university of groningen / faculty of science and engineering / mathematics and applied mathematics

A Comparison of Frequentist and Bayesian Approaches to Variable Selection in Logistic Regression for Heart Disease

Bachelor's Project Mathematics

November 2025

Student: Avram Zanana

First supervisor: Prof. M.A. Grzegorczyk

Second assessor: dr. W.P. Krijnen

Abstract

Variable selection is critical in medical prediction models to identify important risk factors while maintaining interpretability. Two statistical frameworks, frequentist and Bayesian, offer distinct approaches. This study systematically compares the two frameworks in the logistic regression context using heart disease prediction as a case study. We analyzed the Cleveland Heart Disease dataset (n=303, 13 predictors) using both approaches. Both methods show strong agreement, selecting five core predictors (ca, cp, thal, sex, oldpeak) with inclusion probabilities larger than 0.5. The methods disagreed only on two marginal predictors. The choice between methods depends on specific inference goals and computational resources. As computational assessment reveals that the frequentist approach is 3-4x faster, we fit a final model using the frequentist approach.

Contents

1 Introduction

Heart disease is the leading cause of death worldwide, taking an estimated 19.8 million lives each year. This mortality also places a huge economic burden on society, causing hundreds of billions in healthcare costs and lost productivity. Cardiovascular diseases include a large spectrum of conditions like coronary artery disease, heart attacks and heart failure. Accurate prediction and interpretation of cardiac risk factors is crucial, as early detection enables timely interventions that can significantly improve patient outcomes and reduce healthcare costs.

In the modern day, physicians have access to massive amounts of patient data ranging from lab results to stress tests and imaging. For heart disease specifically, these measurements can be cholesterol levels, blood sugar levels, maximum heart rate and so on. However, more data does not automatically mean a better model, and the abundance presents a methodological challenge: determining which predictors are truly informative in terms of making a decision.

Variable selection is critical, especially for interpreting medical data. An example is the fact that doctors need to explain decisions to patients in a summarized way, so patients (who may not be doctors) can have a better understanding of their situation. "You're at high risk because of X,Y,Z" makes sense, while "You're at high risk because of these 47 variables" doesn't work in this context. From a clinical perspective, simpler models are preferable: they are easier to implement in practice, more transparent for medical decision-making, and less prone to overfitting when applied to new patient populations. Variable selection addresses this challenge by identifying a select subset of predictors that maintains predictive accuracy while making sure that interpreting the model is feasible.

The problem of selecting predictors has been studied extensively, although usually within one paradigm.

Two major statistical frameworks offer distinct approaches to variable selection: frequentist and Bayesian inference. While both frameworks have been applied successfully to medical prediction problems, direct comparisons of their performance on the same dataset are relatively rare, particularly in the context of heart disease prediction with mixed continuous and categorical predictors. We used the Cleveland Heart Disease dataset to conduct all of our analyses.

Frequentist variable selection relies on the fact that the regression parameters (β coefficents) are fixed but take unknown values. Methods such as bootstrap resampling combined with information criteria are utilized via assessing variable importance through repeated model fitting on resampled data, giving us inclusion frequencies that quantify selection choices. We use BIC as our information criterion, meaning we penalize complexity in our model. BIC has the nice property of being asymptotically consistent, in other words with increasing sample size it gets better at finding the true model. One problem that arises here is that different samples might pick different models. (Efron, 1979) proposes the "bootstrap" solution to this problem. This way, we can fit the model multiple times on the same resampled dataset, recording the selected variables via BIC each time. The "inclusion frequency" of a variable is then the proportion of bootstrap samples including that variable. Some selection threshold for the variables is then applied to form a final subset of stable variables.

On the other hand, Bayesian variable selection takes a fundamentally different approach. This time we assume that both the parameters and models are random with certain probability distributions. These probability distributions, also called priors, represent the uncertainty concerning which model is correct. Assume that the data contains p predictors. Using Bayes' Theorem, instead of picking one best model, we can consider all 2^p possible models simultaneously, and compute their posterior probabilities. For variable selection, we need to consider the Posterior Inclusion Probability(PIP) of each model, which is equivalent to the sum of posterior probabilities of all models including the variable. However, with a large number of variables, the number of all models 2^p grows very large. How do we explore each model when there are so many of them? The answer is given by (Green, 1995), which introduces Reversible Jump Markov Chain Monte Carlo (RJMCMC). This algorithm tackles the model exploration problem by proposing two types of moves, within model and between model moves. Namely, within model moves update(or don't) the β's according to the Metropolis-Hastings ratio. The between model moves are proposals to add/remove a certain variable from the model. After a certain number of burn-in iterations to come close to the stationary distribution, the remaining iterations determine the PIP's for each variable.

Extensive work exists on variable selection within each paradigm separately. On the frequentist side, (Hosmer, Lemeshow, and Sturdivant, 2013) discusses a comprehensive treatment of logistic regression for

medical data, covering stepwise model selection, information criteria, model diagnostics and so on. (Hastie, Tibshirani, and Friedman, 2009) has sections on theoretical properties of model selection using AIC and BIC. On the Bayesian side, (O'hara and Sillanpää, 2009) presents a comprehensive review of Bayesian variable selection methods.

The key gap motivating our study is the fact that the aforementioned papers majorly stay within the same paradigm. Although rare, there are existing comparison studies such as (Lu, Chow, and Loken, 2017), however the context is different. This paper for example compares methods for factor analysis models in the Psychology context, not logistic regression for clinical risk prediction. Our study aims to address this gap by directly comparing methods from two different paradigms on real-world health disease data.

This study therefore addresses multiple research questions. The foremost question is "Do the methods agree on which variables to select?". We answer this question via computing correlations between the results of both methods. Sample size stability is the next concern, "How does selection stability change when sample size decreases?". Also by assessing computational efficiency, we report method performance. Then we arrive at a practical recommendation: "Given the results, which approach is more practical for building clinically useful models?"

The remainder of this paper is as follows. Section 2 describes the Cleveland Heart Disease dataset, including preprocessing, missing data handling, and exploratory data analysis. Section 3 presents the methodological framework, logistic regression, bootstrap + BIC, and Bayesian RJMCMC. Section 4 reports results, beginning with implementation details and convergence diagnostics, then presenting variable selection results on the full dataset, computational performance, and the final selected model. Section 5 concludes with a discussion.

2 Cleveland Heart Disease Dataset

In this project, we work with the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. Detrano et al., 1989 This dataset is commonly used as a benchmark in many modeling projects, especially in the area of cardiovascular disease prediction. The dataset comprises 303 patient records, each patient is characterized by 13 clinical attributes and one target variable indicating the severity and presence of heart disease.

The predictor variables are a mix of continuous and categorical. Categorical predictors are sex, chest pain type(`cp`), fasting blood sugar(`fbs`), resting electrocardiogram(`restecg`), exercise induced angina(`exang`), slope of the peak exercise ST segment(`slope`), and thalassemia status(`thal`). The target variable represents heart disease status(`heart_disease`). Table 1 shows all variables in detail.

In total the missing data is only 4 observations of `ca`(1.98%) and 2 observations of `thal`(0.66%). We cannot use such rows with missing values. As `thal` is a categorical variable and `ca` is continuous, we use mode imputation for both, where the missing observations are 'imputed' or completed using the most common observation of the remainder of the dataset. This had negligible impact:

Table 2: Impact of Mode Imputation on Distributional Parameters

Variable	Metric	Before	After	Difference	Change (%)
CA	Mean	0.672	0.663	−0.009	−1.32
	SD	0.937	0.934	−0.003	−0.33
THAL	Mean	4.734	4.723	−0.011	−0.24
	SD	1.940	1.938	−0.001	−0.07

Note: CA = number of major vessels colored by fluoroscopy, treated as continuous (4 values imputed with mode = 0); THAL = thalassemia status, categorical (2 values imputed with mode = 3). All distributional changes are negligible (<2%), confirming minimal impact of the imputation strategy.

Variable	Type	Values/Units	Description (preprocessing)
age	numeric	years	Patient age.
sex	categorical	0=Female, 1=Male	Recoded to factor with labels Female/Male.
cp	categorical	1=Typical angina; 2=Atypical angina; 3=Non-anginal pain; 4=Asymptomatic	Chest pain type.
trestbps	numeric	mm Hg	Resting blood pressure (on admission).
chol	numeric	mg/dl	Serum cholesterol.
fbs	categorical	0: $\leq$120 mg/dl; 1: >120 mg/dl	Fasting blood sugar.
restecg	categorical	0=Normal; 1=ST–T abnormality; 2=LVH	Resting ECG.
thalach	numeric	bpm	Maximum heart rate achieved.
exang	categorical	0=No; 1=Yes	Exercise-induced angina.
oldpeak	numeric	ST depression	ST depression induced by exercise.
slope	categorical	1=Upsloping; 2=Flat; 3=Downsloping	Slope of peak exercise ST segment.
ca	integer	0–3	Number of major vessels by fluoroscopy; missing set to 0.
thal	categorical	3=Normal; 6=Fixed; 7=Reversible	Thalassemia; missing set to 3 (Normal).
num	integer	0–4	Original diagnosis code. Not used
heart_disease	binary factor	0=No; 1=Yes	Heart disease status.

Table 1: Cleveland Heart Disease variables with descriptions.

2.1 Exploratory Data Analysis

Following data preprocessing, we examined the distribution of the outcome variable and the relationship between clinical predictors and heart disease presence. Figure 1 presents the distribution of heart disease cases in the dataset and the patterns of key clinical variables by disease status.

The dataset exhibits a relatively balanced class distribution, with 139 cases (45.9%) showing evidence of heart disease and 164 cases (54.1%) with no disease. This balance is favorable for predictive modeling, reducing concerns about class imbalance that could bias model performance.

The clinical variable analysis reveals several notable patterns. Patients with heart disease tend to be older, show higher maximum heart rate (thalach), and exhibit greater ST depression (oldpeak) during exercise. Categorical variables such as chest pain type (cp) and the number of major vessels colored by fluoroscopy (ca) also demonstrate marked differences between disease groups, suggesting their potential importance as predictive features. Figure 2 shows the continuous variables by heart disease status:

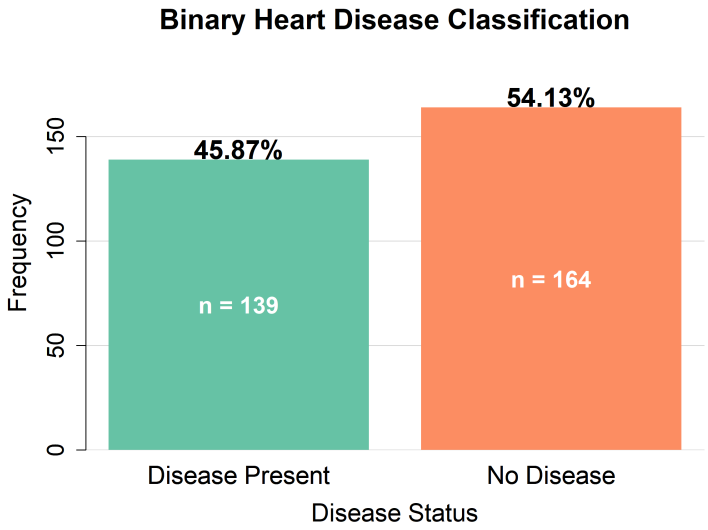

Figure 1: Distribution of heart disease diagnosis in the Cleveland dataset. The outcome shows 139 cases (45.9%) with heart disease and 164 cases (54.1%) without disease, indicating a relatively balanced dataset.

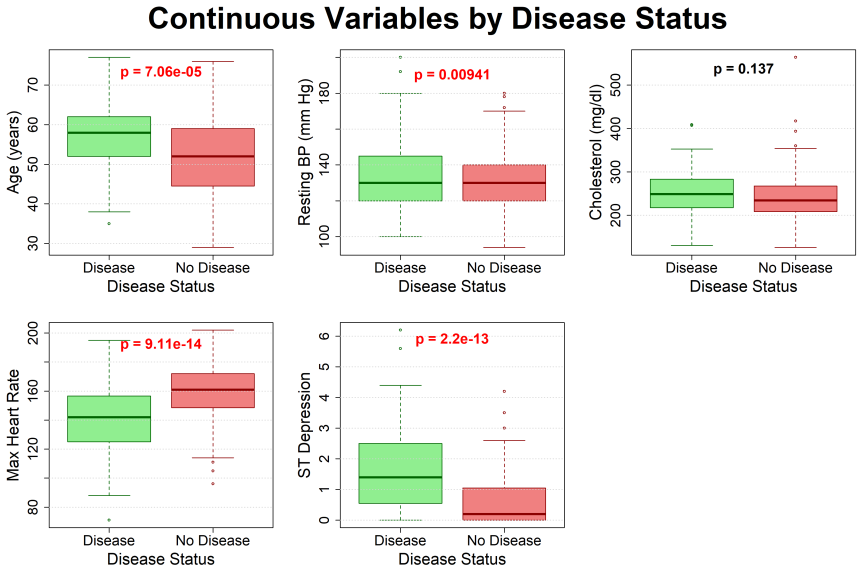

Figure 2: Continuous variables by heart disease status. Age, Max Heart Rate and ST Depression show clear differences on disease presence.

Table 3: Dummy variables extracted from categorical variables. A categorical variable with K levels generates $K-1$ dummy variables.

Variable	Reference Level	Dummies	Dummy Variable Names
sex	Female	1	sexMale
cp	Typical Angina	3	cpAtypical Angina, cpNon-anginal Pain, cpAsymptomatic
fbs	$\leq$120 mg/dl	1	fbs>120 mg/dl
restecg	Normal	2	restecgST-T Abnormality, restecgLV Hypertrophy
exang	No	1	exangYes
slope	Upsloping	2	slopeFlat, slopeDownsloping
thal	Normal	2	thalFixed Defect, thalReversible Defect
Total (7 categorical variables)		**12**	**12 dummy variables**

2.1.1 Categorical Variable Encoding

For logistic regression, categorical variables were encoded using dummy (treatment) coding via R's `model.matrix()` function. Under this scheme, a categorical variable with K levels generates $K-1$ binary indicator variables, with the first level serving as the reference category.

The 7 categorical predictor variables (sex, cp, fbs, restecg, exang, slope, thal) generated 12 dummy variables:

- 3 binary variables $\rightarrow$ 3 dummies
- 3 three-level variables $\rightarrow$ 6 dummies
- 1 four-level variable $\rightarrow$ 3 dummies

Table 3 details the reference levels and dummy variable names for each categorical predictor.

In the Bayesian variable selection analysis (RJMCMC), categorical variables are treated as grouped units: all dummy variables from a single categorical predictor are selected or excluded together. This ensures model interpretability and prevents nonsensical partial inclusion of factor levels. For instance, if `cp` (chest pain type) is selected, all three associated dummy variables (`cpAtypical Angina`, `cpNon-anginal Pain`, `cpAsymptomatic`) are simultaneously included in the model. The frequentist BIC-based variable selection employs the same grouping principle for consistency. Figure 3 shows how the categorical variables are distributed:

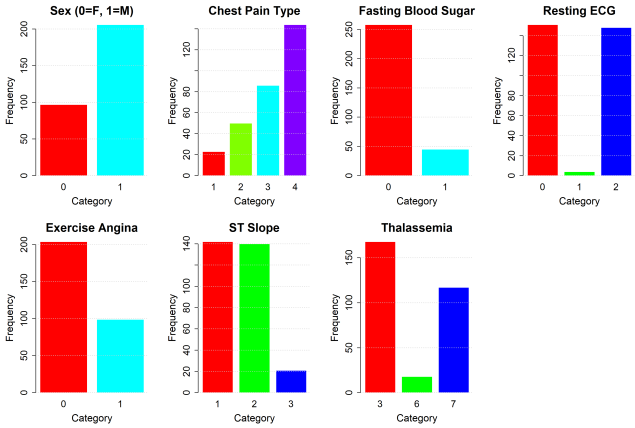

Figure 3: Distribution of the categorical variables. Major Vessels(`ca`) is also included here since it takes integer values. `ca` is treated as continuous nonetheless.

3 Methods

In this section we explain the methods in detail. We first construct a design matrix, concatenating the variables in order to fit our model. In Section 3.1 we introduce the statistical framework we use, Logistic Regression. Section 3.2 details the Frequentist approach to variable selection, while Section 3.3 specifies the Bayesian approach.

Design Matrix Construction

The full model design matrix was constructed as:

$$\mathbf{X} = [\mathbf{1}, \mathbf{X}_{\text{continuous}}, \mathbf{X}_{\text{dummy}}] \tag{1}$$

where $\mathbf{1}$ is an $n \times 1$ intercept vector, $\mathbf{X}_{\text{continuous}}$ contains 6 continuous variables (age, trestbps, chol, thalach, oldpeak, ca), and $\mathbf{X}_{\text{dummy}}$ contains 12 dummy variables from 7 categorical predictors, yielding 19 total columns.

For the Bayesian RJMCMC analysis, we standardized all columns except intercept to improve the MCMC simulation, namely:

$$x_{ij}^{\text{std}} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{2}$$

where x_{ij}^{std} denotes the entry of the standardized design matrix corresponding to row i and column j, and $\bar{x}_j$ and s_j are the sample mean and standard deviation of predictor j, respectively. Standardization ensures that all coefficients are on comparable scales, which is critical for proposing appropriate step sizes in Metropolis-Hastings updates. Otherwise the step sizes corresponding to different variables would not fit at times, due to big variations in the scale of the data. For example; `age` ranges from 29 to 77 while `chol` ranges from 127 to 564. This leads to step sizes at different scales.

3.1 Statistical Framework: Logistic Regression

Given the binary outcome $Y_i \in \{0, 1\}$, we employed logistic regression throughout this study. Logistic regression provides interpretable coefficients, handles both continuous and categorical predictors, and imposes no assumptions on the distribution of parameters. Let p denote the number of predictor variables and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$ the coefficient vector, where β_0 is the intercept and β_j for $j = 1, \ldots, p$ are slope coefficients.

For observation i, let $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^{p+1}$ denote the vector of predictors (including intercept). The model specifies:

$$\log\left(\frac{P(Y_i = 1 \mid \mathbf{x}_i)}{1 - P(Y_i = 1 \mid \mathbf{x}_i)}\right) = \mathbf{x}_i^T \boldsymbol{\beta} \tag{3}$$

Equivalently, the probability of disease presence is:

$$P(Y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})} \tag{4}$$

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the dataset. The log-likelihood function is (Hosmer, Lemeshow, and Sturdivant, 2013, p. 9, Eq. 1.4):

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})) \right] \tag{5}$$

where n is the number of observations. In all implementations, we used numerically stable computation of $\log(1 + \exp(z))$ to avoid overflow when z is large. The standard computation $\log(1 + \exp(z))$ overflows for $z > 700$. We instead use:

$$\log(1+\exp(z)) = \begin{cases} z + \log(1+\exp(-z)) & \text{if } z > 0 \\ \log(1+\exp(z)) & \text{otherwise} \end{cases} \tag{6}$$

When $z > 0$, we compute $z + \log(1 + \exp(-z))$ where $\exp(-z)$ is small and numerically stable. This is equivalent since:

$$z + \log(1+\exp(-z)) = \log(\exp(z)) + \log(1+\exp(-z)) = \log(\exp(z)(1+\exp(-z))) = \log(\exp(z)+1)$$

(Blanchard, D. J. Higham, and N. J. Higham, 2019) computes a generalization of this.

3.2 Variable Selection: Frequentist Approach

The frequentist approach combines bootstrap resampling with BIC-based model selection to assess the importance of predictor variables. Unlike single-model selection, which produces a binary include/exclude decision for each variable, the bootstrap allows us to see how likely a covariate is included in the model over many iterations.

3.2.1 Bootstrap Resampling Framework

By repeatedly drawing samples from $\mathcal{D}$ with replacement, we can simulate the process of drawing new datasets from the same population and thereby assess how variable our results are across hypothetical repetitions of the study.

To assess the stability of variable selection, we employ bootstrap resampling. For $b = 1, \ldots, B$, we generate bootstrap samples $\mathcal{D}^{(b)}$ by sampling n observations from $\mathcal{D}$ with replacement:

$$\mathcal{D}^{(b)} = \{(\mathbf{x}_i^{(b)}, y_i^{(b)})\}_{i=1}^n \tag{7}$$

where each $(\mathbf{x}_i^{(b)}, y_i^{(b)})$ is drawn uniformly from $\mathcal{D}$. Under the empirical distribution, the bootstrap approximates the sampling distribution of any statistic computed from $\mathcal{D}$. We are then able to compute variable inclusion probabilities, as for each resampling a copy of the full model is fitted. By aggregating these models and averaging the predictors over them, we see the frequency of variables being included. (Efron, 1979)

3.2.2 Model Selection via BIC

Having generated bootstrap sample $\mathcal{D}^{(b)}$, we must select a subset of predictors from the p available variables. This requires a criterion to compare models of different sizes. We employ the Bayesian Information Criterion (BIC), which balances model fit against model complexity.

For a given model $M \subseteq \{1, \ldots, p\}$ indexing a subset of predictors, let $\boldsymbol{\beta}_M \in \mathbb{R}^{|M|+1}$ denote the coefficients for the intercept and predictors in M. The maximum likelihood estimator is:

$$\hat{\boldsymbol{\beta}}_M = \arg\max_{\boldsymbol{\beta}_M} \ell(\boldsymbol{\beta}_M; \mathcal{D}) \tag{8}$$

where ℓ is the log-likelihood defined in Section 3.1.

The Bayesian Information Criterion (BIC) for model M is:

$$\text{BIC}(M) = -2\ell(\hat{\boldsymbol{\beta}}_M) + |M| \cdot \log(n) \tag{9}$$

where $|M|$ is the number of predictors in M. The first term measures how well the model fits the data (smaller is better), while the second term penalizes the number of parameters. Models with lower BIC are preferred. (Schwarz, 1978)

For each bootstrap sample $\mathcal{D}^{(b)}$, we perform backward stepwise selection:

1. Initialize with the full model $M_0 = \{1, \ldots, p\}$

2. At step t, given current model M_t, compute[1]:

$$M_{t+1} = \begin{cases} M_t \setminus \{j^*\} & \text{if } \min_{j \in M_t} \text{BIC}(M_t \setminus \{j\}) < \text{BIC}(M_t) \\ M_t & \text{otherwise} \end{cases} \tag{10}$$

where $j^* = \arg\min_{j \in M_t} \text{BIC}(M_t \setminus \{j\})$

3. Stop when $M_{t+1} = M_t$, yielding final model $M^{(b)}$

Let $1(j \in M^{(b)})$ denote the indicator that predictor j is in the final model for bootstrap sample b. The inclusion frequency for predictor j is:

$$\text{IF}_j = \frac{1}{B} \sum_{b=1}^{B} 1(j \in M^{(b)}) \tag{11}$$

This counts the proportion of bootstrap samples selecting predictor j.

The maximum likelihood estimation in step 2 is performed via *iteratively reweighted least squares* (IRLS), as implemented in R's `glm()` function . IRLS solves the score equations of the logistic regression model by iteratively updating the coefficient estimates. (McCullagh and Nelder, 1989) gives us the following algorithm: At iteration t, given current estimate $\boldsymbol{\beta}^{(t)}$, the algorithm computes

$$\boldsymbol{\beta}^{(t+1)} = \left(\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)},$$

where $\mathbf{X}$ is the design matrix, $\mathbf{W}^{(t)}$ is a diagonal weight matrix with elements

$$w_i^{(t)} = \pi_i^{(t)} \left(1 - \pi_i^{(t)}\right),$$

and $\pi_i^{(t)} = \left[1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta}^{(t)})\right]^{-1}$ are the fitted probabilities. The adjusted dependent variable $\mathbf{z}^{(t)}$ has entries

$$z_i^{(t)} = \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)} + \frac{y_i - \pi_i^{(t)}}{w_i^{(t)}}.$$

Iterations continue until convergence of $\boldsymbol{\beta}$ or the log-likelihood.

To reduce computational time, bootstrap replicates were executed in parallel. On Windows systems, we used socket-based parallelization with `parallel::makeCluster()`, allocating up to 4 CPU cores[2]:

```
cores <- max(1, min(parallel::detectCores(), 4))
cl <- parallel::makeCluster(cores)
parallel::clusterSetRNGStream(cl, seed)
res_list <- parallel::parLapply(cl, seeds, bootstrap_one)
```

3.3 Variable Selection: Bayesian Approach

3.3.1 Prior Specification

The Bayesian framework requires prior distributions for both model indicators and coefficients.

Model Prior. We assign equal prior probability to each model in the model space:

$$P(\boldsymbol{\gamma}) = \frac{1}{2^p} \quad \forall \boldsymbol{\gamma} \in \{0,1\}^p \tag{12}$$

This uniform model prior implies that each predictor has marginal prior inclusion probability $P(\gamma_j = 1) = 0.5$, showing no preference for inclusion or exclusion of said covariate.

[1] All dummy columns representing a categorical variable are removed simultaneously when one is excluded.

[2] The seed was changed for every iteration, to ensure independence acrross the bootstrap samples.

Coefficient Priors. For model $\boldsymbol{\gamma}$, the coefficients $\boldsymbol{\beta}_\gamma$ are assigned independent normal priors:

$$\pi(\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}) = \prod_{j:\gamma_j=1} p(\beta_j \mid \mu_0, \sigma_0^2) \tag{13}$$

where $p(\cdot)$ is the density of the $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution, $\mu_0 \in \mathbb{R}$ and $\sigma_0^2 > 0$ are hyperparameters. We set $\mu_0 = 0$ (centering coefficients at zero) and $\sigma_0^2 = 1$ (standard variance), providing a weakly informative prior that allows the data to dominate the posterior distribution later.

The intercept β_0 always receives the same prior specification and is never subject to selection ($\gamma_0 = 1$ always).

Posterior Distribution. Combining the likelihood, coefficient priors, and model prior via Bayes' theorem, the joint posterior is:

$$P(\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma} \mid \mathbf{y}) \propto P(\mathbf{y} \mid \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma})\, \pi(\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma})\, P(\boldsymbol{\gamma}) \tag{14}$$

where $P(\mathbf{y} \mid \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}) = \exp(\ell(\boldsymbol{\beta}_\gamma))$ with ℓ defined in Section 3.1, evaluated using only the predictors with $\gamma_j = 1$. Here, $P(\mathbf{y} \mid \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma})$ represents the likelihood, which can be computed as $\exp(\ell(\boldsymbol{\beta}_\gamma))$, where $\ell(\cdot)$ is the log-likelihood function.

3.3.2 Reversible Jump MCMC Algorithm

To sample from the posterior distribution over both discrete model indicators $\boldsymbol{\gamma}$ and continuous coefficients $\boldsymbol{\beta}_\gamma$, we employ Reversible Jump Markov Chain Monte Carlo (RJMCMC), introduced by Green (1995). RJMCMC extends standard MCMC to handle target distributions with varying dimension, allowing the chain to move between models with different numbers of parameters.

Algorithm Overview. At each iteration t, the algorithm proposes one of two types of moves with specified probabilities:

- **Within-model moves** (probability p): Update coefficient values $\boldsymbol{\beta}_\gamma$ for the current model $\boldsymbol{\gamma}$ using Metropolis-Hastings
- **Between-model moves** (probability $1 - p$): Propose adding or removing a predictor, transitioning to a different model dimension

Within-Model Moves: Metropolis-Hastings Updates. Given the current state $(\boldsymbol{\beta}_\gamma^{(t)}, \boldsymbol{\gamma}^{(t)})$, we update each active coefficient β_j with $\gamma_j^{(t)} = 1$ using a random-walk Metropolis-Hastings step:

1. For each j with $\gamma_j^{(t)} = 1$ (including intercept):
   - Propose: $\beta_j^* \sim \text{Uniform}(\beta_j^{(t)} - \epsilon, \beta_j^{(t)} + \epsilon)$ where $\epsilon > 0$ is the step size
   - Compute log-acceptance ratio:

$$\log \alpha_{\text{MH}} = \min\left(0, \ell(\boldsymbol{\beta}^*) - \ell(\boldsymbol{\beta}_\gamma^{(t)}) + \log \pi(\boldsymbol{\beta}^*) - \log \pi(\boldsymbol{\beta}_\gamma^{(t)})\right) \tag{15}$$

     where $\boldsymbol{\beta}^*$ differs from $\boldsymbol{\beta}_\gamma^{(t)}$ only in the j-th component.
   - Accept with probability $\alpha_{\text{MH}} = \exp(\log \alpha_{\text{MH}})$: set $\beta_j^{(t+1)} = \beta_j^*$ if accepted, otherwise $\beta_j^{(t+1)} = \beta_j^{(t)}$.

Since the proposal is symmetric, the Hastings ratio cancels. The acceptance ratio is then given on the log scale above.

Between-Model Moves. To try different models, we propose dimension-changing moves that add or remove a predictor:

1. Randomly select a predictor $j \in \{1, \ldots, p\}$
2. Propose toggle: $\gamma_j^* = 1 - \gamma_j^{(t)}$
3. If $\gamma_j^{(t)} = 0$:
   - Draw $\beta_j^* \sim p(\beta_j \mid \mu_0, \sigma_0^2)$ from the prior
   - Set $\boldsymbol{\beta}_{\gamma^*}^* = (\boldsymbol{\beta}_{\gamma}^{(t)}, \beta_j^*)$ (update coefficient vector)
4. If $\gamma_j^{(t)} = 1$:
   - Set $\beta_j^* = 0$ and remove from coefficient vector
   - Set $\boldsymbol{\beta}_{\gamma^*}^* = \boldsymbol{\beta}_{\gamma}^{(t)} \setminus \{\beta_j\}$ (reduce dimension, remove covariate)
5. Compute log-acceptance ratio and accept/reject (details below)

Reversible Jump Acceptance Ratio. For the dimension-changing move from $(\boldsymbol{\beta}_{\gamma}^{(t)}, \boldsymbol{\gamma}^{(t)})$ to $(\boldsymbol{\beta}_{\gamma^*}^*, \boldsymbol{\gamma}^*)$, the log-acceptance ratio is:

$$\log \alpha_{\text{RJ}} = \min\left(0, \ell(\boldsymbol{\beta}_{\gamma^*}^*) - \ell(\boldsymbol{\beta}_{\gamma}^{(t)}) + \log \pi(\boldsymbol{\beta}_{\gamma^*}^* \mid \boldsymbol{\gamma}^*) - \log \pi(\boldsymbol{\beta}_{\gamma}^{(t)} \mid \boldsymbol{\gamma}^{(t)})\right) \tag{16}$$

and the acceptance probability is $\alpha_{\text{RJ}} = \exp(\log \alpha_{\text{RJ}})$.
For our implementation:

- The model prior is uniform: $P(\boldsymbol{\gamma}^*) = P(\boldsymbol{\gamma}^{(t)})$, so this ratio equals 1
- For addition moves, we propose β_j^* from the prior $\pi(\beta_j^* \mid \gamma_j^* = 1)$, which exactly matches the prior term in the numerator
- The proposal probabilities for selecting predictor j are symmetric: $q(j) = 1/p$ in both directions
- The Jacobian simplifies to 1 for our proposal scheme

Thus the log-acceptance ratio simplifies to:

$$\log \alpha_{\text{RJ}} = \min\left(0, \ell(\boldsymbol{\beta}_{\gamma^*}^*) - \ell(\boldsymbol{\beta}_{\gamma}^{(t)})\right) \tag{17}$$

for addition moves. For removal moves, the prior on the deleted coefficient appears in the denominator, but since we proposed it from that same prior during addition, the detailed balance condition is maintained.

3.3.3 Grouped Variable Moves for Categorical Predictors

An important consideration needs to be addressed for categorical variables represented by multiple dummy columns in the design matrix $\mathbf{X}$. Allowing individual dummy variables to be selected independently would yield hard to interpret models. For example, including only the dummy for "Atypical Angina" from the chest pain type variable (cp) without the reference category creates a meaningless model.

Grouping Structure. Let $\mathcal{G} = \{G_1, \ldots, G_K\}$ partition the predictor indices $\{1, \ldots, p\}$, where each G_k contains the dummy variable indices for a single categorical predictor (or a singleton set for continuous predictors). In our case, for the Cleveland dataset:

- $G_1 = \{1\}$ (age, continuous)
- $G_2 = \{2\}$ (sex, one dummy)
- $G_3 = \{3, 4, 5\}$ (cp, three dummies for four-level factor)
- ... and so forth

Modified Between-Model Moves. We modify the addition and removal procedure to operate on groups rather than only on individual indices:

1. Randomly select a group $G_k \in \mathcal{G}$ with probability $1/K$
2. For all $j \in G_k$ simultaneously:
   - Toggle: $\gamma_j^* = 1 - \gamma_j^{(t)}$ for all $j \in G_k$
   - If addition ($\gamma_j^{(t)} = 0$ for all $j \in G_k$): draw $\beta_j^* \sim \mathcal{N}(\mu_0, \sigma_0^2)$ independently for each $j \in G_k$
   - If removal ($\gamma_j^{(t)} = 1$ for all $j \in G_k$): set $\beta_j^* = 0$ for all $j \in G_k$
3. Compute acceptance ratio:

$$\alpha_{\text{RJ}} = \min\left(1, \exp\left[\ell(\boldsymbol{\beta}_{\gamma^*}^*) - \ell(\boldsymbol{\beta}_{\gamma}^{(t)}) + \sum_{j \in G_k} \log \pi(\beta_j^*) - \sum_{j \in G_k} \log \pi(\beta_j^{(t)})\right]\right) \tag{18}$$

This grouped approach ensures that categorical variables are either fully included or fully excluded, so we do not encounter problems when interpreting the results.

3.3.4 Convergence and Posterior Inference

Burn-in Period. We discard the first T_{burn} iterations to allow the chain to reach its stationary distribution. Otherwise there are big jumps in the chain, poorly reflecting the stationary distribution. Let $\{(\boldsymbol{\beta}_{\gamma}^{(t)}, \boldsymbol{\gamma}^{(t)})\}_{t=1}^{T}$ denote the full MCMC output. The post-burn-in samples $\{(\boldsymbol{\beta}_{\gamma}^{(t)}, \boldsymbol{\gamma}^{(t)})\}_{t=T_{\text{burn}}+1}^{T}$ approximate draws from the posterior distribution $P(\boldsymbol{\beta}_{\gamma}, \boldsymbol{\gamma} \mid \mathbf{y})$. Section 4.2 demonstrates the relevant diagnostics.

Posterior Inclusion Probabilities. For predictor j, the posterior inclusion probability is estimated by:

$$\text{PIP}_j = P(\gamma_j = 1 \mid \mathbf{y}) \approx \frac{1}{T - T_{\text{burn}}} \sum_{t=T_{\text{burn}}+1}^{T} \gamma_j^{(t)} \tag{19}$$

This represents the proportion of post-burn-in iterations in which predictor j was included in the model. For grouped categorical variables, we define:

$$\text{PIP}_{G_k} = \frac{1}{T - T_{\text{burn}}} \sum_{t=T_{\text{burn}}+1}^{T} 1\left(\sum_{j \in G_k} \gamma_j^{(t)} > 0\right) \tag{20}$$

indicating the frequency with which any dummy from group G_k was active (equivalently, all dummies due to our grouped moves).

4 Results

In this section we present the results. First, in Section 4.1, technical implementation details like the parameters and computer specifications are shown. Section 4.2 presents the relevant convergence diagnostics for our MCMC simulation. We then present the computed inclusion frequencies of variables in Section 4.3. The stability of said inclusion frequencies is then analyzed using smaller sample sizes in Section 4.4. Section 4.5 demonstrates computational performance across our methods. Based on this, Section 4.6 details how we select a final model, directly comparing the two frameworks. Finally, in Section 4.7 we present the final model and relevant information.

4.1 Implementation Details

All analyses were conducted on a Windows 10 system equipped with an AMD Ryzen 7 5700U processor (8 cores) and 16GB RAM, using R version 4.3.2 and Python 3.10.11. For the frequentist approach, we used the `stats` package (**R-base**) for logistic regression via `glm()`, and the `parallel` package (**R-parallel**) for bootstrap parallelization. For the Bayesian RJMCMC, all computations were implemented from (Grzegorczyk, 2024), code used for both approaches are available on GitHub at (Zanana, 2025)

The frequentist bootstrap-BIC and Bayesian RJMCMC procedures were implemented with consistent settings to allow for fair comparison. For the bootstrap-BIC approach, we used $B = 300$ bootstrap samples. Model selection within each resample was based on the Bayesian Information Criterion (BIC), with penalty term $k_M \log(303) = 5.71 \cdot k_M$ for a model containing k_M parameters. The selection process followed a backward elimination strategy starting from the full model. To reduce computation time, bootstrap replicates were executed in parallel using up to 4 CPU cores. The random seed was fixed at 123 to ensure reproducibility.

For the Bayesian approach, the Reversible Jump MCMC (RJMCMC) algorithm was run for $T = 20{,}000$ iterations, with a burn-in period of $T_{\text{burn}} = 10{,}000$ iterations (50%). Within-model Metropolis-Hastings updates were selected with probability $p = 0.8$, while between-model proposals (adding or removing a predictor) were chosen with probability $1 - p = 0.2$. The Metropolis-Hastings step size was set to $\epsilon = 0.1$. Prior hyperparameters were $\mu_0 = 0$ and $\sigma_0^2 = 1$, and the model prior was uniform over the $2^{13} = 8{,}192$ possible models. The random seed was again set to 123 for consistency across procedures.

4.2 MCMC Convergence Diagnostics

To determine an appropriate number of iterations and burn-in period, we ran the RJMCMC algorithm for varying iteration counts (10,000; 20,000; 30,000; 40,000) on the full dataset and monitored convergence. We can use the log posterior to demostrate convergence:

Log Posterior $= \log P(\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma} \mid \mathbf{y}) = \ell(\boldsymbol{\beta}_\gamma) + \log \pi(\boldsymbol{\beta}_\gamma) + \log P(\boldsymbol{\gamma})$

Figure 4 shows the trace of the log posterior over iterations. The log posterior represents how well the current model fits the data plus the prior probability. Hence we are able to see the step sizes the chain takes. Initial iterations show high volatility as the chain explores different models, but the trace stabilizes after approximately 5,000-8,000 iterations, indicating the chain has reached regions of high posterior probability.

Table 4 compares the final posterior inclusion probabilities across different MCMC run lengths. The high consistency across run lengths (maximum difference ≤ 0.05 for all variables) confirms that 20,000 iterations yields stable posterior estimates.

Based on these diagnostics, we conclude that 20,000 total iterations with 10,000 burn-in provides reliable posterior inference while maintaining reasonable computational cost.

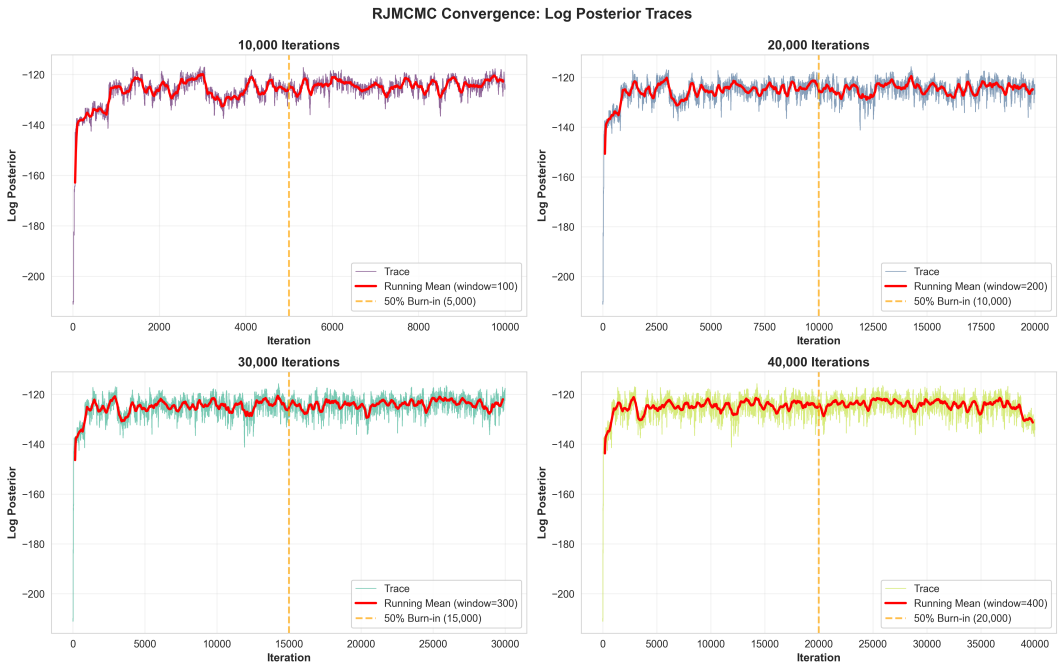

Figure 4: Log posterior trace for RJMCMC runs with different iteration counts. All chains stabilize after approximately 8,000 iterations, with similar plateau values indicating convergence to the same stationary distribution. The consistent behavior across different run lengths confirms that 20,000 total iterations (with 10,000 burn-in) is sufficient for reliable posterior inference.

Table 4: Comparison of Posterior Inclusion Probabilities Across Different MCMC Run Lengths

Variable	**10K iter**	**20K iter**	**30K iter**	**40K iter**
ca	1.000	1.000	1.000	1.000
cp	0.995	1.000	1.000	1.000
thal	0.862	0.868	0.871	0.870
oldpeak	0.815	0.820	0.818	0.822
sex	0.680	0.684	0.687	0.685
exang	0.548	0.552	0.555	0.550
slope	0.368	0.373	0.370	0.375
trestbps	0.355	0.359	0.362	0.358
thalach	0.345	0.349	0.351	0.348
chol	0.071	0.074	0.075	0.073
fbs	0.070	0.073	0.071	0.074
age	0.040	0.042	0.043	0.041
restecg	0.009	0.011	0.010	0.011

4.3 Variable Selection Results: Full Dataset

4.3.1 Frequentist Bootstrap-BIC Results

Table 5 presents the inclusion frequencies from 300 bootstrap samples. Six variables exceeded the 0.5 threshold commonly used for selection: `ca` (0.997), `cp` (0.860), `sex` (0.777), `thal` (0.720), `trestbps` (0.513), and `oldpeak` (0.503).

Table 5: Frequentist variable selection results obtained from 300 bootstrap samples using the BIC-based selection criterion. For each resample, a stepwise backward elimination was applied, and the inclusion frequency represents the proportion of bootstrap replicates in which a variable was retained in the final model. Variables with inclusion frequency greater than 0.5 are marked as selected.

Variable	**Inclusion Frequency**	**Selected (0.5)**
ca	0.997	Yes
cp	0.860	Yes
sex	0.777	Yes
thal	0.720	Yes
trestbps	0.513	Yes
oldpeak	0.503	Yes
thalach	0.443	No
exang	0.410	No
slope	0.397	No
chol	0.193	No
restecg	0.073	No
fbs	0.060	No
age	0.013	No

4.3.2 Bayesian RJMCMC Results

Table 6 presents the posterior inclusion probabilities from RJMCMC. Using the same 0.5 threshold, seven variables were selected: `ca` (1.000), `cp` (1.000), `thal` (0.868), `oldpeak` (0.820), `sex` (0.684), `exang` (0.552), and marginally `trestbps` at the boundary.

Table 6: Bayesian variable selection results obtained via Reversible Jump MCMC (RJMCMC) using 20,000 iterations and a burn-in of 10,000. The posterior inclusion probability indicates the estimated probability that each variable belongs to the true model, marginalizing over all possible subsets of predictors. Variables with inclusion probabilities greater than 0.5 are reported as selected. These probabilities quantify the strength of evidence for inclusion under the posterior distribution, reflecting both model uncertainty and prior regularization effects.

Variable	**Posterior Inclusion Prob.**	**Selected (0.5)**
ca	1.000	Yes
cp	1.000	Yes
thal	0.868	Yes
oldpeak	0.820	Yes
sex	0.684	Yes
exang	0.552	Yes
slope	0.373	No
trestbps	0.359	No
thalach	0.349	No
chol	0.074	No
fbs	0.073	No
age	0.042	No
restecg	0.011	No

4.4 Sample Size Effects

To assess how variable selection stability changes with sample size, we analyzed subsamples of the data: 50% ($n = 152$), 25% ($n = 76$), and 12.5% ($n = 38$).

Figure 5 shows a 2×2 grid comparing frequentist and Bayesian inclusion probabilities across all four data fractions. As sample size decreases, both methods show: Reduced confidence in variable selection (probabilities move toward 0.5), increased variability across replicates, and reservation of ranking for strong predictors (ca, cp, thal remain top-ranked)

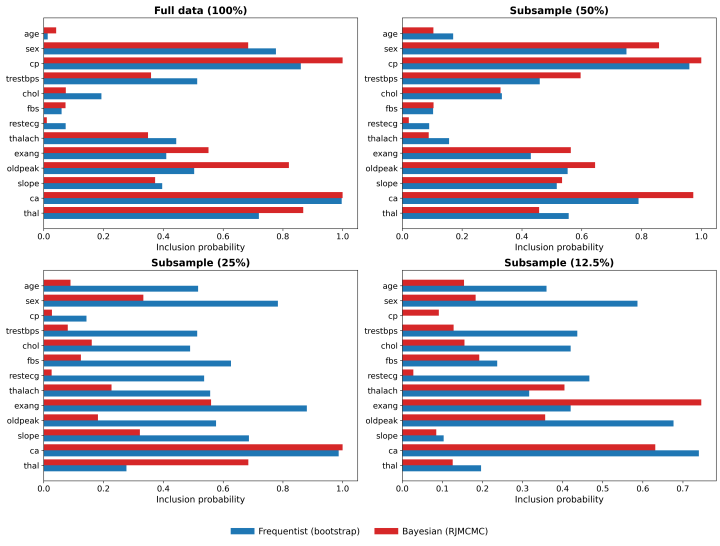

Figure 5: Side-by-side comparison of frequentist (blue) and Bayesian (orange) inclusion probabilities across four data fractions. Each panel shows one subsample size. Strong predictors maintain high inclusion probabilities even with small samples, while marginal predictors show increased uncertainty with reduced sample size.

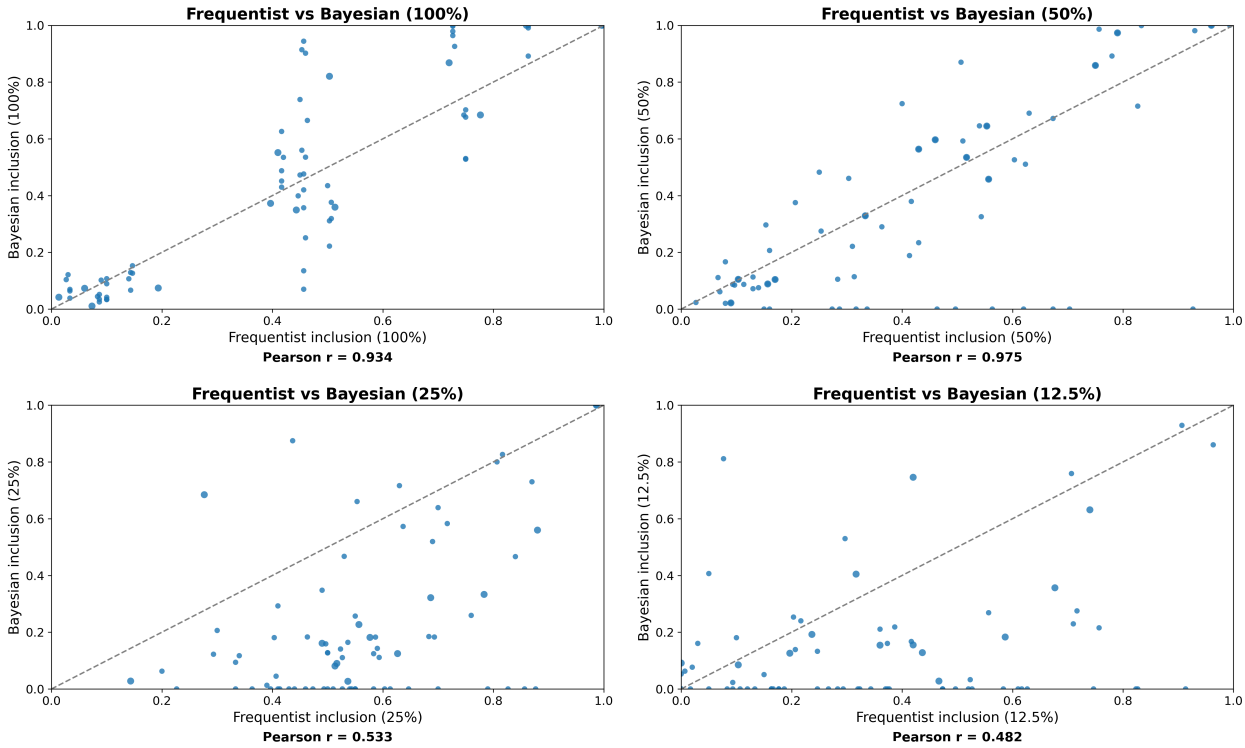

Figure 6: Scatter plots comparing frequentist and Bayesian inclusion probabilities for four data fractions. Each point represents one predictor variable. Pearson correlation coefficients show decreasing trend with smaller sample. The dashed line shows perfect agreement (y = x).

Fraction	Frequentist (s)	Bayesian (s)
100%	38.283 ± 0.180	130.934 ± 2.350
50%	36.553 ± 5.310	105.166 ± 8.887
25%	16.730 ± 0.903	57.891 ± 1.883
12.5%	18.630 ± 0.354	55.905 ± 1.376

Table 7: Wall-clock time (mean ± sd over 5 replicates).

Figure 6 shows scatter plots of frequentist versus Bayesian inclusion probabilities for each data fraction. Pearson correlations generally decrease with the decrasing sample size. This is to be expected as there is less data for inference.

4.5 Computational Performance

Table 7 shows wall-clock execution times across data fractions and methods. The frequentist approach consistently outperforms the Bayesian method by a factor of 3-4×, with both methods showing reduced computation time for smaller datasets.

4.6 Method Selection: Frequentist vs. Bayesian Approach

Table 8 presents a comparison of the frequentist BIC and Bayesian RJMCMC approaches for variable selection. The frequentist method was selected for the final model due to its substantially lower computational cost while producing comparable inclusion probabilities.

Table 8: Comparison of Frequentist and Bayesian Variable Selection Approaches

Variable	Frequentist BIC	Bayesian RJMCMC	Selected (Freq.)	Agreement
ca	0.997	1.000	Yes	✓
cp	0.860	1.000	Yes	✓
thal	0.720	0.868	Yes	✓
sex	0.777	0.684	Yes	✓
oldpeak	0.503	0.820	Yes	✓
trestbps	0.513	0.359	Yes	×
exang	0.410	0.552	No	×
thalach	0.443	0.349	No	✓
slope	0.397	0.373	No	✓
chol	0.193	0.074	No	✓
restecg	0.073	0.011	No	✓
fbs	0.060	0.073	No	✓
age	0.013	0.042	No	✓
Computational Time (Full Dataset):				
Frequentist (mean ± SD)	38.3 ± 0.2 seconds			
Bayesian (mean ± SD)	130.9 ± 2.5 seconds			
Speed advantage	3.4× faster			

Table 9: Final Logistic Regression Model Coefficients

Variable	Coefficient	Std. Error	z-value	p-value
(Intercept)	−3.428	0.652	−5.256	< 0.001
ca	1.124	0.225	4.993	< 0.001
cp (Atypical Angina)	0.554	0.711	0.780	0.435
cp (Non-anginal Pain)	−0.105	0.630	−0.167	0.867
cp (Asymptomatic)	1.724	0.611	2.820	0.005
exang (Yes)	0.978	0.392	2.493	0.013
oldpeak	0.649	0.199	3.264	0.001
thal (Fixed Defect)	0.800	0.700	1.143	0.253
thal (Reversible Defect)	1.843	0.366	5.041	< 0.001
Note: Reference categories are cp (Typical Angina) and thal (Normal)				
BIC = 270.934; Number of bootstrap iterations = 300				

4.7 Final Model Coefficients

The selected model achieved a BIC of 270.934 with 300 successful bootstrap iterations. Table 9 presents the estimated coefficients, standard errors, test statistics, and p-values for the final logistic regression model.

The strongest predictors in the final model were ca (number of major vessels colored by fluoroscopy; $\beta = 1.124$, $p < 0.001$), thal reversible defect ($\beta = 1.843$, $p < 0.001$), and asymptomatic chest pain ($\beta = 1.724$, $p = 0.005$). Note that exang (exercise-induced angina) appears in the best-fit model despite having an inclusion probability of 0.410, indicating it was frequently selected alongside the six core variables.

5 Conclusion

This study systematically compared frequentist bootstrap + BIC and Bayesian RJMCMC variable selection for heart disease prediction using the Cleveland Heart Disease dataset. Overall, we see strong agreement between methods: 85% agreement (11/13 variables) on full data. Upon our computational considerations, we found that due to the possibility of parallelization, the frequentist method proved to be 3 to 4 times faster (38s vs 131s) than the Bayesian method. Correlations between methods remained positive across all sample sizes, even though they decreased dramatically as sample size decreased. The most important predictors remain stable and identifiable.

The agreement between methods suggests that the selected variables' clinical relevance is high. Note that the selected variables span multiple domains, anatomical(ca), symptomatic(cp), hematological(thal), and functional(oldpeak). This aligns with the clinical understanding of heart disease, where factors from multiple domains are known to affect disease presence. Notably, traditional risk factors such as age, cholesterol, blood sugar had low inclusion probabilities in both methods. This does not imply that they are unimportant, instead, it means that they do not add marginal value in this dataset. Direct measures of disease provide more immediate diagnostic information in this setting.

As for the limitations, there are a few considerations. The first one is that we only utilize a single dataset coming from a single institution. The study can be generalized and validated using multiple heart disease datasets with diverse populations. Another point is that even though we compared the standard methods of bootstrap + BIC and RJMCMC, other methods do exist(LASSO, spike-and-slab priors, elastic nets etc.) that may give different results.

Both methods provide valuable frameworks and their agreement on core predictors boosts confidence in clinical relevance. The choice between them depends on the goal: frequentist offers us computational efficiency and simplicity, Bayesian provides us with an explicit uncertainty qualification. Ultimately, it all boils down to a trade-off between predictive accuracy and model interpretability. The value lies not in declaring one method best, but in understanding the best context to use it in.

Bibliography

Blanchard, Pierre, Desmond J. Higham, and Nicholas J. Higham (2019). "Accurate Computation of the Log-Sum-Exp and Softmax Functions". In: *arXiv preprint arXiv:1909.03469*. URL: https://arxiv.org/abs/1909.03469.

Detrano, Robert et al. (1989). *International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease*.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1, pp. 1–26. DOI: 10.1214/aos/1176344552. URL: https://doi.org/10.1214/aos/1176344552.

Green, Peter J. (Dec. 1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4, pp. 711–732. ISSN: 0006-3444. DOI: 10.1093/biomet/82.4.711. eprint: https://academic.oup.com/biomet/article-pdf/82/4/711/699533/82-4-711.pdf. URL: https://doi.org/10.1093/biomet/82.4.711.

Grzegorczyk, Marco (2024). *Lecture 10*. Course materials for Project Statistical Reasoning.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (Feb. 2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer. ISBN: 0387848576.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant (2013). *Applied Logistic Regression*. 3rd. Hoboken, NJ: John Wiley & Sons. ISBN: 9780470582473. DOI: 10.1002/9781118548387.

Lu, Zhiyong H, Sy-Miin Chow, and Eric Loken (2017). "A comparison of Bayesian and frequentist model selection methods for factor analysis models". In: *Psychological Methods* 22.2, pp. 361–381.

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models*. 2nd. London: Chapman and Hall. ISBN: 9780412317606.

O'hara, Robert B and Mikko J Sillanpää (2009). "A review of Bayesian variable selection methods: what, how and which". In: *Bayesian Analysis* 4.1, pp. 85–117.

Schwarz, Gideon (1978). "Estimating the dimension of a model". In: *Annals of Statistics* 6.2, pp. 461–464.

Zanana, Avram (2025). *Variable Selection Methods for Heart Disease Prediction*. https://github.com/azanana/heart-disease-variable-selection. Accessed: November 2025.