Predicting Coronary Heart Disease (CHD) Risk through Logistic Regression Modeling

Ebo Essilfie-Amoah, Siqiao Chen, Zhengrui Huang, Mengxue Zhang

2023-12-13

Introduction

The heart is an indispensable muscular organ ensuring the bodies' optimal functioning [1]. However, the global prevalence of cardiovascular diseases (CVDs) remains a formidable challenge, contributing to nearly one-third of all recorded deaths worldwide. Among the spectrum of CVDs, Coronary Heart Disease (CHD) stands as a predominant threat [2]. A pivotal milestone in cardiovascular research, the Framingham Heart Study (FHS) [3], was initiated in 1948 to identify risk factors associated with coronary heart disease. One noteworthy subset of the Framingham study, accessible on Kaggle, features 4,434 participants meticulously examined at three intervals spanning from 1956 to 1968, with a subsequent 24-year follow-up period. In this report, we construct a logistic regression model to delve into predicting the 10-year risk of coronary heart disease, based on factors such as demographic information, lifestyle and behavioral habits, health conditions, and biomedical indicators.

Methodology

Study population

For the purpose of this study, all data and covariates were taken from the Framingham Heart Study conducted from 1956 to 1968. When exploring the data with our covariates of interest, we decide to work with only the complete cases since some variables have a significant amount of missing data. In such circumstances, imputation could lead to misestimation of the true variable distributions and relationships within the data, which will betray our priority of maintaining the accuracy and integrity of the raw data for prediction.

Predictor(s) of Interest

Our predictors of interest were selected based on a comprehensive search of academic databases using terms like "Coronary heart disease," "Risk factors," and "Cardiovascular health" to find relevant studies. In numerous studies on age and CHD, it consistently highlights advancing age as a significant risk factor for CHD, with the risk increasing substantially after the age of 45–50 years [4]. There is evidence that suggests high blood pressure has been established as a major risk factor for CHD development [5]. Studies by Gallucci et al. (2020) [6] confirms the strong association between smoking and CHD incidence, with smokers having significantly higher CHD risks. Individuals with diabetes have a substantially elevated risk of CHD, as evidenced by studies by Wilson, P. W. (1998) [7].

Outcome

In our study, to evaluate the risk of coronary heart disease, the ten-year risk of coronary heart disease (TenYearCHD) was used as the outcome variable, which is a binary variable. This variable is coded as '0' for individuals who are not expected to develop coronary heart disease within the next ten years and '1' for individuals who are at risk of developing the condition within the same period. This dichotomization is based on a combination of risk factors, including the covariates that we will introduce in the next section.

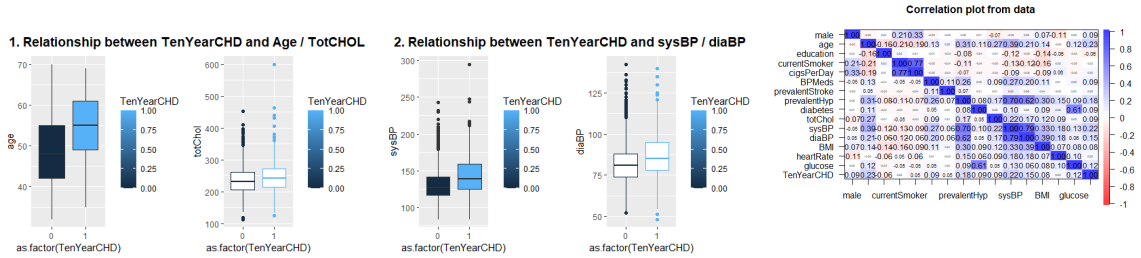
Covariates

In our research into the predictors of the ten-year risk of coronary heart disease (TenYearCHD), we have applied various covariates. Among these covariates, demographic variables, lifestyle factors, clinical measurements and personal health history are included, which are continuous variables, binary variables or categorical variables. Each of these variables has been associated with CHD risk and will be analyzed to understand their influence on the likelihood of developing CHD over a ten-year period.

Exploratory Data Analysis

Based on the complete cases, we performed the exploratory analysis for the data. First, the descriptive statistics for continuous variables such as age, smoking, cholesterol, and blood pressure is calculated. It illustrates mean values and standard deviations, indicating an average age of 49.56 years and 9.02 cigarettes smoked per day, as the example. Further, the descriptive statistics for binary and categorical variables for the same group is obtained, with percentages for gender, education, smoking status, medication use, and health conditions like hypertension and diabetes. For example, 44.37% are male and 55.63% are female, which indicates that the distribution between males and females is relatively balanced. In addition, 15.24% are at risk of coronary heart disease over ten years and 84.67% are not at risk, so the distribution suggests that a greater proportion of the sample is not at risk.

Risk Factor Analysis



The box plots show that individuals who develop CHD in the next 10 years have higher mean ages, cholesterol levels, and mean systolic and diastolic blood pressures. We also analyzed the relationship between other risk factors and CHD. However, only whether current smoker and heart rate were found to be insignificantly correlated with CHD. Correlation analysis is performed to study the strength and direction of the relationships between response variable and potential covariates. It is also a crucial part in feature selection for predictive models, since highly correlated variables might provide redundant information and can be avoided to simplify models and improve their interpretability. From our analysis, there is a strong correlation at 0.77 between cigsPerDay and currentSmoker, and a strong correlation at 0.79 between diaBP and sysBP. As a result, in model building, we will only keep some of those variables.

Logistic regression

Utilizing data from the Framingham Heart Study, logistic regression was employed to identify key risk factors associated with the development of Coronary Heart Disease (CHD). The full model contains all covariates in the dataset.

Full Model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \times \text{male} + \beta_2 \times \text{age} + \beta_3 \times \text{education} + \beta_4 \times \text{currentSmoker} + \beta_5 \times \text{cigsPerDay} + \beta_6 \times \text{BPMeds} + \beta_7 \times \text{prevalentStroke} + \beta_8 \times \text{prevalentHyp} + \beta_9 \times \text{diabetes} + \beta_{10} \times \text{totChol} + \beta_{11} \times \text{sysBP} + \beta_{12} \times \text{diaBP} + \beta_{13} \times \text{BMI} + \beta_{14} \times \text{heartRate} + \beta_{15} \times \text{glucose} + \epsilon_i$$

Potential interaction terms were selected based on literature review and were incorporated in our regression model. The coefficients and their significance were examined in new models. However, none of those coefficients were significant according to p-values. We also compared the models with and without interaction terms using anova( ). The results did not indicate a better fit when interaction terms were added. So, we may conclude that interaction terms should not be added to the model.

Final Model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \times \text{male} + \beta_2 \times \text{age} + \beta_3 \times \text{cigsPerDay} + \beta_4 \times \text{totChol} + \beta_5 \times \text{sysBP} + \beta_6 \times \text{glucose} + \epsilon_i$$

The final model was selected from the models without interaction terms using a combination of forward selection and backward elimination. We introduced the approaches taking p-value as a criteria. In the final model there were only six covariates. Covariates might be significantly associated with the response variable when examined individually, but might not be significant in the full model due to collinearity, interaction, sample size and so on.

Prediction

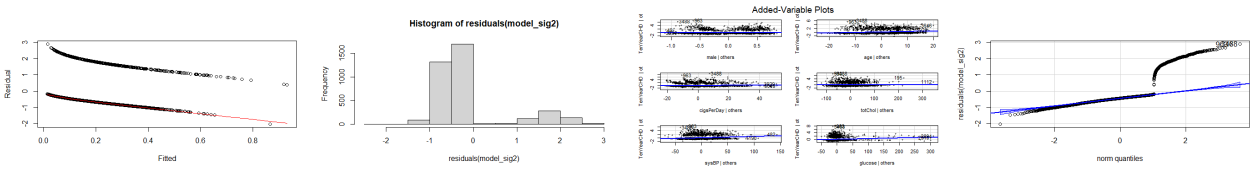
To explore the gender differences in the predictions, we separated the data of males and females. In both of these datasets, we further divided the data into 70% training set and 30% test set, respectively. For the training set in both male and female data, we conduct model selection once more, commencing with the previously established final model and proceeding with backward elimination.
Male Model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{cigsPerDay} + \beta_3 \times \text{totChol} + \beta_4 \times \text{sysBP} + \beta_5 \times \text{glucose} + \epsilon_i$$

Female Model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{cigsPerDay} + \beta_3 \times \text{prevalentHyp} + \beta_4 \times \text{glucose} + \epsilon_i$$

Model Diagnosis



Diagnostic plots depicted above indicate the inherent nature of the residuals following the logistic distribution.

The Residual vs. Fitted plot shows two distinct groups, while the histogram of residuals plot shows that the residuals are clustered around 0 and 1. Due to the nature of binary outcome, we consider this as a normal situation to encounter when doing diagnostics plots for logistic regression. The partial regression plot shows patterns of random scatter for the independent variables, indicating that the relationship between the independent variables and the log-odds of the binary outcome is approximately linear. There is an abrupt jump in the Q-Q plot, suggesting that there might be outliers that cause this deviation from the expected distribution.

VIF was performed to check the collinearity, while we found no VIF values greater than 5. AIC was calculated to assess better fitting among models. Lower AIC was obtained for the final model, compared to the full model, indicating the final model being a better fit.

Model Performance: Accuracy in Male and Female Predictions

This confusion matrix shows the model's prediction against the actual outcomes for male and female data in a binary classification scenario.

Male	Predicted Negative (PN)	Predicted Positive (PP)	Female	Predicted Negative (PN)	Predicted Positive (PP)
Negative (N)	382 (TN)	8 (FP)	Negative (N)	521 (TN)	22 (FP)
Positive (P)	88 (FN)	9 (TP)	Positive (P)	52 (FN)	16 (TP)
Accuracy of Male Model = (TN + TP) / (TP + TN + FP + FN) = 80%			Accuracy of Female Model = (TN + TP) / (TP + TN + FP + FN) = 88%		

Result

In general, 6 covariates are predictive factors of CHD. The logistic regression model yielded insightful results, highlighting significant predictors contributing to CHD risk. Age (p $<<$ 2e-16): Consistently emerged as a strong predictor, indicating a direct association between advancing age and increased CHD risk. Cigarettes per Day (p = 5.7e-12): Showed a statistically significant positive association with CHD occurrence. Systolic Blood Pressure (p = 1.2e-14): Systolic Blood pressure exhibited a moderate yet significant impact on CHD risk. Glucose (p = 0.0018): Glucose demonstrated higher odds of developing CHD compared to non-smokers. For both male and female, age, cigsPerDay and glucose are predictive of CHD. For male, age, cigsPerDay, totChol, sysBP and glucose are predictive of CHD. For female, age, cisgPerDay, prevalentHyp, glucose are predictive of CHD.

Both_Gender	Male	Female
age cigsPerDay glucose	age cigsPerDay totChol sysBP glucose	age cigsPerDay prevalentHyp glucose

Shiny app

To visualize our prediction results, a Shiny app was built. The users can view and interact with our Shiny app via the link https://10yearchdprediction.shinyapps.io/625rshinyapp/. After entering personal information in the dashboard, the prediction of 10 year CHD Risk would appear in the instrument panels. The default interface is shown in figures above.

Conclusion

The final attributes chosen following the elimination process exhibit p-values below 5%, which indicates their potential significance in predicting Heart disease. For Men, an Increase in age, number of cigarettes smoked per day and systolic blood pressure, total cholesterol and glucose show increasing odds of having heart disease in the next 10 years. For women, an increase in age, number of cigarettes smoked per day, prevalence of hypertension and glucose show increasing odds of CHD. The model achieved an accuracy of 88% in its predictions for women and 80% accuracy for men. It leans toward being more specific than sensitive. To enhance the overall model performance, additional data could be beneficial. These findings underscore the importance of targeted interventions and lifestyle modifications to mitigate CHD risk.

Discussion

This study pinpointed key CHD risk factors like age and smoking, using historical data from the Framingham Heart Study. However, the reliance on older data limits its contemporary relevance. Future studies should utilize more current data and broaden their scope to include diverse populations, which would enhance the understanding and applicability of the findings. A longitudinal study design would be particularly valuable, offering insights into how these risk factors evolve over time and their long-term impact on CHD. Additionally, examining lifestyle factors in greater detail would provide a more comprehensive view of CHD prevention and management.

Reference

[1] Buijtendijk, M.F.; Barnett, P.; van den Hoff, M.J. Development of the human heart. Am. J. Med. Genet. Part C Semin. Med. Genet. 2020, 184, 7–22.
[2] Wong ND. Epidemiological studies of CHD and the evolution of preventive cardiology. Nat Rev Cardiol. 2014 May;11(5):276-89. doi: 10.1038/nrcardio.2014.26. Epub 2014 Mar 25. PMID: 24663092.
[3] Framingham Heart Study. (2021). Longitudinal Data Documentation. Retrieved from https://biolincc.nhlbi.nih.gov/media/teachingstudies/FHS_Teaching_Longitudinal_Data_Documentation_2021a.pdf?link_time=2023-12-01_16:41:38.187197
[4] Rodgers JL, Jones J, Bolleddu SI, Vanthenapalli S, Rodgers LE, Shah K, Karia K, Panguluri SK. Cardiovascular Risks Associated with Gender and Aging. J Cardiovasc Dev Dis. 2019 Apr 27;6(2):19. doi: 10.3390/jcdd6020019. PMID: 31035613; PMCID: PMC6616540.
[5] Fuchs, F. D., & Whelton, P. K. (2019). High Blood Pressure and Cardiovascular Disease. Hypertension, 75(2), 285–292. https://doi.org/10.1161/HYPERTENSIONAHA.119.14240
[6] Gallucci G, Tartarone A, Lerose R, Lalinga AV, Capobianco AM. Cardiovascular risk of smoking and benefits of smoking cessation. J Thorac Dis. 2020 Jul;12(7):3866-3876. doi: 10.21037/jtd.2020.02.47. PMID: 32802468; PMCID: PMC7399440.
[7] Wilson, P. W. (1998). Diabetes mellitus and coronary heart disease. American Journal of Kidney Diseases, 32(5), 89–100. https://doi.org/10.1053/ajkd.1998.v32.pm9820468

Contribution

Zhengrui Huang: Data analysis and shiny app building, writing report.Rmd; Siqiao Chen: Writing the report; Ebo Essilfie-Amoah: Model developing. Mengxue Zhang: Model developing and model diagnostics.