**Exploring the risk factors of the 10-year risk of future coronary heart disease by conducting logistic regression modelling.**

**Abstract:**

Coronary heart disease is a specific type of cardiovascular disease that primarily affects the coronary arteries supplying blood to the heart muscle. Cardiovascular diseases are the most common cause of mortality in developed countries [1, 2]. Across the globe, the incidence of death from cardiovascular and circulatory diseases has risen by one third between 1990 and 2010, such that by 2015 one in three deaths worldwide will be due to cardiovascular diseases [3]. Epidemiologic studies played a crucial role in elucidating factors that predispose to cardiovascular disease and highlighting opportunities for prevention. One such historical epidemiological study is Framingham Heart Study (FHS) which is a long term, ongoing cardiovascular cohort study of residents of the city of Framingham, Massachusetts [4].

The objective of the Framingham Heart Study was to identify the common factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed symptoms of CVD or suffered a heart attack or stroke [5]. The Framingham Heart Study participants, and their children and grandchildren, voluntarily consented to undergo a detailed medical history, physical examination, and medical tests every three to five years, creating a wealth of data about physical and mental health, especially about cardiovascular disease.948 with 5,209 adult subjects from Framingham and is now on its third generation of participants [6].

The main purpose of this study is to explore and investigate the risk factors associated with the ten-year risk of future coronary heart disease (TenYearCHD) using data from the Framingham Heart Study. Logistic regression modelling is utilized to examine the various risk factors and their correlation with Ten-year CHD. Descriptive statistics were used to summarize the dataset, followed by fitting simple logistic regression models for individual risk factors. Later, a multiple logistic regression model was constructed using all predictors, and variable selection methods were applied to identify the optimal model. The results show the significant associations between certain risk factors and TenYearCHD, providing insights into the prediction of coronary heart disease risk.

Background:

This study takes data set from Framingham heart study as the background, as it is well known for longitudinal cohortstudy that has been foing from 1940s to till date. This data set is used to explore the different risk factors that are leading to coronary heart disease in the next 10 years.

**Study Design**

**Aim:**

The Study aims to analyze and identify correlated with Ten Year CHD by exploring data set “HeartDisease.csv” from Framingham heart study and using logistic regression modelling, prediction of the outcome which is the 10-year risk of future coronary heart disease.

**Data:**

The data used for this study is from the Framingham Heart Study, which is a cohort study that’s been conducting since 1940s for over 3 generations. Data is received in the form of csv file and data was analysed using RStudio.

**Materials:**

Materials used for this study are data obtained from Framingham heart study and RStudio software.

**Statistical Analysis**

**Methods:**

**Descriptive Statistics:**

Descriptive Statistics are used to summarize the dataset for better understanding about various risk factors in the data. Logistic regression models are used to assess the association between individual risk factors and 10-year CHD. Later, multiple logistic regression model is constructed using all predictors. Variable selection methods are applied to identify the optimal model.

Results:

Using Descriptive Statistics, the data set collected was summarized. A total of 3,656 individuals are studied in the FHS, whereas out of 3, 656 individuals 557 are at risk of CHD within next 10 years. Below table, no means number of individuals that do not get affected from CHD, and yes means number of individuals that will get affected by CHD in future ten years.

|  |  |
| --- | --- |
| no | yes |
| 3099 | 557 |

Based on the prevalentStroke risk factor, a total of 21 individuals are at risk of CHD.

|  |  |
| --- | --- |
| no | yes |
| 3635 | 21 |

Summary of the dataset is described elaborately in the below table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| Male | 0 | 0 | 0 | 0.4437 | 1 | 1 |
| Age | 32 | 42 | 49 | 49.56 | 56 | 70 |
| Current Smoker | 0 | 0 | 0 | 0.4891 | 1 | 1 |
| BPMeds | 0 | 0 | 0 | 0.03036 | 0 | 1 |
| Prevalent Stroke | 0 | 0 | 0 | 0.005744 | 0 | 1 |
| Prevalent Hyp | 0 | 0 | 0 | 0.3115 | 1 | 1 |
| Diabetes | 0 | 0 | 0 | 0.02708 | 0 | 1 |
| Total Cholesterol | 113 | 206 | 234 | 236.9 | 263.2 | 600 |
| Systolic BP | 83.5 | 117 | 128 | 132.4 | 144 | 295 |
| Diastolic BP | 48 | 75 | 82 | 82.91 | 90 | 142.5 |
| BMI | 15.54 | 23.08 | 25.38 | 25.78 | 28.04 | 56.8 |
| Heart Rate | 44 | 68 | 75 | 75.73 | 82 | 143 |
| Glucose | 40 | 71 | 78 | 81.86 | 87 | 394 |
| TenYearCHD | 0 | 0 | 0 | 0.1524 | 0 | 1 |

**Simple Logistic Regression**:

Later, using a simple logistic regression model, the outcome is predicted using one risk factor as variable at a time. For Variable male, analysis showed a statistically significant association with CHD and Odds ratio 1.67, (1. 39 to 2.00 of 95% confidence interval), followed by for age variable, statistically significant with Odds Ratio: 1. 08, (1. 07 – 1.09 of 95% CI) Furthermore, the utilization of blood pressure medications (BPMeds), a history of stroke, prevalent hypertension, and diabetes status exhibited substantial associations with an increased likelihood of developing CHD, with odds ratios ranging from 2.73 to 3.46. Similarly, total cholesterol, systolic and diastolic blood pressure, as well as BMI, demonstrated statistically significant correlations with CHD risk, with odds ratios ranging from 1.01 to 1.05. Conversely, no notable correlation was detected for current smoking status or heart rate. Notably, glucose levels were also found to have a significant association, with an odds ratio of 1.01 (95% CI: 1.01-1.01). These findings underscore the necessity of addressing multiple risk factors when evaluating and managing the ten-year risk of CHD, thereby influencing preventive measures and clinical interventions. Below is the description of results obtained by running simple logistic regression model.

1. Male: Shows a statistically significant association with CHD (Odds Ratio: 1.67, 95% CI: 1.39-2.00).

2. Age: Statistically significant (Odds Ratio: 1.08, 95% CI: 1.07-1.09).

3. Current Smoker: Not statistically significant.

4. BPMeds: Statistically significant (Odds Ratio: 2.91, 95% CI: 1.92-4.33).

5. Prevalent Stroke: Statistically significant (Odds Ratio: 3.46, 95% CI: 1.36-8.25).

6. Prevalent Hypertension: Statistically significant (Odds Ratio: 2.73, 95% CI: 2.27-3.28).

7. Diabetes: Statistically significant (Odds Ratio: 3.18, 95% CI: 2.06-4.82).

8. Total Cholesterol: Statistically significant (Odds Ratio: 1.01, 95% CI: 1.00-1.01).

9. Systolic BP: Statistically significant (Odds Ratio: 1.02, 95% CI: 1.02-1.03).

10. Diastolic BP: Statistically significant (Odds Ratio: 1.03, 95% CI: 1.03-1.04).

11. BMI: Statistically significant (Odds Ratio: 1.05, 95% CI: 1.03-1.08).

12. Heart Rate: Not statistically significant.

13. Glucose: Statistically significant (Odds Ratio: 1.01, 95% CI: 1.01-1.01).

**Multiple Logistic Regression:**

Top of FormSubsequently, a multiple logistic regression model was fitted to predict the response using all of the risk factors. This model considered all of the risk factors to predict the impack of CHD in the next 10 years on the individuals. Based on the risk factors with statistical significance, we rejected null hypothesis. The risk factors which are variables demonstrate a strong link with the likelihood of developing CHD over ten years.

An AUC curve is plotted to assess the model's ability to differentiate between individuals prone to developing CHD and those who are not. The obtained C-index was 0.737, and the AUC was also determined to be 0.7375. These values indicate the model's strong discriminatory power in predicting 10-year CHD risk. A higher AUC value, closer to 1, signifies superior predictive performance, indicating that the model effectively identifies individuals at risk for CHD with both high sensitivity (true positive rate) and specificity (true negative rate).

Top of Form

A graph of a line

Description automatically generated with medium confidence

Image -ROC curve

The image is a graph labeled "ROC Curve," which stands for Receiver Operating Characteristic curve. It is a plot of the true positive rate (Sensitivity) against the false positive rate (1-Specificity) for the different possible cut points of a diagnostic test. The x-axis is labeled "Specificity" and runs from -0.5 to 1.5, while the y-axis is labeled "Sensitivity" and ranges from 0 to 1, which in terms means that the model identifies high sensitivity which is true positives and low specificity which is false positives. This graph depicts that the model is able to identify both true positives and false positives, which helps to avoid misunderstanding of those who are truly at risk of CHD and those who are truly not.

**Model selection and choosing Optimal model to predict response:**

A stepwise selection approach was employed, prioritizing the AIC criterion to select the logistic regression model. This method aimed to strike a balance between model accuracy and simplicity. Important factors strongly associated with the risk of coronary heart disease, like age, gender, and systolic blood pressure, were given precedence. Conversely, variables such as prevalentHyp and prevalentStroke showed less significant p-values. However, despite this, the overall model demonstrated a low AIC value, suggesting it effectively captures the data. Further validation of the model's usefulness and predictive accuracy necessitates testing on diverse datasets.

Below table shows the Coefficients, Stan dard Error, z value of the variables (risk factors)

Coefficients:

Estimate

(Intercept) -8.732794

male 0.627665

age 0.063653

currentSmoker 0.382022

prevalentStroke 0.733815

prevalentHyp 0.227870

totChol 0.002355

sysBP 0.014398

glucose 0.007216

Std. Error z value

(Intercept) 0.524764 16.641

male 0.103172 6.084

age 0.006393 9.957

currentSmoker 0.103918 3.676

prevalentStroke 0.483411 1.518

prevalentHyp 0.135105 1.687

totChol 0.001120 2.104

sysBP 0.002855 5.043

glucose 0.001666 4.331

The model’sAIC value is 2783.1, suggesting the model adequately describes the data, and residual deviance is 2765.1, which depicts an optimal balance. The selected model incorporates predictors such as gender, age, smoking status, cholesterol levels, blood pressure, and glucose levels, all of which are significant in predicting 10-year CHD risk, as per established clinical knowledge. Additionally, factors like prevalent hypertension, though marginally significant, may warrant further investigation for potential interactions with other variables affecting CHD risk.

By employing a stepwise selection approach, the final model ensures it is not overly complex and includes only variables contributing substantially to prediction accuracy. This methodology results in an effective tool for forecasting CHD risk.

**Discussion**:

The study's findings emphasize the critical role of age, smoking status, and blood pressure in determining CHD risk, aligning with established medical knowledge. However, limitations such as potential biases and unmeasured confounders should be acknowledged. Additionally, further research is warranted to explore the complex interplay between different risk factors and their collective impact on CHD risk prediction. Despite these limitations, the multiple logistic regression model provides a valuable tool for clinicians to assess CHD risk comprehensively and tailor preventive interventions accordingly.

**Conclusion:**

In conclusion, this study contributes to our understanding of CHD risk factors by highlighting the significance of various predictors, including age, smoking, and blood pressure. By employing rigorous statistical methods and model selection strategies, we have developed an effective model for predicting CHD risk over a 10-year period. These findings have important implications for public health interventions and underscore the importance of early detection and management of modifiable risk factors to mitigate CHD risk in at-risk populations.

References:

1. Go AS, Mozaffarian D, Roger VL, et al. heart disease and stroke statistics--2013 update: a report from the American Heart Association. *Circulation.*2013;127:e6–e245. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408511/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/23239837)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Circulation&title=Heart+disease+and+stroke+statistics--2013+update:+a+report+from+the+American+Heart+Association&author=AS+Go&author=D+Mozaffarian&author=VL+Roger&volume=127&publication_year=2013&pages=e6-e245&pmid=23239837&)] [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159698/#R1)]
2. *The European health report 2012: charting the way to well-being.* Copenhagen, Denmark: WHO Regional Office for Europe; 2012. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=The+European+health+report+2012:+charting+the+way+to+well-being&publication_year=2012&)] [[Ref list](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159698/#R2)]
3. Mahmood, Levy, Vasan, Wang (2013). [*"The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective"*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159698) *(fee required)*. Lancet. **383** (9921): 999–1008. [*doi*](https://en.wikipedia.org/wiki/Doi_(identifier)):[*10.1016/S0140-6736(13)61752-3*](https://doi.org/10.1016%2FS0140-6736%2813%2961752-3). [*PMC*](https://en.wikipedia.org/wiki/PMC_(identifier)) [*4159698*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159698). [*PMID*](https://en.wikipedia.org/wiki/PMID_(identifier)) [*24084292*](https://pubmed.ncbi.nlm.nih.gov/24084292).
4. [NHLBI](https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs)
5. Wikipedia<https://en.wikipedia.org/wiki/Framingham_Heart_Study>