ADVANCING CARDIOVASCULAR HEALTH: A DATA-DRIVEN EXPLORATION OF CORONARY HEART DISEASE RISK ASSESSMENT

Mohak Khatri [1], Muhammad Imran Khan [2], Mohammad Danish Khan [3], Mukesh Chandra [4]
1,2,3,4 Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India

ABSTRACT:

In an era marked by technological advancements, our research endeavours to harness the power of data-driven insights to revolutionize healthcare. This pioneering project focuses on unravelling the complexities of coronary heart disease (CHD) risk assessment through the lens of predictive modelling and data analysis. At its core lies the iconic Framingham Heart Study, a seminal cardiovascular epidemiological investigation that has shaped modern healthcare by identifying risk factors for cardiovascular diseases across diverse cohorts. The central question driving our inquiry is whether the predictive functions developed within the **Framingham Heart Study** can be extrapolated to diverse populations beyond the initial white middle-class cohort.

INTRODUCTION:

Cardiovascular diseases (CVD) constitute a significant global health concern, encompassing conditions such as coronary artery disease and heart failure. Among these, coronary heart disease (CHD) stands out as a leading cause of morbidity and mortality. The research takes inspiration from the rich legacy of the Framingham Heart Study, a longitudinal exploration that has been instrumental in reshaping our understanding of cardiovascular risk factors.

Our project embarks on a mission to extend the insights derived from the Framingham Heart Study to broader demographics, transcending the limitations of the original cohort. The dataset employed comprises 4240 observations and 16 variables, ranging from demographic information and lifestyle factors to crucial health indicators like blood pressure, cholesterol levels, and diabetes status.

In the pursuit of predictive accuracy, the research integrates machine learning, statistical analysis, and epidemiological insights. The ultimate goal is to develop robust predictive models capable of assessing an individual's 10-year risk of CHD. These models aim to offer valuable insights for healthcare practitioners and individuals alike, fostering informed decision-making regarding cardiovascular health.

The project is characterized by its comprehensive exploration of diverse machine learning algorithms, meticulous feature engineering, and rigorous model evaluation. Beyond merely developing predictive models, the research critically assesses their applicability to populations beyond the original Framingham cohort, thereby contributing to the broader landscape of predictive healthcare analytics.

This paper introduces the fundamental terms and tools essential for navigating the intricate landscape of data science and machine learning, laying the groundwork for a detailed exploration of the methodology, results, and implications. The intersection of Python, NumPy, Pandas, Matplotlib, Seaborn, and machine learning concepts forms the backbone of our approach, symbolizing the interdisciplinary nature of our quest to improve early detection and prevention of heart disease.

LITERATURE REVIEW: ADVANCEMENTS IN NON-LINEAR STROKE RISK ASSESSMENT

Introduction: Stroke, a devastating neurological event, remains a significant global health concern, necessitating accurate risk assessment for effective prevention strategies. Traditional stroke risk assessment tools, often rooted in linear and cumulative models, face challenges in capturing the complexities of novel risk factors and their intricate interactions^ [1^]. This literature review explores recent endeavours to enhance stroke risk prediction through the development of non-linear models, with a particular focus on the Revised Framingham Stroke Risk Score and the emergence of an interactive Non-Linear Stroke Risk Score.

Traditional Models and Limitations: Conventional stroke risk assessment tools predominantly rely on linear and cumulative risk models. While these models have provided valuable insights, their inherent limitations in capturing non-linear relationships and novel risk factors have prompted a paradigm shift in the quest for more accurate predictive models^ [2^]. The inadequacies of these traditional tools underscore the need for innovative approaches capable of unveiling intricate risk patterns.

Non-Linear Approaches: The study by Organouranium et al. introduces a novel Non-Linear Stroke Risk Score, aiming to overcome the limitations of linear models. Leveraging machine learning algorithms, particularly Optimal Classification Trees, the research adopts a two-phase approach^ [2^]. In the first phase, the Framingham offspring cohort serves as the training dataset, enabling the development of a tree-based model that dynamically adjusts splits on independent variables, thus introducing non-linear interactions.

Validation and Multi-Ethnicity Considerations: To validate the proposed non-linear model, a multi-ethnic cohort from the Boston Medical Centre is utilized^ [2^]. The study reveals a

pivotal dichotomy between patients with a history of cardiovascular disease and the rest of the population, highlighting the nuanced nature of stroke risk profiles. Importantly, the non-linear approach not only aligns with established findings but also uncovers 23 unique stroke risk profiles, shedding light on previously unnoticed relationships, such as the impact of T-wave abnormality on electrocardiography and haematocrit levels.

Clinical Implications and Significance: The results of the study suggest that the non-linear approach significantly improves upon baseline models in terms of the c-statistic, both in training and validation sets, even when applied to multi-ethnic populations^ [2^]. The clinical implications of the Non-Linear Stroke Risk Score are profound, emphasizing the prioritization of risk factor modification and the potential for personalized care at the patient level. The newfound precision in identifying stroke risk profiles allows for more targeted interventions, marking a crucial step toward personalized stroke prevention strategies.

Conclusion: This literature review illuminates the evolving landscape of stroke risk assessment, emphasizing the transition from traditional linear models to more sophisticated non-linear approaches. The study by Organouranium et al. contributes substantially to this shift, providing valuable insights into the intricate relationships influencing stroke risk and advocating for a personalized and targeted approach to stroke prevention at the clinical level.

DATA COLLECTION

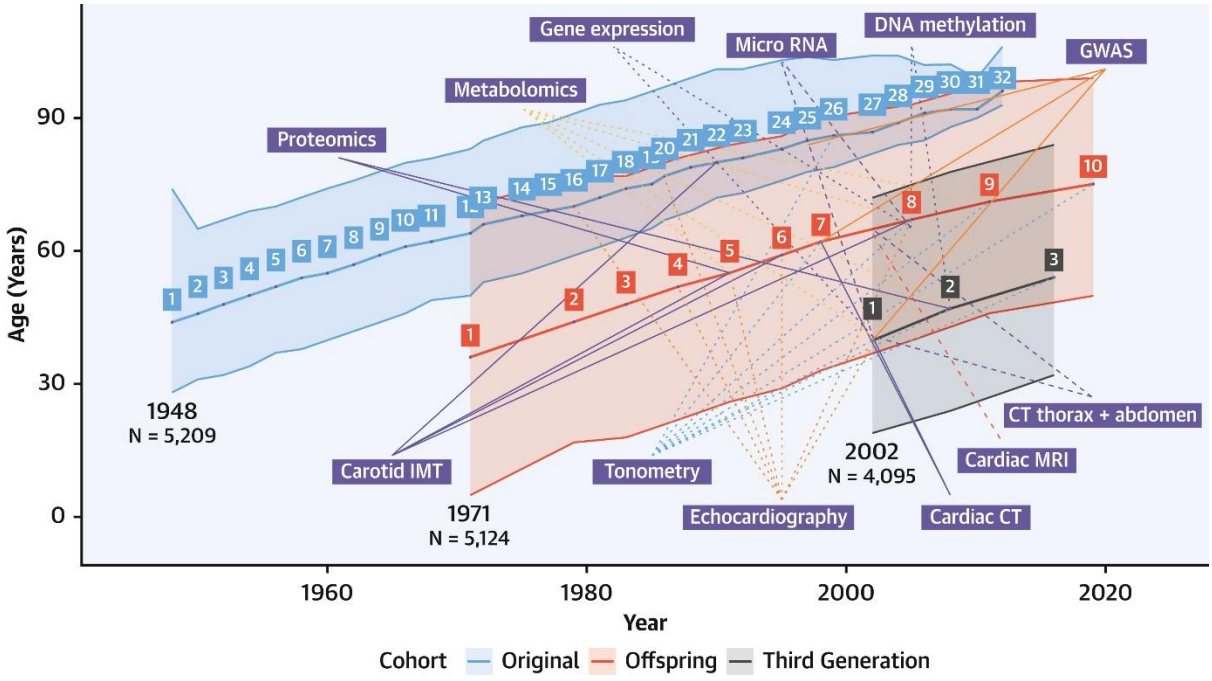

Figure 1, Age Span of the 3-Generation Cohorts of the Framingham Heart Study by Calendar Period and Examination Cycle

In-person visits at the FHS research centre have taken place every 2 years for the original cohort and approximately every 4 to 7 years for the Offspring, Third Generation, and Omni cohorts. During each of these participant visits, a comprehensive interview is performed using standardized medical questionnaires, along with a cardiovascular-focused physical examination, electrocardiogram, bio sample collection (blood and urine), and lifestyle-related questionnaires. Additionally, examination-specific tests have been performed at each FHS visit, as outlined in **Figure 1**. The FHS cohorts are among the most densely phenotype contemporary cardiovascular epidemiological cohorts. Ongoing surveillance of cardiovascular and no cardiovascular endpoints is accomplished by ongoing review of medical records and participant interviews to ensure timely updated data on key outcome events that constitute FHS clinical endpoints. All potential endpoints are adjudicated by review panels comprising internists and cardiologists (for cardiovascular endpoints) and neurologists (for stroke/dementia endpoints), which enhances their validity.

METHODOLOGY

The methodology encompasses two main phases: Dataset 1 for developing a predictive model for coronary heart disease (CHD) risk ([1]) and Dataset 2 for building a classification model for CHD prediction.

Dataset 1: Framingham Heart Disease Prediction Project

The project initiation involves defining objectives, forming a dedicated team, and identifying stakeholders. A feasibility analysis assesses market demand, technical feasibility, financial modelling, and legal compliance ([1]). Data collection focuses on the Framingham Heart Study dataset, with exploratory data analysis (EDA) and preprocessing, including user-defined functions for missing value imputation.

Multivariate analysis explores relationships between variables ([9]), and user-defined functions categorize individuals into age groups and heart rate categories. Visualization aids understanding, and log transformation and normalization enhance variable effectiveness. The dataset is split for training and testing ([14]), and logistic regression models are trained, including a weighted model for class imbalance.

Dataset 2: Classification Model for CHD Prediction

Data exploration involves loading the coronary heart disease dataset, inspection, label encoding, and handling duplicates ([12]). Class imbalance is resolved through oversampling ([5]), and data normalization ensures standard scaling. The dataset is split for training and testing ([16]), and models include a Support Vector Machine (SVM) and a Boost classifier. Model evaluation employs accuracy, confusion matrix, and classification report metrics.

RESULT

Our methodology unfolds in a systematic sequence. Commencing with data exploration and preprocessing, we leveraged the Framingham Heart Study dataset, meticulously examining its 4240 records and 16 features ([1]). A heatmap visualized feature correlations, guiding subsequent analyses. Missing value imputation was executed with a tailored function, 'impute median,' ensuring contextually accurate imputation for key features like 'glucose' and 'heartrate' ([1]). Descriptive statistics and visualizations in the exploratory data analysis (EDA) phase unveiled patterns in numerical features. Feature engineering introduced age and heart rate encodings, enriching the dataset for predictive modelling. Multivariate analysis, through boxen plots, explored gender influences on cardiovascular indicators and the distribution of glucose and total cholesterol across age groups ([9]). A pie chart illustrated the target class distribution, emphasizing class imbalance, addressed through oversampling using SMOTE.

Log transformation and normalization prepared continuous variables for logistic regression modelling. Two logistic regression models were trained—one with default settings and another with balanced class weights, addressing class imbalance ([6]). Rigorous model evaluation employed accuracy, cross-validation scores, classification reports, confusion matrices, Receiver Operating Characteristic (ROC) curves, and Area Under the Curve (AUC) ([2]). Result concludes by emphasizing the importance of continuous monitoring and adaptation for sustained accuracy and relevance in CHD risk assessments, bridging technical aspects with informed healthcare decision-making.

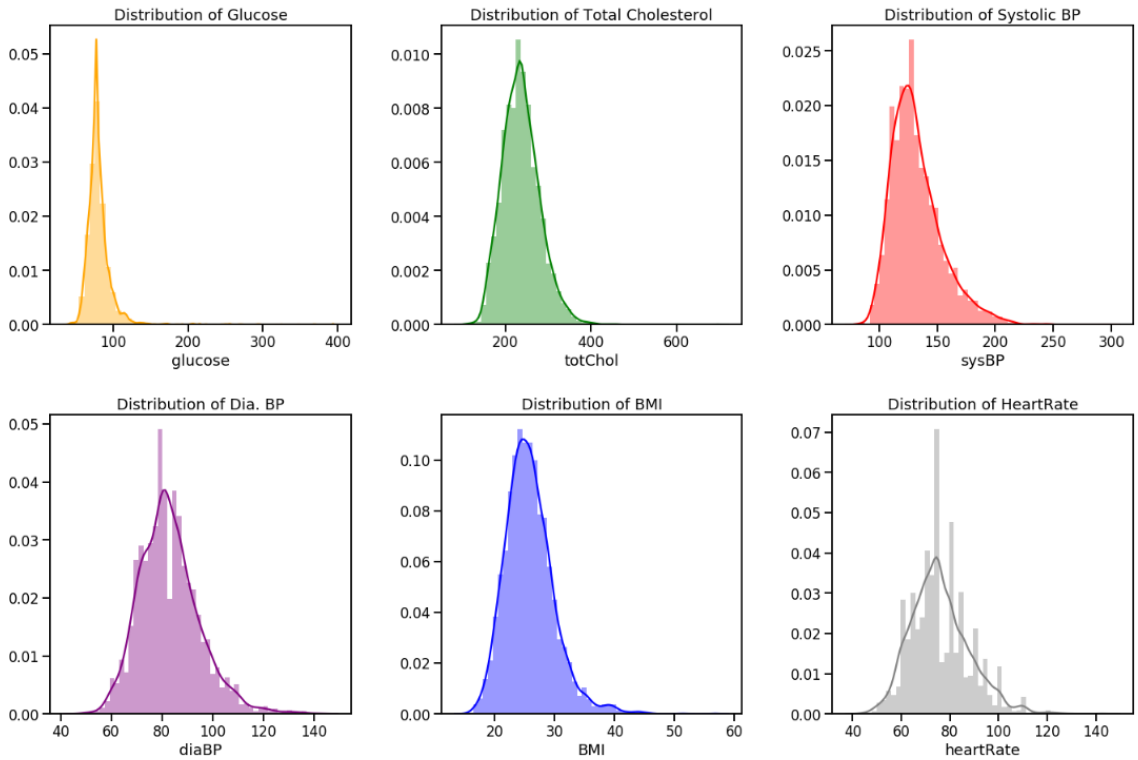

Figure 2, Pair plots

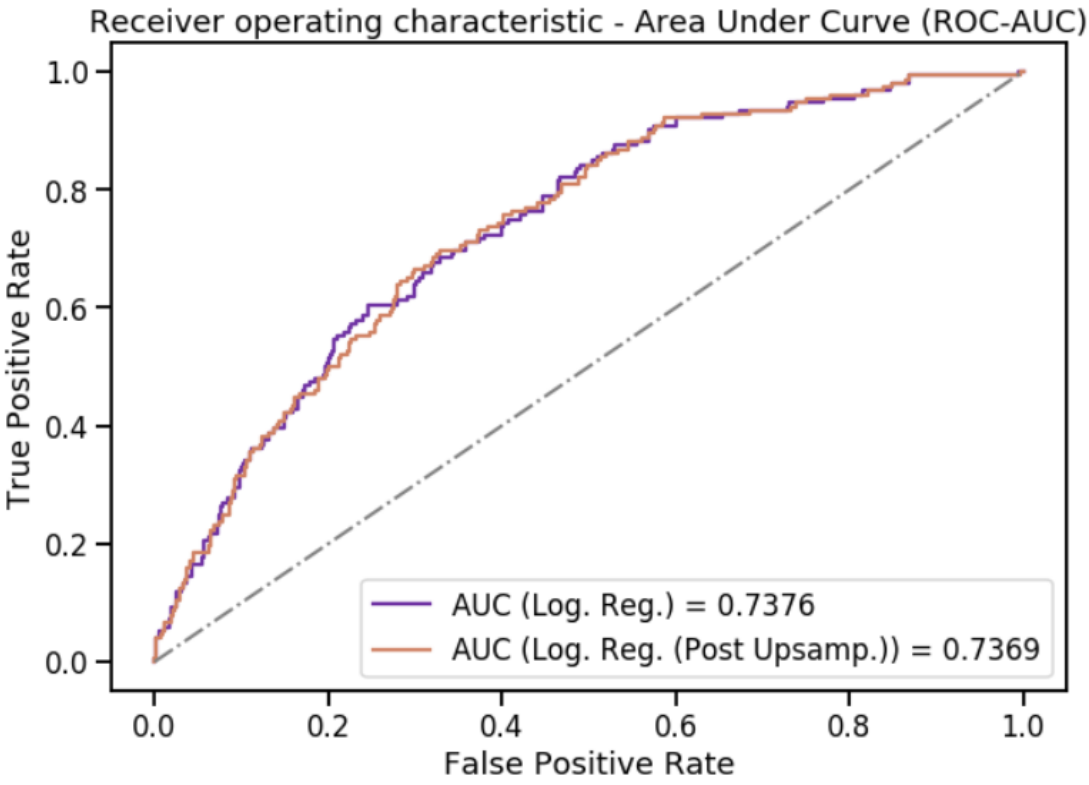

Figure 3, ROC-AUC

CONCLUSION

In wrapping up the Framingham Heart Disease Prediction project, a meticulous strategy unfolds, harnessing the Framingham Heart Study dataset to assess coronary heart disease (CHD) risk. Extensive data preprocessing, exploratory analysis, and sophisticated modelling techniques collectively illuminate critical risk factors. Logistic regression models, validated through cross-validation and classification reports, reveal promising predictive capabilities, highlighting the potential of data-driven tools in cardiovascular health ([15]). The study emphasizes the continuous evolution of research to refine predictive models, enabling proactive CHD management ([3]). This endeavour contributes to the progressive landscape of preventive healthcare, showcasing the pivotal role of data science in delivering personalized risk assessments and fostering informed decision-making in cardiovascular disease prevention ([2]).

The project's holistic methodology, from initial data exploration to model evaluation, underscores the interdisciplinary synergy between healthcare and machine learning. As the healthcare landscape advances, this research provides a solid foundation for future endeavours, emphasizing the ongoing need for innovation and adaptation in the pursuit of enhanced CHD risk assessment and personalized patient care.

FUTURE SCOPE

The Framingham Heart Disease Prediction project ambitiously seeks to propel cardiovascular epidemiology forward by enhancing the generalizability of coronary heart disease (CHD) prediction functions. Embracing a multi-faceted scope, the project envisions a rigorous evaluation of CHD prediction functions derived from the Framingham Heart Study, extending their relevance to diverse populations beyond the original cohort ([1]). Delving into demographic and lifestyle variations, the research aims to unravel novel factors influencing CHD risk that were not previously considered in the Framingham Study ([3]).

Pioneering the use of advanced predictive models through machine learning and statistical techniques, the project takes a holistic approach, considering a broader spectrum of demographic and lifestyle variables ([2]). Ethical and policy considerations are prioritized, with a dedicated focus on examining the implications of extending CHD prediction functions to new populations, addressing fairness, bias, and data privacy issues ([4]).

This forward-looking endeavour aspires to contribute significantly to cardiovascular epidemiology, fostering a deeper understanding of CHD risk assessment and expanding the applicability of prediction functions ([5]). Positioned as a foundational work, the project anticipates serving as a catalyst for future research and collaboration, ultimately enhancing CHD risk assessment and preventive strategies for diverse populations in the realm of healthcare practices ([11]).

REFERENCES

1. Framingham Heart Study: JACC Focus Seminar, 1/8- Charlotte Andersson, Matthew Mayor, Connie W. Tsao, Daniel Levy, and Ramachandran S. Vasan J Am Coll Cardio. 2021 Jun, 77 (21) 2680–2692.
2. Machine learning provides evidence that stroke risk is not linear: The non-linear Framingham stroke risk score Agni Organouranium, Emma Chesley, Christian Caddish, Barry Stein, Amer Nouh, Mark J. Alberts, Dimitris Bertsimas.
3. A Machine Learning Analysis of the FHS, the ARIC Study, and the CHS Laura M. Stevens, Erik Linstead, Jennifer L. Hall and David P. Kao Originally published 9 Feb 2021.
4. Risk Prediction in Patients with Heart Failure with Preserved Ejection Fraction Using Gene Expression Data and Machine Learning, Liye Zhou.
5. Artificial Intelligence and Machine Learning in Cardiovascular Health Care Arman Kilic MD.
6. Gordon T. "Mortality in the United States, 1900-1950". Public Health Rep 1953; 68:441-444.
7. Dawber T.R., Meadors G.F., Moore F.E. "Epidemiological approaches to heart disease: the Framingham Study". Am J Public Health Nations Health 1951; 41:279-281.
8. Mahmood S.S., Levy D., Vasan R.S., Wang T.J. "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective". Lancet 2014; 383:999-1008.
9. Dawber T.R., Moore F.E., Mann G.V. "Coronary heart disease in the Framingham study". Am J Public Health Nations Health 1957; 47:4-24.
10. Kannel W.B., Dawber T.R., Kagan A., Revoke N., Stokes J. "Factors of risk in the development of coronary heart disease—six-year follow-up experience. The Framingham Study". Ann Intern Med 1961; 55:33-50.
11. Doyle J.T., Dawber T.R., Kannel W.B., Heslin A.S., Kahn H.A. "Cigarette smoking and coronary heart disease. Combined experience of the Albany and Framingham studies". N Engl J Med 1962; 266:796-801.
12. Doyle J.T., Dawber T.R., Kannel W.B., Kinch S.H., Kahn H.A. "The relationship of cigarette smoking to coronary heart disease; the second report of the combined experience of the Albany, NY. and Framingham, Mass. studies". JAMA 1964; 190:886-890.
13. Kannel W.B. "Habitual level of physical activity and risk of coronary heart disease: the Framingham study". Can Med Assoc J 1967; 96:811-812.
14. Truett J., Cornfield J., Kannel W. "A multivariate analysis of the risk of coronary heart disease in Framingham". J Chronic Dis 1967; 20:511-524.

15. Fenley M., Kannel W.B., Garrison R.J., McNamara P.M., Castelli W.P. "The Framingham offspring study. Design and preliminary data". Prev Med 1975; 4:518-525.
16. Polansky G.L., Corey D., Yang Q., et al. "The Third-Generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination". Am J Epidemiology 2007; 165:1328-1335.
17. Bromfield S., Muntner P. "High blood pressure: the leading global burden of disease risk factor and the need for worldwide prevention programs". Curr Hypertense Rep 2013; 15:134-136.
18. Kannel W.B., Wolf P.A., Verger J., McNamara P.M. "Epidemiologic assessment of the role of blood pressure in stroke. The Framingham Study". JAMA 1970;214:301-310.
19. Kannel W.B., Castelli W.P., McNamara P.M., McKee P.A., Feinleib M. "Role of blood pressure in the development of congestive heart failure. The Framingham Study". N Engl J Med 1972;287:781-787.
20. Levy D., Larson M.G., Vasan R.S., Kannel W.B., Ho K.K. "The progression from hypertension to congestive heart failure". JAMA 1996;275:1557-1562.