A Logistic Approach to Heart Disease Diagnosis

Ellee Millard
and
Jenna Worthen

December 11, 2024

Abstract

This project highlights two statistical models that can be used to assess the risks of developing heart disease. The first model, termed the simple model, was generated using simple predictors that can be obtained without costly medical testing (age, sex, resting heart rate, and resting systolic blood pressure). With the second model, we included more complicated predictors that require extensive medical testing, thus termed the medical model. We compare the predictive power of each model via cross validation analysis and discuss cutoff optimization tactics for more robust out-of-sample prediction.

1 Introduction

Heart Disease is the number one cause of death in the United States, with over 600,000 deaths in 2023 alone. In the world, cardiovascular diseases (including congestive heart failure, coronary heart disease, peripheral artery disease, etc.) are the leading cause of death. Although heart disease is not a new problem, it is one that is far from solved. Research must underscore the causes and nature of disease so that medical best educate their patients about prevention. In this project, we are interested in using data to find the most significant and most accurate predictors of heart disease.

The UCI heart disease data is a merged dataset collected by medical doctors from hospitals in Ohio, Hungary, Switzerland, and California. 920 patients were measured and diagnosed with one of five stages of heart disease: 0 = no symptoms of heart disease, 1 = no symptoms during normal activity, but doctor has noticed heart weakness , 2 = mild symptoms during exercise, 3 = symptoms with basic activities, 4 = major symptoms even at rest, Besides the heart disease stage diagnosis, there are 15 other variables in the model, and we became fascinated by using them as predictors for heart disease. Explanations of the variables we found important are as follows:

- sex: patient's biological sex
- age: patient's age in years
- max hr: maximum heart rate achieved by patient during intense exercise
- resting bp: patient's systolic blood pressure when at rest.
- exercise chest pain: 1 = patient has exercise- induced chest pain (cp), 0 = no cp
- lvh: where 1 = presence of left ventricular hypertrophy (lvh), 0 = no lvh.
- chol: patient's cholesterol level
- thalassemia rdefect: patient has a reversible blood condition
- thalassemia fdefect: patient has irreversible blood condition
- recovery impairment: measures the slope of the ST segment of heart contraction relative to baseline.

The red variables are easily obtainable measurements from a patient, so we decided to use these variables as the backbone of the "simple model", as we are interested in how accurate the predictions of heart disease would be if a doctor were only to look at a few quick factors. The "medical model" consists of the variables in blue, and require longer, more expensive testing that can only be provided to a patient in a hospital- setting. Creating the two models allows us to compare prediction accuracy between a simpler and more complicated approach to diagnosis.

1.1 Exploratory Data Analysis

We began familiarizing ourselves with the data by plotting relationships among the variables to better gauge what variables are indicative of heart disease. Because each of the possible predictor variables were measured from the same patient, we expected to see high levels of multicollinearity. However, we were pleasantly surprised that the variables shown in the correlation plot below were not correlated. This verified that our models would not have multicollinearity.

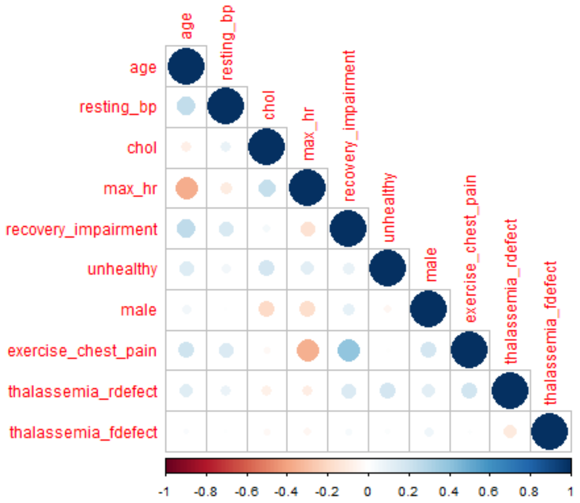


Figure 1: Correlation Matrix Of Predictor Variables

Distribution of recovery impairment across different heart disease stages and age groups are shown below. Each panel presents a violin plot depicting the relationship between heart disease stage and recovery impairment across various age ranges: (a) All ages, (b) Ages 28-47, (c) Ages 48-60, and (d) Ages 61-77. The plots illustrate the variability in recovery impairment across these groups, which will inform the logistic regression model's assessment of how age and heart disease stage impact the likelihood of recovery impairment. These visualizations support the logistic regression model's variables by showing potential patterns and relationships in the data before model fitting.

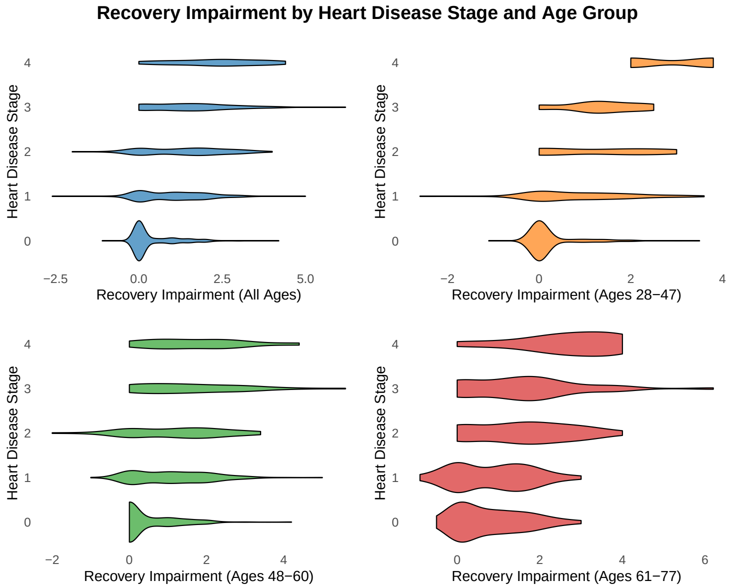


Figure 2: Violin Plots Showing Spread of Recovery Impairment by Age

This plot below illustrates the relationship between age and heart disease stage, with data separated by gender. The jitter plot displays the distribution of heart disease stages (as a continuous variable) for each age group, with points color-coded by gender. The plot helps to visualize any potential gender differences in heart disease stage distribution across the age spectrum. This highlights potential gender-based disparities in heart disease stage which clearly shows us that being male is indeed a risk factor. This analysis and our variable selection (found in Appendix) was crucial for us to set up our simple model, focusing on gender with other factors.

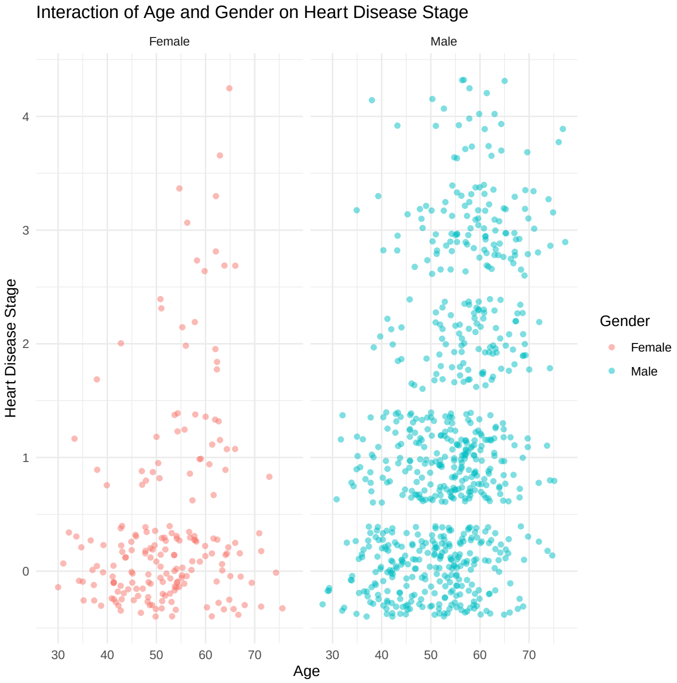


Figure 3: Men and Women Across Stages of Heart Disease

2 Model Selection and Validation

There is much evidence in the medical community that stage one of heart disease is reversible, while stages two through four are considered irreversible. We decided to simplify our model by converting the five stages to a 0 and 1 binary response where 0 = stages 0 and 1, and 1 = stages 2, 3, and 4, essentially a "reversible vs irreversible" diagnosis. Then, we used logistic regression to fit two models based on "simple" predictors or "medical" predictors. Creating the simple model, we included maximum heart rate, age, an indicator variable for if a patient's sex is male, and resting blood pressure.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -4.56 - 0.016 \cdot \text{max\_hr} + 0.078 \cdot \text{age} + 1.16 \cdot I(\text{sex} = \text{male}) + 0.0027 \cdot \text{resting\_bp}$$

$$\pi_i = \Pr(y = 1 \mid x_1, \ldots, x_5) \text{ and } y_i \sim \text{Bernoulli}(\pi_i)$$

For interpretability, we chose to convert the coefficients of the model to odds ratios by raising (e) to the power of each of our beta estimates. Thus, odds ratios of less than one suggest that the variable decreases the likelihood of developing irreversible heart disease while odds ratios greater than one suggest an increase in the likelihood. Thus, as seen in table 1, an increase in max heart rate decreases someone's risk of developing heart disease. On the other hand, an increase in age, resting blood pressure, and if a patient is a male, are risk factors of heart disease.

Predictor Variable	**Odds Ratio**	**95% Confidence Interval**
max hr	0.984	0.977 - 0.99
age	1.08	1.05 - 1.11
Sex = male	3.211	1.95 - 5.54
resting bp	1.01	0.994 - 1.03

Table 1: Table 1: Simple Model Estimates with Error

We note that the confidence interval for resting blood pressure includes 1 in the interval, meaning that there is a chance that blood pressure has no effect on likelihood of developing heart disease. However, including it in the model improved out of sample predictive power as tested by later cross- validation.

With the Medial model, we found the following variables significant diagnostics of heart disease: the presence of chest pain during exercise, presence of left ventricular hypertrophy, cholesterol levels, presence of reversible or fixed thalassemia, and heart recovery impairment.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -1.452 + 0.749 \cdot I(\text{exercise chest pain} = 1) + 0.537 \cdot I(\text{lvh} = 1) - 0.007 \cdot \text{chol}$$

$$+ 1.089 \cdot I(\text{thalassemia rdefect}) + 1.42 \cdot I(\text{thalassemia fdefect})$$

$$\pi_i = \Pr(y = 1 \mid x_1, \ldots, x_5) \text{ and } y_i \sim \text{Bernoulli}(\pi_i)$$

Again, we converted the log odds beta coefficients to odds ratios for increased understandability (see table 2). We were surprised that an increase in cholesterol levels appeared to have a slightly negative effect on developing heart disease. As all other predictors increased, the likelihood that a person develops heart disease strongly increased.

Predictor Variable	**Odds Ratio**	**95% Confidence Interval**
exercise chest pain	2.115	1.44 - 3.12
lvh	1.712	1.08 - 2.703
thalassemia rdefect	2.974	1.97 - 4.49
thalassemia fdefect	4.14	2.001 - 8.523
recovery impairment	2.00	1.69 - 2.398

Table 2: Table 2: Medical Model Estimates with Error

3 Prediction Analysis

Because we were most interested in using our model to predict someone's likelihood that they do or do not have irreversible heart disease, we tested the models through cross-validation testing. We divided the data set into a 'test' and a 'training' set, where we measured how accurately we could predict the training data from only the test data. Originally, we chose 0.5 as the cutoff value for the diagnosis of heart disease. The confusion matrices are shown below in tables 3 through 6:

Prediction/ Reference	**0**	**1**
0	110	39
1	14	13

Table 3: Confusion Matrix for Simple Model

Metric	**Value**
Sensitivity	0.88
Specificity	0.24
Accuracy	0.699

Table 4: Performance Metrics for Simple Model

Prediction/ Reference	**0**	**1**
0	106	30
1	13	22

Table 5: Confusion Matrix for Medical Model

Metric	**Value**
Sensitivity	0.89
Specificity	0.42
Accuracy	0.749

Table 6: Performance Metrics for Medical Model

We were not surprised to see that the medical model was superior in predicting stages of heart disease to the simple model, but it did not vastly outperform. Because cutoff values are malleable to allow for better predictive performance, we looped through all possible cutoff values between 0 and 1 by increments of 0.01. Each possible cutoff was evaluated by sensitivity, specificity, and accuracy. Figure 4 and 5 below shows how each metric changes as the cutoff values range from 0 to 1 for both of the models.

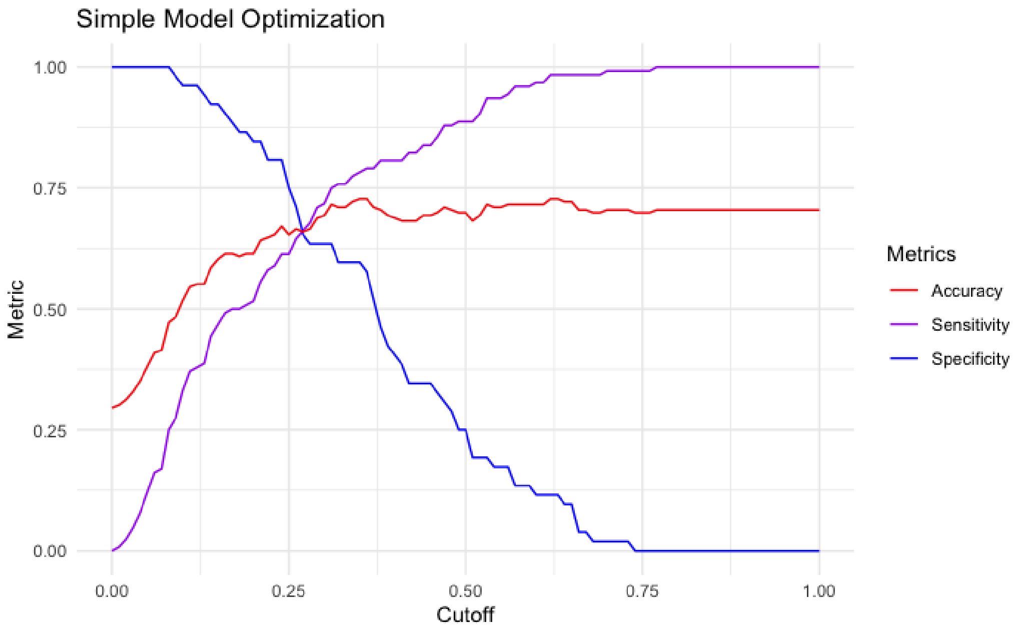


Figure 4: Optimizing Performance Metrics: Simple Model

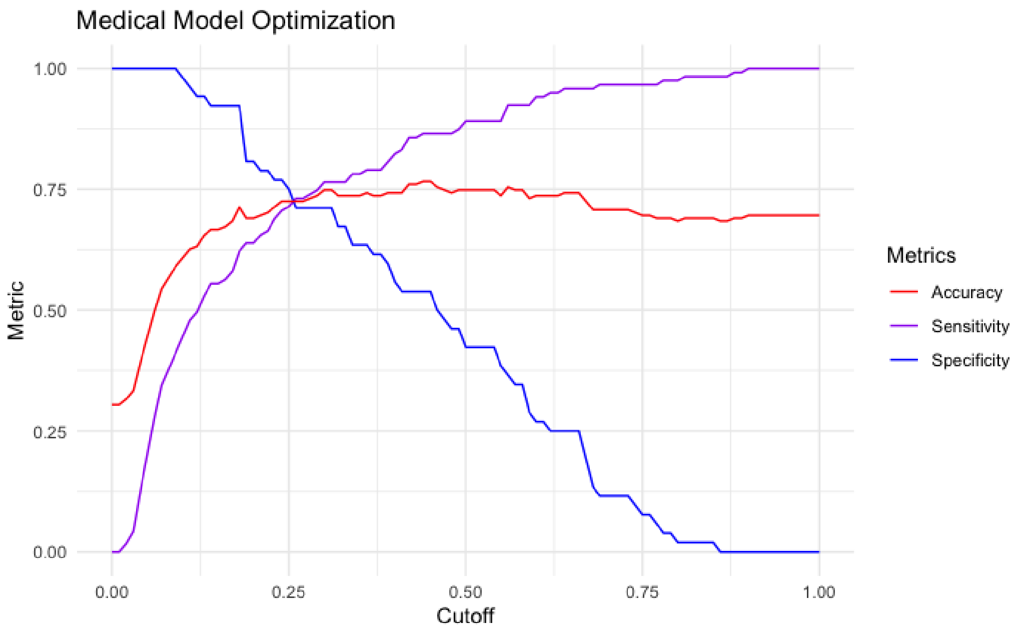


Figure 5: Optimizing Performance Metrics: Medical Model

The using the optimal cutoff value of the simple model at 0.625, the accuracy increased from 0.69 to 0.73. Using the optimal cutoff of the medical model at 0.45 increased the accuracy from 0.75 to 0.77. Adjusting the cutoff values for each of the models also favored the sensitivity. Given the background context of diagnosing heart disease, we postulated that it would be better to falsely diagnose a patient as having heart disease, as opposed to failing to diagnose with heart disease when the patient truly does have the condition.

4 Conclusions

This report highlights key findings. First, the medical model only outperformed the simple model by about 4% in prediction accuracy, suggesting that achieving over 80% accuracy in diagnosis is challenging. We recommend exploring datasets with additional predictor variables, such as heart rhythm and oxygen saturation, and considering a random forest approach to improve predictive accuracy. Additionally, nearly all predictor variables, except max heart rate, increased the likelihood of developing heart disease, limiting our understanding of its causes. Future studies should focus on factors that reduce the risk of heart disease. Finally, we acknowledge that simplifying heart disease stages from 0-4 to 0 and 1 means our models only assess whether the disease is in a reversible or irreversible stage, though stage 2 (irreversible) requires different medical care than stage 4 (irreversible).

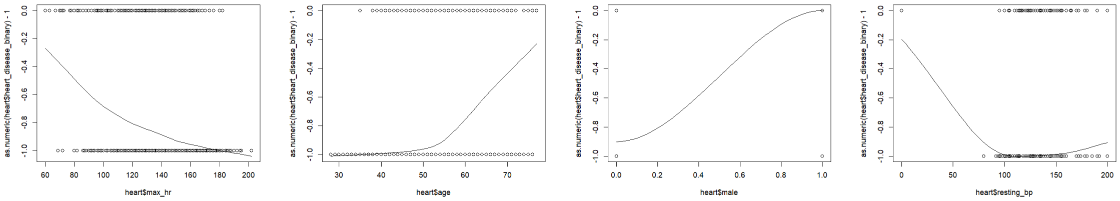
5 Contributions

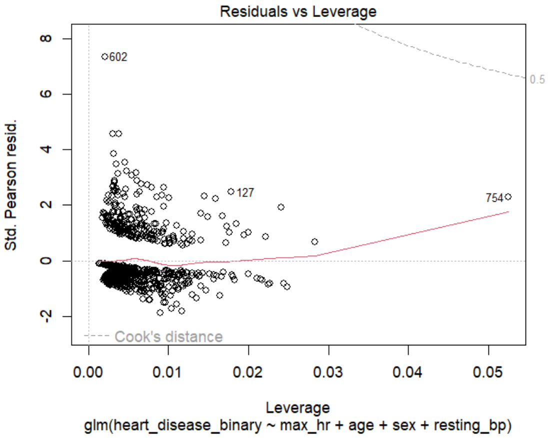
As a team, we met together on Fridays before class to discuss ideas we had for the project and divide the workload. We decided our research questions together as well as the best models to answer our questions. Separately, Jenna worked on the code and graphs to check model assumptions for the project proposal. She also created the graphs and wrote the descriptions for the ”Exploratory Data Analysis” and ”Appendix”. Ellee wrote the abstract and introduction. She wrote the ”Model Selection” and ”Prediction Analysis” sections and designed all the graphs and tables included.

APPENDIX

Simple Model Assumptions

As see in the following figures below, the logistic assumptions for the simple model are met. Although we do have a potential outlying point as seen in Figure 7, it is still greatly within the bounds of Cook's distance so we decided to include it in the model.



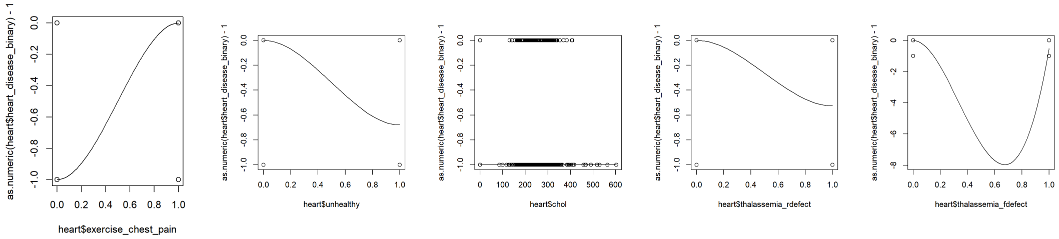


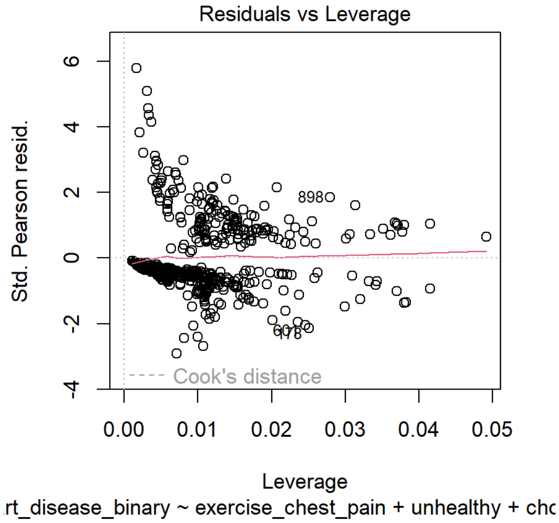
Predictor Variables	VIF
Max Heart Rate	1.060507
Age	1.081535
Sex	1.023852
Resting Blood Pressure	1.036555

Figure 6: Summary of logistic regression (Simple) model assumptions. Top row: The x's vs. log odds are linear (monotone in probability). Bottom row: On the left, No influential observations. On the right, No influential observations and no multicollinearity, as assessed by VIFs.

Medical Model Assumptions

Similarly, as demonstrated in the following figures below, the logistic assumptions for the medical model are also met. Influential observations are less of a concern in this model, which is a possible leading factor, among many, in why this model predicts more accurately compred to the simple model.





Predictor Variables	VIF
Exercise Chest Pain	1.104591
Unhealthy	1.140114
Cholesterol	1.144392
Reservable blood condition	1.096838
Irreversible blood condition	1.047561
Recovery Impairment	1.130174

Figure 7: Summary of logistic regression (Medical) model assumptions. Top row: The x's vs log odds are linear (monotone in probability). Bottom row: On the left, No influential observations. On the right, No influential observations and no multicollinearity, as assessed by VIFs.

Variable Selection Analysis

Before creating our models, we decided to perform a variable selection to understand how all of our many potential predictor variables were interacting with one another. We temporarily removed any categorical (non-binary) variables we did not want to use from our original data set and performed variable selection with the bestglm() function and AIC. As we can see below, being male and an individuals recovery impairment are the most significant. Because of this we wanted to make a model that revolved around this idea which became our simple model.

Coefficient	**Estimate**	**Std. Error**	**z value**	**Pr(>z)**
(Intercept)	-5.76676	3.36098	-1.716	0.0862 .
resting_bp	0.02831	0.01676	1.689	0.0913 .
max_hr	-0.02163	0.01206	-1.793	0.0730 .
recovery_impairment	0.49283	0.24067	2.048	0.0406 *
ca	0.16828	0.29239	0.576	0.5649
male	2.78943	1.09339	2.551	0.0107 *
thalassemia_rdefect	-19.42302	1507.20913	-0.013	0.9897

Table 7: Variable selection the best subset of predictors for your logistic regression model based on criteria like AIC using Bestglm()

Model Summary:

Null deviance: 137.277 on 308 degrees of freedom
Residual deviance: 88.486 on 302 degrees of freedom
(611 observations deleted due to missingness)
AIC: 102.49
Number of Fisher Scoring iterations: 19

Additional Analyses

95% Confidence Intervals (CIs) for the odds ratios of predictors in a logistic regression model predicting heart disease presence. The plot shows the odds ratios along with their 95% confidence intervals, with variables ordered by their odds ratio values. The simple model is displayed on top while the medical model is shown on bottom. An odds ratio greater than 1 indicates a positive association with heart disease, while an odds ratio less than 1 indicates a negative association. In the simple model we see that being male is the most significant variable in predicting with this model as it is the further from including 1 in the interval. Resting blood pressure does cross zero so that is not influential. This was predicted when we performed our variable selection; however we wanted to still predict someone's likelihood of having irreversible heart disease we went ahead with our model as the assumptions were met.

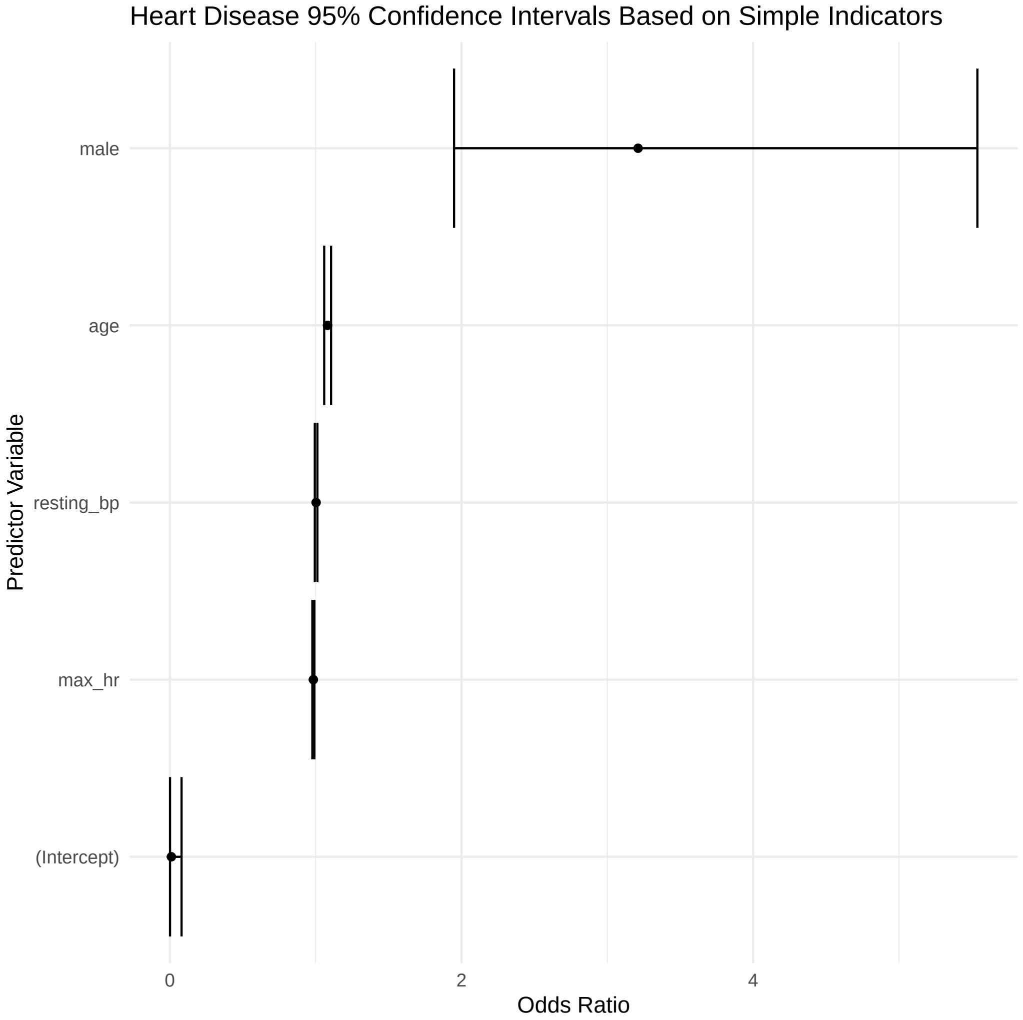


Figure 8: Logistic model using simple predictors for heart disease, and the odds ratio graph to quantify effect.

In the medical model we see that all predictors are statistically significant; however, as depicted in this figure, unhealthy and cholesterol are extremely close to including 1 in their intervals. We conclude that although the all variables are significant, exercise chest pain, reversible/irreversible blood condition, and recovery impairment are all more statistically significant variables in this model.

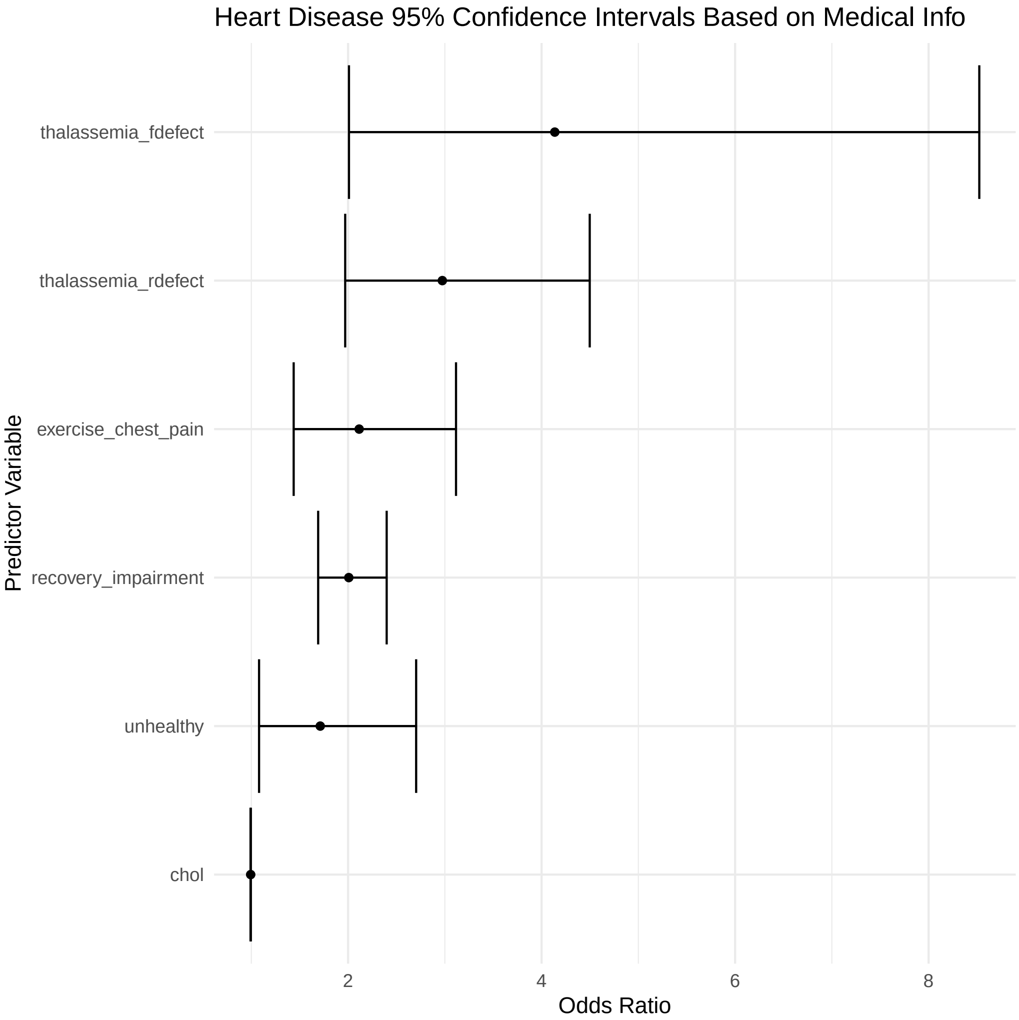


Figure 9: Logistic model using medical predictors of heart disease with similar odds ratio graph.

Below are ROC curve's for both of the model's predicting heart disease presence. The curve illustrates the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate) at different classification thresholds. The AUC value of the medical model is 0.814, suggesting a strong ability of the model to differentiate between patients with and without heart disease. An AUC of 0.814 indicates that the model performs significantly better than random chance, where an AUC of 0.5 would indicate no discriminatory ability. This model performs better than our simple model with an AUC of 0.7481.

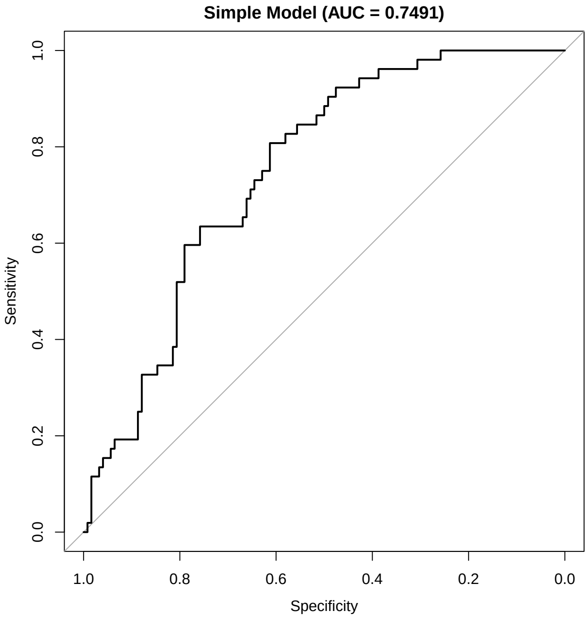


Figure 10: ROC Curve of our Simple Model with an AUC of 0.7481

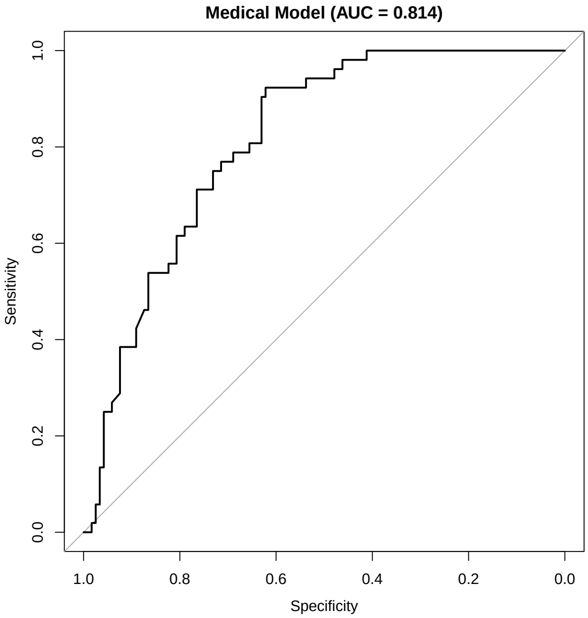


Figure 11: ROC Curve of our Medical Model with an AUC of 0.814