HEART FAILURE PREDICTION USING A BAYESIAN APPROACH

Tiasha Ali, Xunhang (David) Gao, Fezile Mhlabane

April 2025

Contents

1 Introduction

Motivation & Scientific Question

Cardiovascular diseases (CVDs) are still the top cause of death around the world, accounting for about one-third of all deaths each year. Many of these deaths, especially from heart failure, happen to people under 70 and could mostly be avoided with the right help in time. In the year 2021 alone, cardiovascular diseases accounted for 931,578 deaths in the United States, and that is only a fraction of the deaths it accounted for worldwide. Reports also state that heart disease claimed more lives in 2021 in the United States than all forms of cancer and chronic lower respiratory diseases combined, hence the undeniable importance to predict heart disease accurately using common clinical measures and symptoms so that susceptible patients can be timely located and preventive measures could be applied [1]. The main question in this study is: How can we use regular clinical features to effectively predict heart failure using a Bayesian approach?

Background

Heart failure and other forms of cardiovascular disease are often associated with modifiable risk factors such as hypertension, diabetes, hyperlipidemia, and lifestyle behaviors. Recent statistical analysis has shown a promising improvement in diagnostic accuracy. The Bayesian methods allow us to incorporate prior clinical knowledge into the analysis and continuously update the prediction as new data become available.

2 Data Processing & Data Description

HEART FAILURE PREDICTION DATASET:
One potential data source is Heart Failure Prediction Dataset [2]. This study consists of clinical records for individuals screened for heart disease. The target variable is "Heart Disease", indicating the presence (1) or absence (0) of heart disease. The predictors include demographic characteristics, such as "Age" and "Sex", lifestyle and symptom indicators, such as "Chest Pain Type" and "Exercise-included Angina", laboratory results, such as "serum cholesterol" and "fasting blood sugar", and physiological measurements, such as "measure of ST depression". The dataset is well-structured, with no missing values, making it suitable for statistical modeling.

This dataset consists of eleven explanatory variables (clinical features) of which five are numerical and six are categorical to predict heart disease where 1 indicates presence of heart disease and 0 indicates absence of heart disease. It has a total of 918 observations and none of the variables in the dataset have missing values [2]. 508 of the patients have heart disease and 410 do not.

For the numerical variables we have: Age, RestingBP, Cholesterol, MaxHR, Oldpeak (ST depression), in which all are supported by the positive real line with the exception of ST depression which has its support on the real line. A healthy heart has low levels of total cholesterol - less than 200 mg/dL and Resting Blood pressure of 120/80 mmHG. Our data shows the systolic blood pressure, i.e; the pressure in the arteries when the heart beats and pumps blood into them. An ST depression is one of the outcomes of an electrocardiogram (ECG) test, a test for measuring the electrical activity in a person's heart. The normal ST segment is usually isoelectric (i.e., flat on the baseline, neither positive nor negative), but it may be slightly elevated or depressed normally (usually by less than 1 mm).

For the categorical clinical features we have: Sex [M:Male/F:Female], Chest Pain Type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic], Fasting blood sugar level where value is 1 if FastingBS $>$ 120 mg/dL and 0 otherwise, RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria], ExerciseAngina: exercise-induced angina [Y: Yes, N: No] and ST-Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]. Normal Fasting blood sugar is 70 to 99 mg/DL, Prediabetes: 100 to 125 mg/dL and Diabetes: 126 mg/dL or higher. A horizontal or downsloping ST depression is often associated with Myocardial ischemia - a condition where the heart muscle is not getting enough oxygen.

2.1 Exploratory Analysis

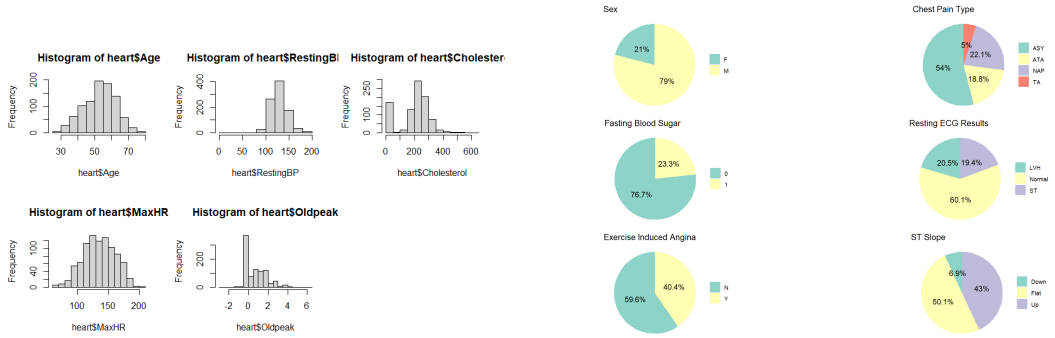


Figure 1: Distribution of numerical (left) and categorical (right) variables in dataset [2]

We see that Age distribution is roughly symmetric, hence can be modeled using the normal distribution. RestingBP is also symmetric with a mean = 132.39651 $\approx$ median = 130. However, it does show to have some outliers. Thus, we could preferably go for a prior with heavier tails such as the t-distribution to allow for a wider rage of possible values. MaxHR also appears to be symmetric and can be modeled using the normal distribution. Cholesterol appears to be bimodal, with a count of 172 for the value zero. We might therefore consider using mixtures of distributions to model it. That is, look in the lines of Zero-inflated Gaussian Mixed Models (ZIGMMs) [3], where

$$y_i = \begin{cases} 0, & \text{with probability } \ p \\ N(y_i|\mu, \sigma^2), & \text{with probability } \ 1-p \end{cases}$$

Oldpeak ST depression is more right skewed. Again, we have a lot of zero values (368) and hence we may consider the zero inflated mixed models here as well.

From the correlation matrix, we see that the features are weakly correlated hence there are no concerns for multi-collinearity between variables.

Correlations

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
Age	1.0000	0.2544	-0.0953	-0.3820	0.2586
RestingBP	0.2544	1.0000	0.1009	-0.1121	0.1648
Cholesterol	-0.0953	0.1009	1.0000	0.2358	0.0501
MaxHR	-0.3820	-0.1121	0.2358	1.0000	-0.1607
Oldpeak	0.2586	0.1648	0.0501	-0.1607	1.0000

Figure 2: Correlation matrix of variables in the dataset [2]

The classes for the categorical data show noticeable imbalance with the exception of ExerciseAngina. For example, from the pie chart 1 we can see that 79% are male. Regarding chest pain type, 54% experienced asymptomatic pain, followed by 22.1% with atypical angina, 18.8% with non-anginal pain, and 5% with typical angina. Fasting blood sugar levels were elevated ($\geq$ 120 mg/dl) in 23.3% of patients, while the rest had normal levels. Resting ECG results indicated that 60.1% had normal readings. Regarding the ST slope, only 6.9% had a downsloping slope. This may lead to low accuracy and we may need to adjust the classification cut-off to better suit the goals of our analysis.

Sampling Model

The Bernoulli probability model is an appropriate model structure for data Y_i [0/1 for disease or no disease].

$$Y_i = \text{Bern}(\pi_i) \qquad \text{with} \qquad E[Y_i|\pi_i] = \pi_i$$

π_i depends on the predictors Xi's, through the logit link function $g(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$. i.e;

$$Y_i|\beta_0, \beta_1 \cdots, \beta_p \sim^{ind} \text{Bern}(\pi_i) \qquad \text{with} \qquad \log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

2.2 Literature Review

Machine Learning based classification of presence or absence of heart disease have been employed before. Uddin et al looked at Machine Learning - based approach to diagnosis of cardiovascular vascular disease using a combined [4]. Their proposed method predicts the chances of heart disease and classifies patient's risk level by using different Machine Learning (ML) algorithm techniques: Support Vector Machine, Random Forest, Decision Tree, K-nearest neighbor (KNN), Multilayer perception neural networks (MLP), Extreme Gradient boost, Gradient boosting, Light gradient boosting machine, Extreme tree classifier. Accuracy and precision was generally high with the lowest being KNN model with 0.982079 for accuracy and 0.982014 for precision. The highest was the decision tree model at an accuracy of 0.991637 and a precision of 0.995122.

On the other hand Gupta et al looked at employing Naive Bayes for heart disease classification, which is outlined on the paper: Heart Disease classification (Naive Bayes) [5]. Here the Naive Bayes model was found to have better performances when compared with with most of the other machine learning models implemented. Naive Bayes and random forest showed improvement in performance when feature selection was applied. As a result Naive Bayes ended up producing an accuracy of 88.16%.

Subarkah et al also looked the Naive Bayes classifier and Correlated Naive Bayes classifier for heart disease classification on their paper: Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classification [6]. Here, the Correlated Naive Bayes Classifier (C-NBC) algorithm had an accuracy of 80.6% which was higher accuracy than the Naive Bayes Classifier (NBC) algorithm at only 67.5%

3 Model Proposal

3.1 Variable selection

Before setting up priors, we conducted variable selection based on the logistic regression model to identify statistically significant variables influencing the occurrence of heart disease, using the likelihood ratio test. First, we considered the full model, including all 11 explanatory variables. From result of likelihood ratio test shown in Table 1, `RestingBP` and `RestingECG` are identified an non-significant variables. Then, we considered the reduced model excluding `RestingBP` and `RestingECG`. And the result from the likelihood ratio test in Table 2, all of the 9 variables in reduced model are identified as significant. Thus, we will use this reduced logistic regression model for further analysis.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
Age	1	75.40	916	1186.74	0.0000
Sex	1	85.16	915	1101.58	0.0000
ChestPainType	3	219.58	912	882.00	0.0000
RestingBP	1	2.08	911	879.92	0.1496
Cholesterol	1	18.85	910	861.07	0.0000
FastingBS	1	20.62	909	840.46	0.0000
RestingECG	2	0.00	907	840.45	0.9982
MaxHR	1	23.74	906	816.71	0.0000
ExerciseAngina	1	67.08	905	749.63	0.0000
Oldpeak	1	39.10	904	710.54	0.0000
ST_Slope	2	116.35	902	594.19	0.0000

Table 1: From this table, `RestingBP` and `RestingECG` are identified as non-significant to heart disease.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
Age	1	75.40	916	1186.74	0.0000
Sex	1	85.16	915	1101.58	0.0000
ChestPainType	3	219.58	912	882.00	0.0000
Cholesterol	1	16.76	911	865.24	0.0000
FastingBS	1	21.55	910	843.68	0.0000
MaxHR	1	23.64	909	820.05	0.0000
ExerciseAngina	1	67.89	908	752.16	0.0000
Oldpeak	1	40.03	907	712.12	0.0000
ST_Slope	2	116.84	905	595.28	0.0000

Table 2: From this table, all of the variables in reduced model are identified as significant.

3.2 Logistic Regression Model

Pooling approach: logistic regression model of Y by a set of predictors X:

$$Y_i|\beta_0, \beta_1 \cdots, \beta_9 \sim^{ind} \text{Bern}(\pi_i) \qquad \text{with} \qquad \log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_9 X_{i9}$$

We use weakly informative priors because of the limited information we have on the variables. Variances on the β coefficients are chosen to be functions of the corresponding variances of variables (for numerical variables only). Same function transformation was used in all the variables. For the categorical variables, the variance was estimated based on how the proportions varies for the classes in that variable:

INTERCEPT: $\beta_0 \sim N(-3, 0.5)$ Since on average 5% of the population has heart disease i.e $\pi \approx 0.05$. We set

the prior mean for β_0 to be given by the log (odds of Heart Disease prevalence (5%)). i.e

$$\log(\frac{\pi}{1-\pi}) = \log(\frac{0.05}{1-0.05}) = -2.944 \approx -3$$

Since we have limited information about the spread so we choose a vague value of 0.5 for variance.

AGE: $\beta_1 \sim N(0, 7)$

CHOLESTEROL: $\beta_2 \sim N(0.18, 14)$
Based on a study conducted in Madrid Spain 2017, 50% of participants who had ideal health, had significant plaque build up in their arteries (an early sign of heart disease) [7]. Data collected from 136 905 patients hospitalized for heart attack between 2000 and 2006 nationwide - America, 54.6% of patients had high density lipoprotein (HDL) cholesterol levels below 40 mg/dl, which is characterized as poor [8]. Thus we set the prior mean for β_3 to be log(odds), i.e;

$$\log(\frac{0.546}{1-0.546}) \approx 0.18$$

MAXHR: $\beta_3 \sim N(0, 11.7)$

OLDPEAK: $\beta_4 \sim N(0, 1)$
A threshold value of 0.2 mV for males more than 40 years old should be considered significant ST elevation. For adult females, the value is 0.15 mV. For both males and females the J-point depression threshold values are -0.1 mV [9]. Hence ST depression for a healthy person follows a generally symmetric distribution centered around zero.

SEX: $\beta_5 \sim N(0, 8.41)$

FASTINGBS: $\beta_6 \sim N(0, 7.3)$
Researchers found that people with elevated blood sugar levels have a 30–50% increased risk of developing heart disease, even if their blood sugar levels are below the threshold for diabetes [10]. Hence mean for β_7 is approximated as:

$$\log(\frac{0.5}{1-0.5}) = 0$$

EXERCISE ANGINA: $\beta_7 \sim N(0, 0.9)$

CHEST PAIN: $\beta_8 \sim N(0, 3.22)$

ST-SLOPE: $\beta_9 \sim N(0, 3.6)$

3.3 Hierarchical Model

We choose Sex as a grouping variable. The motive being that certain medical attributes are associated to one gender group based on hormone composition, physique etc. As a result, we are able to access the degree to which gender introduce variability in the presence/occurrence of Heart Disease between genders. Scientists found that men with increased blood sugar levels below the threshold for diabetes had a 30% greater risk of developing cardiovascular disease. However, women with the same levels had a great risk of between 30–50%. A threshold value of 0.25 mV for males less than 40 years old should be considered significant ST elevation. For males 40 years old and older, this value is 0.2 mV. For adult females, the value is 0.15 mV [9]. Also males are generally more susceptible to heart disease than females [1]. Thus it clear that these groups have slighltly different characteristics, hence can be treated separately.

let Y_{ij} denote whether or not patient i in gender j is diagonised as having a heart disease:

$$Y_{ij} = \begin{cases} 1, & \text{if heart disease present,} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_{ij}|\beta_{0j}, \beta_1 \cdots, \beta_8 \sim \text{Bern}(\pi_{ij}) \qquad \text{with} \qquad \log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta_{0j} + \beta_1 X_{ij1} + \cdots + \beta_8 X_{ij8}$$

where π_{ij} is be the probability that patient i in gender j has heart disease.

$\beta_{0j}|\beta_0, \sigma_0 \sim^{ind} N(\beta_0, \sigma_0^2)$

$\beta_0 \sim N(-3, 0.5)$

$\sigma_0 \sim \exp(1)$

intercept β_{0j} describe the underlying heart disease rates, as measured by the log(odds of heart disease), for each gender. These acknowledge that one gender is inherently more susceptible to heart disease than the other.

σ_0 captures the between-group variability in heart disease rates between genders.

β_1 describes the global relationship between heart disease and age when controlling for all the other variables (Cholesterol, MaxHR, Oldpeak, etc). Similarly, β_2 describes the global relationship between heart disease and cholesterol when controlling for all the other variables. The same applies to the other coefficients (β's).

Here the proirs on the coefficient β, are similar to the ones derived on the logistic regression section. i.e:

AGE: $\beta_1 \sim N(0, 7)$

CHOLESTEROL: $\beta_2 \sim N(0.18, 14)$

MAXHR: $\beta_3 \sim N(0, 11.7)$

OLDPEAK: $\beta_4 \sim N(0, 1)$

FASTINGBS: $\beta_5 \sim N(0, 7.3)$

EXERCISE ANGINA: $\beta_6 \sim N(0, 0.9)$

CHEST PAIN: $\beta_7 \sim N(0, 3.22)$

ST-SLOPE: $\beta_8 \sim N(0, 3.6)$

4 MCMC Samplers

We used the following MCMC samplers for our models:

1. Metropolis - Hastings
2. Metropolis - Within - Gibbs - Sampler
3. Stan Code (applied for both models)

4.1 Metropolis - Hastings:

Model 1: Logistic Regression:

The joint posterior is given by:

$p(\beta_0,\ \beta_1,\ \cdots,\beta_9|y_{1:918}) \propto \{\prod_{i=1}^{N} p^y(1-p)^{1-y}\} \times N(\beta_0 : -3, 0.5) \times N(\beta_1 : 0, 7) \times N(\beta_2 : 0.18, 14) \times N(\beta_3 : 0, 11.7) \times N(\beta_4 : 0, 1) \times N(\beta_5 : 0, 8.41) \times N(\beta_6 : 0, 7.3) \times N(\beta_7 : 0, 0.9) \times N(\beta_8 : 0, 3.22) \times N(\beta_9 : 0, 3.6)$

where

$$p = \frac{\exp(\eta)}{1+\exp(\eta)}, \quad \text{and } \eta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_9 X_{i9}$$

Our proposal distribution was the Multivariate Normal distribution, i.e;

$$(\beta_0^*, \beta_1^*, \cdots, \beta_9^*) \sim MVN\left(\begin{pmatrix}\beta_0^*\\ \beta_1^*\\ \vdots\\ \beta_9^*\end{pmatrix}, \Sigma\right)$$

We employed the log transformation on our joint posterior (i.e log likelihood $\times$ log prior) for more numerically stable results and also used adaptive step-size to help select the best tuning value for our variance.

4.2 Metropolis - Within - Gibbs Sampler:

Model 2: Hierarchical Model: Here the joint posterior distribution is given by:

$$p(\beta_{0j}, \beta_1, \cdots, \beta_8 | y_{1:918} \propto \{\prod_{j=1}^{2} \{\prod_{i=1}^{n_j} p(y_{ij}|\beta_{0j}, \beta_1, \cdots, \beta_8)\} p(\beta_{0j}|\beta_0, \sigma_0^2)\} \times p(\beta_1) \times p(\beta_2) \times \cdots p(\beta_8)$$

And the full conditional distributions up to a normalizing constant are given by:

$$p(\beta_{0j}|y_{1:918}, \beta_0, \sigma_0^2) \propto \{\prod_{i=1}^{n_j} p(y_{ij}|\beta_{0j}, \beta_1, \cdots, \beta_8)\} p(\beta_{0j}|\beta_0, \sigma_0^2)$$

$$p(\beta_k|y_{ij}, \beta_{0j}, \beta_{-k}) \propto \{\prod_{i=1}^{n_j} p(y_{ij}|\beta_{0j}, \beta_{1:8})\} \times p(\beta_k), \qquad \text{for } k = 1, \cdots 8,$$

$$p(\beta_0|y_{1:918}, \beta_0, \sigma_0^2) \propto \; p(\beta_{0j}|\beta_0, \sigma_0^2) \times p(\beta_0| -3, 0.5) \sim N\left(\frac{\frac{\beta_0 1 + \beta_0 2}{\sigma_0^2} + (-6)}{\frac{2}{\sigma_0^2} + 2}, \frac{1}{\frac{2}{\sigma_0^2} + 2}\right)$$

$$p(\sigma_0^2|y_{1:918}, \beta_0, \sigma_0^2) \propto \; p(\beta_{0j}|\beta_0, \sigma_0^2) \times p(\sigma_0^2|1) \sim \text{Inverse Gamma}(2, \tfrac{1}{2}\sum_{j=1}^{2}(\beta_{0j} - \beta_0)^2)$$

Hence we use the Metropolis- within - Gibbs sampler and sample from the given distributions each iteration for β_0 and σ_0^2, then sample β_{0j} and β_k, from the embedded 1- sted Metropolis - Hastings each iteration.

Similarly, the target distribution for the β_{0j}'s and β_k's respectively are:

$$p(\beta_{0j}|y_{1:918}, \beta_0, \sigma_0^2) \propto \{\prod_{i=1}^{n_j} p(y_{ij}|\beta_{0j}, \beta_1, \cdots, \beta_8)\} p(\beta_{0j}|\beta_0, \sigma_0^2)$$

and

$$p(\beta_k|y_{ij}, \beta_{0j}, \beta_{-k}) \propto \{\prod_{i=1}^{n_j} p(y_{ij}|\beta_{0j}, \beta_{1:8})\} \times p(\beta_k)$$

With proposal distributions being:

$$(\beta_{01}^*, \beta_{02}^*) \sim MVN\left(\begin{pmatrix}\beta_{01}^* \\ \beta_{02}^*\end{pmatrix}, \Sigma\right)$$

and

$$(\beta_1^*, \beta_2^*, \cdots, \beta_8^*) \sim MVN\left(\begin{pmatrix}\beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_8^*\end{pmatrix}, \Sigma\right)$$

5 Results:

5.1 Model 1: Logistic Regression

Estimates:

	Mean	SD	$CI_2.5$	$CI_97.5$
(Intercept)	-0.05	0.01	-0.06	-0.04
Age	0.05	0.01	0.04	0.06
Cholesterol	-0.00	0.00	-0.00	-0.00
MaxHR	-0.02	0.00	-0.02	-0.01
Oldpeak	0.29	0.02	0.25	0.32
SexM	0.25	0.01	0.24	0.27
FastingBS	-0.01	0.01	-0.04	-0.00
ExerciseAnginaY	0.18	0.01	0.16	0.19
ChestPainTypeATA	-0.25	0.01	-0.26	-0.24
ChestPainTypeNAP	-0.24	0.01	-0.25	-0.23
ChestPainTypeTA	-0.02	0.01	-0.04	-0.01
ST_SlopeFlat	0.20	0.01	0.19	0.23
ST_SlopeUp	-0.35	0.02	-0.39	-0.32

Traceplot Discussion: For some of the variables, the step-size tends to be small even after applying adaptive variance tuning (acceptance rate = 0.275). **Model Evaluation:**

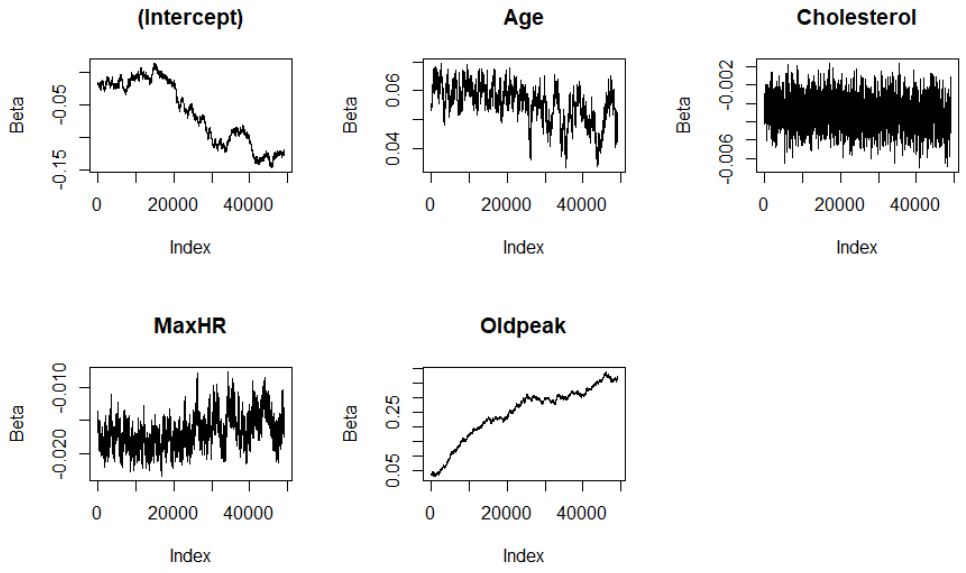


Figure 3: Trace plots of numerical variables

Our model gave an accuracy of 78.1%, Specificity = 82.1% and Sensitivity = 73.17 %. We chose a cut-off point in our classification which will increase our Specificity. That is, we are more concerned about reducing type 1 error, reason being that we do not want to not correctly diagnose a sick patient as being sick as this might be fatal.

```
          Reference
Prediction   0   1
         0 300  91
         1 110 417
```

Figure 4: Confusion Matrix

5.2 Model 1 - Stan

Estimates:

term	estimate	std.error	conf.low (80 %)	conf.high (80 %)
Intercept	-1.178985820	1.441265134	-3.007998e+00	0.681367117
Cholesterol	-0.004241922	0.001123521	-5.687248e-03	-0.002816117
Oldpeak	0.390087665	0.119918622	2.390911e-01	0.545412735
Age	0.017031473	0.013275377	3.860409e-05	0.034175776
MaxHR	-0.004399067	0.005104355	-1.098112e-02	0.002139032
FastingBS	1.171782366	0.281657271	8.162106e-01	1.535497672
ST-SlopeFlat	1.467050976	0.428905016	9.104070e-01	2.023406529
ST-SlopeUp	-1.050543647	0.454508364	-1.634329e+00	-0.469128803
SexM	1.508486654	0.287015563	1.146678e+00	1.870750429
ExerciseAnginaY	0.924571329	0.246267303	6.022080e-01	1.236300662
ChestPainTypeATA	-1.874374643	0.331612508	-2.306838e+00	-1.452553471
ChestPainTypeNAP	-1.730357371	0.264450028	-2.079245e+00	-1.384606230
ChestPainTypeTA	-1.529044585	0.449547657	-2.108653e+00	-0.961052158

Our model gave an accuracy of 86.38%, Specificity = 86.1% and Sensitivity = 86.61 %.

```
$confusion_matrix
 y   0   1
 0 353  57
 1  68 440
```

Figure 5: Confusion Matrix

5.3 Model 2: Hierarchical Model

	Mean	SD	CI_2.5	CI_97.5
beta0_global	-2.31	0.69	-3.72	-1.05
sigma0	1.73	1.09	0.67	4.46
beta0_female	-1.63	0.19	-2.00	-1.25
beta0_male	-0.18	0.17	-0.51	0.11
Age	0.06	0.02	0.04	0.08
Cholesterol	-0.00	0.00	-0.00	-0.00
MaxHR	-0.02	0.01	-0.03	-0.01
Oldpeak	0.24	0.03	0.20	0.30
FastingBS	0.16	0.03	0.13	0.20
ExerciseAngina	-0.08	0.04	-0.13	0.01
ChestPainTypeATA	0.20	0.08	0.05	0.28
ChestPainTypeNAP	0.01	0.02	-0.01	0.06
ChestPainTypeTA	-0.07	0.03	-0.11	-0.04
ST_SlopeFlat	0.15	0.03	0.11	0.20
ST_SlopeUp	-0.04	0.02	-0.06	0.00

Convergence Diagnostics: We see that the trace plots mixes well hence there are no issues on convergence.

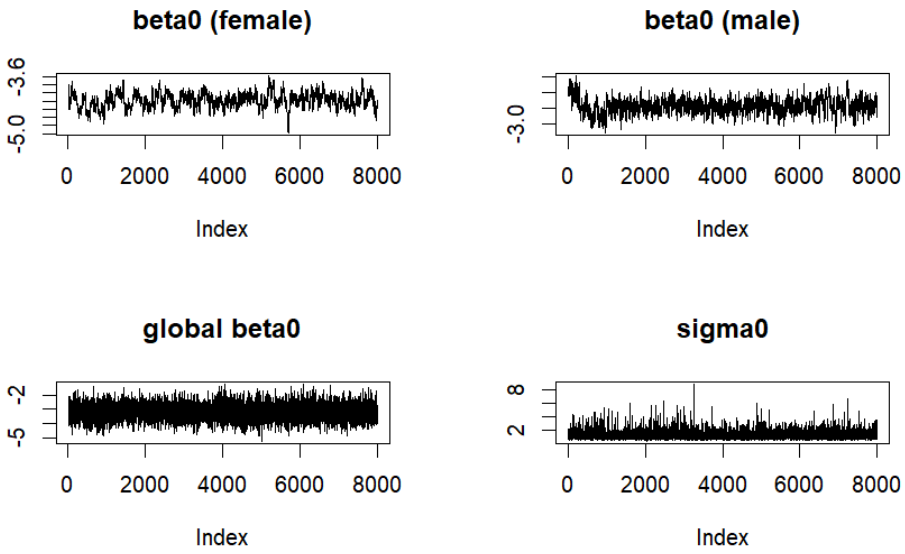


Figure 6: Trace plots

Model Evaluation: Here we used the confusion matrix to see how our model performs. We chose a cut-off point in our classification which will increase our Specificity. That is, we are more concerned about reducing type 1 error in this problem, reason being that we do not want to not correctly diagnose a sick patient as this might be fatal.
Our model gave an accuracy of 65.9%, Specificity = 80.12% and Sensitivity = 48.29 %.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 198 101
         1 212 407
```

Figure 7: Confusion Matrix

5.4 Model 2 - Stan

term	estimate	std.error	conf.low (80 %)	conf.high (80 %)
Intercept	-0.644194011	1.458699037	-2.650520644	1.176358505
Cholesterol	-0.003950739	0.001053792	-0.005300564	-0.002605604
Oldpeak	0.395447526	0.118538232	0.244657931	0.547911520
Age	0.020463830	0.012640941	0.004296074	0.036422618
MaxHR	-0.003745881	0.004942287	-0.010123797	0.002672614
FastingBS	1.165089602	0.278503805	0.814624317	1.521562483
ST-SlopeFlat	1.455535480	0.431021691	0.892447038	2.008007422
ST-SlopeUp	-1.027428478	0.456210328	-1.628211672	-0.453801243
ExerciseAnginaY	0.921455063	0.248284516	0.607331323	1.240383794
ChestPainTypeATA	-1.867277412	0.329177705	-2.299286797	-1.451548687
ChestPainTypeNAP	-1.717787313	0.265904715	-2.061128039	-1.373942777
ChestPainTypeTA	-1.469778603	0.434706190	-2.029074678	-0.910241508

Convergence Diagnostics: We see that the trace plot mixes well. All 4 chains are well mixing

Figure 8: Trace plots

Model Evaluation:
Below, we show a posterior predictive check of the hierarchical logistic regression model of being diagnosed with heart disease. The histogram displays the proportion of patients that were diagnised of having Heart Disease in each of 100 posterior simulated datasets. The vertical line represents the observed proportion of patients that were diagnosed of heart disease in the data. We see that the proportions from the posterior simulated datasets are relatively close to the observed proportion of people diagnosed of heart disease, showing that our model can generalize well on unseen data.

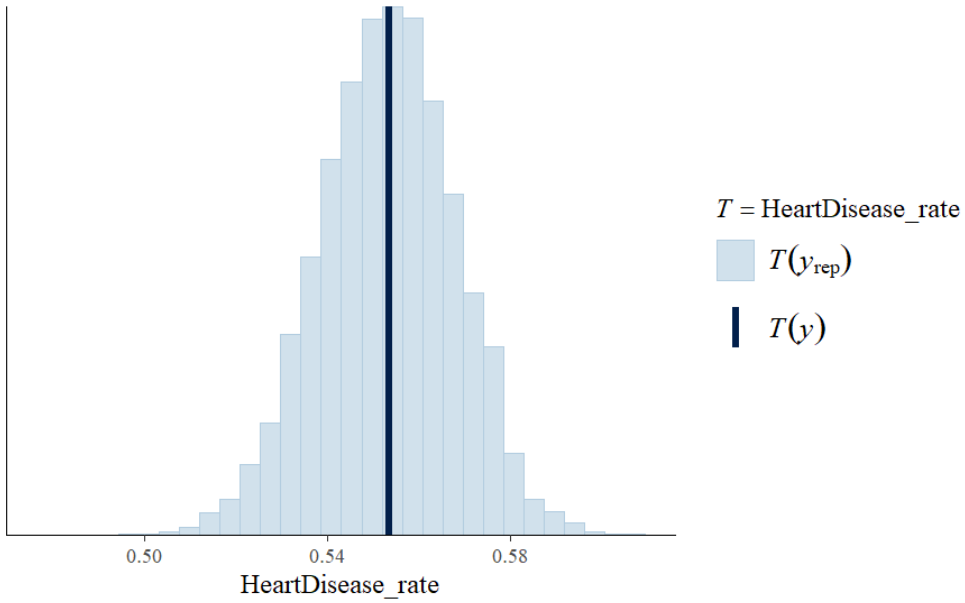


Figure 9: proportion of patients that were diagnised of having Heart Disease in 100 posterior simulated datasets

Our model gave an accuracy of 86.5%, Specificity = 86.59% and Sensitivity = 86.42 %.

```
$confusion_matrix
 y   0   1
 0 355  55
 1  69 439

$accuracy_rates
```

Figure 10: Confusion Matrix

Conclusion: Both Models: The Logistic Regression Model and Hierarchical Model perform relatively the same based on the accuracy, specificity and sensitivity results.

6 Bayesian Model Comparison:

6.1 Bayes Factor:

Model 1:
The marginal likelihood for Model 1 is given by:

$$p(y|M_1) = \int \left[\prod_{i=1}^{918} \left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)} \right)^{y_i} \left(\frac{1}{1+\exp(\eta_i)} \right)^{1-y_i} \right] \times \prod_{j=0}^{9} N(\beta_j|\mu_j, \sigma_j^2) \, \mathrm{d}\boldsymbol{\beta}$$

which we will estimate using Laplace approximation since we do not have a closed form solution. Log marginal likelihood (Laplace approximation) = -320.0024
Model 2:

$$p(y|M_2) = \int \{ \prod_{j=1}^{2} \{ \prod_{i=1}^{n_j} p(y_{ij}|\beta_{0j}, \beta_1, \cdots, \beta_8) \} p(\beta_{0j}|\beta_0, \sigma_0^2) \} \times p(\beta_0) \times p(\beta_1) \times p(\beta_2) \times \cdots p(\beta_8) \times p(\sigma_0^2) \, \mathrm{d}\theta$$

for $\theta = (\beta_{01}, \beta_{02}, \beta_1, \cdots, \beta_8, \beta_0, \sigma_0^2)$.

Log Marginal Likelihood: -320.7133
Now the Bayes Factor is given by:

$$\text{Bayes Factor} = \frac{p(y|M_1)}{p(y|M_2)} = \exp(\log p(y|M_1) - \log p(y|M_2)) = \exp(-320.0024 + 320.7133) = 2.035823$$

Our strength of evidence is barely woth mentioning that Model 1 is better compared to Model 2. That is Model 1 and Model 2 explain the data relatively the same.

6.2 Bayes Rule:

Here we choose the model with the highest Posterior Model Probability (PMP). Now

$$p(M_i|y) = \{1 + \sum_{j \neq i} \pi_{ji} B_{ji}\}^{-1}$$

where:
$\pi_{ji} = \frac{p(M_j)}{p(M_i)}$ and B_{ji} is the Bayes Factor of M_j over M_i.

Assuming equal prior model probability, we have:

$$p(M_2|y) = (1 + B_{12})^{-1} = (1 + 2.035823)^{-1} = 0.3294$$

Showing that Model 1 is preferred over Model 2 since it has a high PMP.

6.3 Bayesian Variable Selection (BVS):

For each of our two models, we have a linear regression model form

$$\mathcal{M}_\gamma : y|\alpha, \beta_\gamma, \gamma, \sigma^2 \sim \mathcal{N}_n(\mathbf{1}_n\alpha + X_\gamma\beta_\gamma, \sigma^2 I_{n\times n})$$

For Model 1, we have nine predictors we want to include in the model: Sex, ChestPainType, FastingBS, ExerciseAngina, ST-Slope, Age, Cholesterol, MaxHR, Oldpeak, which we denote by γ, i.e; $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_9)$, whilst for Model 2 we have eight predictors we want to include in the model, which are the same as for Model 1 with the exception of Sex as we will use that for grouping. Hence for Model 2: $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_8)$. Where:

$$\gamma_j = \begin{cases} 1, & \text{if the } j^{th} \text{ predictor is included in the model} \\ 0, & \text{otherwise} \end{cases}$$

$p_\gamma = \sum_{j=1}^{p} \gamma_j$ is the number of predictors in Model γ
X_γ is the $n \times p_\gamma$ matrix of predictors with $\gamma_j = 1$ and
β_γ is the corresponding vector of predictors with $\gamma_j = 1$.

Steps for BVS:

Model 1 - Logistic Regression:

1. Likelihood under Model γ:

$$y_i \mid \boldsymbol{\beta}_\gamma, M_\gamma \sim \text{Bernoulli}(\pi_i), \quad \text{logit}(\pi_i) = \beta_0 + \mathbf{x}_{i\gamma}^\top \boldsymbol{\beta}_\gamma$$

2. Prior for γ:
   We assume uniform prior:

$$p(M_\gamma) = \frac{1}{p}$$

   where p is total number of possible predictors. Alternatively, we can let $\gamma_j \sim \text{Bern}(\theta)$ with $\theta \sim \text{Beta}(a, b)$ Then:

$$p(M_\gamma) = \int \prod_{j=1}^{p_\gamma} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j} \cdot \text{Beta}(a, b)\, d\theta$$

3. Priors for model parameters:

$$\boldsymbol{\beta}_\gamma \sim \mathcal{N}(\boldsymbol{\mu}_\gamma, \Sigma_\gamma)$$

   where

   $\boldsymbol{\mu}_\gamma$ and Σ_γ are as defined previously under model specification.

4. Marginal Likelihood:

$$p(y \mid M_\gamma) = \int p(y \mid \boldsymbol{\beta}_\gamma, M_\gamma) \cdot p(\boldsymbol{\beta}_\gamma \mid y, M_\gamma)\, d\boldsymbol{\beta}_\gamma$$

   Where:
   $p(y \mid \boldsymbol{\beta}_\gamma, M_\gamma) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$ and
   $\pi_i = \text{logit}^{-1}(x_{i\gamma}^\top \beta_\gamma)$

   This integral does not have a closed form, hence we approximate it using Laplace approximation.

Model 2 - Hierarchical Logistic Regression Model:

1. Likelihood:
$$y_i \sim \text{Bernoulli}(\pi_i), \quad \text{logit}(\pi_i) = \beta_{0,s_i} + \mathbf{x}_{i\gamma}^\top \boldsymbol{\beta}_\gamma$$
Where:
$\beta_{0,1}, \beta_{0,2}$: random intercepts for sex groups (Male, Female) and
$\boldsymbol{\beta}_\gamma = \{\beta_j : \gamma_j = 1\}$: included fixed-effect coefficients

2. Prior for γ:
We assume uniform prior:
$$p(M_\gamma) = \frac{1}{p}$$

3. Priors for model parameters:
For the random intercepts we have:
$$\beta_{0j} \sim \mathcal{N}(\beta_0, \sigma_0^2), \quad j = 1, 2$$
$$\beta_0 \sim \mathcal{N}(\mu_0, \tau_0^2), \quad \sigma_0 \sim \text{Exponential}(\lambda)$$
Priors on fixed-effect coefficients are given by:
$$\boldsymbol{\beta}_\gamma \sim \mathcal{N}(\boldsymbol{\mu}_\gamma, \Sigma_\gamma)$$
where
$\boldsymbol{\mu}_\gamma$ and Σ_γ are as defined previously under model specification.

4. Marginal Likelihood:
$$p(y \mid M_\gamma) = \int\int\int\int p(y \mid M_\gamma, \boldsymbol{\beta}_\gamma, \beta_{01}, \beta_{02}) \cdot p(\boldsymbol{\beta}_\gamma \mid y, M_\gamma) \cdot p(\beta_{01}, \beta_{02}, \beta_0, \sigma_0)\, \mathrm{d}\boldsymbol{\beta}_\gamma\, \mathrm{d}\beta_{01:02}\, \mathrm{d}\beta_0\, \mathrm{d}\sigma_0$$
This integral is high-dimensional and non-conjugate, hence we approximate it using Laplace approximation.

Posterior Model Probabilities:

$$p(M_\gamma \mid y) = \frac{p(y \mid M_\gamma)p(M_\gamma)}{\sum_{\gamma'} p(y \mid M_{\gamma'})p(M_{\gamma'})}$$

Posterior Inclusion Probabilities (PIP):

$$\text{PIP}_j = \sum_{\gamma:\gamma_j=1} p(M_\gamma \mid y)$$

The sum of the posterior model probabilities of all models that include predictor γ_j.

Laplace approximation, Logistic Regression Model:

We integrate
$$p(y \mid M) = \int p(y \mid \boldsymbol{\beta})\, p(\boldsymbol{\beta})\, d\boldsymbol{\beta}$$
Using Laplace approximation as shown below:

1. Setup for the Full Model:
   Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_9)^\top$ with Design matrix $X \in \mathbb{R}^{n \times 9}$, where

$$y_i \mid \boldsymbol{\beta} \sim \text{Bernoulli}(\pi_i), \quad \text{where } \pi_i = \frac{e^{x_i^\top \boldsymbol{\beta}}}{1 + e^{x_i^\top \boldsymbol{\beta}}}$$

2. Prior: independent normals

$$\beta_j \sim \mathcal{N}(\mu_j, \sigma_j^2) \quad \Rightarrow \quad p(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu})\right)$$

3. Now for Laplace Approximation Formula:
   Given

$$I = \int \exp(h(\boldsymbol{\beta}))\, d\boldsymbol{\beta}, \quad \text{where } h(\boldsymbol{\beta}) = \log p(y \mid \boldsymbol{\beta}) + \log p(\boldsymbol{\beta})$$

   Laplace's method gives:

$$p(y \mid M) \approx (2\pi)^{d/2} |H|^{-1/2} \exp(h(\hat{\boldsymbol{\beta}}))$$

   where,
   $\hat{\boldsymbol{\beta}}$: posterior mode (MAP estimator) obtained by maximizing $h(\boldsymbol{\beta})$.i.e;

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \{\log p(y \mid \boldsymbol{\beta}) + \log p(\boldsymbol{\beta})\}$$

   $H = -\nabla^2 h(\boldsymbol{\beta})\big|_{\hat{\boldsymbol{\beta}}}$: negative Hessian of log posterior at the mode

   $d = 9$: dimension of $\boldsymbol{\beta}$

Hierarchical Model: For the Hierarchical Model, we follow the same procedure, however ensuring that we take into account the random intercepts for sex group.i.e;

$$p(y \mid M) \approx (2\pi)^{(p+2)/2} |H|^{-1/2} \exp(h(\hat{\boldsymbol{\theta}})), \qquad \text{for } \theta \in \mathbb{R}^{p+2}$$

Bayesian model averaging (BMA):

The posterior distribution for the quantity of interest Δ is:

$$p(\Delta|y) = \sum_{M \in \mathcal{M}} p(\Delta|y, M) p(M|y) = \sum_{M \in \mathcal{M}} p(\Delta|y, M) \hat{\prod}_M$$

where

$$\hat{\prod} = p(M_\gamma \mid y) = \frac{p(y \mid M_\gamma) p(M_\gamma)}{\sum_{\gamma'} p(y \mid M_{\gamma'}) p(M_{\gamma'})}$$

Results

Logistic Regression Model:

We obtained the below results:

Variable	PIP
Age	0.8603775
Cholesterol	0.9955968
MaxHR	0.1460744
Oldpeak	0.9998424
Sex	1.0000000
FastingBS	0.9999980
ExerciseAngina	0.9999805
ChestPainType	1.0000000
ST-Slope	1.0000000

MaxHR had the smallest PIP and was thus removed from the model.

PMP = 0.96748567

Hierarchical Logistic Regression Model:

We obtained the below results:

Variable	PIP
Age	0.04122508
Cholesterol	1.00000000
MaxHR	0.99999985
Oldpeak	1.00000000
Sex	0.95877507
FastingBS	1.00000000
ExerciseAngina	1.00000000
ChestPainType	1.0000000
ST-Slope	1.0000000

Age had the smallest PIP and was thus removed from the model.

PMP = 0.03251433

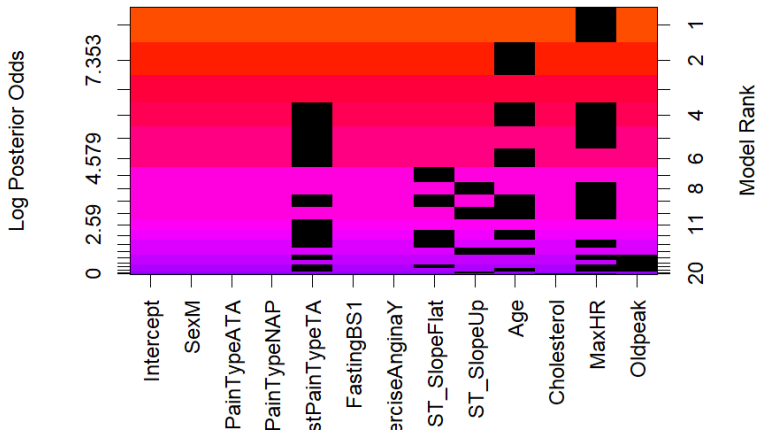
We see that Model 1 has a higher PMP than Model 2, hence helps explain the data better thus preferred.

BVS with Zellner's g-prior:
Also from using Zellner's g prior, we see that MaxHR and Age indeed have low PIP values compared to the other predictors.

```
Marginal Posterior Summaries of Coefficients:

Using  BMA

Based on the top  2048 models
                 post mean   post SD     post p(B != 0)
Intercept         0.5533769   0.0107798   1.0000000
SexM              0.1589193   0.0278134   0.9999930
ChestPainTypeATA -0.2506531   0.0339239   1.0000000
ChestPainTypeNAP -0.2302503   0.0296378   1.0000000
ChestPainTypeTA  -0.1800564   0.0698692   0.9324157
FastingBS1        0.1327101   0.0273623   0.9999778
ExerciseAnginaY   0.1415213   0.0279550   0.9999153
ST_SlopeFlat      0.1613603   0.0488938   0.9931099
ST_SlopeUp       -0.2167847   0.0545409   0.9947707
Age               0.0021629   0.0016561   0.7452409
Cholesterol      -0.0005114   0.0001070   0.9998231
MaxHR            -0.0002269   0.0004620   0.2650095
Oldpeak           0.0498784   0.0128103   0.9985135
```



Conclusion: The Bayes Factor showed that Model 1 and Model 2 perform relatively the same. From this, we may conclude that gender does not have much impact on heart disease, of course such an impact might have been overshadowed by the fact that we have imbalanced data between the genders. From the Bayes Rule and Bayesian Variable Selection we saw that the Logistic Regression Model performed better than the Hierarchical Logistic Regression Model.

7 Contributions:

Tiasha: Searching for potential datasets, conducting exploratory data analysis, selecting appropriate models, searching for priors, trying to write the necessary code for model comparison and results, reviewing groupmates' code, making presentation slides, writing and proofreading groupmates' writing.

Xunhang (David): Searching for potential datasets; selecting appropriate models; write part of the theoretical model specifications related to MCMC estimation, Bayesian variable selection (BVS) and model comparison; write the R code for likelihood ratio tests, MCMC estimation models, and BVS and model comparison, write part of the results of MCMC estimation and BVS.

Fezile: Searching for potential datasets. selecting appropriate models. Sort of sifted through some literature to help with prior formulation. Assisted with MCMC simulation, assisted with Bayesian model comparison. Assisted with the write-up and general structure of the work. Tried to coordinate most of the activities done.

As a group we had weekly meetings both in person and online to discuss progress and a way forward.

References

[1] A. H. Association, "2024 heart disease and stroke statistics update fact sheet," 2024. [Online]. Available: https://www.heart.org/-/media/PHD-Files-2/Science-News/2/2024-Heart-and-Stroke-Stat-Update/2024-Statistics-At-A-Glance-final_2024.pdf

[2] A. Janosi, W. Steinbrunn, M. Pfisterer, and Robert Detrano, "Heart failure prediction dataset," 2021. [Online]. Available: https://www.kaggle.com/fedesoriano/heart-failure-prediction

[3] X. Zhang, B. Guo, and N. Yin, "Zero-inflated gaussian mixed models for analyzing longitudinal microbiome data," 2020. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7652264/

[4] K. M. Mohi Uddin, R. Ripa, N. Yeasmin, N. Biswas, and S. K. Dey, "Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset," *Intelligence-Based Medicine*, vol. 7, p. 100100, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666521223000145

[5] A. Gupta, L. Kumar, R. Jain, and P. Nagrath, "Heart disease prediction using classification (naive bayes)," in *Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)*, P. K. Singh, W. Pawłowski, S. Tanwar, N. Kumar, J. J. P. C. Rodrigues, and M. S. Obaidat, Eds. Singapore: Springer Singapore, 2020, pp. 561–573.

[6] P. Subarkah, W. R. Damayanti, and R. A. Permana, "Comparison of correlated algorithm accuracy naive bayes classifier and naive bayes classifier for heart failure classification," *Ilkom Jurnal Ilmiah*, vol. 14, no. 2, pp. 120–125, 2022.

[7] C. News, "Half of patients with ideal cholesterol have underlying heart risks," 2017. [Online]. Available: https://www.cardiosmart.org/news/2017/12/half-of-patients-with-ideal-cholesterol-have-underlying-heart-risks

[8] U. Health, "Most heart attack patients' cholesterol levels did not indicate cardiac risk," 2009. [Online]. Available: https://www.uclahealth.org/news/release/most-heart-attack-patients-cholesterol-levels-did-not-indicate-cardiac-risk

[9] A. Kashou, H. Basit, and A. Malik, "St segment," 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK459364/

[10] M. Today, "High blood sugar may raise heart disease risk even if you don't have diabetes," 2023. [Online]. Available: https://www.medicalnewstoday.com/articles/high-blood-sugar-heart-disease-risk