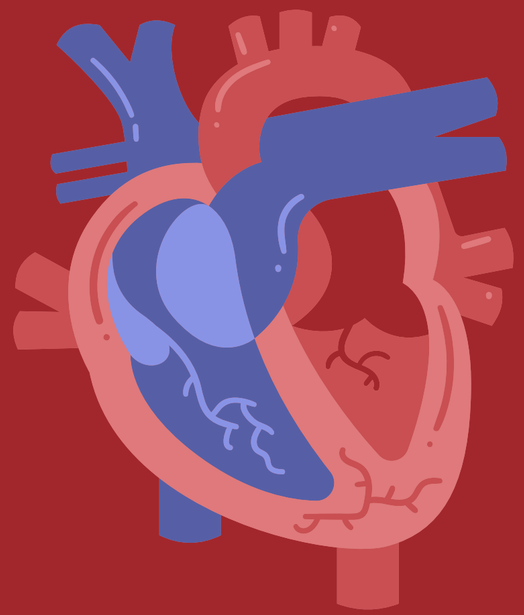
# HEART DISEASE RISK PREDICTION



Bayesian Logistic Regression

# BACKGROUND

17.8 MILLION have a heart disease

in 2023, 9 million people affected

## WHAT CAN CAUSE ?

# STATEMENT OF THE RESEARCH PROBLEM

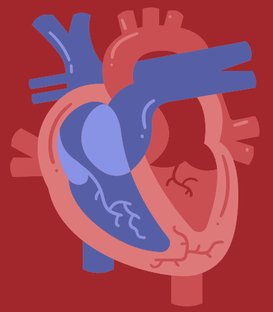
Heart disease is one of the most important diseases to be aware of and prevent due to the high number of deaths it causes. Most countries with the highest number of heart disease patients are low- and middle-income countries, as people lack information about the dangers of this disease and the preventive steps to avoid it. Unhealthy lifestyle like being sedentary, poor diet, and smoking, contribute to heart disease, increasing the risk of diabetes and obesity.

Limited access to healthcare, especially in low-income and remote areas, makes prevention and treatment difficult. High treatment costs also discourage many from seeking care.

# OBJECTIVES OF THE STUDY

The purpose of this project is to identify the risk factors that can cause someone to have cardiovascular disease.

# THE DATASET



contains several predictor variables related to cardiovascular disease risk

| age <int> | sex <int> | cp <int> | trestbps <int> | chol <int> | fbs <int> | restecg <int> | thalach <int> | exang <int> | oldpeak <dbl> |
|---|---|---|---|---|---|---|---|---|---|
| 52 | 1 | 0 | 125 | 212 | 0 | 1 | 168 | 0 | 1.0 |
| 53 | 1 | 0 | 140 | 203 | 1 | 0 | 155 | 1 | 3.1 |
| 70 | 1 | 0 | 145 | 174 | 0 | 1 | 125 | 1 | 2.6 |
| 61 | 1 | 0 | 148 | 203 | 0 | 1 | 161 | 0 | 0.0 |
| 62 | 0 | 0 | 138 | 294 | 1 | 1 | 106 | 0 | 1.9 |
| 58 | 0 | 0 | 100 | 248 | 0 | 0 | 122 | 0 | 1.0 |
| 58 | 1 | 0 | 114 | 318 | 0 | 2 | 140 | 0 | 4.4 |
| 55 | 1 | 0 | 160 | 289 | 0 | 0 | 145 | 1 | 0.8 |
| 46 | 1 | 0 | 120 | 249 | 0 | 0 | 144 | 0 | 0.8 |
| 54 | 1 | 0 | 122 | 286 | 0 | 0 | 116 | 1 | 3.2 |

1-10 of 1,025 rows | 1-10 of 14 columns

Previous 1 2 3 4 5 6 ... 100 Next

| chol <int> | fbs <int> | restecg <int> | thalach <int> | exang <int> | oldpeak <dbl> | slope <int> | ca <int> | thal <int> | target <int> |
|---|---|---|---|---|---|---|---|---|---|
| 212 | 0 | 1 | 168 | 0 | 1.0 | 2 | 2 | 3 | 0 |
| 203 | 1 | 0 | 155 | 1 | 3.1 | 0 | 0 | 3 | 0 |
| 174 | 0 | 1 | 125 | 1 | 2.6 | 0 | 0 | 3 | 0 |
| 203 | 0 | 1 | 161 | 0 | 0.0 | 2 | 1 | 3 | 0 |
| 294 | 1 | 1 | 106 | 0 | 1.9 | 1 | 3 | 2 | 0 |
| 248 | 0 | 0 | 122 | 0 | 1.0 | 1 | 0 | 2 | 1 |
| 318 | 0 | 2 | 140 | 0 | 4.4 | 0 | 3 | 1 | 0 |
| 289 | 0 | 0 | 145 | 1 | 0.8 | 1 | 1 | 3 | 0 |
| 249 | 0 | 0 | 144 | 0 | 0.8 | 2 | 0 | 3 | 0 |
| 286 | 0 | 0 | 116 | 1 | 3.2 | 1 | 2 | 2 | 0 |

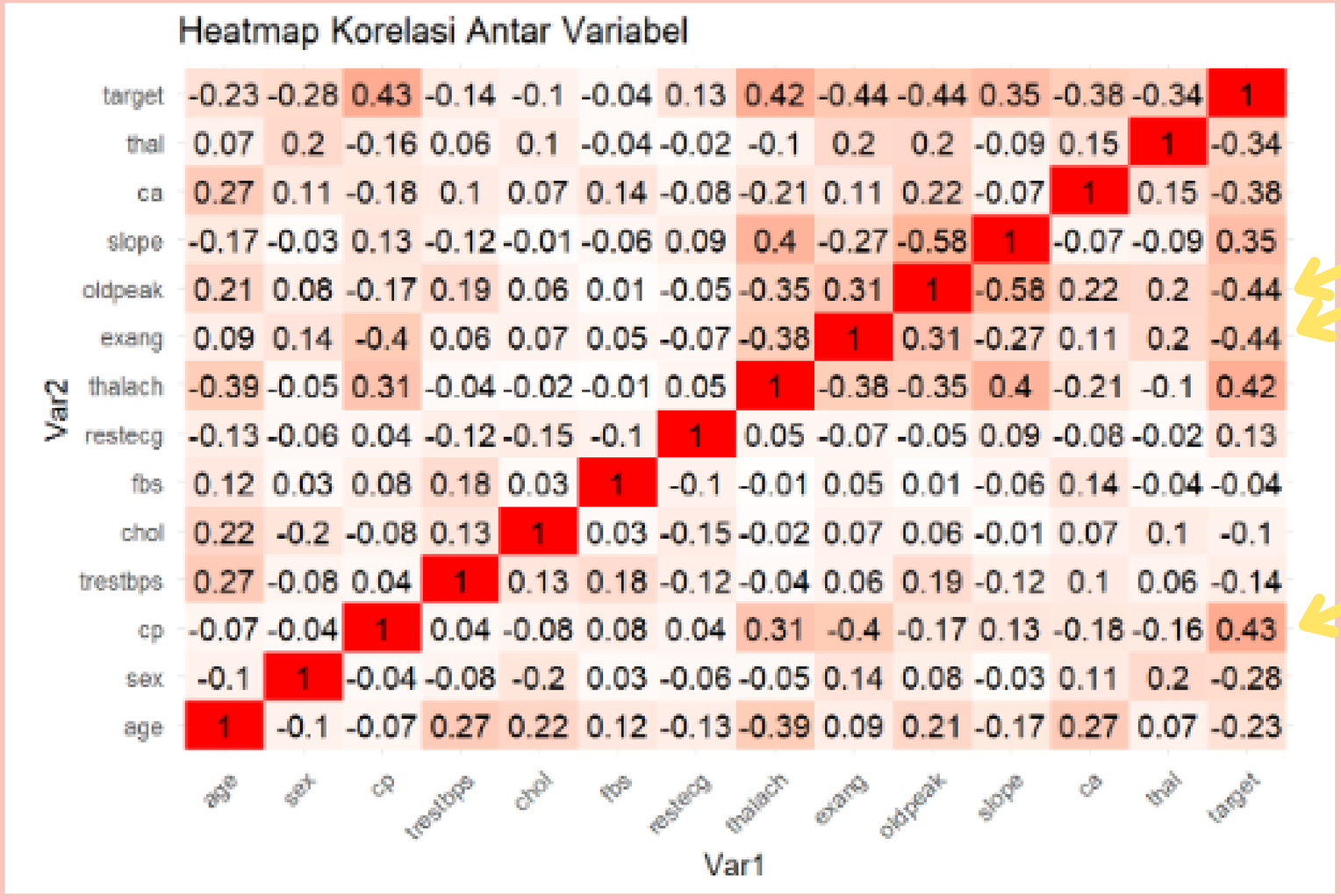
1-10 of 1,025 rows | 5-14 of 14 columns

Previous 1 2 3 4 5 6 ... 100 Next

source: kaggle

# HEATMAP

shows the correlation between variable. Variable exang, oldpeak, and cp have the strongest correlation with the target. Therefore, they are used as the predictors.



# MODEL USED

## Bayesian Logistic Regression Model

**Handles Data Complexity**
Bayesian Logistic Regression effectively manages complex data, including non-linear relationships and missing information.

**Identifies Key Variables**
Helps determine which variables have the strongest relationship with the outcome.

**Accurate Prediction**
Delivers precise likelihood predictions for target outcome based on the given data.

**Prevents Interpretation Errors**
Reduces errors in interpreting data with multiple predictors, such as age, gender, and health indicators.

# MODELLING APPROACH

**Prior Distribution**

- Weakly informative priors, allowing the data to guide posterior estimates without being overly restrictive.

**Likelihood**

- For binary outcomes (e.g., presence or absence of disease), a Bernoulli likelihood is used.
- Using logit function, links predictors to probability, ensuring values remain between 0 and 1.

$$P(Yi = 1 \mid x_i; \theta) = \frac{1}{1+e^{(-\theta^T xi)}}$$

**Atau**

$$P(Yi = 0 \mid x_i; \theta) = 1 - \frac{1}{1+e^{(-\theta^T xi)}}$$

$y_i$ : Outcome variable (presence/absence of cardiovascular disease).

$x_i$ : Predictor variables (age, gender, cp, trestbps, etc).

θ: Model parameters

# PARAMETER ESTIMATION

$$P(\theta|data) = \frac{P(data|\theta)\,P(\theta)}{P(data)}$$

- P(θ | data): Posterior distribution.
- P(data | θ): Likelihood function.
- P(θ): Prior distribution.
- P(data): Marginal likelihood (normalizing constant).

- Posterior distributions are often complex, so we use **Markov Chain Monte Carlo** (**MCMC**), with **Gibbs Sampling** for breaking data into simpler conditional distributions.
- To ensure reliable MCMC results, **convergence diagnostics** like
- Gelman-Rubin (result near 1 indicates convergence)
- Geweke (result near 0 indicates convergence) are applied.
- posterior distributions are summarized using the mean, median, or credible intervals, which represent Bayesian confidence levels.

# MODEL VALIDATION AND COMPARISON

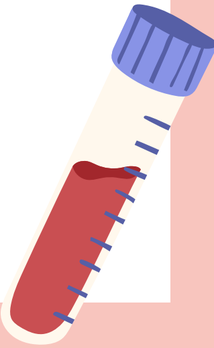
Posterior Predictive Check (PPC) was used as model validation to assess how well the model fits the observed data. Whereas model comparison, WAIC (Watanabe-Akaike Information Criterion), DIC (Deviance Information Criterion), and Bayes Factor were used:

- WAIC (Watanabe-Akaike Information Criterion): Balances model fit and complexity.
- DIC (Deviance Information Criterion): a model selection that used the complexity of a Bayesian model.
- Bayes Factors: Compare the models that we use.

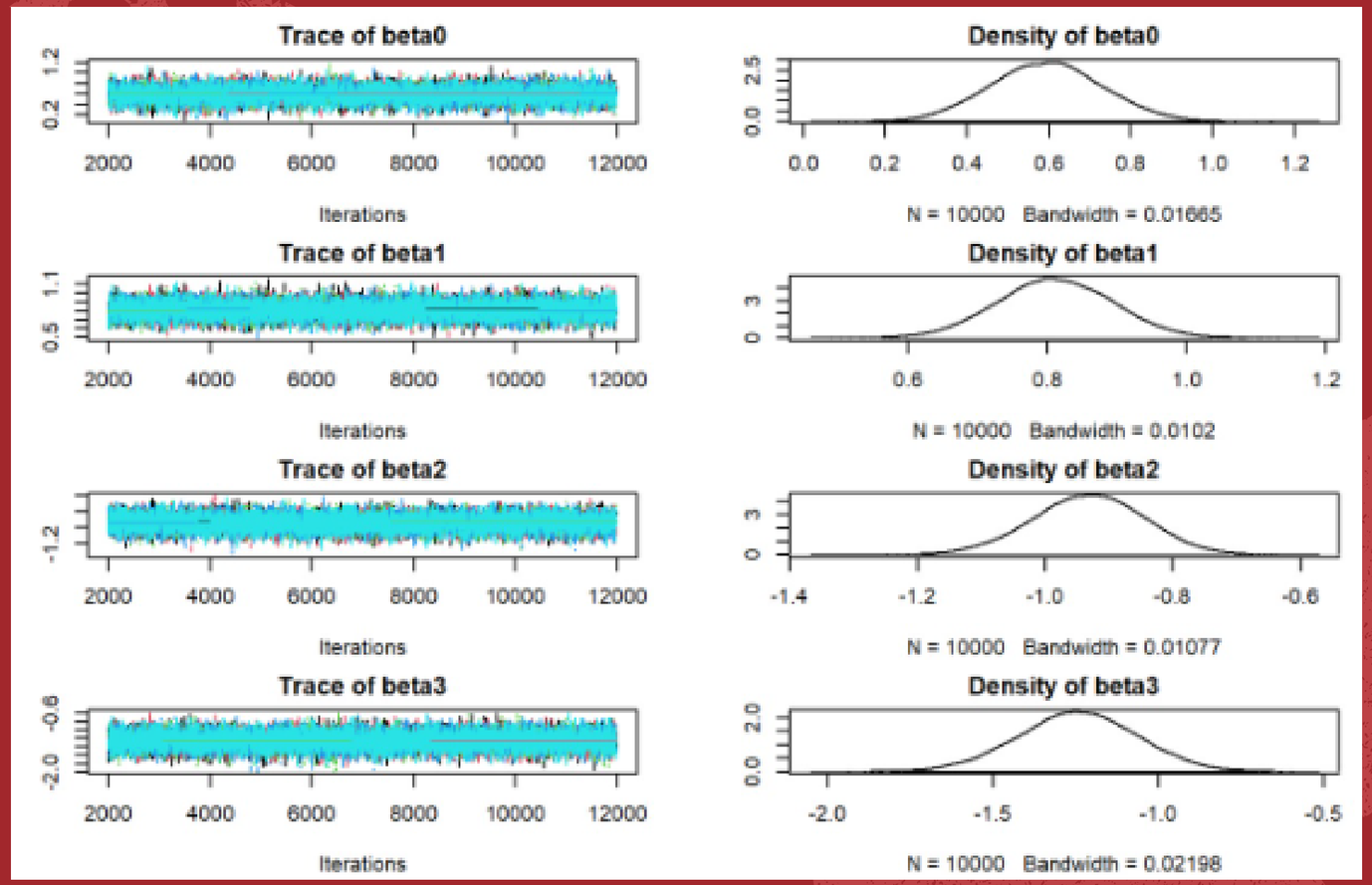
# SUMMARY OF POSTERIOR

A positive mean value for β1 indicates a positive relationship between the variable and the log-odds of the response, meaning an increase in cp (β1) raises the probability of heart disease, while a decrease in oldpeak (β2) and exang (β3) lowers the probability. β1 has a smaller standard deviation and credible interval, indicating better precision, supported by its smallest standard error. In contrast, β3 (exang) has the largest standard deviation and credible interval, showing higher variability and uncertainty.

```
Iterations = 2001:12000
Thinning interval = 1
Number of chains = 5
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean      SD  Naive SE Time-series SE
beta0  0.5912 0.13641 0.0006101      0.0015709
beta1  0.8103 0.08341 0.0003730      0.0008068
beta2 -0.9333 0.08811 0.0003940      0.0008069
beta3 -1.2433 0.18110 0.0008099      0.0014703

2. Quantiles for each variable:

          2.5%     25%     50%     75%   97.5%
beta0  0.3236  0.4990  0.5910  0.6826  0.8584
beta1  0.6483  0.7540  0.8098  0.8661  0.9754
beta2 -1.1086 -0.9922 -0.9318 -0.8736 -0.7644
beta3 -1.6002 -1.3655 -1.2434 -1.1210 -0.8914
```



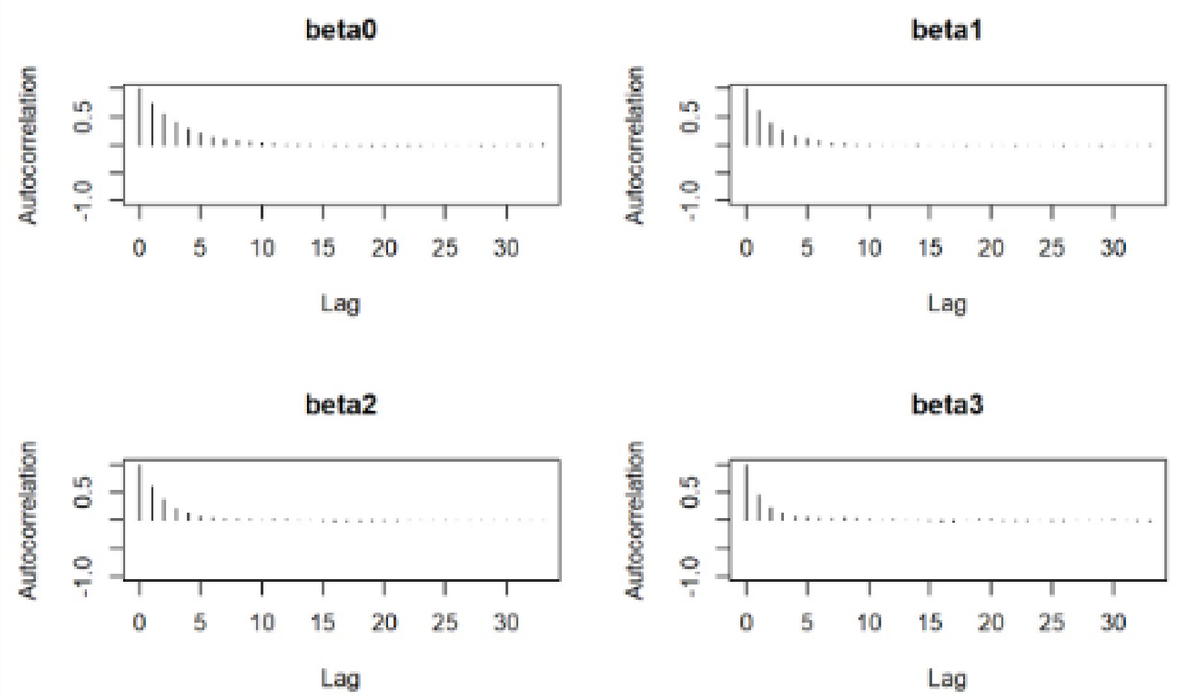
# TRACE AND DENSITY PLOT




β0 is around 0.59, β1 is around 0.81, β2 is around -0.93, and β3 is around -1.25
0.01665 for β0, 0.0102 for β1, 0.01077 for β2, and 0.02198 for β3

The MCMC sampling process has run smoothly and the model has reached convergence, making the results trustworthy. The density plots indicate strong confidence in the parameter values, with β1 having a positive effect on the target, while β2 and β3 have negative effects. Overall, the model has evaluated the data well and provided reliable conclusions.

# AUTOCORRELATION PLOT




The autocorrelation plot shows that the correlation between samples decreases after a few iterations, indicating that the samples become independent of each other. This suggests that the sampling process is effective, with the chain exploring parameters well without getting stuck on any single value. Overall, the samples are independent enough for further analysis, supporting the quality of the MCMC sampling process.

# GEWEKE DIAGNOSTIC

**to evaluate the convergence of the MCMC's chain to stationer distribution**

- There is **variation** for each chain's convergence.
- Only chain **2 and 4** indicates **full convergence**.
- Chain **1, 3, and 5** still have parameters that **haven't converged.**

```
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
 0.7531 -0.8539  0.4963 -1.4046


[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
-0.2514  0.5930 -0.2860 -0.6033


[[3]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
-0.6249  0.2648  0.2263  1.1662


[[4]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
 0.1777  0.2879 -0.6577 -0.3497


[[5]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
 1.9221 -1.5251 -2.1261 -0.3147
```

# GELMAN-RUBIN DIAGNOSTIC

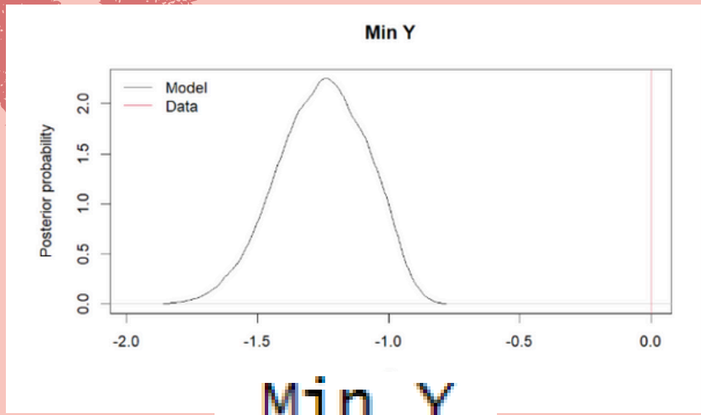
**also known as Potential Scale Reduction Factor (PSRF)**

The markov chain **has converged** effectively as the result is 1 for all parameter.

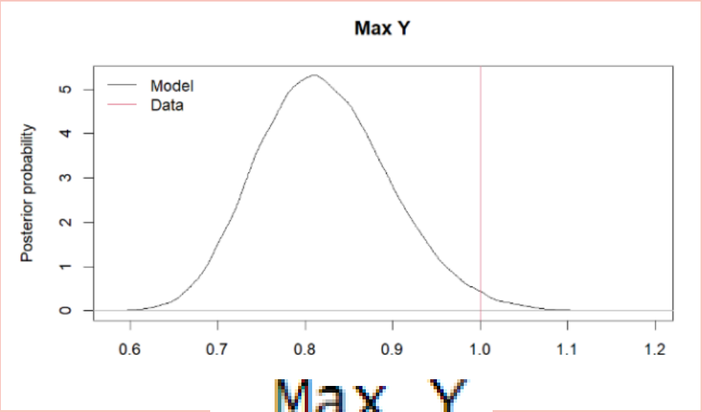
```
Potential scale reduction factors:

      Point est. Upper C.I.
beta0          1          1
beta1          1          1
beta2          1          1
beta3          1          1

Multivariate psrf

1
```

# PPC

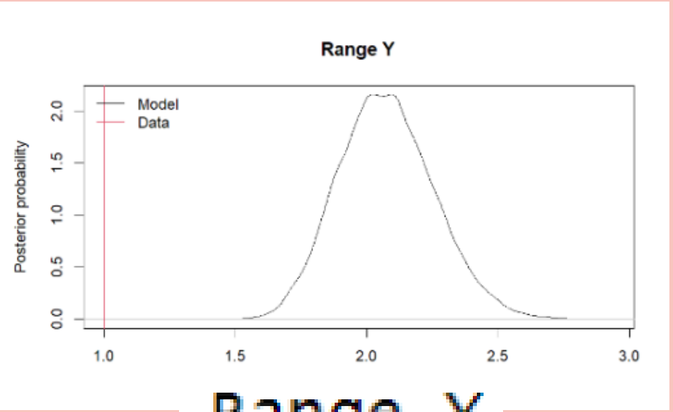
## Posterior Predictive Check




Min Y
0.0000




Max Y
0.0146




Range Y
1.0000

.Really rare.

Quite high but not too far.

High variability.

**Overall, the model predicts mostly small probabilities with a narrow range.**

# BAYES FACTOR

```
Iterations = 2001:12000
Thinning interval = 1
Number of chains = 5
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
     0.0531814      0.0617155      0.0002760      0.0003423

2. Quantiles for each variable:

    2.5%      25%      50%      75%    97.5%
-0.06903  0.01172  0.05301  0.09495  0.17438
```

[1] 1.00064

- The mean of the tested parameters is close to zero with a narrow distribution.
- There is variability from negative to positive as seen from the quantiles.
- There is considerable uncertainty for the alternative model.
- No significant difference in support for both the full and null models.

# DIC

## Posterior Predictive Check

```
Rata-rata DIC untuk Model Logistik: 0.953033
Rata-rata DIC untuk Model Binomial: 0.9530761
```

- Both models fit the data similarly well.
- Both models can be considered equally good in terms of the balance between complexity and fit.
- Model selection can be based on other needs or preferences.

# WAIC

```
WAIC: 981.0214

Iterations = 2001:12000
Thinning interval = 1
Number of chains = 5
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

           Mean      SD  Naive SE Time-series SE
beta[1]  0.5924 0.13862 0.0006199      0.0016190
beta[2]  0.8103 0.08375 0.0003745      0.0008104
beta[3] -0.9347 0.08888 0.0003975      0.0008234
beta[4] -1.2434 0.18118 0.0008103      0.0015339

2. Quantiles for each variable:

           2.5%     25%     50%     75%   97.5%
beta[1]  0.3175  0.4995  0.5923  0.6852  0.8648
beta[2]  0.6471  0.7538  0.8093  0.8657  0.9773
beta[3] -1.1131 -0.9938 -0.9329 -0.8742 -0.7649
beta[4] -1.6021 -1.3639 -1.2433 -1.1207 -0.8879
```

The WAIC value of 981.0214 indicates that this model is suitable for comparison with other models using the same dataset. The posterior means for all parameters show consistent estimates with low variability, and the 95% confidence intervals exclude zero, confirming the statistical significance of all parameters. Positive coefficients for β1 and β2 suggest a positive association with the response variable, while negative coefficients for β3 and β4 indicate a negative association. These findings show that all parameters significantly contribute to the model and can be interpreted based on their relationships.

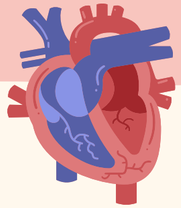
According to the Journal of Medical Sciences from the Zainoel Abidin General Hospital, Banda Aceh, chest pain is one of the most common reasons people visit the emergency department.

# COMPARISON WITH EXISTING LITERATURE

According to the National Library of Medicine, a case was observed where a 56-year-old man came to the hospital after experiencing chest tightness during physical activity for two weeks.

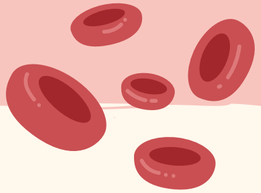
Cross-Sectional Study, which previously showed that 19 to 66 percent of patients who had a heart attack also had mental health disorders, particularly depression and anxiety

# IMPLICATION OF THE FINDINGS



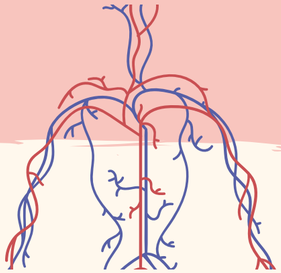
## Chestt Pain

often caused by various conditions, is particularly dangerous when linked to heart disease and should never be underestimated due to its life-threatening potential.



## Exang

chest pain occurring after exercise due to narrowed coronary arteries, indicates insufficient oxygen to the heart muscle during physical activity, requiring careful adjustment of exercise routines for heart disease patients.



## Old peak

common in cardiovascular patients, increase the risk of sudden heart failure and cardiac arrest, highlighting the need for mental health support, especially after heart bypass surgery.

# CONCLUSION

The study concludes that heart disease can be effectively predicted using the variables chest pain type (cp), oldpeak, and exercise-induced angina (exang). These predictors significantly influence the likelihood of heart disease, supporting the model's utility in clinical risk assessment.

# LIMITATIONS

**Dataset**

Relatively **small** dataset

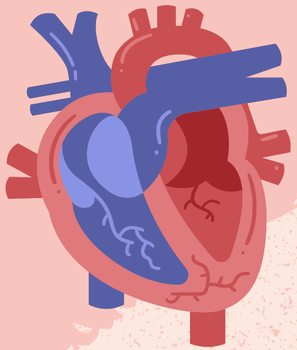
**Instability**

Parameter β2 and β3 show indications of **instability**.

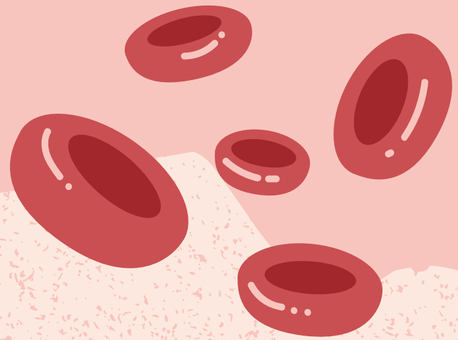
**Resources**

**Lack of sufficient resources** as references.

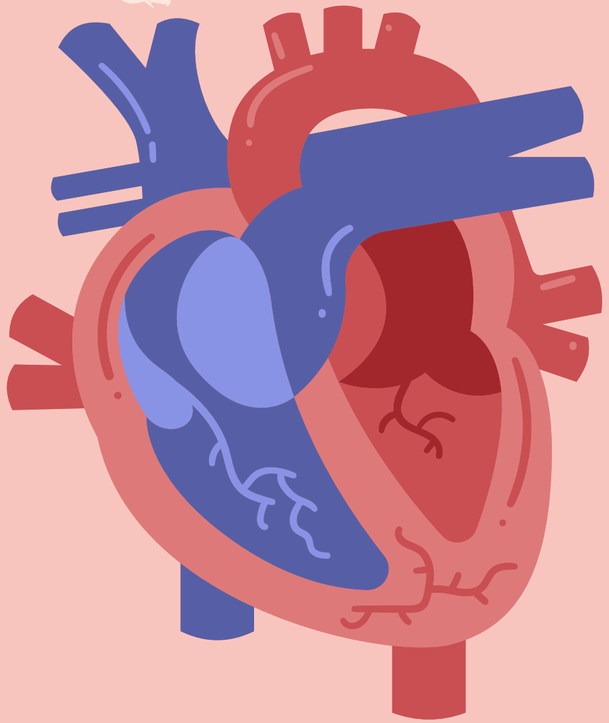
# FOR FURTHER RESEARCH



larger and more diverse dataset in order to analyze more factors



consider geographical factors to broaden the context and increase the precision



# THANK YOU !!