Predicting Heart Disease With a Binary Logistic Regression Model

Jonathan King[1]
[1]UCLA Fielding School of Public Health, Dept. of Biostatistics

I. INTRODUCTION

Heart disease is the leading cause of death in the United States. From 2016 to 2017, heart disease costed the United States an estimated $219.6 billion, including $103.2 billion in direct costs (hospital services, home health care, prescribed medicine, and healthcare professional costs) and $116.4 billion in indirect costs (lost productivity due to premature mortality).[1] In 2020, heart disease caused over 690,000 deaths nationwide, which was nearly 5% increase compared to the year prior and was double the number of deaths attributed to COVID-19.[2]

There are many well-established modifiable and non-modifiable risk factors for heart disease. The American Heart Association specifically tracks seven modifiable risk factors in the United States, also known as “Life’s Simple 7,” to assess the heart health of the nation: smoking, physical inactivity, malnutrition, obesity, high cholesterol, diabetes, and high blood pressure.[1] Non-modifiable risk factors including age, sex, and ethnicity, with age being the biggest cause for concern due to the aging population.[3]

The aim of this study is to assess the associations of different risk factors with developing heart disease and to build a robust predictive model that determines whether an individual will develop heart disease based on the presence or absence of these risk factors. For this study, the specific risk factors that will be explored and considered for the model are age, sex, resting blood pressure, cholesterol level, maximum heart rate, presence of exercise-induced angina, and presence of fasting blood sugar level over 120 mg/dl.

II. METHODS & MATERIALS

A. Data Set

The dataset used for this analysis, “Heart Disease Dataset” was from Kaggle, an online community for data scientists that hosts free, downloadable datasets. The dataset dates compiles patient information from four hospital databases that are located in Cleveland, Hungry, Switzerland, and Long Beach . Although the original data contains seventy-six attributes, the published dataset only contains a subset of fourteen attributes. The data includes dummy-coded categorical variables such as age, sex, and presence of heart disease and numerical variables such as cholesterol level and resting blood pressure.

B. Study Population

The study population consists of patients at the four database-contributing hospitals in 1988. There are a total of 1,025 patient records analyzed for this study. Table 1 shows the mean and standard deviation for all the numerical variables of interest by presence of heart disease. Figure 1 offers stacked bar graphs by presence of heart disease for each categorical variable of interest to help visualize the distribution of these variables for the study population.

Table 1: Mean and Standard Deviation for Age (years), Resting Blood Pressure (mm Hg), Maximum Heart Rate (bpm), and Serum Cholesterol (mg/dl)

Statistic	No Heart Disease	Heart Disease
n_patients	499	526
m_age	56.57	52.41
sd_age	7.91	9.63
m_restbp	134.11	129.25
sd_restbp	18.58	16.11
m_max heart_rate	139.13	158.59
sd_max_heart_rate	22.57	19.10
m_cholesterol	251.29	240.98
sd_cholesterol	49.56	53.01

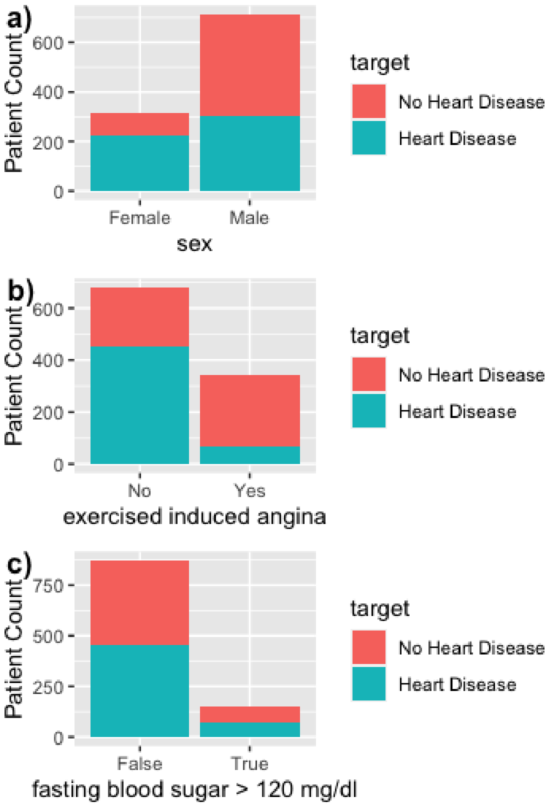

Figure 1: Stacked Bar Charts by Presence of Heart Disease for **a)** Sex, **b)** Exercised Induced Angina, and **c)** Fasting Blood Sugar >120 mg/dl

C. Statistical Methods

A binary logistic regression model was created from this dataset to predict the presence of heart disease. The dataset was initially split into two, with 80% (820 records) used for training the model and 20% (205 records) used for testing the model. While the original predictors included in the creation of the model were age, sex, resting blood pressure, cholesterol level, maximum heart rate, presence of exercised induced angina, and fasting blood sugar, fasting blood sugar was removed from the model to minimize Akaike Information Criterion (AIC) score, generating a model that best fits the data. Model coefficients were connected to the log odds of an individual having heart disease using Equation 1 to determine the expected change in log odds of heart disease for every one unit increase in the corresponding variable. After assumptions for a binary logistic regression were checked, predicted probabilities for the testing dataset were generated, and a histogram was generated to visualize the distribution of these predicted values.

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1(age) + \beta_2(resting\ bp) + \beta_3(max\ heart\ rate) + \beta_4(cholesterol) + B_5(sex\ male) + \beta_6(exercise - induced\ angina) \qquad (1)$$

The performance of the model was initially evaluated using a confusion matrix that was created using the threshold value of 0.5, meaning that the model predicts heart disease if the predicted probability is greater than 0.5. There are four values in the confusion matrix: true positives (TP), false positives (FP), true negatives (TP), and false negatives (FP). True positives and true negatives represent the actual outcome being the same as the predicted outcome for heart disease and no heart disease, respectively. False positives are when there was no heart disease but the model predicted heart disease, while false negatives are the opposite error. The metrics of accuracy, sensitivity, and specificity were calculated based on the confusion matrix using equations 2-4. To further evaluate the performance of the model, a Receiver Operating Characteristic (ROC) curve was generated to determine how well the model performed at all classification thresholds, which is reflected by the Area Under the Curve (AUC).

R Studio Version 1.4.1717 running R version 4.1.0 was used for all statistical methods conducted.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

$$sensitivity = \frac{TP}{TP+FN} \qquad (3)$$

$$specificity = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad (4)$$

III. RESULTS

Table 2: Coefficient Table for Binary Logistic Regression Model

	Estimate	Std. Error	*p*-value
(Intercept)	1.87	1.24	0.13
age	-0.03	0.01	1.5 x 10^{-3}
resting_bp	-0.02	0.01	5.6 x 10^{-4}
max_heart_rate	0.04	0.01	3.41 x 10^{-13}
cholesterol	6.3 x 10^{-3}	1.8x10^{-3}	3.2 x 10^{-4}
sexMale	-1.83	0.22	4.33 x 10^{-17}
exercise_anginaYes	-1.63	0.20	1.03 x 10^{-15}

As predicted by the model, male patient experienced a decrease in the log odds of having heart disease by 1.83 compared to females, and patients with exercise-induced angina experienced a decrease in the log odds of having heart disease by 1.63 compared to patients without exercise angina (Table 2). Both results are statistically significant with a p-value well less than 0.001.The other risk factors also did statistically significant influence the log odds of heart disease but to a lesser degree, with every year increase in age decreasing log odds of having heart disease by 0.03, each mm Hg increase in resting blood pressure decreasing the log odds of having heart disease by 0.02, every bpm increase in maximum heart rate increasing log odds of having heart disease by 0.04, and every mg/dl increase in cholesterol increasing the log odds of having heart disease by 6.3 x 10^{-3}.

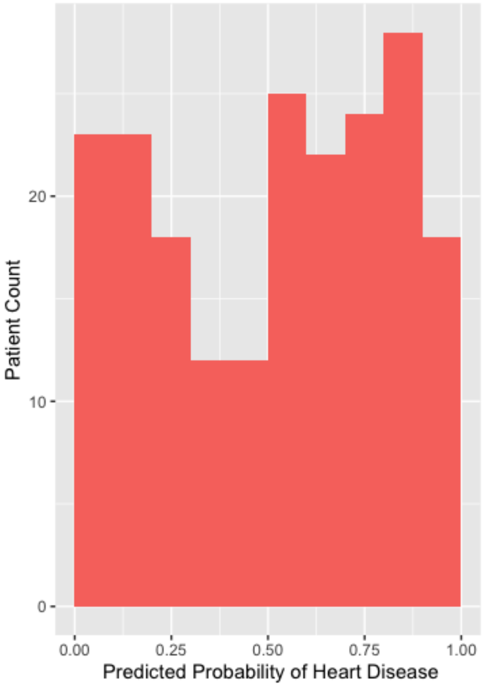

Figure 2: Histogram of Predicted Probabilities From Testing Dataset

Most predicted probabilities for the testing dataset are greater than 0.5 (Figure 2), indicating that the model predicted that the majority of patients had a greater than 50% chance of developing heart disease, but there is a noticeable peak from 0 to 0.20, indicating that the model predicted that approximately 50 patients had between a 0 to 20% chance of developing heart disease.

Table 3: Confusion Matrix (Decision Threshold = 0.5)

	Predicted No	Predicted Yes
No Heart Disease	71	35
Heart Disease	17	82

Accuracy: 0.746 Sensitivity: 0.828 Specificity: 0.670

At the 0.5 decision threshold, the model correctly predicted 82.8% of patients who had heart disease to have heart disease, while 67.0% of patients who did not have heart disease were correctly predicted to not have heart disease (Table 3). Overall, 74.6% of patients were correctly classified by the model.

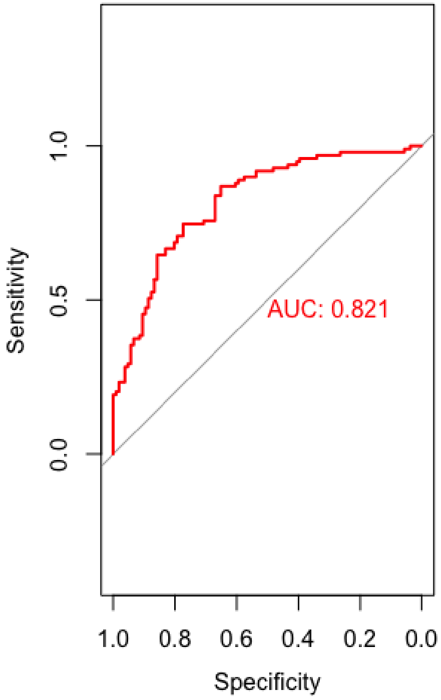

Figure 3: ROC Curve for Model Performance

The area under the ROC curve is 0.821, indicating that the model performed much better than a random classifier at all decision thresholds but did not perform so well that overfitting could be suspected (Figure 3).

IV. CONCLUSIONS

It can be concluded that based on the binary logistic regression model created to predict heart disease, sex and presence of exercised-induced angina were strong predictors for heart disease, while age, resting blood pressure, maximum heart rate, and cholesterol level were weak predictors. The second main conclusion that can be made is that the model was found to be a good classifier of heart disease presence.

V. DISCUSSION

According to the constructed binary logistic regression model, female sex and not having exercised-induced angina significantly increased the odds of developing heart disease. A similar study by Correia et al. on coronary artery disease (CAD), the most common heart disease, directly contradicts these results, finding that male gender was a positive predictor of CAD but angina and other chest pains were not independently associated with CAD.[4] These contradictions can be explained by several major limitations of this study.

One main limitation of this study is that the study population only represented patients from four hospitals scattered throughout the world, so the scope of inference is limited to that population. Figure 1a presents this problem visually by establishing that there was an unusually high proportion of female patients who had developed heart disease in this dataset compared to the general population. Another related limitation is that the dataset analyzed is from 1988, so the generated model is outdated. Therefore, a new model would have to be trained on current data from the databases to accurately predict heart disease for the current population. Ideally, this modern model would be regularly maintained to avoid having to build new models from scratch every few years. Finally, the dataset used for this study does not include information about other known risk factors for heart disease such as smoking history, weight, and ethnicity and was too old to account for new potential risk factors such as COVID-19 infection, so the created model does not account for all possible predictors.[5]

Overall, although the findings of this study cannot be extended to a general population, the same statistical methods used in this study can be applied to future studies to create a more representative model for predicting heart disease. It is recommended that a more representative dataset be generated by limiting the population of interest to a single country (e.g., the United States) and randomly sampling medical records from different hospitals throughout the country. This process would ensure that future logistic regression models created would be useful to public health professionals and policymakers when determining the optimal strategy for preventing heart disease.

REFERENCES

1. Virani, S. S., Alonso, A., Aparicio, H. J., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Carson, A.P., Chamberlain, A. M., Cheng, S., Delling, F. N., Elkind M. S.V., Evenson, K. R., Ferguson, J. F., Gupta, D. K., Khan, S. S., Kiseela,. B. M., Knutson, K. L., Lee, C. D., Lewis, T. T… Tsao, C. W., “Heart Disease and Stroke Statistics—2021 Update,” *Circulation*, *143*(8), pp. e254-743, 2021.
2. Ahmad F. B., & Anderson, R. N., “The Leading Causes of Death in the US for 2020,” *JAMA, 325*(19), pp. 1829-1830, 2021.
3. Rodgers, J. L., Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L.E., Shah, K., Karia, K., & Panguluri, S. K., “Cardiovascular Risks Associated with Gender and Aging,” *Journal of Cardiovascular Development and Disease*, *6*(2), p. 19, 2019.
4. Correia, L. C. L., Cerqueira, M., Carvalhal, M., Ferreira, F., Garcia, G., de Silva, A. B., de Sá, N., Lopes, F., Barcelos, A. C., & Noya-Rabelo, M., “A Multivariable Model for Prediction of Obstructive Coronary Disease in Patients with Acute Chest Pain: Development and Validation,” *Arquivos Brasileiros de Cardiologia*, *108*(4), pp. 304-314, 2017.
5. Chung, M.K., Zidar, D.A., Bristow, M.R., Cameron, S. J., Chan, T., Harding III, C.V., Kwon, D. H., Singh, T., Tilton, J.C., Tsai, E.J., Tucker, N.R., Barnard, J., & Loscaizo, J, “COVID-19 and Cardiovascular Disease,” *Circulation*, *128*(8), pp. 1214-1236, 2021.

Appendix

I. Source Data File

Dataset:
https://www.kaggle.com/johnsmith88/heart-disease-dataset

II. Modeling Data Set

Final Dataset was uploaded to CCLE under the name “final_dataset.csv.

III. R Code

R Code was uploaded to CCLE under the name “actual_final_project.R.”