PREDICTION OF HEART DISEASE USING A LOGISTIC REGRESSION MODEL BASED ON FRAMINGHAM HEART STUDY

Apoorv Saraogee
Northwestern University, Practical Machine Learning

May 22, 2022

Abstract

Heart disease is one of the leading causes of mortality among people across the world. Different demographic, behavioral and physiological factors are studied as indicators for coronary heart disease in the Framingham Heart Study (FHS). Predictive classification models can help in the early detection of heart disease as well as inform lifestyle prevention techniques at an affordable cost. This study analyzes 3658 observations of 16 variables from the FHS using a binary classification logistic regression model. Subset models for males and females are evaluated using ten-fold cross-validation with receiver operator characteristic area under the curve (AUC ROC) scores of 0.65 and 0.67 respectively.

1. Introduction

Coronary heart disease (CHD) is one of the leading causes of human mortality (Ali et al. 2021). Various demographic, behavioral and physiological variables had been explored with the Framingham Heart Study (FHS) in Framingham, Massachusetts. Since the FHS started in 1948 (Andersson et al. 2019), hypertension treatment, cholesterol reduction and smoking cessation have contributed to a 50-year decline in cardiovascular deaths (Memon and Khoja 2019). Early detection of CHD can further reduce this toll. Prediction of CHD using predictive analytics tools such as logistic regression offer an affordable pathway to early detection and prevention of disease (Ali et al. 2021).

This study's central research topic is to analyze a dataset from the FHS of 3658 individuals using a binary classification logistic regression model. Coronary incident after 10 years or TenYearCHD is the output variable with categorical variables (sex, education, currentSmoker, BPmeds, prevakentStroke, diabetes) and numerical variables (age, cigsPerDay, totChol, sysBP, diaBP, BMI, heartRate, glucose) as explanatory variables. Research questions include exploring interactions between variables, comparisons of explanatory variables and models between men and women. Other factors such as age and cigarette smoking are also further explored. The suitability of the model is estimated using the ROC AUC scores.

2. Literature Review

Predictive analytic models have been employed by others on the FHS dataset with TenYearCHD. The literature includes methods that use non-linear predictions and other complicated machine learning methods such as random forest, decision tree, neural networks, Naïve Bayes, and support vector (Obasi and Omair Shafiq 2019; Beunza et al. 2019; Ali et al. 2021; Pathan et al. 2022; Mangathayaru et al. 2020). These researchers are often comparing these methods with other study datasets as well (Mangathayaru et al. 2020; Pathan et al. 2022; Ali et al. 2021). This investigation also uses a logistic regression models and found an AUC score of ~0.65 as in (Mangathayaru et al. 2020). However, this investigation uses more simple logistic regression models for predictions on subsets males and females compared to the more complicated models used in the literature. Solely logistic regression methods are explored in the textbook and other literature (Memon and Khoja 2019; Christensen 1997; Rahman and Tabassum 2020).

3. Methods

This research will be conducted on the bodyfat data in Jupyter notebook using kernels for both R as well as Python3. The data will be visualized in R and then analyzed with various models using the sklearn package in Python3. Real FHS data is used to train the logistic regression model and each model is evaluated using their ROC AUC scores. Education, a multi-category variable, will be split into 4 columns as binary categorical variables. Keeping currentSmoker and cigsPerDay as separate variables instead of encoding them as a single binary factor based on a cutoff for the continuous variable retains more overall information for the model. The key objectives include evaluating and comparing a traditional main effects regression model (logistic regression with ten-fold cross validation), an interactions effects regression model (second degree), and subset models (male and female). The odds ratios will be calculated using a two-way contingency table for cigarette smoking and compared with the odds ratio calculated from the relevant logistic regression model coefficients. Other explanatory variables will also be compared using the model coefficients and calculated odds ratios. The probability of heart disease with age will be compared for males and females using a fixed set of explanatory variables using the average values across the dataset.

3. Results

The ten-fold cross-validated males and females datasets had ROC-AUC scores of 0.65 and 0.67 respectively. We see a slight reduction in the ROC-AUC scores with 0.63 and 0.64 respectively for the males and females datasets indicating a lack of interaction in the datasets. Odds ratios of 1.215 and 0.803 for males and females are calculated based on two-way contingency tables for smoking and heart disease in Table 1. The odds ratios are slightly different when calculated using the logistic regression model at 0.937 and 0.818 for males and females respectively (shown in Table 2). These odds ratios are more reliable than the contingency tables because they control for all of the other explanatory variables in the model. Most of the binary factors similarly have even odds, that is it is equally likely for each of the factors to contribute to an incident of heart disease. Hypertension was the most important binary factor with an odds ratio of 1.96 for the females dataset and 1.68 for the males dataset. When comparing the model coefficients that include all of the variables in Table 3, we see few differences between the males and females datasets. The higher education values have negative coefficients indicating lower risk for developing heart disease. Finally, we compare males and females datasets based on age by plotting the probability of a developing CHD with a fixed set of explanatory variables (mean) as shown in Figure 4. We see a steady increase in probability with age and a marked difference between males and females with females having a much overall probability of developing heart disease. From these results, I learned that health issues such as heart disease can be complex with many contributing variables. All of the chosen explanatory variables have similar model coefficient absolute values.

4. Conclusions

Within this study, the binary classification models using logistic regression had AUCs of 0.65 and 0.67 for the males and females datasets respectively with 17 explanatory variables. This is quite poor in obtaining accurate predictions for the CHD as you would like to be above 0.7 (Memon and Khoja 2019). Insignificant change in the AUCs after adding interaction variables suggests minimal interaction in the FHS dataset. Finding hypertension as the most important factor, increasing risk with age and the marked difference in probabilities between men and women is supported in the literature. However, more variables from the FHS dataset could be included to obtain a more accurate model of TenYearCHD. In addition, others have had success with more complex classification models such as random forest models (Obasi and Omair Shafiq 2019). Other binary classification models could be employed to obtain a better AUC score close to 1.

5. Appendices

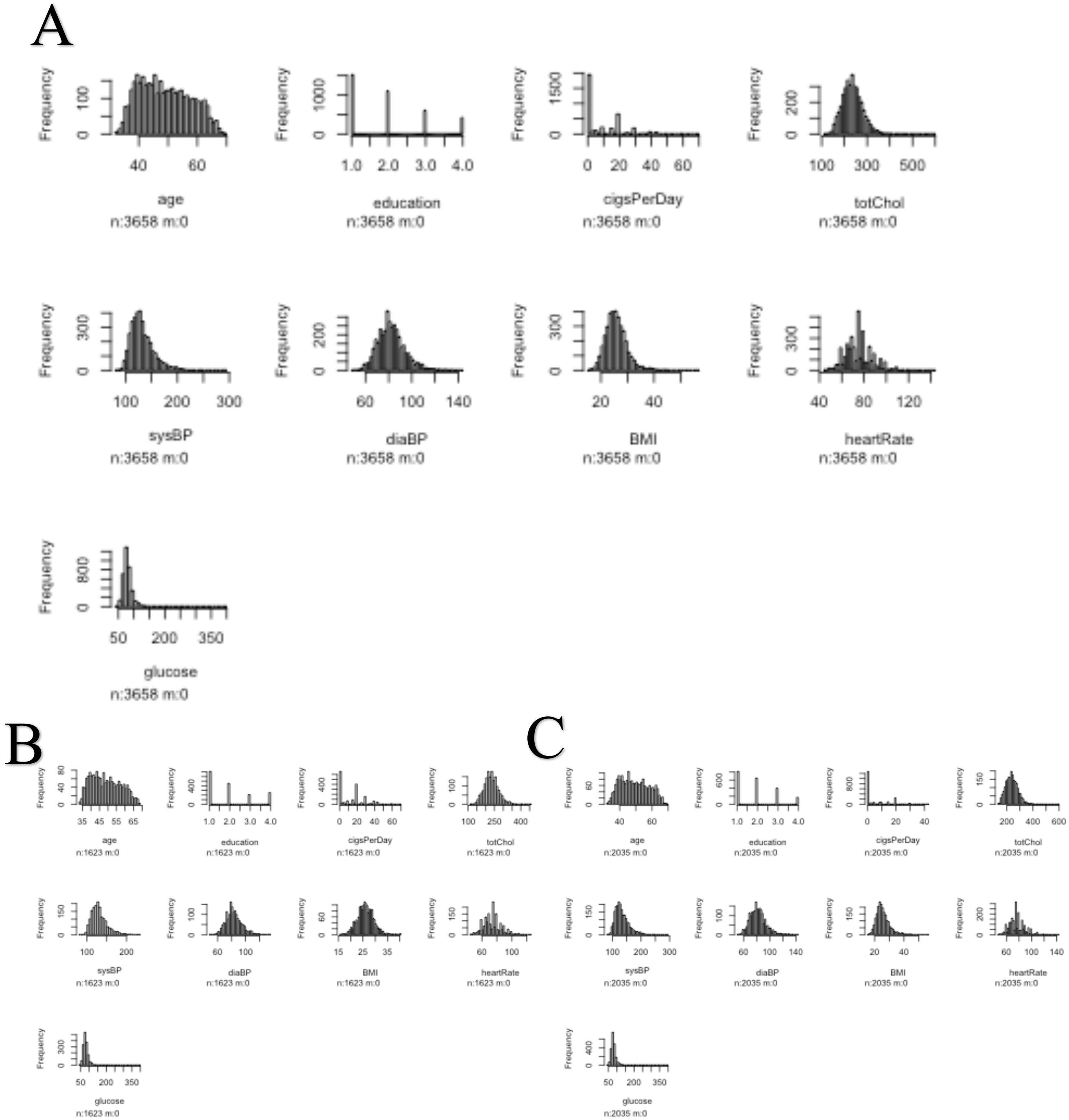


Figure 1. (A) Histograms of explanatory variables for full dataset (B) males only and (C) females only, using R

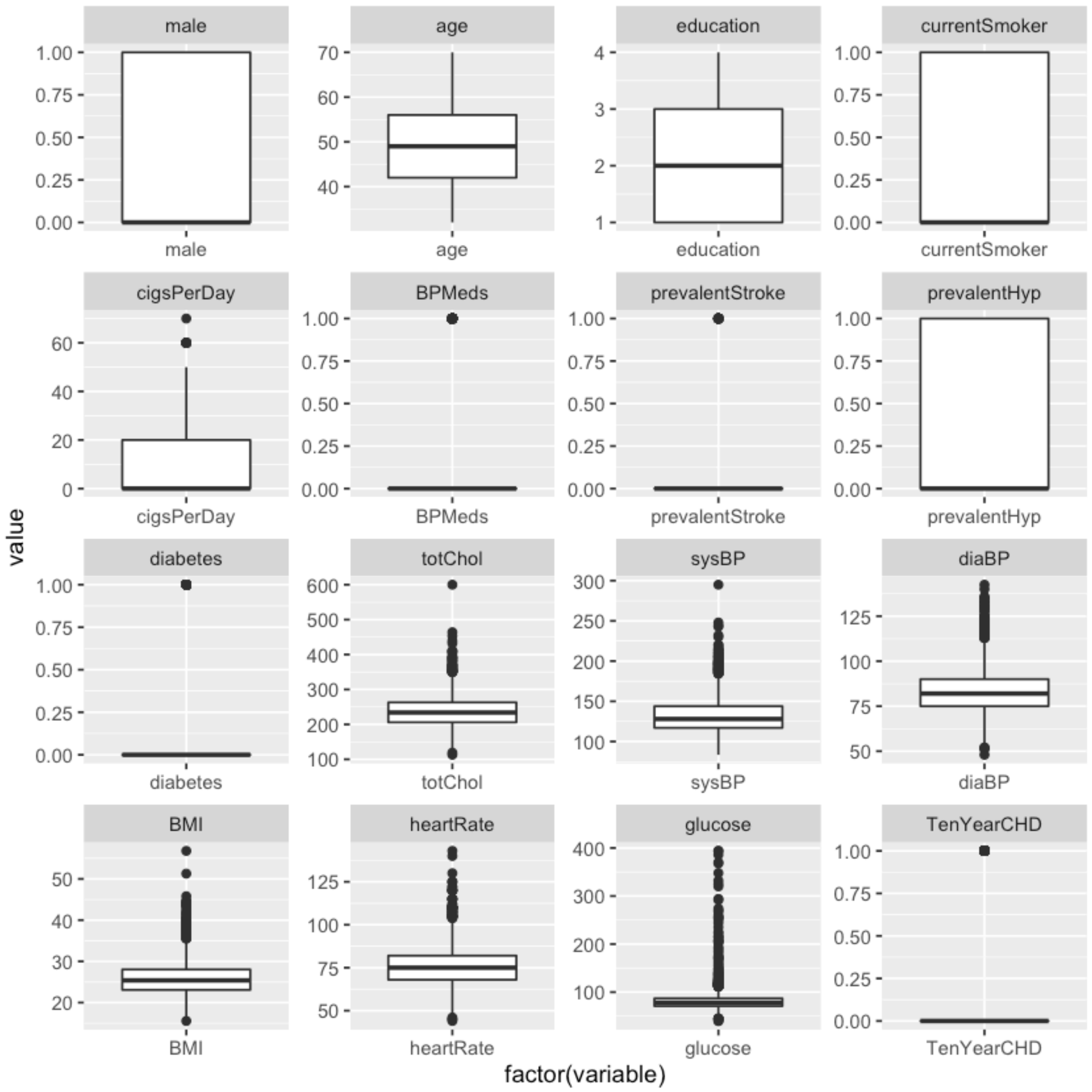


Figure 2. Box plots of explanatory variables of FHS data, using R

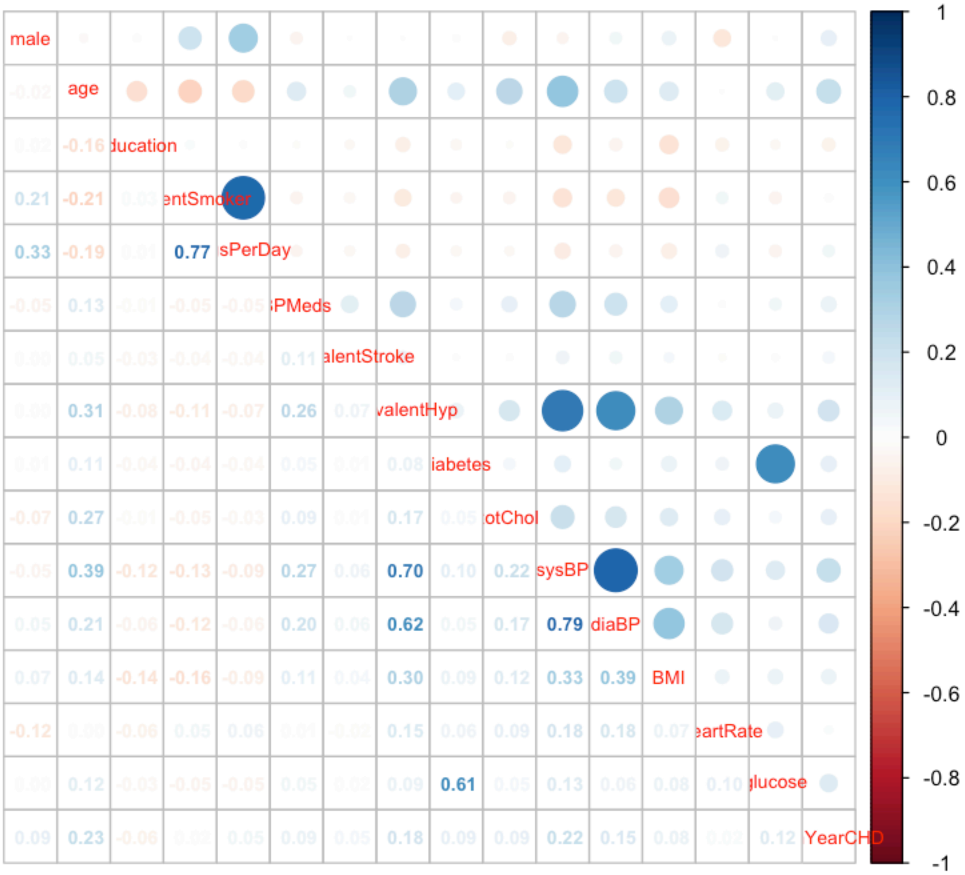


Figure 3. Correlation plot of explanatory variables for FHS data, using R

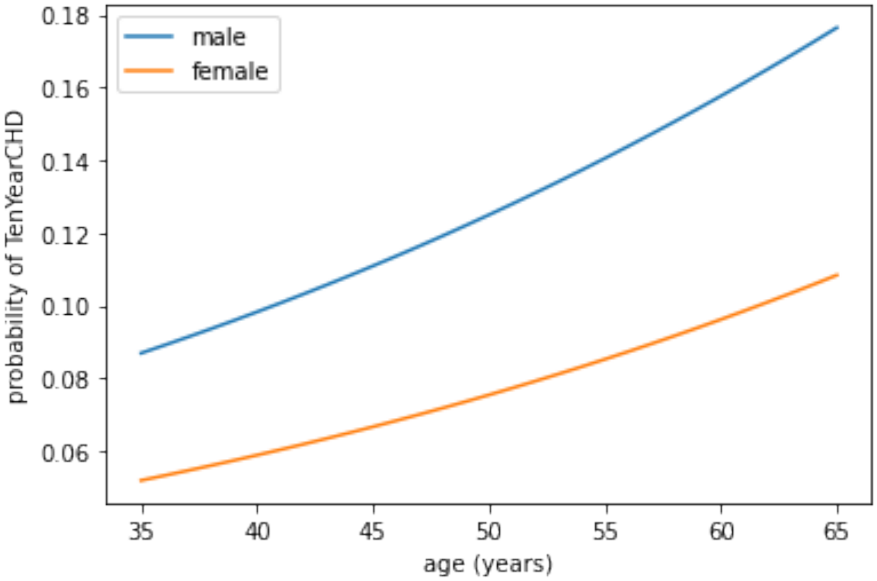
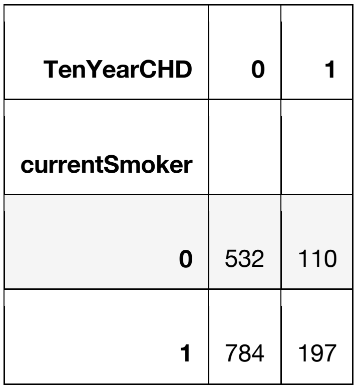


Figure 4. Probability of a coronary incident (CHD) calculated for male and females based on logistic regression model prediction coefficients for a fixed set of explanatory variables (mean) compared to age.

Table 1: Two-way contingency table for (A) males dataset and (B) females dataset

A



TenYearCHD	0	1
currentSmoker		
0	532	110
1	784	197

B

TenYearCHD	0	1
currentSmoker		
0	1065	162
1	720	88

Table 2: Odds ratio for different binary factors compared between males and females datasets calculated using a logistic regression model

	male	female
currentSmoker	0.936733	0.817650
BPMeds	1.091402	1.172306
prevalentStroke	1.055553	1.046149
prevalentHyp	1.681934	1.964095
diabetes	1.189329	1.102967
Education__1	1.025964	1.187471
Education__2	0.809069	0.817745
Education__3	0.852386	0.885750
Education__4	0.997989	0.836805

Table 3: Logistic regression model coefficients for males and females datasets

	male	female
age	0.027047	0.026590
currentSmoker	-0.065356	-0.201320
cigsPerDay	0.012087	0.021436
BPMeds	0.087463	0.158973
prevalentStroke	0.054065	0.045116
prevalentHyp	0.519944	0.675031
diabetes	0.173390	0.098004
totChol	-0.000103	-0.002279
sysBP	0.018997	0.018673
diaBP	-0.023257	-0.032733
BMI	-0.088926	-0.024786
heartRate	-0.017336	-0.029418
glucose	0.004273	0.004121
Education__1	0.025632	0.171826
Education__2	-0.211871	-0.201205
Education__3	-0.159716	-0.121321
Education__4	-0.002013	-0.178164

Supporting Files

- Assignment2.html
- Assignment2.py
- Assignment2.ipynb
- framingham.csv

References

Ali, Mamun, Bikash Kumar Paul, Kawsar Ahmed, Francis M Bui, Julian M W Quinn, and Mohammad Ali Moni. 2021. “Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison.” *Computers in Biology and Medicine* 136: 10–4825. https://doi.org/10.1016/j.compbiomed.2021.104672.

Andersson, Charlotte, Andrew D. Johnson, Emelia J. Benjamin, Daniel Levy, and Ramachandran S. Vasan. 2019. “70-Year Legacy of the Framingham Heart Study.” *Nature Reviews Cardiology 2019 16:11* 16 (11): 687–98. https://doi.org/10.1038/s41569-019-0202-5.

Beunza, Juan-Jose, Enrique Puertas, Ester García-Ovejero, Gema Villalba, Emilia Condes, Gergana Koleva, Cristian Hurtado, and Manuel F Landecho. 2019. “Comparison of Machine Learning Algorithms for Clinical Event Prediction (Risk of Coronary Heart Disease).” https://doi.org/10.1016/j.jbi.2019.103257.

Christensen, Ronald. 1997. *Log-Linear Models and Logistic Regression*. 2nd ed. Springer.

Mangathayaru, Nimmala, B. Padmaja Rani, V. Janaki, Sai Mohan Gajapaka, Shilhora Akshay Patel, and B. Lalith Bharadwaj. 2020. “An Imperative Diagnostic Model for Predicting CHD Using Deep Learning.” *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*, November. https://doi.org/10.1109/INOCON50539.2020.9298423.

Memon, Qurban A. (Qurban Ali), and Shakeel Ahmed Khoja. 2019. *Data Science : Theory, Analysis, and Applications*. https://www.routledge.com/Data-Science-Theory-Analysis-and-Applications/Memon-Khoja/p/book/9781032240244.

Obasi, Thankgod, and M. Omair Shafiq. 2019. “Towards Comparing and Using Machine Learning Techniques for Detecting and Predicting Heart Attack and Diseases.” *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, December, 2393–2402. https://doi.org/10.1109/BIGDATA47090.2019.9005488.

Pathan, Muhammad Salman, Avishek Nag, Muhammad Mohisn Pathan, and Soumyabrata Dev. 2022. “Analyzing the Impact of Feature Selection on the Accuracy of Heart Disease Prediction.” *Healthcare Analytics*, May, 100060. https://doi.org/10.1016/J.HEALTH.2022.100060.

Rahman, Azizur, and Arifa Tabassum. 2020. “Model to Assess the Factors of 10-Year Future Risk of Coronary Heart Disease among People of Framingham, Massachusetts.” *International Journal of Public Health Science (IJPHS)* 9 (3): 259–66. https://doi.org/10.11591/ijphs.v9i3.20469.