Impact of Age and Sex on Heart Failure Risk

October 21, 2024

Contents

1 Introduction

Cardiovascular diseases (CVDs) are a significant global health challenge, causing 31% of all deaths worldwide. Among these, heart attacks and strokes are the leading causes, accounting for four of every five CVD-related deaths. Cardiovascular disease is particularly prevalent among the older and ageing population as compared to the general population (Rodgers et al., 2019). Age is an independent risk factor for cardiovascular disease (CVD) in adults. However, this susceptibility is confounded by other factors, such as obesity and diabetes (Lu et al., 2014).

Many risk factors contribute to the development of CVD. However, sex has emerged as a potential risk factor among older adults, with studies indicating that women have a comparable risk of CVD to men of similar age groups. According to the American Heart Association's 2022 Heart Disease and Stroke Statistical Update, the incidence of CVD was 77.5% in males and 75.4% in females aged 60 to 79 years. Furthermore, the incidence of CVD was reported to be 89.4% in males and 90.8% in females in adults over the age of 80 (American Heart Association, 2022). Regardless of gender, the risk of CVD rises with age, and outcomes between men and women are primarily due to sex hormones and their receptors. These hormonal fluctuations influence the prevalence and severity of CVD among different genders as people age (Garcia et al., 2016). Hence, our research is based on a better understanding of the impact of age and gender on the risk of heart failure.

The primary objective of this analysis is to investigate how age and sex impact the likelihood of heart failure occurrence. Additionally, we aim to explore the interaction effects between age, sex, and other variables to assess if any change in age or sex directly impacts some of the other variables, as previous research has suggested (Madonna, 2019; Maas AH, 2010). Specifically, we will examine the interaction effects between Age and RestingBP, Age and ChestPainType, Sex and FastingBS, and Sex and RestingBP, while considering all other variables in the dataset. Furthermore, this analysis will examine the variability in the risk of heart failure attributable to different locations, considering location as a random effect in the modeling process.

By exploring these relationships, we seek to enhance our understanding of the complex dynamics of heart failure onset, thus facilitating more effective prevention and management strategies. In our analysis, we utilize the UCI machine learning repository Heart Disease (Janosi and Detrano, 1988) and Statlog (Heart) (dat) data set to investigate the

influence of age and sex on heart failure. Employing the logistic regression model within the Bayesian framework allows us to assess these relationships effectively.

The subsequent section offers a comprehensive overview of the dataset, including definitions of variables and information on data quality. Statistical analysis methods are presented and explained in the third section. In the fourth section, the analysis and interpretation of the results are presented. Finally, in the fifth section, the key findings are summarized.

2 Problem Statement

2.1 Dataset and Data Quality

This report is based on the UCI machine learning repository heart disease data sets, comprising five distinct heart datasets featuring 14 common features. The contributing datasets include Cleveland (303 observations), Hungarian (294 observations), Switzerland (123 observations), Long Beach, VA (200 observations), and the Stalog (Heart) Data Set (270 observations), amounting to a total of 1190 observations. An extra variable called location was created to account for the source of the data; two of the 14 variables contained over 90% missing data, which led to removing them. For the remaining 12 variables, the rows of the categorical variables containing missing data were dropped. In contrast, for the numeric variables, the missing values are replaced by the group mean of their respective location. The processed data contains 303 observations from Cleveland, 100 from Hungary, 270 from Statlog, 46 from Switzerland and 92 from Long Beach VA, making 811 observations.

This dataset includes 12 independent variables and 1 dependent variable. Adjustments were made to improve the dataset's properties, such as centering and standardizing numerical variables to ensure scale uniformity and facilitate parameter interpretation. The dependent variable, Heart Disease, is binary, with values representing the presence of heart disease (1) or absence (0) of heart disease.

The independent variables include a variety of factors related to cardiovascular health. Age represents the patient's age in years, while sex denotes the individual's gender (M: Male, F: Female). ChestPainType delineates the type of chest pain experienced, with categories including Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NAP), and Asymptomatic (ASY). RestingBP signifies resting blood pressure in mm Hg,

and Cholesterol denotes serum cholesterol levels in mm/dl. FastingBS indicates fasting blood sugar, with values of 1 indicating levels exceeding 120 mg/dl and 0 indicating otherwise. RestingECG reflects the results of a resting electrocardiogram, categorized as Normal, ST-T wave abnormality (ST), or displaying left ventricular hypertrophy (LVH). MaxHR represents the maximum heart rate achieved, measured in beats per minute. ExerciseAngina indicates whether exercise-induced angina was present (Y) or absent (N). Oldpeak refers to ST depression, measured numerically, while the slope of the peak exercise ST segment (ST-Slope) is categorized as upsloping (Up), flat (Flat), or downsloping (Down). The location represents the source of the data.

Among these variables, Age, RestingBP, Cholesterol, MaxHR, and Oldpeak are numerical, while Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST-Slope and Location are nominal variables (Fedesoriano, 2021).

2.2 Project Objectives

The primary goal of this report is to examine the individual contributions of Age and Sex as potential risk factors for heart failure. Additionally, we aim to explore the interaction effects between Age, Sex, and other variables such as Age and RestingBP, Age and ChestPainType, Sex and FastingBS, and Sex and RestingBP. Furthermore, this analysis will examine the variability in the risk of heart failure attributable to different locations, considering location as a random effect in the modelling process. The statistical method section will go over the methods used to meet the objectives of this report in detail.

3 Statistical Methods

This section presents several statistical methods that will be used to analyze the dataset based on the objectives of this report. All analyses and visualizations were performed using the statistical software R (Version 4.3.0, R Core Team, 2023), utilizing the R packages brms(Bürkner, 2021), dplyr(Wickham et al., 2023), scaler(Tyre, 2024), and xtable (Scott, 2019).

The UCI machine learning repository heart disease data set has undergone thorough exploration and analysis using various classical machine learning techniques. Previous methods employed include XGBoost, PCA, SVM, KNN, Neural Networks, and Decision Trees. This study uses the logistic regression model, a generalized linear model (GLM)

implemented within the Bayesian framework. This approach seeks to complement and extend the insights from previous analyses conducted with classical machine learning methods.

3.1 Generalized Linear Model

Under the Generalized Linear Model (GLM) framework, we have models in the form:

$$Y_i \sim \text{Dist}(\mu_i, \tau)$$

where Y_i represents the response variable for the ith observation and $Dist$ denotes the probability distribution (Normal, Binomial, Bernoulli, etc) μ_i represents the mean parameter of the distribution for the ith observation. τ represents the dispersion parameter of the distribution. The link function $g(\mu_i)$ transforms the linear predictor η_i as follows:

$$g(\mu_i) = \eta_i$$

where:

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots$$

is a linear combination of predictor variables. $\beta_0, \beta_1, \beta_2, \dots$ are the coefficients associated with the predictor variables $X_{1i}, X_{2i}, \dots$ (Dobson and Barnett, 2018, p. 45-51).

3.2 Logistic Model

The logistic model in the Bayesian framework provides a flexible and powerful tool for modelling binary outcomes, such as the presence or absence of heart disease, based on predictor variables. It allows for the estimation of probabilities and uncertainty quantification, facilitating informed decision-making in clinical and research settings.

Binary Logistic Regression: For $i = 1, \dots, n$, $y_i \sim \text{Bernoulli}(\pi_i)$, where pi_i represents the probability of success(event occurrence) for the $i - th$ observations. The Bernoulli distribution is commonly used for modelling binary outcomes. The probability is modelled using a linear predictor η_i, transformed using the logistic function $g^{-1}(.)$

$$\mu_i = g^{-1}(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}.$$

The Logit Function:

$$\eta_i = \text{logit}(\mu_i) = \log\left(\frac{1-\mu_i}{\mu_i}\right) = x_i^T\beta$$

The logit link converts probabilities to log odds. $g^{-1}(\cdot)$ is called the inverse link function. The model is generally presented using either the link function or the inverse link function (Dobson and Barnett, 2018, p. 45-51).

In our analysis, the logistic regression model, the log odds of the probability of occurrence of an event, denoted as $Pr(y = 1|x)$ allows for the estimation of the likelihood of an outcome (e.g., heart disease) based on predictor variables. Here, the linear predictor, denoted as $X\beta$, is transformed using the logit link function to derive probabilities.

Mathematically, the relationship is represented as:

$$Pr(\text{having heart disease}|x) = g^{-1}(\eta) = g^{-1}(X\beta)$$

Where $X\beta$ represents the linear predictor, and $g(X\beta)$ is defined as:

$$g(X\beta) = \log\left(\frac{Pr(y=1|x)}{1-Pr(y=1|x)}\right)$$

3.3 Likelihood and Priors

In Bayesian approaches, prior beliefs and observed data are combined to derive posterior distributions, distinguishing them from frequentist methods (Dobson and Barnett, 2018, p. 233). Within this framework, prior knowledge about parameters can be integrated, and posterior distributions can be inferred using methods like Markov Chain Monte Carlo (MCMC) sampling. This facilitates a probabilistic framework for inference, enabling the quantification of uncertainty and robust estimation of parameters.

In logistic regression, coefficients (β) can exhibit significantly different behaviour compared to those in normal regression. Therefore, heavy-tailed distributions, such as the Cauchy or t-distribution, are suggested for a better fit. Recent discussions suggest that priors with small degrees of freedom, such as the t-distributions, strike a balance between heavy tails and efficiency in MCMC sampling (Gelman et al., 2008)

In our study, we address the following components:

Likelihood: The response variable (y) in our dataset is binary, where individuals with heart disease are represented as 1 and 0 otherwise. The likelihood in this context follows a Bernoulli distribution, modelling each observation's probability of success (1) or failure (0).

Priors: The predictors in our dataset include both numerical and categorical variables. We assign normal priors to these variables, with specific distributions as follows:

- Intercept $\sim$ Student-t distribution with parameters (1, 0, 2.5)
- Age $\sim$ Normal distribution with mean 0.5 and unit variance (0.5, 1)
- SexM (Male) $\sim$ Normal distribution with mean 0.5 and unit variance (0.5, 1)
- RestingBP $\sim$ Normal distribution with mean 0.5 and unit variance (0.5, 1)
- FastingBS1 $\sim$ Normal distribution with mean 0.5 and unit variance (0.5, 1)
- RestingECGST $\sim$ Normal distribution with mean 0.5 and unit variance (0.5, 1)
- ExerciseAnginaY $\sim$ Normal distribution with mean 0.5 and unit variance (0.5, 1)

For all other parameters, we assign a Normal distribution with mean 0 and unit variance.

4 Statistical Analysis

The variables considered for this analysis are Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak and location. Female is the reference category in the sex variable

4.1 Effects of Age and Sex in Heart Failure

To understand the effects of age and sex on heart failure, several logistic regression models using the Bayesian approach with location as a random effect are fitted. Firstly, a model with only age and sex parameters is fitted, and the output is presented in Figure 1 and Table 1. Figure 1 shows the trace plots and the distribution of the parameters; the trace plot indicates proper mixing of the Markov chain Monte Carlo(MCMC) and that the model converges with Rhat values of 1.00 for both Age and Sex (see Figure 1).

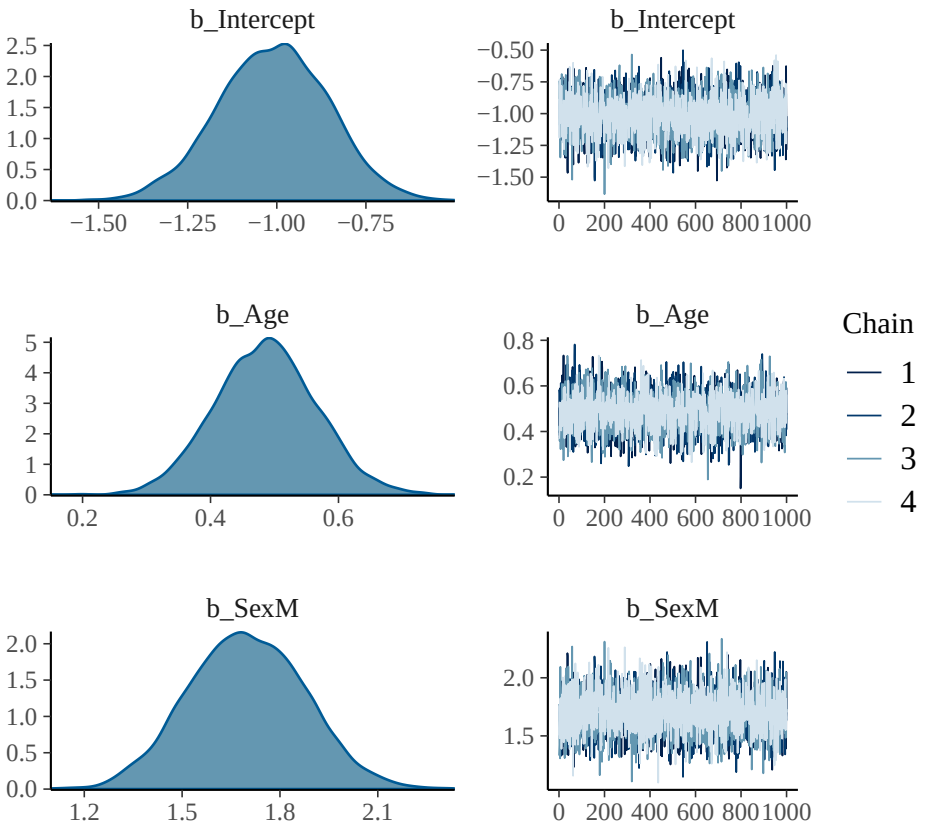


Figure 1: Trace plot of the model with age and sex

Table 1: Posterior mean estimates for model (Sex and Age)

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-1.01	0.15	-1.32	-0.72	1.00	3772.72	2834.68
Age	0.49	0.08	0.33	0.64	1.00	3009.98	2698.61
SexM	1.70	0.18	1.36	2.05	1.00	3756.18	3371.99

Table 1 shows that the Age and SexM parameters have positive estimates. At the same time, the intercept, which represents the effect of SexF, is negative, indicating that as age increases, the log odds of males at risk of heart failure increase more when compared to females. All posterior point estimates are within their credible intervals.

To examine how age and sex affect the risk of heart failure when other factors are considered, a model with all the defined parameters is fitted, and the result is presented in Table 2. Figure 3 and Figure 4 on the Appendix page show the trace plot of the model and the MCMC samples mixed properly.

Table 2: Posterior mean estimates for model with all predictors

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat
Intercept	-0.62	0.42	-1.43	0.25	1.00
Age	0.17	0.11	-0.05	0.39	1.00
SexM	1.61	0.25	1.13	2.10	1.00
ChestPainTypeATA	-1.30	0.30	-1.89	-0.73	1.00
ChestPainTypeNAP	-1.72	0.25	-2.20	-1.24	1.00
ChestPainTypeTA	-1.86	0.37	-2.58	-1.14	1.00
RestingBP	0.27	0.11	0.06	0.48	1.00
Cholesterol	-0.23	0.12	-0.47	-0.00	1.00
FastingBS1	0.03	0.27	-0.47	0.55	1.00
MaxHR	-0.44	0.13	-0.70	-0.19	1.00
RestingECGNormal	-0.34	0.21	-0.75	0.07	1.00
RestingECGST	-0.29	0.37	-1.00	0.44	1.00
ExerciseAnginaY	0.73	0.22	0.28	1.16	1.00
Oldpeak	0.58	0.13	0.33	0.84	1.00
ST_SlopeFlat	0.84	0.36	0.11	1.53	1.00
ST_SlopeUp	-0.07	0.39	-0.84	0.69	1.00

Table 2 shows the posterior mean estimates for the model with all variables while Table 8 on the appendix page shows the Bulk_ESS and Tail_ESS. Here, the effect of age was reduced compared to the model with just age and sex; the credible interval(CI) contains zero, which indicates that other variables influence the risk of heart failure than age. The influence of sex also reduces, but it is still positive, and its CI does not contain zero; all levels of chest pain type, restingBP, RestingECG that are normal, MaxHR, ExerciseAnginaY and Oldpeak are significant, while Cholesterol, fastingBS, and RestingECG have zero, in their CI.

All levels of chestPainType, Cholesterol, MaxHR, all levels of RestingECG and level up of ST_Slope have negative effects on the log odds of risk of heart failure while age, sexM, RestingBP, FastingBP1, ExerciseAnginaY, Oldpeak and ST_SlopeFlat have positive effects which means that an increase in any of these variables will cause an increase in the log odd of risk of heart failure.

4.2 Effect of Variability Due to Locations

Since the dataset used for this analysis comes from five different locations, the variability in the parameter estimate due to different locations is examined. To account for these variations in the latter models, a model with sex and age as predictors and one with all variables in the data set as predictors with location as a random effect was fitted; the outputs of the two models are presented in Table 3 and Table 4 respectively. Table 3 shows that the risk of having heart failure varies randomly across the five locations. The variation in the estimates due to Location is 1.69, and it is within credible intervals; it accounts for the differences in the heart failure risk between the locations that are not explained by age and sex.

Table 3: Posterior mean estimates for model1 (Sex and Age) with Location as random effect

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	BulkESS	TailESS
Intercept	-0.29	0.77	-1.98	1.19	1.00	1197.18	1389.37
Age	0.50	0.09	0.33	0.67	1.00	3114.32	2444.68
SexM	1.53	0.19	1.17	1.91	1.00	2651.09	2345.26
Group-Level Effects Location							
sd(Intercept)	1.69	0.74	0.71	3.51	1.00	1101	1653

Table 4, shows the posterior mean estimates for the model with all variables with Location as a random effect and the Table 9 on the appendix page shows the Bulk_ESS and Tail_ESS. All the estimates still have the same effects as the model, which does not account for the different locations, but it reduces the effect for some of the variables and increases it for others. The variation in the estimates due to Location is 1.63, and within credible intervals, it accounts for the differences in the heart failure risk between the locations that the predictors do not explain (see Table 4).

Table 4: Posterior mean estimates for the model with all variables and Location as random effect

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat
Intercept	-0.25	0.86	-1.92	1.55	1.01
Age	0.18	0.12	-0.05	0.42	1.00
SexM	1.65	0.25	1.19	2.14	1.00
ChestPainTypeATA	-1.26	0.30	-1.86	-0.69	1.00
ChestPainTypeNAP	-1.69	0.25	-2.17	-1.21	1.00
ChestPainTypeTA	-1.84	0.36	-2.56	-1.14	1.00
RestingBP	0.23	0.11	0.02	0.46	1.00
Cholesterol	0.05	0.15	-0.25	0.35	1.00
FastingBS1	0.05	0.27	-0.47	0.58	1.00
MaxHR	-0.36	0.14	-0.65	-0.10	1.00
RestingECGNormal	-0.34	0.22	-0.78	0.08	1.00
RestingECGST	-0.44	0.42	-1.30	0.37	1.00
ExerciseAnginaY	0.73	0.23	0.28	1.18	1.00
Oldpeak	0.66	0.13	0.42	0.93	1.00
ST_SlopeFlat	0.82	0.37	0.10	1.54	1.00
ST_SlopeUp	0.01	0.40	-0.75	0.80	1.00
Group-Level Effects Location					
sd(Intercept)	1.63	0.85	0.45	3.65	1.01

4.3 Interaction Effects of Age and Sex with Other Variables

Assess if any change in age or sex directly impacts some of the other variables, as previous research has stated. We consider the interaction effects between age and RestingBP, age and ChestPainType, Sex and FastingBS, and Sex and RestingBP, given all the other variables. Table 5 shows the posterior mean estimates of this model while Table 10 on the appendix page shows the Bulk_ESS and Tail_ESS. The model converges with Rhat values of 1.00 for all estimates in the model. Including the interaction effects reduces the impact of age to 0.05, and the CI also contains zero, indicating that other important variables influence the risk of heart (see Table 5).

Table 5: Posterior mean estimates for the model with all variables with interaction effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat
Intercept	-0.69	0.42	-1.50	0.13	1.00
Age	0.05	0.14	-0.23	0.33	1.00
SexM	1.80	0.25	1.31	2.31	1.00
ChestPainTypeATA	-1.28	0.30	-1.85	-0.70	1.00
ChestPainTypeNAP	-1.76	0.25	-2.25	-1.30	1.00
ChestPainTypeTA	-1.84	0.36	-2.52	-1.12	1.00
RestingBP	0.72	0.22	0.31	1.19	1.00
Cholesterol	-0.24	0.12	-0.48	-0.01	1.00
FastingBS1	0.77	0.50	-0.22	1.74	1.00
MaxHR	-0.42	0.13	-0.69	-0.16	1.00
RestingECGNormal	-0.33	0.21	-0.73	0.10	1.00
RestingECGST	-0.29	0.38	-1.03	0.44	1.00
ExerciseAnginaY	0.65	0.23	0.20	1.09	1.00
Oldpeak	0.59	0.13	0.33	0.85	1.00
ST_SlopeFlat	0.80	0.35	0.10	1.48	1.00
ST_SlopeUp	-0.20	0.40	-0.97	0.58	1.00
Age:ChestPainTypeATA	0.51	0.30	-0.07	1.11	1.00
Age:ChestPainTypeNAP	0.29	0.24	-0.18	0.76	1.00
Age:ChestPainTypeTA	-0.07	0.33	-0.70	0.59	1.00
Age:RestingBP	0.03	0.12	-0.20	0.26	1.00
SexM:FastingBS1	-0.88	0.54	-1.95	0.15	1.00
SexM:RestingBP	-0.61	0.24	-1.10	-0.15	1.00

Including the interaction effects reduces the impact of age to 0.05, and the CI also contains zero, indicating that other important variables influence the risk of heart failure in the data set. The effect of sexM is positive, and also the CI does not include zero, which means sexM is an important factor in heart failure risk, given all the other variables. The interaction effects of age and all levels of ChestPainType aside from the reference class and RestingBP contain zero between the upper and lower CI, which means they are not significant. That is, the change in age to these variables does not impact heart failure risk. The interaction effects between SexM and FastingBS1 is negative, but the CI contains zero, making it insignificant but the upper and lower CI of the estimate of the interaction effect between SexM and RestingBP does not contain zero, making it impactful in heart failure risk. It has negative effects, which means that Males with higher RestingBP have a lower chance of having heart failures compared to Females, given all other variables held constant.

Table 6, shows the posterior mean estimates for the model's parameters while Table 11 on the appendix page shows the Bulk_ESS and Tail_ESS, including interaction effects with Location as random effects to explain the variation in the estimate due to different locations. This model does not increase the number of significant estimates.

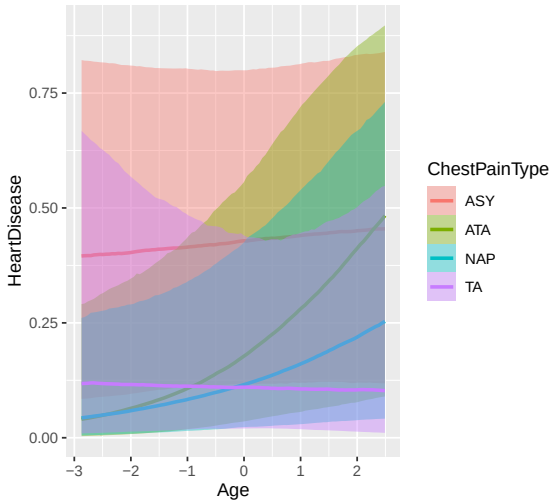
Table 6: Posterior mean estimates for the model with all variables and interaction effects and Location as random effect

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat
Intercept	-0.29	0.87	-2.01	1.38	1.00
Age	0.05	0.15	-0.25	0.33	1.00
SexM	1.82	0.26	1.31	2.34	1.00
ChestPainTypeATA	-1.24	0.31	-1.83	-0.64	1.00
ChestPainTypeNAP	-1.73	0.25	-2.23	-1.25	1.00
ChestPainTypeTA	-1.80	0.37	-2.53	-1.09	1.00
RestingBP	0.68	0.22	0.26	1.11	1.00
Cholesterol	0.04	0.16	-0.28	0.34	1.00
FastingBS1	0.71	0.49	-0.27	1.66	1.00
MaxHR	-0.34	0.14	-0.62	-0.07	1.00
RestingECGNormal	-0.34	0.22	-0.77	0.09	1.00
RestingECGST	-0.47	0.43	-1.31	0.37	1.00
ExerciseAnginaY	0.64	0.24	0.18	1.12	1.00
Oldpeak	0.67	0.14	0.40	0.95	1.00
ST_SlopeFlat	0.79	0.37	0.06	1.53	1.00
ST_SlopeUp	-0.13	0.41	-0.96	0.66	1.00
Age:ChestPainTypeATA	0.54	0.31	-0.07	1.12	1.00
Age:ChestPainTypeNAP	0.33	0.25	-0.15	0.82	1.00
Age:ChestPainTypeTA	-0.07	0.35	-0.75	0.63	1.00
Age:RestingBP	-0.00	0.12	-0.25	0.24	1.00
SexM:FastingBS1	-0.80	0.54	-1.86	0.25	1.00
SexM:RestingBP	-0.62	0.24	-1.10	-0.14	1.00
Group-Level Effects Location					
sd(Intercept)	1.59	0.87	0.31	3.76	1.00

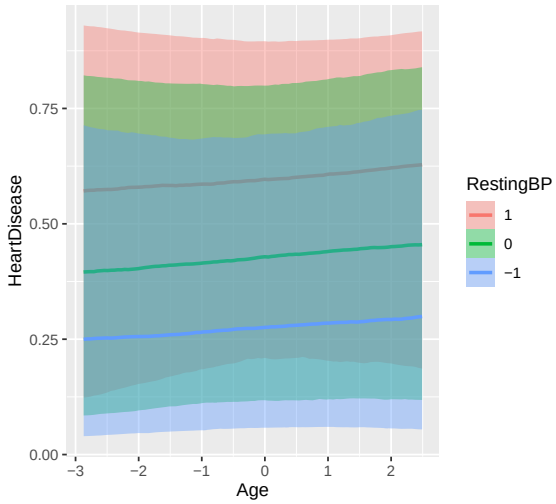
The effects of the interaction between Age and RestingBP become zero, which means that by accounting for the change in variation of the estimates due to location, there is no relationship between Age and RestingBP.

4.4 Conditional Effects for Interaction Effects

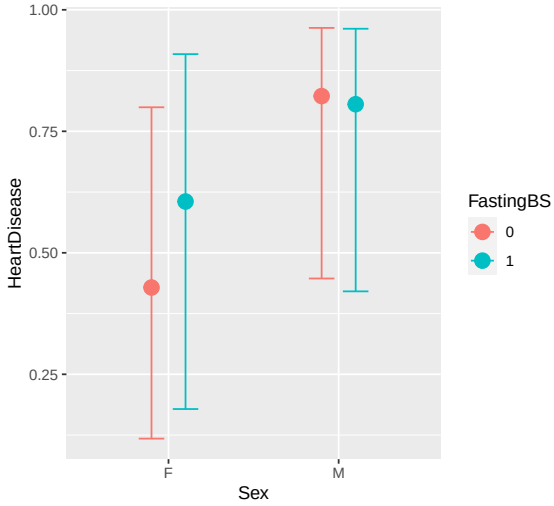
Figure 2(a) - (d) shows the conditional effect of the interaction effects in the model of Table 6. Figure 2(a) shows the interaction effects of ChestPainType and Age. Even though the plot shows interaction effects between the levels of ChestPainType and Age, the credible intervals are too wide to be significant. Figure 2(b) shows no interaction between RestingBP and Age.



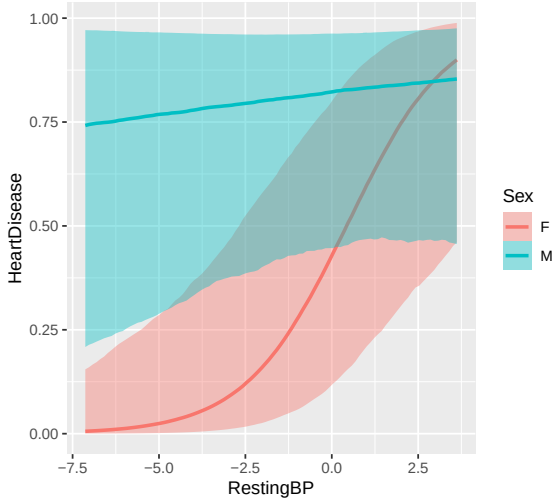
(a) Conditional effect of Heart disease against ChestPainType and Age



(b) Conditional effect of Heart disease against RestingBP and Age



(c) Conditional effect of Heart disease against FastingBS and Sex



(d) Conditional effect of Heart disease against RestingBP and Sex

Figure 2: Conditional effects of all four interaction effects

Figure 2(c) shows the conditional effects of Heart failure against FastingBS and sex; it shows that the way FastingBS affects males and females at risk of heart failure is different. Males have higher fastingBS compared to females in Heart failure risk, but the credible interval is large, affecting the interaction's significance. Figure 2(d) shows the conditional effects of Heart disease against RestingBP and Sex and shows an interaction between the two variables. The rate of heart failure in males increases as RestingBP increases, while females are always at higher risk than males even if RestingBP does not increase (see Figures 2c and 2d).

4.5 Model Comparison

To select the model whose parameters best explain the risk of having heart failure, the six fitted models are compared, that is, the model with just sex and Age as the only predictors(model_Age_Sex), the model with sex and Age as predictors and Location as a random effect(model_Age_Sex_location), the model with all the variables in the data set(model_All), model with all the variables in the data set and Location as a random effect(model_All_location), the model with all the variables and interaction effects and model with all the variables(model_All_interaction), interaction effects and Location as random effects(model_All_interaction_location). The result is presented in Table 7 which shows that model_Age_Sex performs the worst while model_All_interaction_location has zero elpd_diff and se_diff, making it the better model of all the six models.

Table 7: Model comparison for all models

	elpd_diff	se_diff
model_All_interaction_location	0.0	0.0
model_All_location	-2.0	3.4
model_All_interaction	-4.6	3.5
model_All	-7.3	4.9
model_Age_Sex_location	-130.6	14.6
model_Age_Sex	-165.5	15.7

5 Summary

Heart failure is a global health challenge that is more common among the aged population, and the risk of having heart failure varies from men to women. Previous research has shown that some causes of heart failure, like blood pressure and blood sugar, affect men and women differently. Also, high blood pressure and chest pain are more common among the older population. The data used for this report's analysis is a combination of five different datasets extracted from the UCL machine learning repository.

The datasets include Cleveland, Hungary, Switzerland, Long Beach VA, and the Stalog (Heart) Data Set. The dataset originally contained 14 variables each. An extra variable called Location was created to account for the source of the data; two of the 14 variables contained over 90% missing data, which led to removing them. From the remaining 12 variables, the rows of the categorical variables containing missing data were dropped, while the numeric ones were replaced with the mean of their respective location. The processed data contains 303 observations from Cleveland, 100 from Hungary, 270 from Statlog, 46 from Switzerland and 92 from Long Beach VA, making a total of 811 observations with 13 variables with 12 of them being the explanatory variable(Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak and Location) and heart failure being the target variable with two levels 0 and 1.

The main objective of this report is to investigate how age and sex impact the risk of heart failure. Additionally, it also aims to explore the interaction effects between sex and blood pressure, sex and blood sugar level and the interaction effects between age and blood pressure, age and chest pain. To achieve these objectives, six different general linear models with the logit function as the link function using the Bayesian approach were fitted, namely, the model with just sex and age as the only predictors, the model with sex and age as predictors and Location as a random effect, the model with all the variables in the data set, the model with all the variables in the data set and Location as a random effect, the model with all the variables and interaction effects and model with all the variables, interaction effects and Location as random effects.

Conditional effects of the interaction effect are visualized, and elpd_diff se_diff were used to compare the model and select the one with the posterior mean estimates that best describe the risk of heart failure. These models used different variations of the normal

distribution as priors of the predictors, while the intercept uses a student-t distribution prior.

The model with only age and sex shows that as age increases, the log odds of males being at risk of heart failure is higher when compared to females when the effect of variation due to location that may not be accounted for in the estimate is included in the model the effects of age increase but the effects of sex reduces. The model with all the variables in the dataset shows that there are variables that can better explain the risk of heart failure than age because the posterior mean estimate becomes small and also contains zero as in its credible intervals. Sex variables remain relevant, including chestpaintype, restingBP, RestingECG that are normal, MaxHR, ExerciseAnginaY and Oldpeak, while the others are not.

Including the Location as a random effect reduces the effect of the posterior mean estimate of some of the variables. The model with interaction effects shows similar effects for the posterior mean estimate of the model with all variables. The posterior mean estimate for age becomes 0.05, with zero in the credible intervals. The interaction effects between age, chestPainType, and Age and RestingBP are not credible. Also, the interaction effects between Sex and FastingBS are not credible, but the interaction effect between Sex and RestingBP is credible; the parameter estimate for SexM:RestingBP is negative, which indicates that the rate of heart failure in males increases gradually as the RestingBP increases while females are always at higher risk than males.

These six models were compared, and the one model with all the variables, interaction effects, and location as a random effect had the lowest elpd_diff and se_diff. In conclusion, based on the dataset used for the analysis of this report and the resulting analysis, age and sex affect the risk of heart failure differently. Other variables are more capable of explaining the risk of heart failure than sex. Females with high blood pressure are more at risk of high blood pressure than men. Sex is an important factor in heart failure risk.

Bibliography

Statlog (Heart). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C57303.

American Heart Association. 2022 heart disease & stroke statistical update fact sheet: Older americans & cardiovascular diseases, 2022. `https://professional.heart.org/-/media/PHD-Files-2/Science-News/2/2022-Heart-and-Stroke-Stat-Update/2022-Stat-Update-factsheet-Older-Americans-and-CVD.pdf` (visited on 23th February 2024).

Paul-Christian Bürkner. Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54, 2021. doi: 10.18637/jss.v100.i05.

Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, New York, 3th edition, 2018. ISBN 9781315182780. doi: 10.1201/9781315182780. url = https://doi.org/10.1201/9781315182780.

Fedesoriano. Heart failure prediction dataset, September 2021. Retrieved [23th February 2024] from `https://www.kaggle.com/fedesoriano/heart-failure-prediction`.

Marla Garcia, Sharon L Mulvagh, Noel C Merz, Julie E Buring, and JoAnn E Manson. Cardiovascular disease in women: Clinical perspectives. *Circ Res*, 118:1273–1293, 2016. doi: 10.1161/CIRCRESAHA.116.307547.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), 2008. ISSN 1932-6157. doi: 10.1214/08-aoas191. url=http://dx.doi.org/10.1214/08-AOAS191.

Steinbrunn William Pfisterer Matthias Janosi, Andras and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.

Yu Lu, Kaveh Hajifathalian, Majid Ezzati, Mark Woodward, Eric B Rimm, Goodarz Danaei, Randi Selmer, Bjørn Heine Strand, Xin Fang, et al. Metabolic mediators of the effects of body-mass index, overweight, and obesity on coronary heart disease and stroke: A pooled analysis of 97 prospective cohorts with $1 \cdot 8$ million participants. *Lancet*, 383:970–983,

2014. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3959199/ (visited on 23th February 2024).

Appelman YE. Maas AH. Gender differences in coronary heart disease. *Journal of clinical medicine*, 18(12):598–602, 2010. doi: 10.1007/s12471-010-0841-y. url=https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3018605, PMID: 21301622; PMCID: PMC3018605 visited: 24th/Feb/ 2024.

Balistreri C. R. De Rosa S. Muscoli S. Selvaggio S. Selvaggio G. Ferdinandy P. De Caterina R. Madonna, R. Impact of sex differences and diabetes on coronary atherosclerosis and ischemic heart disease. *Journal of clinical medicine*, 98(8(1)), 2019. url=https://doi.org/10.3390/jcm8010098.

Jennifer L Rodgers, Jarrod Jones, Samuel I Bolleddu, Sahit Vanthenapalli, Lydia E Rodgers, Kinjal Shah, Krishna Karia, and Siva K Panguluri. Cardiovascular risks associated with gender and aging. *J Cardiovasc Dev Dis*, 6(2):19, 2019. doi: 10.3390/jcdd6020019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616540/B6-jcdd-06-00019 (visited on 23th February 2024).

David Scott. *xtable: Export Tables to LaTeX or HTML*, 2019. https://cran.r-project.org/package=xtable, (visited on 2nd July 2023).

Andrew Tyre. *scaler: Center and Scale Data Frames*, 2024. R package version 0.0.0.9000.

Version 4.3.0, R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023. URL https://CRAN.R-project.org/package=dplyr. R package version 1.1.2.

Appendix

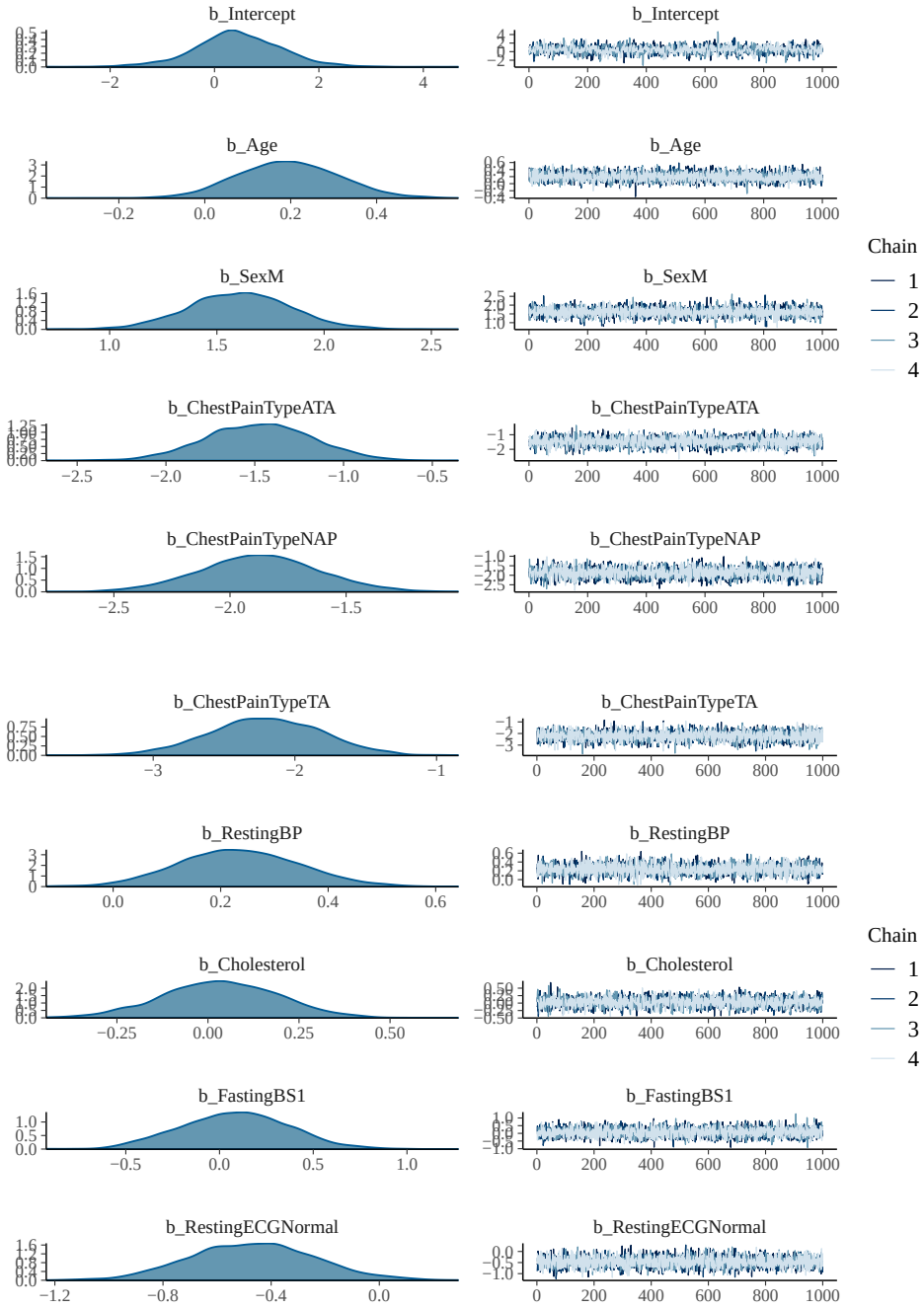


Figure 3: Plot of model with all variables

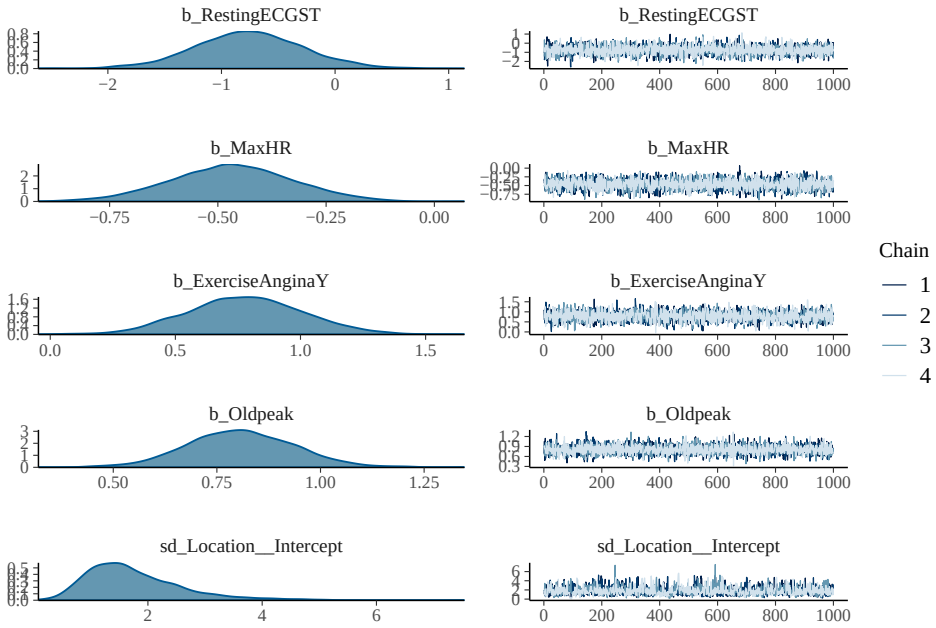


Figure 4: Plot of model with all variables

Table 8: Table showing Bulk_ESS and Tail_ESS model with all predictors

	BulkESS	TailESS
Intercept	4185.75	2650.66
Age	4750.72	3061.62
SexM	5186.46	2976.90
ChestPainTypeATA	4869.94	3102.64
ChestPainTypeNAP 4351.19	3060.43	
ChestPainTypeTA	4975.75	3174.15
RestingBP	4483.68	3368.49
Cholesterol	4772.05	3115.53
FastingBS1	5640.57	3218.59
MaxHR	4456.63	2911.16
RestingECGNormal	4751.72	3032.99
RestingECGST	4715.35	3280.38
ExerciseAnginaY	4805.31	3245.57
Oldpeak	4299.06	2602.26
ST_SlopeFlat	3538.64	3220.99
ST_SlopeUp	3032.62	3016.14

Table 9: Table showing Bulk_ESS and Tail_ESS of the model with all variables and Location as random effect

	Bulk_ESS	Tail_ESS
Intercept	1652.08	2100.34
Age	4727.64	3100.16
SexM	5208.97	2558.65
ChestPainTypeATA	4942.11	3036.80
ChestPainTypeNAP	4664.76	3329.71
ChestPainTypeTA	5060.67	2820.66
RestingBP	4031.57	2863.35
Cholesterol	2870.26	2327.83
FastingBS1	5072.05	2704.31
MaxHR	4660.72	3075.49
RestingECGNormal	4384.54	2893.05
RestingECGST	5035.15	2999.66
ExerciseAnginaY	4838.76	3009.93
Oldpeak	3773.33	3212.44
ST_SlopeFlat	3310.31	3072.87
ST_SlopeUp	3090.25	3104.40
Group-Level Effects Location		
sd(Intercept)	1091	1473

Table 10: Table showing Bulk_ESS and Tail_ESS of the model with all variables and interaction effects

	Bulk_ESS	Tail_ESS
Intercept	3304.67	3006.94
Age	2960.00	2902.57
SexM	4148.75	3244.88
ChestPainTypeATA	4599.29	3294.60
ChestPainTypeNAP	4472.43	2890.96
ChestPainTypeTA	4414.98	3274.89
RestingBP	2670.98	2772.33
Cholesterol	3962.40	3266.98
FastingBS1	3000.82	3024.43
MaxHR	4386.70	3060.97
RestingECGNormal	4348.50	2868.20
RestingECGST	3989.33	2825.09
ExerciseAnginaY	4316.47	2810.36
Oldpeak	3885.76	2962.19
ST_SlopeFlat	2622.99	2994.77
ST_SlopeUp	2689.67	2846.39
Age:ChestPainTypeATA	3862.97	3085.27
Age:ChestPainTypeNAP	3589.50	3013.65
Age:ChestPainTypeTA	4429.66	3244.63
Age:RestingBP	4392.91	3211.18
SexM:FastingBS1	2948.06	3284.87
SexM:RestingBP	2669.37	2932.90

Table 11: Table showing Bulk_ESS and Tail_ESS of the model with all variables and interaction effects and Location as random effect

	Bulk_ESS	Tail_ESS
Intercept	1605.72	1923.18
Age	3740.47	3125.30
SexM	5292.87	2942.42
ChestPainTypeATA	5624.97	3062.28
ChestPainTypeNAP	5212.47	3164.91
ChestPainTypeTA	5223.32	3041.01
RestingBP	3077.80	3070.82
Cholesterol	2344.29	1816.89
FastingBS1	3498.99	3274.99
MaxHR	4527.27	2815.51
RestingECGNormal	6063.87	3156.24
RestingECGST	5254.68	3071.09
ExerciseAnginaY	6168.56	2992.13
Oldpeak	4119.96	3240.08
ST_SlopeFlat	2860.53	2862.55
ST_SlopeUp	2883.67	2991.27
Age:ChestPainTypeATA	4669.21	2935.64
Age:ChestPainTypeNAP	4303.91	3338.98
Age:ChestPainTypeTA	5919.64	3586.61
Age:RestingBP	5459.24	2977.50
SexM:FastingBS1	3428.23	2976.20
SexM:RestingBP	3169.05	3247.21
Group-Level Effects Location		
sd(Intercept)	871	644