Stat 632 Final Project:

Predicting Coronary Artery Disease

By: Michael Najarro and Cheuk Tam

5/14/19

Introduction

According to the World Health Organization, Cardiovascular diseases are the leading global cause of death (Miao 2016). The CDC estimated that 31% of all deaths in 2016 were caused by cardiovascular diseases (Heron, 2016). Within this family of diseases, Coronary Artery Disease (CAD) is the most common and is defined as the accumulation of plaque and cholesterol on the interior walls of the arteries surrounding the heart. This accumulation can narrow and harden arteries, thereby reducing blood flow and needed oxygen to the heart. In advanced stages, CAD can lead to a myocardial infarction (heart attack) when a blockage bursts, leading to a clot in the artery that deprives the heart tissue oxygen and nutrients.

CAD and Myocardial Infarction symptoms among the population are variable, leading to many undetected yet avoidable deaths. Thus, predicting the presence CAD using statistical models has been a relevant approach to promoting preventive coronary health care before the onset of a massive myocardial infarction. Detrano et al. (1989) attempted to predict the presence or absence of CAD in a global population of patients. He first built a logistic regression model upon a set of predictors from patients from a Cleveland Medical center (Detrano 1984). He then applied his training model to predict CAD in global distribution of patients from Hungary, Switzerland, and Long Beach, California, with limited success.

In this report we attempted to predict CAD in patients using Detrano's original data set. However, we applied modern machine learning approaches in hopes of improving CAD prediction. We also applied domain knowledge and compared four different models that contained possible combinations of two clinically relevant predictors, thallium imaging ('thal') and the slope of the ST segment of an Electrocardiograph (ECG, 'slope').

Data Description

We downloaded the original data collected by Detrano et al. (1984, 1989) from the UC Irvine Machine Learning Repository (Aha, n.d.). The entire data set was a 920 x 14 data frame consisting of numeric, factorial (from 2 to 4 levels), and dummy variables. Originally, the response variable 'num,' was a factorial variable that indicated the presence and number of arteries suffering from CAD, with values ranging from 0 to 4. For our analysis, we modified the response variable to be a strict dummy variable with 0 and 1 for the absence or presence of CAD. The 'ca' variable was not included for analysis because most patients (66%; 611/920) did not have cardiac fluoroscopy data (Fig. 1A). Since fluoroscopy provides information similar to thallium imaging, we excluded 'ca' from the analysis, yielding 371 complete records for 12 predictors, after removing all rows containing an NA value (Fig. 1B).

Several of our variables were categorical variables with 2 to 4 levels (Figure 1C). To reduce computational load and to simplify our model, we converted these variables to continuous, numeric variables. Summary statistics are displayed in Fig. 1C for categorical variables

Figure 1C describes all variables within the data set. Three of the variables, 'restecg,' 'oldpeak,' and 'slope,' are derived from electrocardiography (ECG), which is a record of the electrical activity of the heart (Fig. 2). The three variables focus on the ST segment, which corresponds to the contraction of the heart to eject blood to rest of the body ("Electrocardiography", n.d.). Two variables, 'thal' and 'ca,' correspond to imaging of blood flow to coronary arteries. Thallium imaging('thal') uses a radioisotope of thallium and a radiation detector while cardiac fluoroscopy ('ca') uses a contrast agent and continuous X-Ray imaging to track blood flow (McKillop 1981 ; FDA, n.d.).

72% (268/371) of the data came from male patients, yet was fairly balanced between those with (54%; 200/371) and without (46%; 171/371) CAD. Density plots of the normal predictors indicated approximately normally distributed data (Fig. 3). Density plots on cholesterol, however, displayed a slight a bimodal distribution due to 48 patients having 0 mg/dL of cholesterol, which we had not discovered prior to the development of the model. Post-modeling analyses on cholesterol without these values had no effect on model performance. The distribution of ST depression induced by exercise (‘oldpeak’) had a preponderance of zeros, which indicated normal values. We did not transform neither the response nor any of the predictors. The median age for the patients of the data set was 56 years (Fig. 4). A pairwise scatterplot matrix indicated reasonably linear relationships (Fig. 4) and pairwise correlation calculations (Fig. 6) did not show significant correlation between quantitative variables.

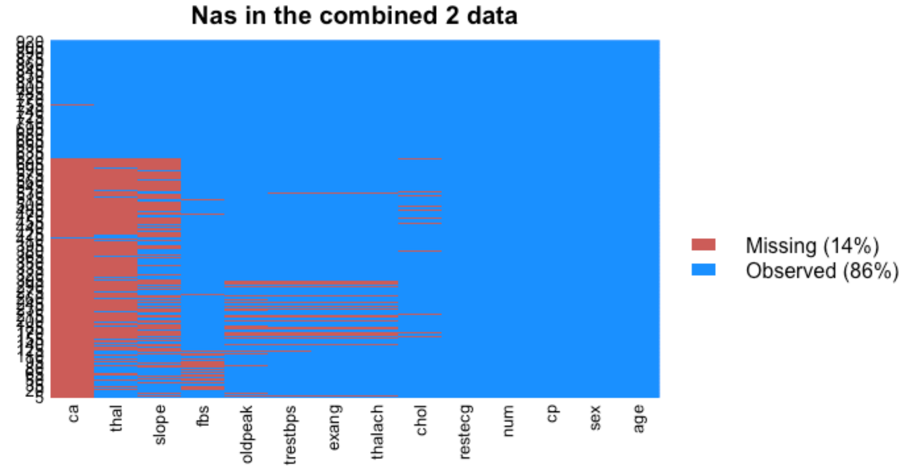


Figure 1.A Missingness Map with All Variables Present (n = 920)

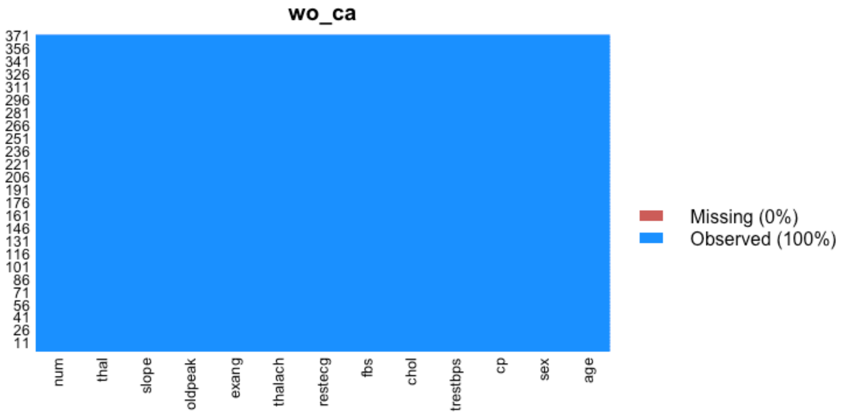
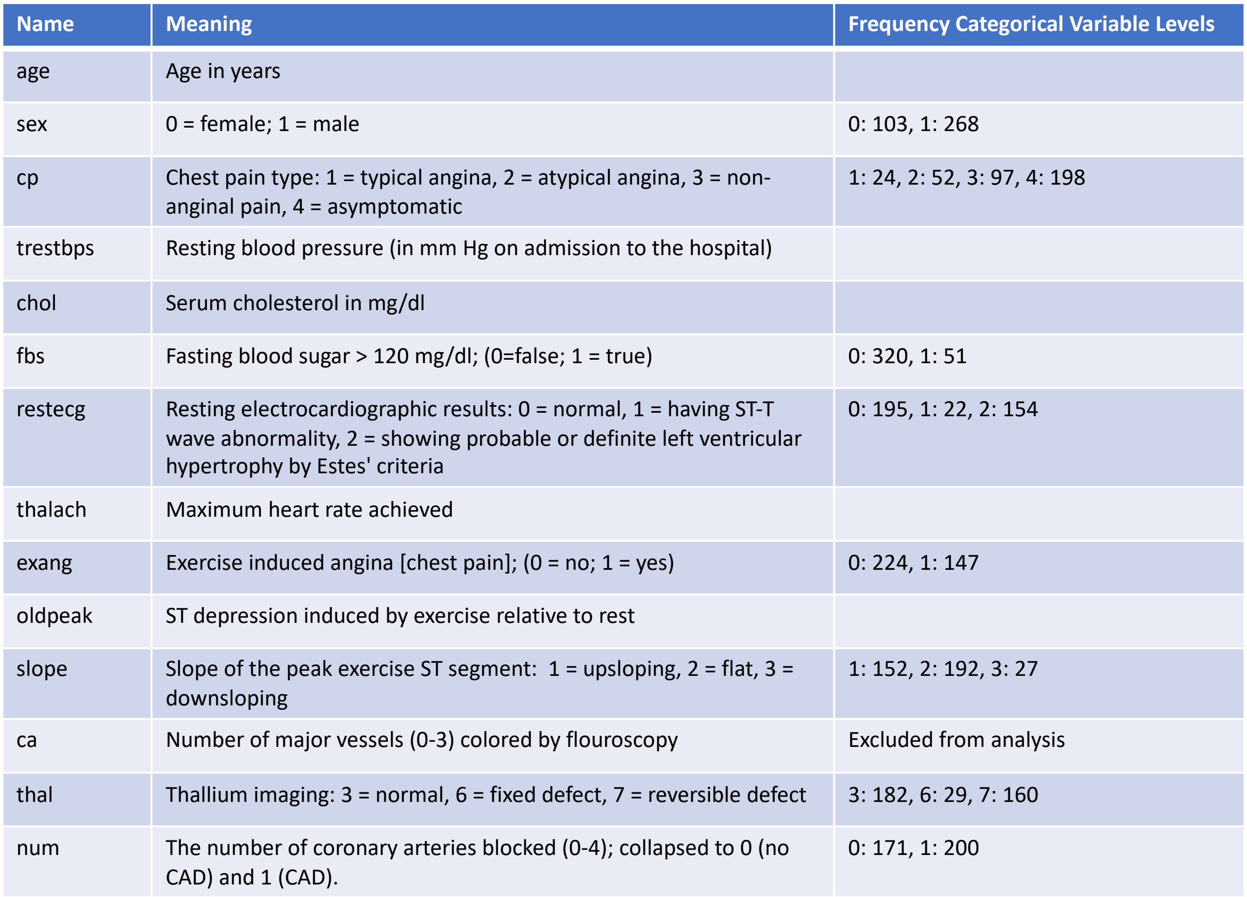


Figure 1.B: Missingness Map with ‘ca’ Column and Empty Cells Dropped (n = 371)



Name	Meaning	Frequency Categorical Variable Levels
age	Age in years	
sex	0 = female; 1 = male	0: 103, 1: 268
cp	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic	1: 24, 2: 52, 3: 97, 4: 198
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	
chol	Serum cholesterol in mg/dl	
fbs	Fasting blood sugar > 120 mg/dl; (0=false; 1 = true)	0: 320, 1: 51
restecg	Resting electrocardiographic results: 0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria	0: 195, 1: 22, 2: 154
thalach	Maximum heart rate achieved	
exang	Exercise induced angina [chest pain]; (0 = no; 1 = yes)	0: 224, 1: 147
oldpeak	ST depression induced by exercise relative to rest	
slope	Slope of the peak exercise ST segment: 1 = upsloping, 2 = flat, 3 = downsloping	1: 152, 2: 192, 3: 27
ca	Number of major vessels (0-3) colored by flouroscopy	Excluded from analysis
thal	Thallium imaging: 3 = normal, 6 = fixed defect, 7 = reversible defect	3: 182, 6: 29, 7: 160
num	The number of coronary arteries blocked (0-4); collapsed to 0 (no CAD) and 1 (CAD).	0: 171, 1: 200

Figure 1.C: Explanation of Variables and Frequencies of Categorical Variable Levels

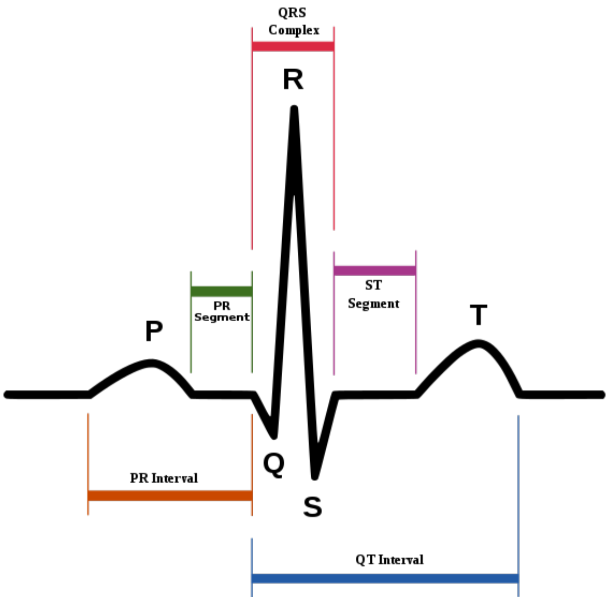


Figure 2: Typical ECG of a Normal Heart

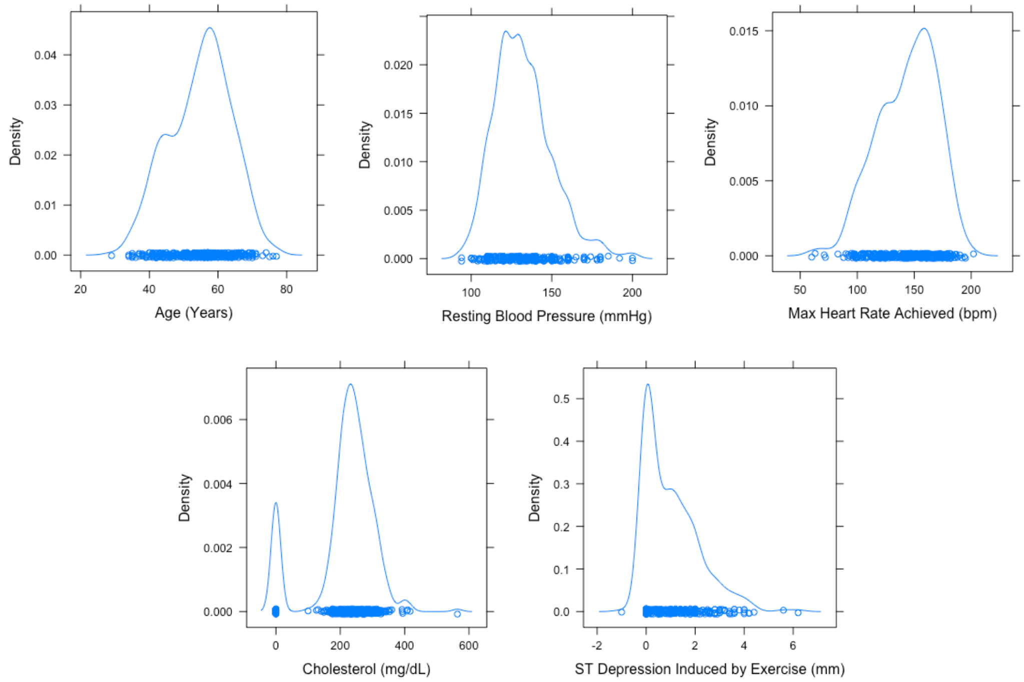


Figure 3: Density Plots of Quantitative Variables

Variable	Min.	1^{st} Q	Median	Mean	3^{rd} Q	Max	IQR
age	29	48	56	54.7	61	77	13
trestbps	94	120	130	132.1	140	200	20
chol	0	197	233	215.5	270.5	564	73.5
thalach	60	125	147	143.7	163	202	38
oldpeak	-1	0	0.8	1.014	1.6	6.2	1.6

Figure 4. Summary Statistics for Quantitative Variables

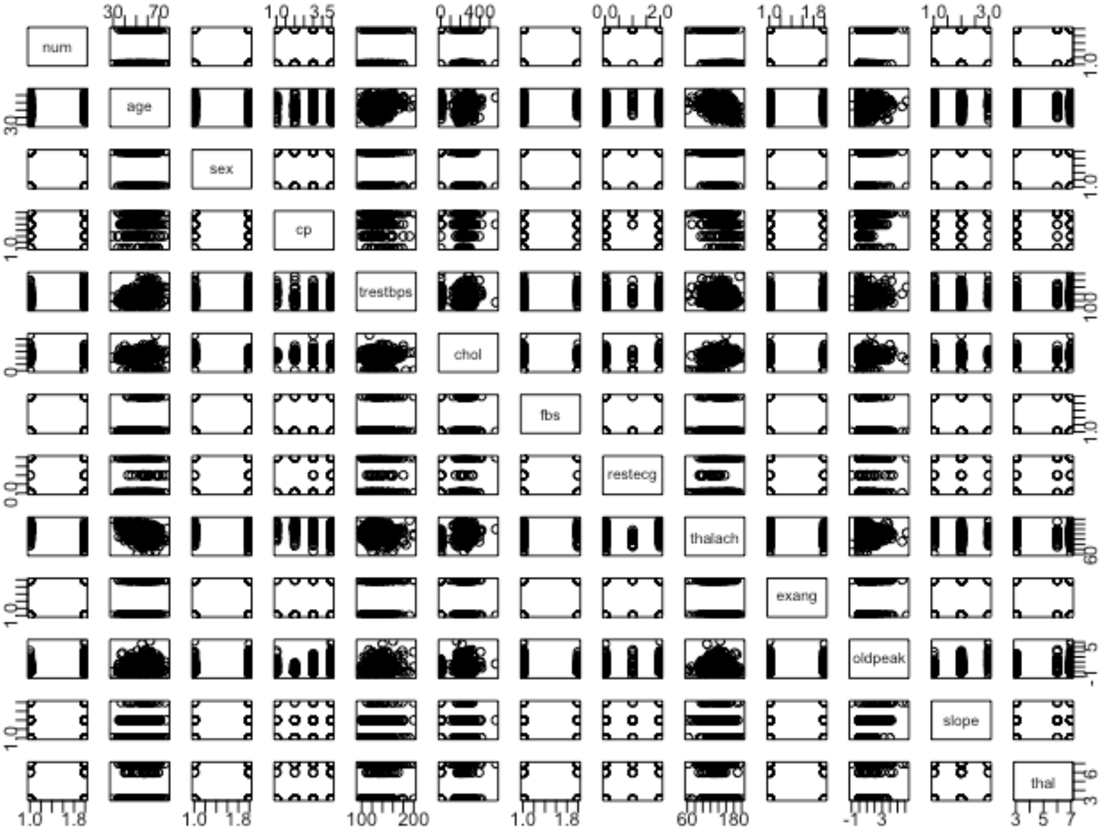


Figure 5. Pairwise Scatterplots of Predictor Variables

	age	trestbps	chol	thalach	oldpeak
age	1.00	0.30	-0.01	-0.36	0.17
trestbps	0.30	1.00	0.04	-0.08	0.16
chol	-0.01	0.04	1.00	0.37	0.15
thalach	-0.36	-0.08	0.37	1.00	-0.20
oldpeak	0.17	0.16	0.15	-0.20	1.00

Figure 6. Correlation among Quantitative Variables

Methods and Results

We first split the data into a 70% training portion and a 30% test set. We then built a logistic regression model to classify patients' CAD status using the training set. Despite its right skew, the 'oldpeak' variable was not transformed because natural choices such as log and square root transformations introduced bimodality (figures not shown). Given our insight on the importance of CAD classification with regards to the predictors of 'thal' and 'slope', we were uncertain how each predictor would respond to the presence or absence of these two predictors leading to high variability in CAD classification. We developed four full models that differed in their combination of the presence or absence of 'thal' and 'slope': 'thal' and 'slope', 'thal' only, 'slope' only, and a model containing neither predictor. We then applied a backwards stepwise procedure upon the remaining 10 predictors while excluding 'thal' and/or 'slope' from the reduction procedure. The backwards stepwise procedure showed that the model including ‘thal’ and ‘slope’ had the lowest AIC (192.88), suggesting that the two variables play important roles in the model (Fig. 7).

Model	AIC
Both thal, slope	192.88
slope, no thal	210.7
thal, no slope	193.16
neither thal nor slope	213.77

Figure 7. AIC Values of Stepwise Reduced Models with Different Combinations of ‘thal’ and ‘slope’ Variables.

The full logistic model with 12 predictors yielded an AIC of 199.56. The AIC improved to 192.88 after backwards stepwise variable selection reduced the full model to 8 predictors. The stepwise model created is described as follows:

$$\boldsymbol{P(Y = 1|x) = \frac{1}{1 + e^{-4.4+1.1sex+cp+.02trestbps+.4restecg-.04thalach+.93exang+.53slope+.47thal}}}$$

To classify patients, we considered a predicted probability >0.5 to be positive for CAD. Using this threshold, we assessed the model’s performance on the test set using a confusion matrix (Fig. 8). The model yielded a sensitivity of 0.72, a specificity of 0.77, and an accuracy of 0.74. The accuracy is significantly better than a random guess based on the data set’s original proportion for CAD (positive for CAD, 54%).

	Actual No CAD	Actual CAD	Sum
Pred No CAD	36	18	54
Pred CAD	11	46	57
Sum	47	64	111

Figure 8. Confusion Matrix of Reduced Model Applied to Test Set

Discussion

We attempted to improve Detrano et al.’s classification schema for CAD using more modern classification approaches within the context of medical and clinical domain knowledge.

Given that the variables of thallium scintigraphy('thal') and slope of peak exercise ST segment ('slope') correspond to relevant, non-invasive, clinical procedures for identifying CAD (Leppo 1989), we predicted these terms would be in our model; our prediction turned out true.

Interpretation of our logistic model can be described as follows, using the “divide by 4” rule on our 'slope' predictor: an increase in the categorical level of the ST segment of a patient’s ECG by 1 unit is associated with at most, a 0.53/4 =.1325decrease in the probability that the patient has CAD. Unfortunately, Detrano et al. (1989) did not publish their resulting methodology and so we could not compare our models.

The most significant predictors in our model were 'thalach', ‘cp’, and 'thal', followed by ‘exang1.’ Intuitively the predictive power of these terms makes sense as 'thal' (thallium scintigraphy, Wald z-test, p = 1.375*10-5), 'thalach' (maximum heart rate, Wald z-test, p = 6.44*10-5) and ‘cp’ (chest pain, Wald z- test, p = 2.10*10-5) are all common indicators of CAD (Leppo 1989, Cassar 2009). As CAD develops, the diameter of the arteries decreases and systolic blood pressure increases, which is associated with increasing CAD risk (Vasan et al. 2001, Franklin et al. 2001).

The data set used represents the logistical, technological, and analytical capabilities standard to the time from which it was collected. We presume that both classification methodologies and medical technologies were limited and non-standardized across medical facilities, resulting in data pocked with missing values (NAs); We suspect this is the reason why Detrano et al. trained their model on the most complete data (Cleveland data) and tested it on the three remaining data sets.

Others have attempted to analyze the Detrano data set for CAD. Miao et al (2016) Also applied machine learning algorithms to the same data set, using ensemble learning classification and adaptive boosting algorithms to each of the four populations of patients. Accuracy measurements for each population were as follows: 80.14% for Cleveland, 89.12% for Hungary, 77.78% for Long beach, and 96.72% for SUH (Miao et al. 2016). Although the methodology of Miao et al. is not fully comparable, their results support our accuracy prediction for CAD being greater than 50%.

While our model was accurate in predicting CAD and had high sensitivity, many of the variables selected as strong predictors are rather obvious predictors and may not share new insights on better understanding CAD. We hope future researchers consider data collection on lifestyle habits, diet measurements, economic standing, and mental well-being as ways to better predict CAD.

References

Aha, D. Heart Disease Data Set. (n.d.). Retrieved May 13, 2019, from http://archive.ics.uci.edu/ml/datasets/heart disease

Cassar, A., Holmes, D. R., Jr, Rihal, C. S., & Gersh, B. J. (2009). Chronic coronary artery disease: diagnosis and management. *Mayo Clinic proceedings*, *84*(12), 1130–1146. doi:10.4065/mcp.2009.0391

Detrano R, Yiannikas J, Salcedo EE, Rincon G, Go RT, Williams G, Leatherman J. (1984). Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. *Circulation. 1984 Mar*;69(3):541-7.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64,304--310.

Dey, S., Flather, M. D., Devlin, G., Brieger, D., Gurfinkel, E. P., … Steg, P. G. (2009). Sex-related differences in the presentation, treatment and outcomes among patients with acute coronary syndromes: the Global Registry of Acute Coronary Events. *Heart*, *95*(1), 20 LP-26. https://doi.org/10.1136/hrt.2007.138537

Electrocardiography. (2019, May 10). Retrieved May 13, 2019, from https://en.wikipedia.org/wiki/Electrocardiography

FDA Center for Devices and Radiological Health. (n.d.). Fluoroscopy. Retrieved May 13, 2019, from https://www.fda.gov/radiation-emitting-products/medical-x-ray-imaging/fluoroscopy

Heron, M. (2018). Deaths: Leading Causes for 2016. *National Vital Statistics Reports*, 67(6).

Leppo, J. (1989). Dipyridamole-Thallium Imaging: The Lazy Man's Stress Test. *Journal of Nuclear Medicine*, 30(3): 281-287.

McKillop, J. (1981). Thallium 201 Scintigraphy. *Western Journal of Medicine,133*(1), 26-43. Retrieved April 29, 2019.

Miao, K., Miao, J., Miao, G. (2016). Diagnosing Coronary Heart Disease Using Ensemble Machine Learning. *International Journal of Advanced Computer Science and Applications*, 7(10): 30-39.

Ramachandran S. Vasan, M.D., Martin G. Larson, Sc.D., Eric P. Leip, M.S., Jane C. Evans, Ph.D., Christopher J. O'Donnell, M.D., M.P.H., William B. Kannel, M.D., M.P.H., and Daniel Levy, M.D., (2001) Impact of High-Normal Blood Pressure on the Risk of Cardiovascular Disease. *N Engl. J. Med.* 345:1291-1297\

Stanley, F. S., G., L. M., A., K. S., D., W. N., P., L. E., B., K. W., & Daniel, L. (2001). Does the Relation of Blood Pressure to Coronary Heart Disease Risk Change With Aging? . *Circulation*, *103*(9), 1245–1249. https://doi.org/10.1161/01.CIR.103.9.1245