# HEART DISEASE STUDY

BY: Albert Alva, Harinder Bhasin, Jeff Chandler, and Matthew Thacker

# Introduction

The dataset we found was used to analyze coronary heart disease through observation of patient data. Our dependent variable, heart disease, is caused when major vessels in the heart are narrowing in patients. 0 indicates minimal narrowing while 1 indicates greater narrowing, which is a negative outcome for the patient. The dependent as well as the various independent variables will be discussed in greater detail below.

Our data was found at the UC Irvine Machine Learning Repository. The original datasets included information from the Hungarian Institute of Cardiology, Budapest, the University Hospital, Zurich, Switzerland, the University Hospital, Basel, Switzerland, as well as the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

The Cleveland clinic data is the most complete repository of the data sets from the 6 treatment centers. It was identified as the only dataset used for machine learning to date. The datasets originally contained 76 attributes from all aspects of heart health, but for machine learning purposes, it was narrowed down to a subset of 14 attributes which were found to be the most predictive for risk of coronary artery disease.

Robert Detrano, M.D., Ph.D., was the primary investigator who collected data for the Cleveland database. The data collection was derived from the clinical and noninvasive test results of 303 patients undergoing angiography at the Cleveland Clink in Cleveland, Ohio. The dataset contains a mix of categorical, continuous, binomial and integer data.

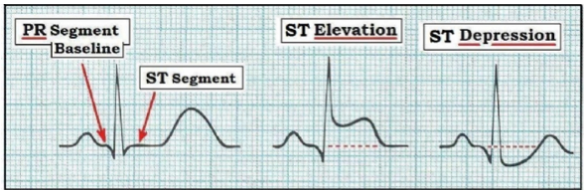
We reduced the dataset to eliminate rows with missing attributes and analyzed the balance of the variables for impossible scores. Though some scores were very high, there was no reason found to eliminate them from the study.

# Descriptive Statistics

The following is a list of the data variables and their description:

1. **Age**: patients age
2. **Gender**: (binomial) Female (0), Male (1)
3. **CP:** (Categorical) represents chest pain type. Angina is a common medical term to label a condition marked by severe pain in the chest, often also spreading to the shoulders, arms, and neck, caused by an inadequate blood supply to the heart.
    a. 1 = Typical angina: the presence of substernal chest pain or discomfort that was provoked by exertion or emotional stress and was relieved by rest and/or nitroglycerin.
    b. 2 = Atypical angina: often doesn't cause pain, but you may feel a vague discomfort in your chest, experience shortness of breath, feel tired or nauseous, have indigestion, or pain in your back or neck.
    c. 3 = Non-angina pain: This is pain unrelated to the cardiac system.
    d. 4 = Asymptomatic: This is when a patient is a carrier for a disease or infection but experiences no symptoms.
4. **Trestbps:** represents resting blood pressure. If your blood pressure is high, it is putting extra strain on your arteries and on your heart. ... This may also cause a heart attack or stroke. The measurement given was taken on admission to the hospital.

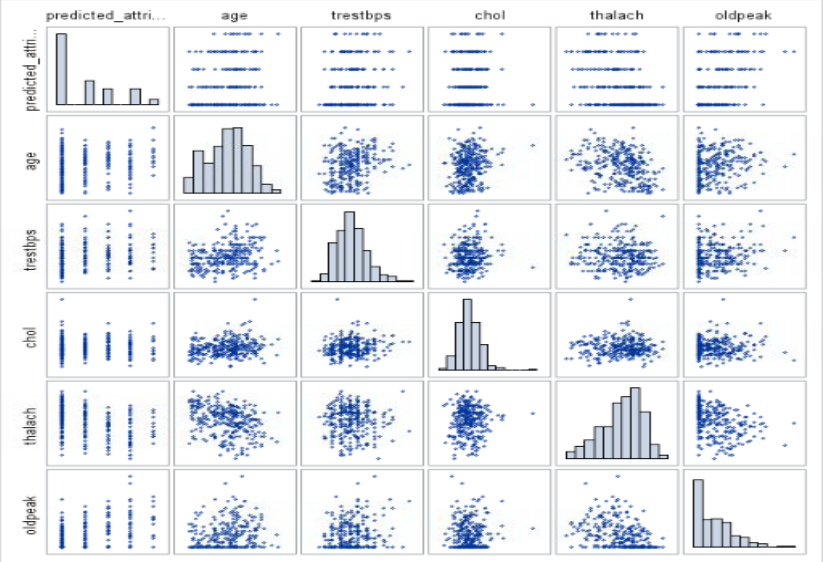
5. **Chol:** represents Serum Cholesterol found in the bloodstream. Cholesterol needs to be combined with fats and proteins to be carried in the bloodstream. If the liver makes too much cholesterol there will be unnecessary fats and proteins clogging up your bloodstream and the risk of heart disease is increased.
6. **FBS:** represents fasting blood sugar. A normal FBS should be under 100 mg/dl while those with diabetes range from 80-130. If the rate is greater than 120 mg/dl it is identified by a 1, otherwise the patient receives a 0 for normal levels.
7. **Restecg:** (Categorical) represents resting electrocardiographic results. This is the electrical activity of the heart as it undergoes excitation (depolarization) and recovery (polarization) to initiate each beat of the heart. The regular pattern of the ECG allows analysis to determine whether there is any abnormality in a patient. Once again 0 indicates normality while 1 indicates some abnormality and 2 indicates probable to definite enlargement of an organ or tissue from the increase in size of its cells.
8. **Thalach:** represents the maximum heart rate achieved during thallium stress test. During the procedure, a radioisotope (nuclear liquid) is administered through an IV. A Nuclear imaging test then shows how well blood flows into the heart while you're exercising or at rest.
9. (binomial) Exang represents whether exercise induces cardiac (angina) pain. 0 means no and 1 indicates yes.
10. **OldPeak**: represents the ST segment prior to exercise. Stress Testing depression refers to a measure on an electrocardiogram (ECG) that compares exercise relative to a resting rate. This helps to measure Coronary Artery Disease (CAD) which is a blockage of arteries that supply blood to the heart. See image of healthy vs unhealthy ST depression.




11. **Slope**: (Categorical) relates to the slope of the peak exercise ST Segment. 1 indicates upsloping, 2 indicates flat, and 3 indicates down sloping.
12. **CA**: represents No. of major vessels colored by fluoroscopy. Fluoroscopy means that a continuous X-ray beam is passed through the body part being examined. The beam is transmitted to a TV-like monitor so that the body part and its motion can be seen in detail. With the heart this enable the doctor to see the flow of blood through the coronary arteries in order to evaluate the presence of arterial blockages. The greater the number of arteries that are blocked (colored) the worse the condition. The range is from 0-3.
13. **Thal**: (Categorical) represents the Thallium stress test which is a nuclear imaging test that shows how well blood flows into the heart while you're exercising or at rest. During the procedure, a radioisotope (nuclear liquid) is administered through an IV. The radioisotope will flow through your blood stream and end up in your heart. Once the radiation is in your heart, a special camera called a gamma camera can detect the radiation and reveal any issues your heart muscle is having.
    The results listed include (3=normal, 6=fixed defect, or 7=reversible defect)
14. **Num**: (predicted_attribute) represents the heart disease status: number of major vessels with >50% narrowing (0, 1, 2, 3, or 4)

## Graphical Descriptive summary

The histograms below show the attributes that are not categorical or binomial before any modification, the exception being the categorical dependent variable. The trestbps appears to have two overlapping patterns, but we believe this to be due to many healthy people having an average resting beats per second of 120, 130, 140 indicated by the straightened lines through the scatter plot. Also note the high number of 0s shown for the oldpeak histogram. We believe this is because prior to exercise, a healthy individual at rest has a univariate ST segment.




## Numerical Descriptive Summary

The follow list of attributes provides the mean and Standard Deviation respectively:

**Sex:** 0.67676768, 0.46849997

**Trestbps:** 131.693603, 17.7628064

**Chol:** 247.350168, 51.9975825

**Fbs:** 0.14478114, 0.35247393

**Cp4:** 0.48398577, 0.50063508

**Restecg1:** 0.01423488, 0.11866912

**Restecg2:** 0.50533808, 0.50086351

**Slope (flat):** 0.46619217, 0.49974574

**Slope (downward):** 0.07117438, 0.25757462

**Ca (1):** 0.22775801, 0.42013385

**Ca (2):** 0.13523132, 0.3425806

**Ca (3):** 0.07117438, 0.25757462

**Thalach (fixed defect):** 0.06405694, 0.24529119

**Thalach (reversible defect):** 0.39145907, 0.48894749

**Exang:** 0.33096085, 0.47139853

**Oldpeak:** 1.06441281, 1.16417372

**Predicted_attribute (pa):** 0.95729537, 1.23561371

# Analysis

Predicting heart disease has been the goal of many studies. This study will join the previous studies in creating a model that can predict heart disease before it develops

The Cleveland Clinic patient data consisted of patient information including those diagnosed of heart disease. There were 297 observations used, 201 men and 96 women, ranging in age 29 to 77 that were being seen at the Cleveland Clinic.

The hypotheses being tested are first there is a correlation between the explanatory variables and the predictor variable, Ho = r = 0 $H_1$ = r ≠0. Secondly, testing the strength of the relation, Ho = mu1 = mu1, $H_1$ = mu1 ≠ mu2. The third hypothesis evaluated if a restricted model could explain the data as well as the full model. Ho = mu Full Model = mu Restricted model adjusted R. The final goal of the study is to be able to predict heart disease with the regression model developed using measures that are known to be risks of heart disease.

The final goal of the study is to be able to predict heart disease with the regression model developed using measures that are known to be risks of heart disease.

The data was recoded to be consistent with the type of data suitable for regression. The categorical variables with more than two levels were recoded with a dummy code on one level and transformed into a binomial numbers. The predictor variable was also transformed into a binomial number, 0 for the absence of heart disease and 1 for presence of heart disease.

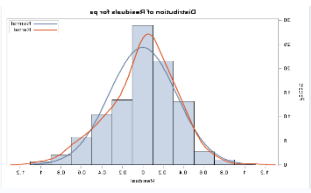
The prediction model was built using the multiple linear regression procedures in SAS 9.4. The 14 variables provided in the data were used in the analysis. When recoded to create binary variables, 19 variables became available for evaluation in the in the full model (see table 2).

The full model containing 19 variables resulted in both a high correlation and significant ANOVA. The correlation = 75.51, r2 = 0.5702 and adjusted r2 = 0.547. The full model explains 57% of the variance in the diagnosis of heart disease. The ANOVA showed the predictor variable differed significantly in their relation to the explanatory variables two tailed ANOVA, F= 19.34 df19 ,277, p <0.001 .

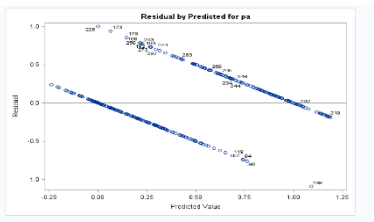
The diagnostic graphs of the data were viewed to ensure all variables and the model meet the assumptions of the regression procedure. None of the variables show correlations higher than .3642 but the heart defect measure (thal_reversable_defect) was negatively correlated with SexMale r=- .3642 and positively correlated with thal_fixed_defect, r = - .355. The two chest pain levels non-anginal pain and asymptomatic are also related, r = .3321. Oldpeak showed higher relations with two levels of the slope of the peak exercise ST, slope _flat(r=.3513) and slope_upsloping(r=.4842).

The histogram of the residuals (graph 1) shows the distribution to be positively skewed and slightly leptokurtic but still fairly normally distributed. All of the other diagnostic plots are showing outliers in both groups. The positive group has outliers on the high end and a table of the estimates shows at least 21 outliers 4 or more standard deviations from the mean. The result is the model loses sensitivity on the low end for the presence of heart disease and on the high end of the absence group. The outlier and Leverage graph shows that none of the outliers are leverage points and none of the leverage points are outliers. The outliers do not appear to be in points of leverage.

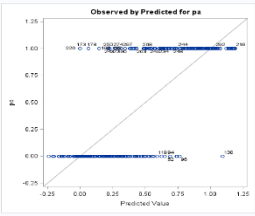
Graph 1



Graph 2

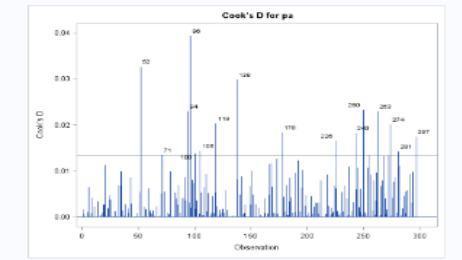



Graph 3

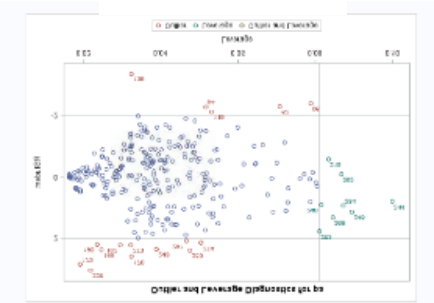


Cooks D confirms the outliers and their influence to the model. The outliers illustrated in the Outliers and Leverage graph were shown to be influential to the results of the model. The cooksD did not surpass .026.

Graph 4




Graph 5



Possible issues with collinearity were a concern with two of the variables in the model, oldpeak and the blood sugar measure. These variables where removed from the model in the auto selection or by intention since the auto selection did not eliminate it in the model. The data suffers from many outliers in both of the predictive variable levels. In review of the data, there was no reason to conclude the outliers were in error nor part of different populations, thus included in the sample.

The model was then tested to see if a more elegant model could result in as good a prediction as does the full model. Multiple model selection methods were employed to build a model that minimizes correlated predictors, lowers VIF and maximizes the AIC for all variables while maintaining or improving on the original model's performance. This includes r2, standard error of the estimates and overall accuracy of the prediction.

The model selection started by viewing the variables in the model, then heuristic tests were performed first using forward then backward selection then stepwise selection with $p<.05$ for entry or elimination of the variables. This was followed by the stepwise selection method using both Proc Reg and Proc GLMSelect. The first model selection method was Proc Reg followed by GLMselect in SAS.

The Proc Reg forward selection method yielded a 17 variable model with r2 = .5932, F = 21.38, df(17,279)p<.0001, though the model was significant, it was not much better than the full model. While the GLM forward model yielded an 8 variable model with adjusted r2 = .5416, F=42.53, df(8,288),P<.001, both models decreased the adjusted r2. The stepwise procedures yielded similar results.

After the heuristic approaches were used, the Akaike information criterion (AIC) approach was employed. The AIC approach selects models based on how well they explain the data as compared to all other models, the AIC should be maximized by the model. At the same time, the model should not increase error nor decrease the R2. Proc Reg uses the adjusted r2 selection method to compile a table with all possible models. The models were then ranked by AIC, R2 and SSE to determine the best model. The system selected a model with 12 variables, adjusted r2 = .5412, SSE = .32.48. After the auto selection presented the aforementioned model, one more model was tested that had one less variable then the auto selection but retained all of the power to find differences. The model eliminated the blood sugar measure and was still able to explain more of the variance than did the full model 54.04% compared to 53.92% of the variance in diagnosis (adjusted r2= .5404), the SSE did increase but less than one hundredth of a point(32.66 compared to 32.05). The AIC did decrease by 8 points with the reduction of the model (- 631.22 compared to - 623.23).

## Hypothesis Results and The Model

The results of testing the 3 hypothesis were favorable. There is strong evidence to suggest there is relation between heart disease and the explanatory variables r=.7521. We reject the null in favor of the alternative hypothesis. The correlation between the variables is not 0. The explanatory variables are related to heart disease. The relation shown by the correlation is statistically significant, two tailed ANOVA yielded, F=21.38, DF=17,279 P<.0001. There is strong

evidence to support the statement that the two levels of heart disease differ. We reject the null hypothesis in favor of the alternative hypothesis. The groups are not equal. Patients with heart disease have different responses to the variables as do the patients without heart disease. The final hypothesis that no restricted model can explain the data as well as the full model is rejected. A restricted model was developed that had higher R than the model with more measures. It is able to predict more accurately than the full model.

The following variables were selected as the best predictors of heart disease among the measures collected. A gender shows men at higher risk than females, men are at .154 higher risk then women, t= 3.40, df = 1 p = .0008. Chest pain was also a risk factor, asymptomatic chest pain increased risk by .235, t=4.95, df = 1 p<.0001. Resting blood pressure adds risk, .00237, t=2.07, df = 1 p=.0391. As resting blood pressure increases, risk increases. Resting electrocardiogram showing probable or definite left ventricular hypertrophy by Estes' criteria adds .0453 to the risk score, t=1.12, df1, p.2646. Though, the measure does not reach significance alone, it adds to the predictability of the model. Maximum heart rate achieved is negatively associated with heart disease. As maximum heart rate increases, risk decreases by -.00919. For every increase of max heart rate, risk decreases by .00919 t=-1.79, df = 1 p=.0742. Exercise induced angina also signals risk. The presence of exercise induced angina increases risk by .0927, t=1.85 df = 1 p=.066. Treadmill segment of stress tests indicates that if any changes occur during the upsloping phase of the test vs flat or downward sloping, risk is decreases by .1579, t=3.47, df=1 p=0.0006.

The strongest indicators of risk were the number of vessels colored during a fluoroscopy. Any vessels showing up on the fluoroscope adds risk with 2 vessels adding the most risk, followed by one and three vessels. One vessel increases risk, .257, two vessels increase risk .308 and three colored vessels during the fluoroscopy increase risk by .293. The final risk factor is a reversible heart defect. The presence of a reversible defect increases risk by .216.

The risk factors culminate into a model that can predict heart disease 86.9% of the time as compared to the full model at 0.86.5% of the time. The regression model is:

Heart Disease = Intercept+ sexmale(0.15362)+cp4(0.23533)+ trestbps(0.00237)+ restecg2(0.04529)+thalach(-0.00191)+exang(0.09274)+slope_upsloping(-0.15787) +ca1(0.25733)+ca2(0.33996)+ca3(0.30831)+thal_reversable_defect(0.21152)

Example

**Heart Disease: actual status Heart Disease Positive, 1.07629** = 0.0373+ 1(0.15362)+1(0.23533)+ 130(0.00237)+ 0(0.04529)+115(-0.00191)+1(0.09274)+0(-0.15787) +1(0.25733)+0(0.33996)+0(0.30831)+1 (0.21152)

## Conclusion

There are many people that would benefit from being able to predict heart disease. It is a challenge to be able to develop a model that will work for anyone. The current study focused on the model process since the data collection was out of scope for the project. Though the model does predict heart disease more than 85% of the time, more work needs to be done to increase the accuracy of the model. Currently, due to outliers, this model underestimates positive heart disease. Though 85% is a large percentage, predicting heart disease must accurately predict the disease closer to 99% of the time. More research is needed to understand the risks of heart disease.