# Prediction of Heart Diseases Based on Chest Pain Types

**Zhiyuan Yu**

## I. Introduction

Heart disease is defined as a type of disease that affects the heart or blood vessels. There are many types of heart diseases people are suffering nowadays such as coronary heart disease (CHD), heart failure, and atrial fibrillation.[1] According to CDC (Centers for Disease Control and Prevention) and WHO (World Health Organization), there were over 700000 people in the United States died because of heart diseases in 2022 [2] and ischaemic heart disease was the top cause of death globally in 2021.[3] There are many indicators such as obesities, hypertensions [4], and high cholesterol levels suggest a person may under the risk of having heart diseases, and other factors including heart rates and blood pressures may also have impacts on whether a person is likely to get heart diseases. Among all these indicators, chest pain, a symptom that can be easily detected by the patients, is routinely investigated as a sign of coronary artery disease [5] and it is the characteristic symptom of ischemic heart disease [6]. The purpose of the study was to generate prediction models on whether a person was likely to develop heart diseases based on the chest pain types using six factors – age, resting blood pressure, cholesterol level, resting electrocardiogram results, maximum heart rate achieved, and exercise-induced angina – and compare which model was most efficient in prediction. By doing the investigation, we might be able to determine effective ways to foresee the occurrence of heart diseases and provide insightful advice for people to reduce their risks of having heart problems.

## II. Methods and Materials

### *A. Dataset*

The dataset was obtained from the Kaggle dataset *Heart Failure Prediction Dataset* donated by David W. Aha from the University of California, Irvine. The data were retrospectively collected from four different places – Cleveland, Long Beach, Hungary, and Switzerland by Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. It included 1190 observations with 12 parameters – age, sex, chest pain types, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, oldpeak, the slope of the peak exercise ST segment, and whether the subjects had heart disease. We managed the dataset by excluding duplicated observations so that accurate and representative results could be created.

### *B. Study Population*

918 distinct individuals that were included in the study were separated into four groups based on the types of chest pain they had – Atypical Angina, Typical Angina, Non-Anginal Pain, and Asymptomatic – they suffer. As shown in Table 1a, age, resting blood pressure, cholesterol level, and maximum heart rate achieved were measured across the groups. The mean and median for these four parameters across the groups were relatively close while the standard deviations of cholesterol level in Asymptomatic and Non-Angina Pain groups were significantly larger than the other two groups. Table 1b outlined the resting electrocardiogram results and whether the subjects had exercise-induced Angina across the four groups. As shown in Figure 1, the frequency of the subjects having heart disease in Asymptomatic group was much higher than the those for other groups, indicating that certain types of chest pain might suggest a higher chance of developing heart diseases.

Table 1a. Baseline continuous characteristics of the study population. Note: “Age” = “Age”, “Cholesterol” = “Cholesterol Level”, “RestingBP” = “Resting Blood Pressure”, “MaxHR” = “Maximum Heart Rate Achieved”

| ChestPainType | Variable | N | Mean | Std Dev | Median | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| Asymptomatic | Age | 496 | 54.95968 | 8.763468 | 56.0 | 31 | 77 |
| Asymptomatic | Cholesterol | 496 | 186.64516 | 122.058634 | 220.5 | 0 | 603 |
| Asymptomatic | MaxHR | 496 | 128.47782 | 23.483317 | 128.0 | 60 | 186 |
| Asymptomatic | RestingBP | 496 | 133.22984 | 18.580961 | 130.0 | 80 | 200 |
| Atypical Angina | Age | 173 | 49.24277 | 9.259754 | 51.0 | 28 | 74 |
| Atypical Angina | Cholesterol | 173 | 233.04624 | 69.266406 | 237.0 | 0 | 468 |
| Atypical Angina | MaxHR | 173 | 150.20809 | 22.281042 | 152.0 | 93 | 202 |
| Atypical Angina | RestingBP | 173 | 130.62428 | 16.861711 | 130.0 | 98 | 192 |
| Non-Anginal Pain | Age | 203 | 53.31034 | 9.608023 | 53.0 | 33 | 76 |
| Non-Anginal Pain | Cholesterol | 203 | 197.43842 | 103.869348 | 218.0 | 0 | 564 |
| Non-Anginal Pain | MaxHR | 203 | 143.23645 | 25.608501 | 147.0 | 70 | 194 |
| Non-Anginal Pain | RestingBP | 203 | 130.96059 | 19.412878 | 130.0 | 0 | 200 |
| Typical Angina | Age | 46 | 54.82609 | 11.449026 | 59.0 | 30 | 74 |
| Typical Angina | Cholesterol | 46 | 207.06522 | 83.383292 | 229.0 | 0 | 308 |
| Typical Angina | MaxHR | 46 | 147.89130 | 23.126923 | 145.0 | 98 | 190 |
| Typical Angina | RestingBP | 46 | 136.41304 | 19.059644 | 140.0 | 95 | 178 |

Table 1b. Baseline discrete characteristics of the study population

Summary Table

RestingECG, ExerciseAngina by Chest Pain Types

| Resting Electrocardiogram Results | ExerciseAngina | Count_ASY | Count_ATA | Count_NAP | Count_TA | Col %_ASY | Col %_ATA | Col %_NAP | Col %_TA |
|---|---|---|---|---|---|---|---|---|---|
| LVH | No | 43 | 20 | 38 | 14 | 37.39 | 17.39 | 33.04 | 12.17 |
| LVH | Yes | 59 | 3 | 9 | 2 | 80.82 | 4.11 | 12.33 | 2.74 |
| Normal | No | 120 | 112 | 94 | 19 | 34.78 | 32.46 | 27.25 | 5.51 |
| Normal | Yes | 164 | 11 | 29 | 3 | 79.23 | 5.31 | 14.01 | 1.45 |
| ST | No | 36 | 24 | 20 | 7 | 41.38 | 27.59 | 22.99 | 8.05 |
| ST | Yes | 74 | 3 | 13 | 1 | 81.32 | 3.30 | 14.29 | 1.10 |

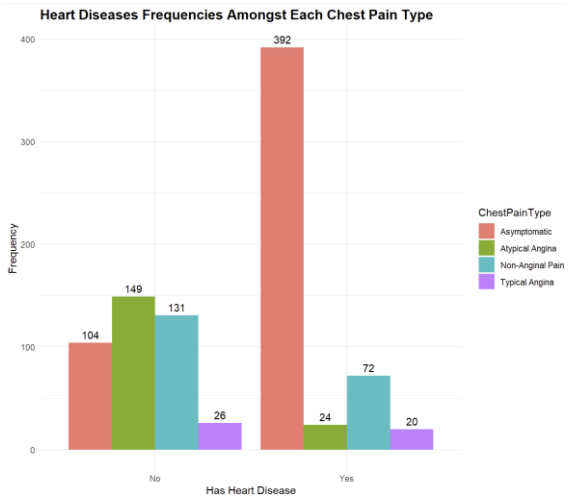


Figure 1. Bar Chart of Heart Disease Frequency

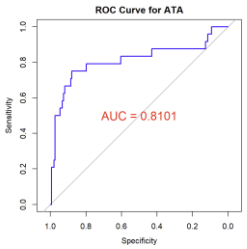


Figure 2a

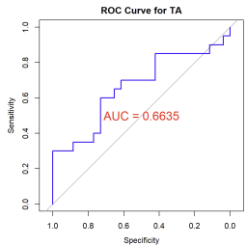


Figure 2b

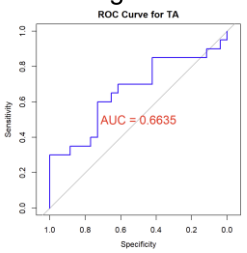


Figure 2c

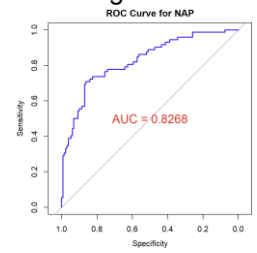


Figure 2d

### C. *Statistical Methods*

We implemented binary logistic regression model to conduct the multiple regression analysis by using the glm model in R to predict the occurrence of heart diseases utilizing both continuous measurements - age, resting blood pressure, cholesterol level, and maximum heart rate achieved - and discrete measurements - electrocardiogram results and whether the subjects have exercise-induced Angina. We optimized the model using Fisher's scoring algorithm. We applied the pROC package and roc.test() function to apply one-sided DeLong Test for comparing AUCs at the significance level of 0.05 [7] to see if one model predicted heart disease better than the others. The null hypothesis was that there was no difference in AUCs between the two models being compared, and the alternative hypothesis was that the AUC of the first model was greater than the second model. Six pairwise comparison were conducted: the AUC of Atypical Angina group versus the AUC of Typical Angina group; the AUC of Atypical Angina group versus the AUC of Non-Anginal Pain group; the AUC of Atypical Angina group versus the AUC of Asymptomatic group; the AUC of Typical Angina group versus the AUC of Non-Anginal Pain group; the AUC of Typical Angina group versus the AUC of Asymptomatic group; the AUC of Non-Anginal Pain group versus the AUC of Asymptomatic group. We utilized R version 4.4.2 for this study.

## III. Results

As shown in Figure 2a – d, we obtained the ROC curves and the areas under the curve (AUC) of 0.8101, 0.6635, 0.8268, and 0.7868 for the logistic model fitted to the Atypical Angina, Typical Angina, Non-Anginal Pain, and Asymptomatic group respectively.

We then calculated the correctness, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for a range of probability cutoffs for each of the chest pain type groups. We used the probability cutoff that maximized the sum of sensitivity and specificity, which were 0.192, 0.471, 0.426, and 0.750 for the Atypical Angina, Typical Angina, Non-Anginal Pain, and Asymptomatic group respectively.

As shown in Table 2a, for the cutoff probability of 0.192, the Atypical Angina logistic model predicted 18 cases of heart disease correctly, 6 cases of heart disease incorrectly, 131 non-heart disease correctly, and 18 cases of non-heart disease incorrectly. Thus, the correctness, sensitivity, specificity, PPV, and NPV were 86.127%, 50%, 95.620%, 75%, and 87.920% respectively.

Table 2a. Confusion matrix of fitted logistic model in Atypical Angina group.

| Table of Heart Disease Predicted | | | |
|---|---|---|---|
| Heart Disease | Predicted | | |
| | Yes | No | Total |
| Yes | 18 | 6 | 24 |
| No | 18 | 131 | 149 |
| Total | 36 | 137 | 173 |

As shown in Table 2b, for the cutoff probability of 0.471, the Typical Angina logistic model predicted 12 cases of heart disease correctly, 7 cases of heart disease incorrectly, 19 cases of non-heart disease correctly, and 8 cases of non-heart disease incorrectly. Thus, the correctness, sensitivity, specificity, PPV, and NPV were 67.391%, 63.158%, 70.370%, 60.000%, and 73.077% respectively.

Table 2b. Confusion matrix of fitted logistic model in Typical Angina group

| Table of Heart Disease Predicted | | | |
|---|---|---|---|
| Heart Disease | Predicted | | |
| | Yes | No | Total |
| Yes | 12 | 8 | 20 |
| No | 7 | 19 | 26 |
| Total | 19 | 27 | 46 |

As shown in Table 2c, for the cutoff probability of 0.426, the Non-Anginal Pain logistic model predicted 51 cases of heart disease correctly, 18 cases of heart disease incorrectly, 113 cases of non-heart disease correctly, and 21 cases of non-heart disease incorrectly. Thus, the correctness, sensitivity, specificity, PPV, and NPV were 80.788%, 73.913%, 61.081%, 70.833%, 86.260% respectively.

Table 2c. Confusion matrix of fitted logistic model in Non-Anginal Pain group

| Table of Heart Disease Predicted | | | |
|---|---|---|---|
| Heart Disease | Predicted | | |
| | Yes | No | Total |
| Yes | 51 | 21 | 72 |
| No | 18 | 113 | 131 |
| Total | 69 | 134 | 203 |

As shown in Table 2d, for the cutoff probability of 0.750, the Asymptomatic logistic model predicted 310 cases of heart disease correctly, 32 cases of heart disease incorrectly, 72 cases of non-heart disease correctly, and 82 cases of non-heart disease incorrectly. Thus, the correctness, sensitivity, specificity, PPV, and NPV were 77.016%, 90.643%, 46.753%, 79.082%, and 69.231% respectively.

Table 2d. Confusion matrix of fitted logistic model in Asymptomatic group

| Table of Heart Disease Predicted | | | |
|---|---|---|---|
| Heart Disease | Predicted | | |
| | Yes | No | Total |
| Yes | 310 | 82 | 392 |
| No | 32 | 72 | 104 |
| Total | 342 | 154 | 496 |

The p-values for each of the six hypothesis tests were 0.0851, 0.594, 0.365, 0.960, 0.913, and 0.160 respectively. Since all p-values were greater than 0.05, the differences between the AUC values of the comparing groups were not significant.

## IV. Conclusions

This study aimed to compare the four models based on chest pain types using six parameters including age, resting blood pressure, cholesterol level, resting electrocardiogram results, maximum heart rate achieved, and exercise-induced angina. Although clear differences could be observed in the correctness, sensitivities, specificities, PPVs, and NPVs among the groups, all the p-values of the six hypothesis tests on AUCs were greater than 0.05, suggesting that there was no statistical difference in the AUCs, which indicated that there was no significant evidence that the outcomes of the models were different. Therefore, we conclude that the effectiveness in predicting heart disease for these four models was not statistically different.

## V. Discussion

The Binary regression model was applied to predict heart disease based on the chest pain types the subjects had with six predictor variables. Four models based on chest pain types were generated for regression analysis and the correctness of prediction of the models were 86.127%, 67.391%, 80.788%, and 77.016% respectively. We also found that the difference in chest pain type did not make statistical differences on the predictions. There were other studies focusing on heart disease prediction applied similar methodologies. Kavya et al. [8] conducted a research focusing on predicting heart disease using logistic regression model. In addition to age and heart rates, the researchers included 12 other parameters into the prediction model such as obesity and had accuracy of 88.149% (662 correct predictions out of a total of 751 cases). Because of this, we believe that the logistic regression model could provide valuable insights into heart disease prediction based on chest pain types.

While the methodology was appropriate, the study had some limitations, and one of them was that the large differences in sample sizes among the four groups. While the Asymptomatic group had the largest sample size (496 subjects), there were only 46 observations within the Typical Angina group. The unevenness in sample sizes could cause negative impacts to the prediction model. For instance, it could lead to overfitting to the patterns of the Asymptomatic group which had the largest sample size and generalized poorly to the Typical Angina group which had the fewest subjects, making the result biased. Another limitation of the study was that although logistic regression model had the potential of generating reliable results in predicting heart disease, the prediction could be further optimized when combined with other methods. Zulkiflee & Rusiman [9] initialized a project on comparing the effectiveness of heart disease among three methods - binary logistic regression models, BLR models with Least Quartile Difference (LQD) method and BLR models with Median Absolute Deviation (MAD) method They found that the binary logistic regression with the applied MAD model had the highest

accuracy of 86.6% in prediction. This result showed that binary logistic regression models could perform better when combined with other methods, and utilizing the combination of binary regression models and other models to predict the development of heart diseases would be topics that are worth investigating in the future.

## VI. Acknowledgements

We would like to offer special thanks to our instructor Dr. Grace Kim and teaching assistant Joaquim Teixeira for their guidance throughout the quarter. Their instructions were insightful for this project.

We also want to thank Andras Janosi from Hungarian Institute of Cardiology, Budapest, William Steinbrunn from the University Hospital, Zurich, Switzerland, Matthias Pfisterer from University Hospital, Basel, Switzerland, and Robert Detrano from V.A. Medical Center, Long Beach and Cleveland Clinic Foundation for collecting the data and David W. Aha from the University of California, Irvine for contributing the dataset *Heart Failure Prediction Dataset*. We will never be able to start and finish the project without such valuable information.

## References

1. Kokubo, Yoshihiro, and Chisa Matsumoto. "Hypertension is a risk factor for several types of heart disease: review of prospective studies." *Hypertension: from basic research to clinical practice* (2017): 419-426.
2. "Leading Causes of Death." *Centers for Disease Control and Prevention*, National Center for Health Statistics, https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. Accessed 11 Dec. 2024.
3. World Health Organization. "The Top 10 Causes of Death." *World Health Organization*, www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed 12 Dec. 2024.
4. Kannel, William B. "Coronary heart disease risk factors in the elderly." *The American journal of geriatric cardiology* 11.2 (2002): 101-107.
5. Alkhatatbeh, Mohammad J., Noor A. Amara, and Khalid K. Abdul-Razzak. "Association of 25-hydroxyvitamin D with HDL-cholesterol and other cardiovascular risk biomarkers in subjects with non-cardiac chest pain." *Lipids in health and disease* 18 (2019): 1-10.
6. Cordero, Alberto, et al. "Low levels of high-density lipoproteins cholesterol are independently associated with acute coronary heart disease in patients hospitalized for chest pain." *Revista Española de Cardiología (English Edition)* 65.4 (2012): 319-325.
7. Concato, John, and John A. Hartigan. "P values: from suggestion to superstition." *Journal of Investigative Medicine* 64.7 (2016): 1166-1171.
8. Kavya, S. M., et al. "Heart Disease Prediction Using Logistic Regression." *Journal of Coastal Life Medicine* 11 (2023): 573-579.
9. Zulkiflee, Nor Fatihah, and Mohd Saifullah Rusiman. "Heart Disease Prediction Using Logistic Regression." *Enhanced Knowledge in Sciences and Technology* 1.2 (2021): 177-184.

## Appendix

### I. Source Data File

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

### II. Dataset

The dataset contained 1190 observations with 12 parameters. 272 of the observations were duplicated and being excluded. The managed dataset included 918 distinct observations.

### III. R Codes

Logistic regression models were applied to predict the heart disease. One-sided DeLong test for comparing AUCs at the 0.05 significance level was utilized to compare the effectiveness of each model in heart disease prediction.