterior probability for the T statistic generated from the model compared to the statistic observed in the data. A model is suspected if the `p-value` is near 0 or 1, typically <0.01 or >0.99 and a `p-value` between 0.05 to 0.95 indicates a good fit.

Figure 10, **Figure 11** and **Figure 12** plot the posterior predictive distribution of the replicates, $T(y^{rep(s)}, \theta^{(s)})$ (bar plot) and the observed discrepancy measure, $T(\boldsymbol{y}, \theta^{(s)})$ (dotted blue line) for S = 1000 replicates and $\theta^{(s)}$ is the probability of success generated from the Beta posterior distribution. The observed T falls in the most frequent bar of the posterior predictive distribution where $T(y^{rep(s)}, \theta^{(s)} > 1$ for class MR and class MS models. Moreover, the `p-value` for these two models is approximately between 0.5 and 0.7, implying the Bernoulli model is a good fit for these two classes. Class None model on the other hand has a `p-value` close to 1, hence some improvement could be done to improve the fit of the model.

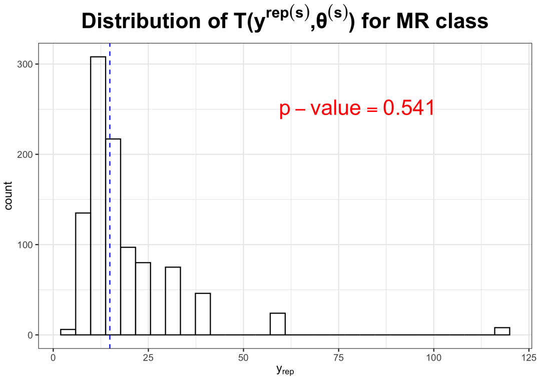


Figure 11. *Posterior predictive check for MR class model using MCMC method (S = 1000 replicates)*

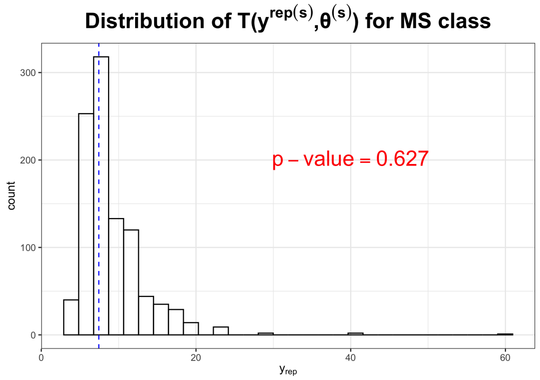


Figure 12. *Posterior predictive check for MS class model using MCMC method (S = 1000 replicates)*

5. *ROC-AUC curves for the multinomial logistic regression model via Bayesian approach*

Next we compute the ROC-AUC curves for “None", “MR" and “MS" classes for the Type of Condition outcome using a confusion matrix. The steps can be documented as follows,

1 Convert the predictors from the test data to dummy variables. For $i = 1, 2, \dots, m$, $m = 24$ observations, $j = 1, 2, 3$ levels corresponding to None, MR and MS, multiply the dummy predictors and the posterior means of β parameters together, i.e.

$$\hat{\pi}_j = \begin{pmatrix} X_{i1} & X_{i2} & X_{i3} & X_{i4} & X_{i5} \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{m1} & X_{m2} & X_{m3} & X_{m4} & X_{m5} \end{pmatrix} \begin{pmatrix} \beta_{j1} \\ \beta_{j2} \\ \beta_{j3} \\ \beta_{j4} \\ \beta_{j5} \end{pmatrix}$$

As a results, three $m \times 1$ matrices of the obtained probabilities for each of the three classes were produced.

2 Combine the three matrices together, assign an observation to the class with the highest probability, i.e.

$$\begin{pmatrix} \text{None} & \text{MR} & \text{MS} \\ 0.03 & 0.02 & 0.93 \\ 0.24 & 0.01 & 0.11 \\ \vdots & \vdots & \vdots \\ \pi_{m1} & \pi_{m2} & \pi_{m3} \end{pmatrix} \rightarrow \begin{pmatrix} \text{None} & \text{MR} & \text{MS} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & x_{m3} \end{pmatrix}$$

3 Compare this to the outcome variable from the test data and produce a confusing matrix.

	A	*B*	*C*	**TP**	**TN**	**FP**	**FN**
A	0	1	10	0	11	11	2
B	0	4	2	4	11	2	1
C	2	0	5	5	5	2	12

Table 21. *Confusion matrix with TP, TN, FP & FN for each class*

This method of calculating the confusing matrix is however not suitable to apply to the null model (i.e. the best model) for $\pi_j = \frac{exp(\beta_{j0})}{1+exp(\beta_{j0})}$. As the null model indicates that the test data won't have any influence on π_j, i.e. π_j will be the same across all row hence resulting in a ROC curve with AUC = 0.5 (classifying all cases as one class and ignoring the other two classes).

Metrics	Macro-avg	Micro-avg
○ NPV or	68.6%	64.3%
Specificity	68.7%	64.3%
○ Precision or PPV	46.0%	37.5%
○ Recall or Sensitivity	36.5%	37.5%
○ F1-score	41.3%	37.5%

Table 22. *Key performance metrics of the second best model (Model 1)*

Accuracy for the second best model is extremely low for a value of 37.5%. **Table 22** describes the performance metrics as maro-average and micro-average. Both results show that F1-score is also low due to a relatively high specificity/low sensitivity (recall) and small precision values. In terms of micro-average, this is means it is good to identify around 64.3% patients who do not belong to a class (but belongs to either of the other two classes) but not so good when it comes to identifying which class all patients belong to.

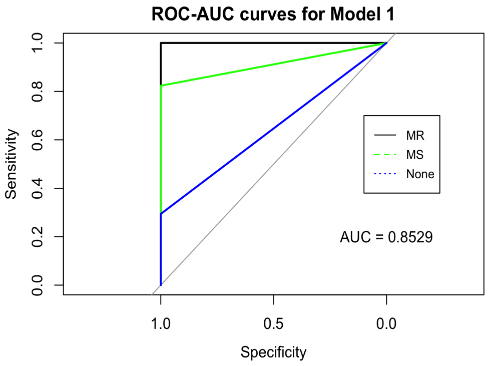


Figure 13. *ROC curves for the second best model (Model 1) with AUC = 0.8529*

The ROC curves for the Bayesian model perform particularly well but AUC is still smaller than that for the Random Forest method with an AUC of 0.8627. Similar to Random Forest, the model is able to classify patients with MR condition to the other patients with high accuracy (black line) and classify patients with None condition to the other patients with low accuracy (blue line).

4.2.2 Results of the chosen model

The results from fitting a multinomial logistic model for the multi-class Type of Condtion outcome (denoted as y) having class “None" as the baseline using the information from **Table 19**, **Item (i)** and **Item (ii)**, are presented in this section. Since this paper focuses mainly on the association between Type of implant and Type of condition in patients with heart disease, only the interpretations of the results with respect to Type of implant predictor and ignore that for Age are documented.

Table 23 reports the posterior means of the β parameters and their highest posterior density lower (HDPL) and upper (HDPU) intervals [3]. This is also known as the credible interval, providing a probabilistic statement in terms of parameter β_{jm}. For example, the credible interval of β_{21} is approximately (-1.15, 4.11). There is a 95% chance that β_{21} falls between -1.15 and 4.11, given that we observed 8 patients with MR condition in a random sample size of 119.

The log odds of having MR versus not having neither condition increases by 1.45 if moving from no

	Mean	**sd**	**95% HPDL**	**95% HPDU**	$\hat{R}$	S_{eff}
β_{20}	-3.59	1.33	-6.16	-1.07	1	6000
β_{21}	1.45	1.37	-1.15	4.11	1	6135
β_{22}	-0.62	1.49	-3.48	2.32	1	6207
β_{23}	1.13	0.83	-0.50	2.70	1	6000
β_{24}	-2.54	2.15	-6.86	1.33	1	6000
β_{25}	-1.46	2.39	-6.20	2.94	1	6000
β_{30}	-3.50	1.30	-6.06	-1.09	1	5482
β_{31}	-0.93	1.59	-4.06	2.12	1	6364
β_{32}	2.05	1.33	-0.46	4.65	1	6076
β_{33}	1.17	0.63	-0.05	2.42	1	6151
β_{34}	-3.93	1.80	-7.71	-0.89	1	5955
β_{35}	-2.92	1.99	-6.81	0.67	1	6000
OR (β_{21})	13.08	42.27	0.09	46.22	1.03	6000
OR (β_{22})	1.92	8.73	0.00	6.88	1.12	5445

Table 23. *Posterior means of β parameters and 95% credible intervals for MR and MS clasess; posterior mean of MR condition odds ratios and 95% credible intervals*

implant to Band implant. The log odds of having MR versus not having neither condition decreases by 0.62 if moving from no implant to Ring implant.

$$\log\left(\frac{Pr(y = MR)}{Pr(y = None)}\right) = -3.59 + 1.45X_1 - 0.62X_2 + 1.13X_3 - 2.54X_4 - 1.46X_5$$

The log odds of having MS versus not having neither condition decreases by 0.93 if moving from no implant to Band implant. The log odds of having MR versus not having neither condition increases by 2.05 if moving from no implant to Ring implant.

$$\log\left(\frac{Pr(y = MS)}{Pr(y = None)}\right) = -3.50 - 0.93X_1 + 2.05X_2 + 1.17X_3 - 3.93X_4 - 2.92X_5$$

The posterior mean and 95% quantile-based intervals for the odds ratio for Band implant is 13.08 (0.09, 46.22). This indicates that Band implant is associated with higher odds of a patient exposed to MR condition compared to not having a Band implant. On the contrary, the posterior mean and 95% quantile-based intervals for the odds ratio of Ring implant is 1.92 (0, 6.88). Since the 95% interval include 0, the odds ratio for Ring Implant is not significant hence Ring implant is not associated to the clinical outcome of MR.

5 Summary and Key results

In order to determine whether there is a significant relationship between the clinical outcome of mitral regurgitation (MR) relating to the long term side effect of annuloplasty Band in 18-year-old patients, we experimented three different statistical methods that could be used for analysing multi-class response variable. These methods are *contingency table analysis*, *Machine learning Random Forest classifier* and *Multinomial Regression using a Bayesian approach*.

Contingency table analysis relies on the calculation of *odds*, *odds ratio* and *relative risks* to describe the strength of the association between two categorical variables, using count data. **Section 2.2** explains how to do so using a *3x3* contingency table, in the context of *marginal*, *joint* and *conditional probabilities*. The results from this section show that there exists a strong relationship between patients with Band implant and the chance of having MR condition, with the odds of having MR condition for those with Band implant are estimated to be 8.1 times higher than that for those with Ring implant.

The second method introduced is a machine learning algorithm, *Random Forest* classification for multi-class problems. Data was split into training and test data prior to fitting the model, in order to check if this model would still be relevant for predicting new data, i.e. test data. The data however is highly imbalanced, as seen in **Table 1**, where the majority of the observations belong to "None" class. For this reason, *oversampling* and undersampling methods were used to make the data less imbalanced. We then obtained the *3x3 confusion matrices* to calculate (manually) and compare the performance metrics for oversampled, undersampled and imbalanced data. The purpose of the performance metrics was to evaluate how well the model fits the data, looking at *accuracy* (percentage of number of correctly classified instances among all other instances) and *F1-score* (the harmonic mean of *precision* and recall taking both metrics into account). *Receiver Operating Characteristic (ROC)* curve and *Area under the ROC curve (AUC)* were computed using the One vs All methodology to describe the *sensitivity* and *specificity* trade offs. We found that random forest fitted both undersampled and oversampled data better (higher F1-score and higher AUC). On the contrary, the same model used for imbalanced data was under-performing and had no discrimination ability with only an AUC of 0.5. Lastly, we looked at the significance for **Type of implant** and **Age group** predictors in the model using *Mean Decrease Accuracy* and *Mean Decrease Gini* measures. The reported measures showed that both predictors have a marked influence on predicting the outcome and hence should be included in the final model.

Section 4.1 introduces Bayes' theorem and outlines the *prior probability* calculated under the Bayesian framework, which is essentially the same as the conditional probability calculated from the contingency table analysis. Using this conditional probability, the results can then be interpreted as a probabilistic statement. In particular, the chance that the patient diagnosed with MR condition is 0.176, given that this patient has a Band implant. Likewise, the chance that the patient diagnosed with MR condition is 0.026, given that this patient has a Ring implant. It is then safe to say that the chance that the patient with a Band implant to be diagnosed with MR condition is higher than the chance that the same patient with a Ring implant to be diagnosed with a Ring implant.

Section 4.2 further analyses the data by fitting several multinomial logistic regression models under the Bayesian framework on training data. In particular, these models were constructed using a Bernoulli likelihood (or sampling model) and two choices of non-informative prior distributions, Normal(0,10) and Beta(2,20). The best model was then selected amongst a set of candidate model based on *Deviance Information Criterion (DIC)* values, where the model with the smallest DIC value is considered as the best fit to the data. An advantage of DIC over other criteria in the case of Bayesian model selection is that the DIC is easily calculated from the samples generated by a *Markov chain Monte Carlo (MCMC)* simulation. Although candidate models are penalised by both $\bar{D}$ and p_D, DIC tends to select over-fitted ones. The full model using the Normal(0, 10) with two predictor variables, Type of Implant and Age group, was reported to be the second best model and the null model with the same prior was reported to be the best model. The latter means neither predictor has an impact on the outcome variable. For the purpose of demonstrating the calculation of the odds ratio of Type of Implant variable and how it relates to Type of condition outcome, we referred to the second best model to interpret the final results.

The results from the second best model found that the odds of having MR condition is higher in patients with Band implants, in comparison to those with a Ring implant. This key finding is consistent with the results obtained from the multinomial logistic regression model constructed in a frequentist sense [22]. We also performed MCMC diagnostics of the estimates for the unknown β_{jm} parameters to check for convergence to the target distribution. Next, assess to the goodness-of-fit of the sampling model was looked into using *posterior predictive checks* method, by generating replicates from the Beta posterior distribution and compare these to the observed data, basing on discrepancy measure. Finally, we set up a number of algorithms in order to derive the confusion matrix from the test data. In doing so, we were able to obtain the ROC-AUC curves for the second best model. The results found that the AUC for this model (AUC = 0.8529) is lower than that for the Random forest classification (AUC = 0.8627) The former also has significantly low accuracy and F1-score.

6 Conclusion and Future Research Direction

All 3 methods (contingency table, Bayes multiple logistic regression for multiclass outcome based on the second best model, and random forest) concluded the same results which answer to the research question. That is, there exists a strong relationship between annuloplasty Band Implant and the clinical outcome of Mitral Regurgitation (MR) in patients with Mitral Valve disease in a postive manner, i.e. patients with Band implant are more likely to be exposed to MR condition. However, since DIC selected the null model as the most fitted (possibly due to the small sample size of the training set, $n_{train} = 95$, we can't simply rely on the results of the second best model with confidence. In the future, it is highly desirable to obtain R^2 statistics and other criteria beside DIC to compare among candidate models, such as AIC and BIC. Moreover, we wish to further investigate the following:

1. Produce ROC-AUC graphs for the `rjags` mulinomila logistic regression models ☑

This study currently does not have the ability to compare the performance between multinomial regression logistic model under the Frequentist framework, multinomial regression logistic model via Bayesian approach and Random Forest classification. In the near future, we wish to obtain ROC curves and AUC measures for the latter and last models.

2. Fit the Bayes multinomial logistic model using other priors other than the current priors □

It may be appropriate to fit the Bayesian models using different values of σ^2 a, and b for the Normal(0,σ^2) and Beta(a,b) prior distributions as this can change the model performance drastically.

3. Perform Random Forest classifier in Python ☑

When it comes to machine learning algorithms, both R and Python have their own advantages. Still, Python seems to be more flexible and cover more desirable features necessary for improving the performance of the model. Implementing the Random Forest classifier in Python helps increase the options for adjusting the threshold of the ROC curves and tuning the model parameters.

4. Investigate other independent variables in the data that might relate to the outcome variable and include interaction terms in the model □

It is important to examine other factors in the data that might have an influence on the outcome variable and improve the regression models' performances overall. By taking other factors into account, we can understand better what might cause the relationship between patients with and without the implantation of an annuloplasty device and the long term side effect of Mitral Regurgitation (MR) to change in a positive or a negative manner. This should be followed by testing the correlation between features if more variables were to be included in the model.

Acknowledgement

I would like to sincerely thank **Dr. Hung Dung Van** for providing the data, materials and valuable technical support much needed for this research work. Additionally, I appreciate his useful advice and assistance in keeping my progress on schedule.

References

[1] Ryan Admiraal. *STAT452 Bayesian Inference Lecture*. School of Mathematics, Statistics and Operations Research, Victoria University of Wellington.

[2] Alan Agresti. *An Introduction to Categorical Data Analysis, Second Edition*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2007. doi:10.1002/0470114754.

[3] Alan Agresti and Yongyi Min. Frequentist performance of bayesian confidence intervals for comparing proportions in 2 x 2 contingency tables. *Biometrics*, 61:515–523, June 2005. doi:10.1111/j.1541-0420.2005.031228.x.

[4] Geráld Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13, 2012. URL: https://www.jmlr.org/papers/volume13/biau12a/biau12a.pd.

[5] Brendon J. Brewer. *Introduction to Bayesian Statistics*. URL: https://www.stat.auckland.ac.nz/~brewer/stats331.pdf.

[6] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian data analysis*. third edition.

[7] Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin Cawley. Model selection: Beyond the bayesian/frequentist divide. *Journal of Machine Learning Research*, 11, 2010. URL: https://www.jmlr.org/papers/volume11/guyon10a/guyon10a.pdf.

[8] James A. Hanley and McNeil Barbara J. The meaning and use of the area under a receiver operating characteristic (roc) curve. April 1982. URL: http://www.med.mcgill.ca/epidemiology/hanley/software/Hanley_McNeil_Radiology_82.pdf.

[9] Mateusz Lango. Tackling the problem of class imbalance in multi-class sentiment classification: An experimental study. *Foundation of computing and decision sciences*, 44, 2019. doi:10.2478/fcds-2019-0009.

[10] Minh Chau Van Nguyen. Finite mixuture model-based approach for clustering linguistics data. February 2020. URL: https://drive.google.com/file/d/1QBTAUTxXp_C3B6fx2Ah9_81BhHMoNgPf/view.

[11] Anna Pernesta and Nyberg Mattias. Using prior information in bayesian inference – with application to fault diagnosis. January 2007. doi:10.1063/1.2821293.

[12] Martyn Plummer. Jags version 3.4.0 user manual, 2013. URL: http://www.stats.ox.ac.uk/~nicholls/MScMCMC15/jags_user_manual.pdf.

[13] Thomas Ratschiller, Thomas Guenther, Ralf Guenzinger, Christian Noebauer, Victoria Kehl Dr. rer. nat., Ralph Gertler, and Ruediger Lange. Early experiences with a new three-dimensional annuloplasty ring for the treatment of functional tricuspid regurgitation. *Ann Thorac Surg*, 98:2039–2044, 2014. URL: https://www.annalsthoracicsurgery.org/article/S0003-4975(14)01520-3/pdf.

[14] R.H. Riffenburgh. Chapter 17 - bayesian statistics. In R.H. Riffenburgh, editor, *Statistics in Medicine*, pages 355–364. Academic Press, San Diego, third edition, 2012. URL: http://www.sciencedirect.com/science/article/pii/B9780123848642000172, doi:https://doi.org/10.1016/B978-0-12-384864-2.00017-2.

[15] Rens Schoot, David Kaplan, Jaap Denissen, Jens Asendorpf, Franz Neyer, and Marcel Aken. A gentle introduction to bayesian analysis: Applications to developmental research. *Child development*, 85, 10 2013. doi:10.1111/cdev.12169.

[16] Nokuthaba Sibanda. *STAT432 Computation Statistics Part I*. School of Mathematics, Statistics and Operations Research, Victoria University of Wellington.

[17] Nokuthaba Sibanda. *STAT452 Bayesian Inference Lecture*. School of Mathematics, Statistics and Operations Research, Victoria University of Wellington.

[18] David J. Spiegelhalter, Best Nicola G., Carlin Bradley P., and Angelika van der Linde. Bayesian measures of model complexity and fit. *Royal Statistical Society*, 64, 2002. doi:https://doi.org/10.1111/1467-9868.00353.

[19] Lang Sun, Lina Tang, Guofan Shao, Quanyi Qiu, Ting Lan, and Jinyuan Shao. A machine learning-based classification system for urban built-up areas using multiple classifiers and data sources. *Remote Sensing*, 2019.

[20] E. M.M van der Heide, Roel F. Veerkamp, M. L. van Pelt, Claudia Kamphuis, Ioannis N. Athanasiadis, and Bart J. Ducro. Comparing regression, naive bayes, and random forest methods in the prediction of individual survival to second lactation in holstein cattle. *Journal of Dairy Science*, 102. doi:https://doi.org/10.3168/jds.2019-16295.

[21] P.H.A.J.M Van Gelder and Noel van Erp. Bayesian logistic regression analysis. August 2013. doi:10.1063/1.4819994.

[22] Minh Chau Van Nguyen. Icmr* in patients under 18 years old. July 2020. URL: https://minhchauvannguyen.github.io/ICMR/Report.html.

Appendices

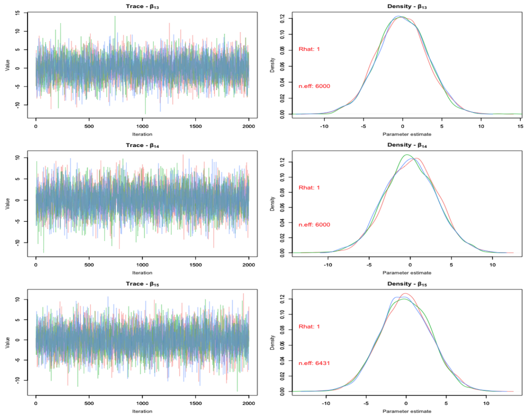


Figure 14. *Trace plots and density plots for β_{13}, β_{14}, β_{15} posterior estimates of None category*

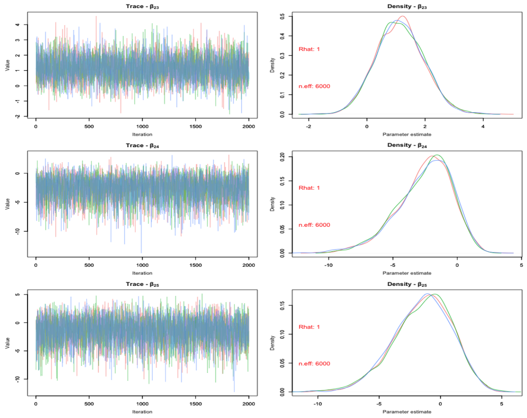


Figure 15. *Trace plots and density plots for β_{23}, β_{24}, β_{25} posterior estimates of MR category*

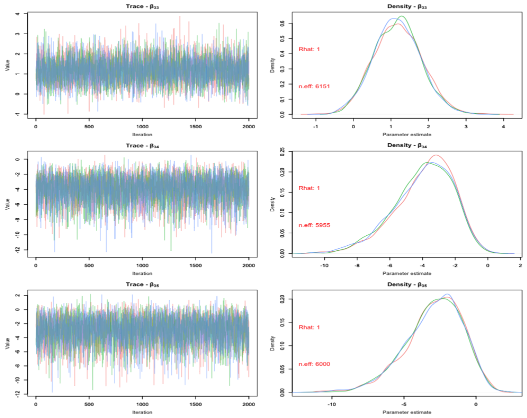


Figure 16. *Trace plots and density plots for β_{33}, β_{34}, β_{35} posterior estimates of MS category*

Table 24. *Report key research categories and subcategories*

Descriptive Statistics	
Summary table	• Contingency table
Statistical Inference	
Specific tests	• Chi-square • Fisher Exact's test • Model selection • DIC
Frequentist Inference	• Likelihood function
Bayesian Inference	• Bayesian Probability (prior, posterior) • credible interval
Regression Analysis	
Generalize Linear Model (GLM)	• Multiple multinomial logistic regression
Statistical Classification	
Multi-class classification	• Machine Learning • Random Forest
Application	
Biostatistics	• Medical statistics

Source from https://en.wikipedia.org/wiki/Bayesian_inference