Significant predictors of heart diseases

Double O 5: Viraj Acharya, Jacqueline Rodriguez, Raymond Xiong, Alina Yin

2023-11-09

Introduction and Data

Introduction

Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels, and they are a critical health issue. They are the leading cause of death globally, accounting for 32% of all global deaths in 2019 (WHO, 2021), and they have also consistently ranked among the top 2 causes of death in the United States since 1975 (Benjamin et al., 2018). Furthermore, the economic toll of CVDs is substantial, with annual indirect costs of $237 billion (Lopez et al., 2023). Consequently, effective solutions are urgently needed.

Understanding the significant predictors of CVDs is crucial to mitigating the impact of heart diseases. It empowers individuals to take proactive steps and adopt healthier lifestyles, thereby preventing the onset of CVDs and reducing morbidity and mortality rates. Models that identify people with high risk of CVD could also be better constructed, which can lead to early intervention and adequate treatment, thus preventing premature deaths (WHO, 2021). A comprehensive understanding of the predictors also enables better resource allocation, more targeted prevention, and reduced economic and social costs.

In this research project, we aim to focus on heart diseases, a subcategory of CVDs, and unravel their significant predictors, contributing to the global fight against CVDs. Our main research question is: **"What are the significant predictors of the occurrence of heart diseases?"** We hypothesize that these predictors of heart diseases include indicator variables of CVD-related symptoms like chest pain or discomfort as well as demographic variables like age and gender.

Data description

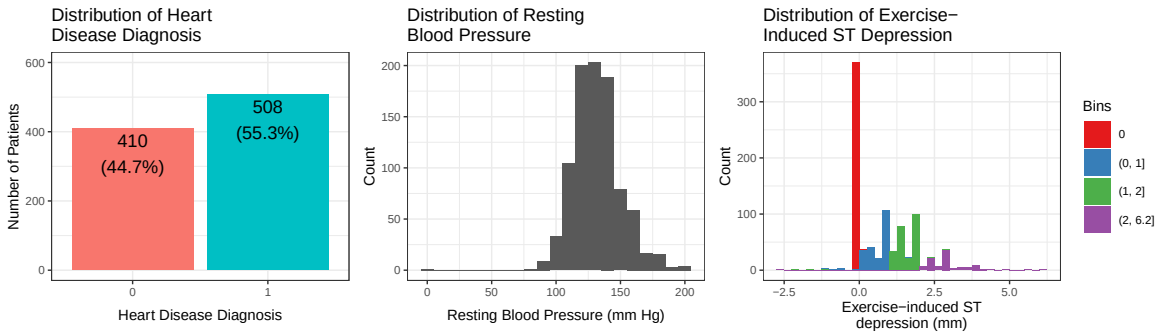
The data set is retrieved from Kaggle and combines the Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog (Heart) data set from the UC Irvine Machine Learning Repository. The data were originally collected from 918 patients between 1981 and 1987 at the Cleveland Clinic in Cleveland, OH; the Long Beach Veterans Administration Medical Center in Long Beach, CA; the Hungarian Institute of Cardiology in Budapest, Hungary; and university hospitals in Zurich and Basel, Switzerland. A more detailed description of when and how the data were originally collected is provided in the original paper (Detrano et al.) (See full citations in Appendix 4). The data set we will use is an excerpted version, with 12 variables out of the original 76 variables. As a note, we consider all the variables in the dataset except for Cholesterol. We decide not to include Cholesterol in our modeling process and analysis because there are two many missing values for Cholesterol (172 observations, or around 18% of the observations, have missing values for Cholesterol), which may lead to low accuracy and precision in statistical analysis.

The codebook for all variables' definitions can be found in the README file for the data. The variables we'll focus on include:

- `HeartDisease`: Response variable, whether the patient has heart disease (1: Yes; 0: No)
- `Age`: Age of patient (numeric value in years)
- `Sex`: Sex of patient (M: Male; F: Female)

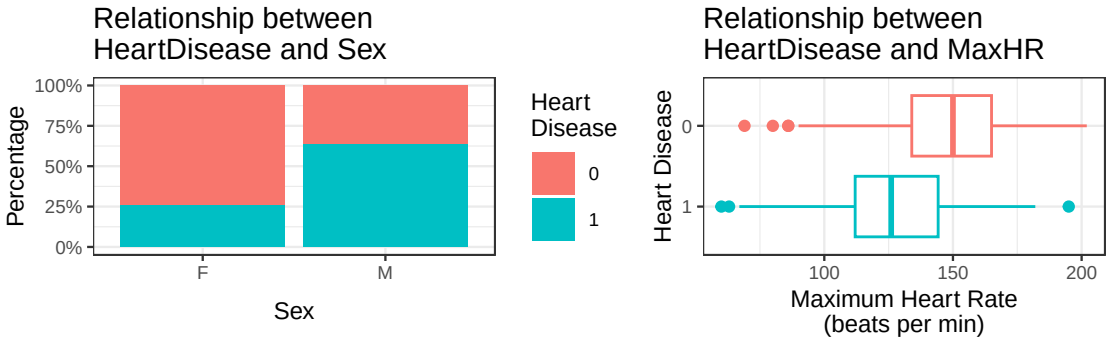
- `ChestPainType`: Chest pain type (TA: Typical Angina; ATA: Atypical Angina; NAP: Non-Anginal Pain; ASY: Asymptomatic)
- `RestingBP`: Resting blood pressure (numeric value in mm Hg)
- `FastingBS`: Whether the patient has high fasting blood sugar (1: if fasting blood sugar > 120 mg/dL; 0: otherwise)
- `RestingECG`: Resting electrocardiogram results (Normal: Normal; ST: having ST-T wave abnormality; LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria)
- `MaxHR`: Maximum heart rate achieved (numeric value in times per minute)
- `ExerciseAngina`: Whether the patient experiences exercise-induced angina (Y: Yes; N: No)
- `Oldpeak`: ST depression induced by exercise relative to rest (numeric value in mm)
- `ST_Slope`: Slope of the peak exercise ST segment (Up: up-sloping; Flat: flat; Down: down-sloping)

Exploratory Data Analysis

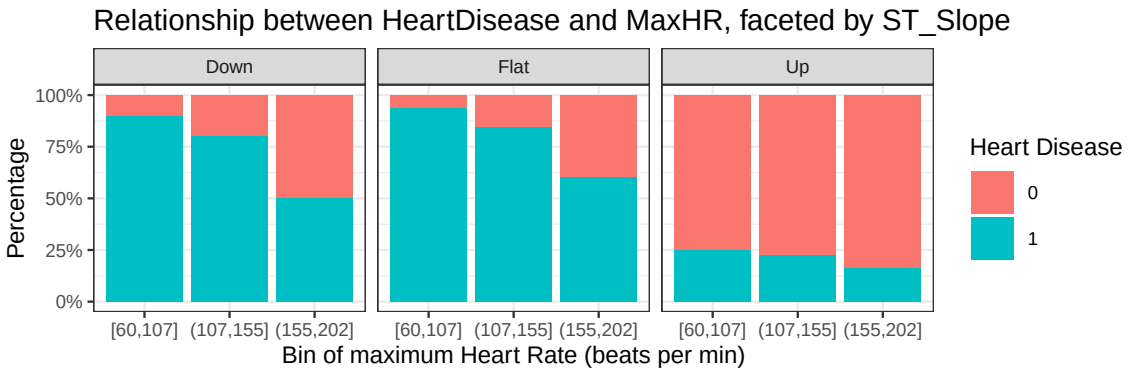


For our response variable, we use "1" to denote diagnosed for heart disease and "0" to denote not diagnosed for heart disease. As shown in the bar plot, 508 patients (55.3% of all observations) are diagnosed with heart disease, and 410 patients (44.7% of all observations) are not.

We determined the data cleaning steps for the predictor variables through univariate EDA results. We used mean imputation to impute the missing values (0s in the original data) for the `RestingBP` variable since the variable follows an approximately normal distribution. Moreover, we discretized the `Oldpeak` variable by their absolute values into bins of 0, $(0, 1]$, $(1, 2]$, and $(2, 6.2]$ (where 6.2 is the maximum absolute value). Our motivation was that the variable has a peculiar distribution, with the value 0 (which indicates that the ST segment is on the baseline) being exceptionally frequent, as shown in the histogram above, while also retaining the deviation of the ST segments from the baseline.



All potential predictors are significantly associated with heart disease occurrence. For example, there seems to be an apparent correlation between sex and heart disease diagnosis: the proportion of male patients diagnosed with heart disease (around 63.2%) is significantly higher than that for female patients (around 25.9%). Another example is maximum heart rate: the median maximum heart rate for patients diagnosed with heart disease (around 126 BPM) is significantly lower than that for patients not diagnosed with heart disease (around 150 BPM). These findings indicate that being male or having a higher heart rate might be correlated with higher chances of being diagnosed with heart disease, which motivates us to further investigate the associations between the response variable and our predictor variables when modeling.



There are also potential interaction effects between predictor variables. An example is the interaction effect between maximum heart rate and the slope of the peak exercise ST segment. To produce the figure shown above, `MaxHR` was temporarily discretized into 3 bins; for later model analysis, it will still be used as a numeric predictor. As shown in the figure, there are significantly greater differences in the proportion of patients with heart diseases between the smallest and largest `MaxHR` bins for those with downward-sloping ST segments ($90.0\% - 50.0\% = 40.0\%$) and those with flat ST segments ($93.9\% - 60.3\% = 33.6\%$), compared to patients with upward-sloping ST segments ($25.0\% - 16.1\% = 8.9\%$). This suggests that the association between maximum heart rate and heart disease diagnosis differs with respect to ST Slope. We are motivated to explore similar interaction effects between predictors when modeling as well.

Methodology

Type of Model and Predictor variables Considered

In regards to our regression model technique, we plan on using logistic regression. Specifically, our logistic model will be in the form of $p = 1/(1 + e^{-x})$ where x represents a regression equation with multiple variables and their corresponding coefficients. After using our data to find the suitable coefficients for our regression equation with multiple variables, this can then be substituted into our logistic model to find a value between 0 and 1 for p, where p represents a "probability" or a "score" for having a true heart disease diagnosis. For example, we plan on potentially using a threshold of 0.5 for p, where p greater than or equal to 0.5 signifies a prediction of 1 that indicates a patient has heart disease. Likewise, a value for p that is less than 0.5 indicates a prediction of 0 that a patient does not have heart disease. In the end, our logistic regression model will be able to predict, even for new data, if a patient has heart disease based on a collection of the predictor variables listed above that make it into our final model.

To make the final decision of whether logistic regression is proper for analyzing our dataset, we check the model conditions for logistic regression:

- **Linearity:** The linearity condition is satisfied. As demonstrated in the plots for our quantitative predictors of Age and Maximum Heart Rate (see Appendix 5 for the empirical logit plots), they both have a linear relationship with the empirical logit.

- **Randomness:** The randomness condition is satisfied. We do not have reason to believe that the participants of this study have a systematic difference from individuals in their general country populations when it comes to their personal health characteristics and their likelihood of having a heart disease.
- **Independence:** The independence condition is satisfied. It is reasonable to conclude that the subjects' health characteristics and the conducted testing are all independent of one another.

Since all three conditions are satisfied, we can reasonably confirm that it is suitable to analyze our dataset with the logistic regression model.

As for the predictor variables we plan on considering for our model, we want to analyze the influence of the collection of variables as stated in the Data Description section above, which includes both quantitative and categorical predictors. The potential quantitative predictors are age, resting blood pressure, and maximum heart rate achieved; Potential categorical predictors are sex, chest pain type, fasting blood sugar, resting ECG results, exercise induced angina, ST slope, and old peak intervals. Along with these initial predictors, we are interested in looking into interaction terms that can allow us to create a better model for heart disease. After we identify the main terms to include in the model, we investigate interaction terms between the selected main terms that stand out to us. Specific interaction terms that we consider include ST_Slope and MaxHR, ChestPainType and Sex, MaxHR and Sex, as well as Oldpeak and ST_Slope, which we will explain in detail in the Modeling Process section below.

Modeling Process

First, we transform the dataset and variables based on our EDA. We mean center the numerical predictors RestingBP and MaxHR to make the interpretation more meaningful. We also re-level categorical variables Oldpeak (from interval representing smallest to interval representing largest values), ST_Slope (set the baseline to be "Up"), ChestPainType (set the baseline to be "ASY", which represents no chest pain), and RestingECG (set the baseline to "Normal") to improve the interpretability of the results. As a note, we transform the dataset and at the same time create a corresponding recipe for the variable transformation process so that the procedure can be easily transferred to other datasets.

We then split our transformed dataset into training set (75%) and testing set (25%). This helps us to evaluate the performance and generalizability of our model and prevent overfitting.

Next, we conduct model selection in two main steps. We first consider only the main effects; We then identify interaction terms that stand out to us based on the resulting main terms in the model, employ drop-in-deviance tests to assess the significance of adding each term to the model, and incorporate the ones that have significant effects to the model.

- **Main Effects:** To select the optimal model considering only the main effects, we conduct forward and backward selection using the AIC as the standard. We choose to use the stepwise selection procedures, specifically the stepAIC method in R, since we want to effectively obtain the model that optimally predict the response variable HeartDisease (in our case, the model with the lowest AIC value) among all potential models given the 10 potential predictors. While forward and backward selection should theoretically give similar results, we decide to conduct both and compare the two models to account for potential differences within the two approaches or due to how R operates. We choose to use AIC for evaluating our model because it allows us to balance the model's accuracy with complexity. Specifically, AIC is an estimator of prediction error and relative quality of models that can support us to select the model that best predict Heart Disease in the context of our study. Compared to other possible estimators, we choose AIC as the main selection criteria since it strikes for a balance between underfitting and overfitting that is consistent with our Research goal. Specifically, AIC penalizes excessive complexity with terms in our model, which is crucial in the context of our study as overfitting can lead to misleading predictions. At the same time, since our research question asks for significant predictors of the occurrence of heart diseases, parsimony is not our highest priority in modeling. AIC is thus more ideal than BIC in this context. In general, through achieving a quite parsimonious model with lower AIC, our goal is to obtain a final model that captures the

essential relationships between the predictor variables and the likelihood of having heart disease, which will allow us to increase generalizability to new, unseen patient data.

The output of the forward selection and backward selection process is shown in Appendix 1. Comparing the output, we find the two approaches resulting in the same model, coefficients, and AIC value (471.4). The VIF values of the model are small (See Appendix 3), indicating no major issue of multicollinearity and ensuring interpretability. This motivates us to adopt this model as the final model considering only main terms. Therefore, the main terms we include are ST_Slope, ChestPainType, Sex, FastingBS, ExerciseAngina, Oldpeak, and MaxHR. The final model considering only main terms is displayed below:

term	estimate	std.error	statistic	p.value
(Intercept)	-2.315	0.385	-6.020	0.000
ST_SlopeDown	1.137	0.505	2.249	0.025
ST_SlopeFlat	2.523	0.291	8.674	0.000
ChestPainTypeATA	-2.044	0.396	-5.166	0.000
ChestPainTypeNAP	-1.721	0.301	-5.726	0.000
ChestPainTypeTA	-1.418	0.489	-2.901	0.004
SexM	1.658	0.320	5.180	0.000
FastingBS1	1.496	0.320	4.680	0.000
ExerciseAnginaY	1.127	0.284	3.963	0.000
Oldpeak(0, 1]	-0.439	0.317	-1.386	0.166
Oldpeak(1, 2]	0.103	0.340	0.303	0.762
Oldpeak(2, 6.2]	1.339	0.535	2.501	0.012
MaxHR	-0.008	0.005	-1.655	0.098

- **Interaction terms:** As we examine potential interactions between the predictor variables in the model above, the followings stand out to us based on the context of the data: **a).** ST_Slope and MaxHR (as explored in EDA) **b).** ChestPainType and Sex (Our intuitive and research suggests that the syndromes and effects of different chest pain types may vary across male and female. We would like to explore if the effect of ChestPainType in predicting heart disease would differ based on Sex.); **c).** MaxHR and Sex (The ranges of "normal" heart rate tend to differ between males and females in our daily lives. Thus, the intuition is that effect of maximum heart rate on the log-odds of having heart disease might be affected by sex, which we would like to explore.); **d).** Oldpeak and ST_Slope (Oldpeak and ST_Slope describe the extent of exercise-induced ST depression (deviation from the baseline) and the slope of the ST segment, respectively. Both are indicators of cardiac stress during physical activity. We would like to explore whether certain combinations of the magnitude (Oldpeak) and trend (ST_Slope) lead to higher odds of having heart disease.)

  We use the drop-in-deviance tests to assess the significance of adding each term to the model. We choose drop-in-deviance tests since they can compare nested models and assess whether the more complex model has a significant improvement in performance compared to the simpler model, thus supporting us to select interaction terms to add into the model in consistent with our research goal. We start by comparing our final model with only main terms to that model with an interaction term added. We then update (or keep) our final model based on the results of the tests, and compare it to that model with another interaction term added, until we conduct tests on all interaction terms we are interested in. Specifically, if p-value $<$ 0.05, which indicates that the data provide sufficient evidence that the coefficient of the term is not equal to 0, we add the term to the model; On the other hand, if p-value $>$ 0.05, we do not add the term.

  We start by comparing the final model considering only main effects to the same model adding terms for interaction between a). ST_Slope and MaxHR. Since the p-value is around 0.015, smaller than 0.05, we add the interaction term to the model. The output is displayed below:

term	df.resid	resid.devi	df	devi	p-val
HeartDisease ~ ST_Slope + ChestPainType + Sex + FastingBS + ExerciseAngina + Oldpeak + MaxHR	675	445.397	NA	NA	NA
HeartDisease ~ ST_Slope + ChestPainType + Sex + FastingBS + ExerciseAngina + Oldpeak + MaxHR + ST_Slope * MaxHR	673	436.977	2	8.42	0.015

This new model with interaction terms is then used as the baseline model to be compared with in the drop-in-deviance tests. Since the p-value for b) ChestPainType and Sex and c) MaxHR and Sex are larger than 0.05, we end up not including these terms. The p-value when adding interaction terms for d) Oldpeak and ST_Slope is approximately 0 (much smaller than 0.05), so we add these interaction terms to our model. The output for b) ChestPainType and Sex and d) Oldpeak and ST_Slope is displayed below, repsectively. See Appendix 2 for output of c) MaxHR and Sex.

term	df.resid	resid.devi	df	devi	p-val
HeartDisease ~ ST_Slope + ChestPainType + Sex + FastingBS + ExerciseAngina + Oldpeak + MaxHR + ST_Slope * MaxHR	673	436.977	NA	NA	NA
HeartDisease ~ ST_Slope + ChestPainType + Sex + FastingBS + ExerciseAngina + Oldpeak + MaxHR + ST_Slope * MaxHR + ChestPainType * Sex	670	435.728	3	1.249	0.741

term	df.resid	resid.devi	df	devi	p-val
HeartDisease ~ ST_Slope + ChestPainType + Sex + FastingBS + ExerciseAngina + Oldpeak + MaxHR + ST_Slope * MaxHR	673	436.977	NA	NA	NA
HeartDisease ~ ST_Slope + ChestPainType + Sex + FastingBS + ExerciseAngina + Oldpeak + MaxHR + ST_Slope * MaxHR + Oldpeak * ST_Slope	667	393.816	6	43.161	0

Thus, we end up adding the interaction effects of ST_Slope and MaxHR, along with Oldpeak and ST_Slope. The output for this final model is shown below:

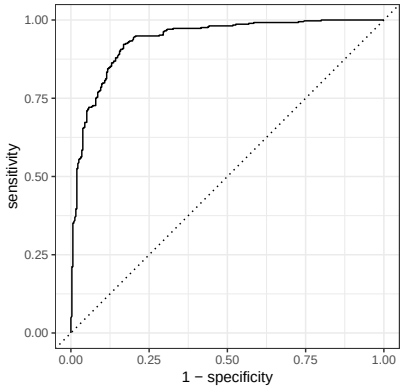
term	estimate	std.error	statistic	p.value
(Intercept)	-3.190	0.459	-6.950	0.000
ST_SlopeDown	1.835	1.913	0.959	0.337
ST_SlopeFlat	5.132	0.610	8.415	0.000
ChestPainTypeATA	-2.040	0.431	-4.729	0.000
ChestPainTypeNAP	-2.004	0.336	-5.964	0.000
ChestPainTypeTA	-1.640	0.516	-3.180	0.001
SexM	1.720	0.346	4.975	0.000
FastingBS1	1.210	0.338	3.576	0.000
ExerciseAnginaY	1.309	0.301	4.348	0.000
Oldpeak(0, 1]	1.192	0.438	2.720	0.007
Oldpeak(1, 2]	1.163	0.578	2.012	0.044
Oldpeak(2, 6.2]	18.323	828.776	0.022	0.982
MaxHR	0.009	0.008	1.091	0.275

term	estimate	std.error	statistic	p.value
ST_SlopeDown:MaxHR	-0.006	0.018	-0.362	0.717
ST_SlopeFlat:MaxHR	-0.036	0.012	-2.880	0.004
ST_SlopeDown:Oldpeak(0, 1]	-1.137	2.156	-0.528	0.598
ST_SlopeFlat:Oldpeak(0, 1]	-4.103	0.751	-5.461	0.000
ST_SlopeDown:Oldpeak(1, 2]	-0.714	2.083	-0.343	0.732
ST_SlopeFlat:Oldpeak(1, 2]	-2.919	0.816	-3.577	0.000
ST_SlopeDown:Oldpeak(2, 6.2]	-16.990	828.778	-0.020	0.984
ST_SlopeFlat:Oldpeak(2, 6.2]	-19.281	828.776	-0.023	0.981

We observe that the deviance of a few terms (one level for Oldpeak and two levels for interaction between ST_Slope and Oldpeak) are unusually high. This may indicate potential multicollinearity issues. To prevent potential multicollinearity issue, we further calculate the VIF (See output in Appendix 3), which is consistent with the unusual deviance. Specifically, the VIF values for Oldpeak (2, 6.2], ST_SlopeDown:Oldpeak(2, 6.2] and ST_SlopeFlat:Oldpeak(2, 6.2] exceeds 10^6 (a million), indicating the existence of strong multicollinearity between Oldpeak and the interaction terms for ST_Slope and Oldpeak. To ensure reliability and interpretability, we decide to drop the interaction terms for ST_Slope and Oldpeak. The VIF values for the updated final model (see Appendix 3) are all smaller than 2, thus no longer implying potential multicollinearity issues. This will be our final model.

Lastly, we update the recipe and construct the workflow based on the result above, fit the final model based on the training set, and evaluate the performance of the model based on the training set. We display the final model in the Result section. We use the ROC curve and AUC to evaluate the model performance on the training set, the results displaying below.

.metric	.estimator	.estimate
roc_auc	binary	0.9365973



The ROC curve is quite close to the upper left corner of the graph and the AUC value is approximately 0.94, both of which indicate that our final model fits the data quite well. We will further evaluate the model using the testing set in the result section for further conclusion of model performance.

Results

Final model output:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.425	0.394	-6.161	0.000	-3.221	-1.676
MaxHR	0.005	0.007	0.664	0.507	-0.009	0.020
ST_Slope_Down	1.218	0.496	2.454	0.014	0.257	2.215
ST_Slope_Flat	2.610	0.298	8.755	0.000	2.041	3.213
ChestPainType_ATA	-2.115	0.401	-5.275	0.000	-2.936	-1.357
ChestPainType_NAP	-1.775	0.306	-5.790	0.000	-2.389	-1.185
ChestPainType_TA	-1.254	0.481	-2.606	0.009	-2.218	-0.324
Sex_M	1.666	0.325	5.131	0.000	1.044	2.320
FastingBS_X1	1.483	0.321	4.616	0.000	0.866	2.128
ExerciseAngina_Y	1.117	0.290	3.857	0.000	0.552	1.690
Oldpeak_X.0..1.	-0.380	0.321	-1.182	0.237	-1.020	0.243
Oldpeak_X.1..2.	0.127	0.345	0.369	0.712	-0.554	0.802
Oldpeak_X.2..6.2.	1.394	0.530	2.631	0.009	0.394	2.488
ST_Slope_Down_x_MaxHR	-0.006	0.017	-0.368	0.713	-0.041	0.026
ST_Slope_Flat_x_MaxHR	-0.031	0.011	-2.793	0.005	-0.053	-0.010

Given our model meet the conditions for inference, we can interpret the final output to investigate the statistically significant (p-value < 0.05) predictors of heart disease in patients.

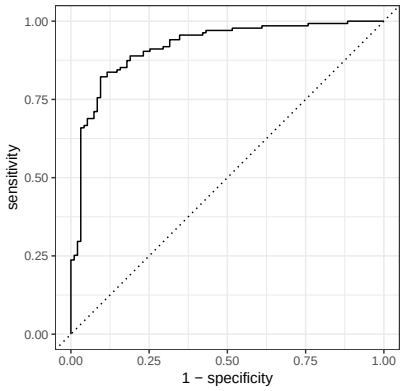
The odds of having heart disease for patients with a down ST slope are expected to be 2.832 times the odds of patients with an up ST slope, holding all else constant. The odds of having heart disease for patients with a flat ST slope are expected to be 13.531 times the odds of patients with an up ST slope, holding all else constant. The odds of having heart disease for patients who are the male sex are expected to be 5.286 times the odds of patients who are the female sex, holding all else constant. The odds of having heart disease for patients with a fasting blood sugar greater than 120 mg/dL are expected to be 4.586 times the odds of patients who do not, holding all else constant. The odds of having heart disease for patients with a exercise-induced angina are expected to be 2.992 times the odds of patients without exercise-induced angina, holding all else constant. The odds of having heart disease for patients with a ST depression induced by exercise relative to rest where the absolute value of the deviation of the ST segments from the baseline is 2 to 6.2 mm is expected to be 4.084 times the odds of patients with an ST depression induced by exercise relative to rest of 0 mm, holding all else constant.

By looking at the final model output, predictors that are statistically significant (p-value < 0.05) that increase the odds of heart disease are a flat ST slope, down ST slope, an asymptomatic chest pain type, being of the male sex, fasting blood sugar greater than 120 mg/dL, having exercise-induced angina, and a ST depression induced by exercise relative to rest where the absolute value of the deviation of the ST segments from the baseline is 2 to 6.2 mm.

Testing our model on testing data:

Using the testing data set, the ROC curve is still quite close to the upper left corner of the graph and the AUC value is approximately 0.92, both of which indicate that our final model fits new data quite well and maintains a high accuracy of predicting heart disease.

.metric	.estimator	.estimate
roc_auc	binary	0.917037



To further evaluate our model, we created a confusion matrix with a cutoff probability of 0.25. We choose a cut-off probability of 0.25 because we seek to prioritize sensitivity. Since our model is predicting whether a patient has heart disease, we want to minimize our false negative rate, because predicting a patient doesn't have heart disease when they do is more dangerous than predicting a patient has heart disease when they don't.

	Does not have heart disease	Has heart disease
Classified as heart disease	30	126
Not classified as heart disease	65	9

In our final model, the false negative rate is 6.67 % and the false positive rate is 31.58 %. The final model's sensitivity rate is 93.33 % and the specificity rate is 68.42 %. Additionally, the final accuracy produced by the model was 83.04%. Based on these metrics, we can conclude our final model is a strong predictor of whether a patient has heart disease.

An interesting find in our model is how an asymptomatic chest pain type in a statistically significant predictor of heart disease, holding all else constant. This finding coincides with the phenomenon of being asymptomatic, particularly in terms of chest pain, yet still having heart disease or a heart attack. This shows that extreme chest pain is not a definitive sign of heart disease.

Discussion + Conclusion

Summary of our findings:

When looking at our results on the testing data, our model produced an extremely high sensitivity at 93.33 % along with a relatively high specificity at 68.42 %—furthermore, the model produced a high accuracy of 83.04%. In the context of our data, it is imperative for our model to be able to predict that one has heart disease with the information that they do truly have heart disease because of the value of human life, therefore indicating our preference to prioritize sensitivity. While maximizing specificity is not as crucial compared to sensitivity in this situation, our outputs for these rates and our accuracy rate indicate that our model is relatively reliable for predicting the occurrence of heart disease and that there is still room for improvement.

In regards to our research question in finding the best predictors for heart disease, our methodology in utilizing AIC as a performance metric along with drop-in-deviance tests for interaction terms allowed us to hone in on the optimal predictors we sought after. More specifically, our final model includes the following statistically significant (p-value < 0.05) predictors that increase the odds of heart disease: a flat and down ST slope, has an asymptomatic chest pain type, male sex, fasting blood sugar > 120 mg/dL, has exercise-induced angina,

has an ST depression induced by exercise relative to rest where the absolute value of the deviation of the ST segments from the baseline is 2 to 6.2 mm (full model in Results output). We can conclude that patients that have these statistically significant characteristics, are more likely to have heart disease than those who do not. Thus, in relation to our research question, our significant predictors of an occurrence of heart disease are these statistically significant characteristics. This is vital in the context of the real world, as it indicates how cardiologists and patients should focus on analyzing these characteristics to both diagnose and mitigate the likelihood of heart disease occurring.

Lastly, as we saw with our AUC value of 0.92, we can see that our final model fits our testing data quite well. This indicates that our model with the aforementioned predictors is likely to perform well on new, unseen data in the real world for predicting heart disease in patients where all these characteristics (especially those concluded to be statistically significant) are collected.

Limitations of our data/analysis and ideas for future work:

During our analysis, we ran into several limitations that pertain to our data and our analysis. In regards to data, we were not provided a location-specific variable for each observation. Our data set was curated by combining patient data from several locations, thus it is possible for location to potential affect the model, but we had no means of evaluating this. Likewise, while we did justify that serial correlation is not a problem and that all individual subjects' health characteristics were independent of one another, it would have been useful to have data related to the date of a patient observation to conduct further analysis on this condition or even for our model (i.e. evaluating if a time period has an influence on heart disease occurrence).

Meanwhile, in regards to our analysis, one limitation we encountered was our time limitation to utilize other performance metrics aside from AIC. Despite thoroughly using AIC as an evaluation benchmark, it may have been interesting to consider other metrics (i.e. BIC or even allowing for more complexity) to see if there is a more optimal model or if our model is corroborated by other metrics. Furthermore, we had a limitation on time to check every possible interaction term that may be incorporated in our model. Our methodology consisted of testing interaction terms that made intuitive sense to have some sort of relationship, but it is possible that we missed certain interaction terms that could have increased the efficiency of our model.

Lastly, we did not encounter much of an issue regarding the reliability and validity of our data, especially given the source of this data set. Our statistical analysis has highly appropriate applications to the real world, especially with the ultimate goal in determining the best possible predictors that allow for one to determine if a patient has heart disease. However, we would like to point out a limitation regarding the application of our model and data in today's society. Given that the data was collected about 40 years ago, there is the possibility of predictors for heart disease to have changed in general, especially in regards to new technologies that are available, changes in foods that are regularly eaten, and other daily living practices that have become more commonplace. Thus, while we do believe that our conclusions on certain predictors could change with more recent data (something which we would like to investigate in future work), our model nonetheless still provides valuable insight on various predictors that need to be considered when checking for heart disease.

Therefore, our ideas for future work would be to address some of our limitations, especially those related to analysis. With more time on this project, we would test more interaction terms and perform corresponding drop-in-deviance tests to get a holistic idea of all possible interaction terms that could optimize the model. Furthermore, we would use another fit evaluation metric such as BIC to see how the best BIC model compares to our model evaluated with AIC. In the event of a new, optimal model using BIC, it is possible we discover further predictors that are crucial for cardiologists in predicting heart disease.

Appendix

Appendix 1: Output of forward and backward selection

The output of the forward selection process:

```
Call:
glm(formula = HeartDisease ~ ST_Slope + ChestPainType + Sex +
    FastingBS + ExerciseAngina + Oldpeak + MaxHR, family = "binomial",
    data = heart_train)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.315093   0.384590  -6.020 1.75e-09 ***
ST_SlopeDown      1.136656   0.505377   2.249  0.02450 *
ST_SlopeFlat      2.523388   0.290899   8.674  < 2e-16 ***
ChestPainTypeATA -2.044073   0.395651  -5.166 2.39e-07 ***
ChestPainTypeNAP -1.720892   0.300532  -5.726 1.03e-08 ***
ChestPainTypeTA  -1.418007   0.488733  -2.901  0.00372 **
SexM              1.657501   0.320002   5.180 2.22e-07 ***
FastingBS1        1.496363   0.319714   4.680 2.86e-06 ***
ExerciseAnginaY   1.126686   0.284321   3.963 7.41e-05 ***
Oldpeak(0, 1]    -0.438830   0.316556  -1.386  0.16567
Oldpeak(1, 2]     0.103126   0.340441   0.303  0.76195
Oldpeak(2, 6.2]   1.339115   0.535427   2.501  0.01238 *
MaxHR            -0.008473   0.005121  -1.655  0.09802 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 948.88  on 687  degrees of freedom
Residual deviance: 445.40  on 675  degrees of freedom
AIC: 471.4

Number of Fisher Scoring iterations: 6
```

The output of the backward selection process:

```
Call:
glm(formula = HeartDisease ~ Sex + ChestPainType + FastingBS +
    MaxHR + ExerciseAngina + Oldpeak + ST_Slope, family = "binomial",
    data = heart_train)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.315093   0.384590  -6.020 1.75e-09 ***
SexM              1.657501   0.320002   5.180 2.22e-07 ***
ChestPainTypeATA -2.044073   0.395651  -5.166 2.39e-07 ***
ChestPainTypeNAP -1.720892   0.300532  -5.726 1.03e-08 ***
ChestPainTypeTA  -1.418007   0.488733  -2.901  0.00372 **
```

```
FastingBS1        1.496363   0.319714   4.680 2.86e-06 ***
MaxHR            -0.008473   0.005121  -1.655  0.09802 .
ExerciseAnginaY   1.126686   0.284321   3.963 7.41e-05 ***
Oldpeak(0, 1]    -0.438830   0.316556  -1.386  0.16567
Oldpeak(1, 2]     0.103126   0.340441   0.303  0.76195
Oldpeak(2, 6.2]   1.339115   0.535427   2.501  0.01238 *
ST_SlopeDown      1.136656   0.505377   2.249  0.02450 *
ST_SlopeFlat      2.523388   0.290899   8.674  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 948.88  on 687  degrees of freedom
Residual deviance: 445.40  on 675  degrees of freedom
AIC: 471.4

Number of Fisher Scoring iterations: 6
```

Appendix 2: Output for selecting interaction terms using drop-in-deviance tests

b). MaxHR and Sex

term	df.resid	resid.devi	df	devi	p-val
HeartDisease ~ ST_Slope + ChestPainType + Sex + FastingBS + ExerciseAngina + Oldpeak + MaxHR + ST_Slope * MaxHR	673	436.977	NA	NA	NA
HeartDisease ~ ST_Slope + ChestPainType + Sex + FastingBS + ExerciseAngina + Oldpeak + MaxHR + ST_Slope * MaxHR + MaxHR * Sex	672	434.070	1	2.907	0.088

Appendix 3: VIF values for each candidate model

a). The final main effects model

	vif
ST_SlopeDown	1.351
ST_SlopeFlat	1.437
ChestPainTypeATA	1.123
ChestPainTypeNAP	1.201
ChestPainTypeTA	1.164
SexM	1.098
FastingBS1	1.070
ExerciseAnginaY	1.220
Oldpeak(0, 1]	1.345
Oldpeak(1, 2]	1.493
Oldpeak(2, 6.2]	1.314
MaxHR	1.113

b). The model with the two interaction terms (interactions between ST_Slope and MaxHR and interactions between ST_Slope and Oldpeak)

	vif
ST_SlopeDown	19.632
ST_SlopeFlat	5.548
ChestPainTypeATA	1.165
ChestPainTypeNAP	1.279
ChestPainTypeTA	1.239
SexM	1.171
FastingBS1	1.082
ExerciseAnginaY	1.240
Oldpeak(0, 1]	2.623
Oldpeak(1, 2]	4.067
Oldpeak(2, 6.2]	3036989.017
MaxHR	2.396
ST_SlopeDown:MaxHR	1.406
ST_SlopeFlat:MaxHR	2.052
ST_SlopeDown:Oldpeak(0, 1]	5.310
ST_SlopeFlat:Oldpeak(0, 1]	4.681
ST_SlopeDown:Oldpeak(1, 2]	10.077
ST_SlopeFlat:Oldpeak(1, 2]	5.723
ST_SlopeDown:Oldpeak(2, 6.2]	1384262.740
ST_SlopeFlat:Oldpeak(2, 6.2]	1782621.578

c). The final model (after dropping the interaction between ST_Slope and Oldpeak)

	vif
ST_SlopeDown	1.353
ST_SlopeFlat	1.472
ChestPainTypeATA	1.133
ChestPainTypeNAP	1.220
ChestPainTypeTA	1.174
SexM	1.104
FastingBS1	1.074
ExerciseAnginaY	1.232
Oldpeak(0, 1]	1.373
Oldpeak(1, 2]	1.514
Oldpeak(2, 6.2]	1.326
MaxHR	2.199
ST_SlopeDown:MaxHR	1.275
ST_SlopeFlat:MaxHR	1.866

Appendix 4: References

Benjamin, Emelia, et al. "Heart disease and stroke statistics—2018 update: A report from the American Heart Association." *Circulation*, vol. 137, no. 12, 2018, https://doi.org/10.1161/cir.0000000000000558.

"Cardiovascular Diseases (Cvds)." *World Health Organization*, World Health Organization, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 1 Dec. 2023.

Detrano, Robert, et al. "International application of a new probability algorithm for the diagnosis of coronary artery disease." *The American Journal of Cardiology*, vol. 64, no. 5, 1989, pp. 304–310, https://doi.org/10.1016/0002-9149(89)90524-9.

Lopez, Edgardo, et al. "Cardiovascular Disease." *StatPearls*, updated 22 Aug. 2023, StatPearls Publishing, 2023 Jan. Web. Available from: https://www.ncbi.nlm.nih.gov/books/NBK535419/

Appendix 5: Output of emplogit plots to check linearity condition

The output of the emplogit function: