Prediction of Coronary Heart Disease with Multiple Regression Method

Elda Maraj[1], Aida Bendo[2,*]

[1]Polytechnic University of Tirana, Faculty of Mathematical Engineering and Physical Engineering, Department of Mathematical Engineering, Tirana, Albania

[2]Sports University of Tirana, Faculty of Physical Activity and Recreation, Department of Movement and Health, Tirana, Albania

Emails: e.maraj@fimif.edu.al; abendo@ust.edu.al

Abstract

Coronary heart disease, a prevalent cardiovascular condition, affects coronary arteries, causing progression over time. Factors include diabetes, hypertension, inactivity, and tobacco use. Treatment includes medications and surgery, while maintaining a balanced diet and regular physical activity can prevent it. This research aimed to develop and validate a predictive model for CHD occurrence, leveraging the power of multiple regression while considering a range of predisposing variables. This study uses a quantitative, retrospective design utilizing multiple regression analysis to predict the likelihood of coronary heart disease (CHD). The study involved 130 participants aged 24-85, with health history data on cardiovascular risk factors, blood pressure, cholesterol, smoking, BMI, and family history of heart disease. Multiple regression analysis was utilized to determine the significant predictors of CHD diagnosis. Significant relationships between responder variables and predictor factors in a multiple linear function are identified using multiple regression analysis. Our model discovered that a higher risk of coronary heart disease (CHD) was closely associated with both total cholesterol and BMI. The model included factors like systolic blood pressure, diabetes, physical activity, and smoking, but they had lower contributions to the prediction equation, despite cholesterol and BMI being the best predictors. This study successfully developed a multiple regression-based prediction model for CHD that can contribute to a more informed and potentially proactive approach to cardiac healthcare. Further work should focus on refining model accuracy and real-world clinical application.

Keywords: Coronary heart disease; Multiple linear regression; Predictors; Explanatory model

1. Introduction

Coronary heart disease (CHD) is a common type of cardiovascular illness that mostly affects coronary arteries, which are the major blood channels that provide blood to the heart. In most cases, CHD is also known as ischemic heart disease, and is the most prevalent, contributing significantly to morbidity and mortality worldwide [1], which develops gradually over time. It is the most common and a major global cause of illness and mortality. This illness, which can cause angina, myocardial infarction in approximately, heart failure, or sudden cardiac death, occurrence and disability worldwide [2], places a significant strain on economies and healthcare systems and other government services for their effectiveness programs in preventing, detecting, and treating coronary heart disease, resulting in direct health care cost estimates [3]. It is responsible for about 50% cardiovascular mortality, which accounts for 30-50% of all deaths in developed nations [4].

Genetic, behavioral, and environmental variables interact intricately in the multifactorial etiology of congenital heart disease. About 80% of congenital heart disease (CHD) is multifactorial and arises through various combinations of genetic and environmental contributors [5]. Research shows traditional risk factors like smoking, obesity, diabetes; dyslipidemia, hypertension, and physical inactivity are linked to cardiovascular risk factors like

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

dyslipidemia, type 2 diabetes, hypertension, and sleep disorders. The modifiable risk factors that can be controlled are high blood pressure, cholesterol levels; smoking; diabetes; overweight or obesity; lack of physical activity; unhealthy diet and stress [6]. Physical inactivity and sedentary behavior are significant risk factors for cardiovascular disease, but new risk factors include environmental stressors, climate change, psychological stress, and chronic inflammation.

Epidemiological studies have made extensive use of this technique to create risk prediction models for several illnesses, including CHD. Several methods are used to assess improvement in risk prediction models, which have been widely applied for the prediction of long-term incidence of disease [7]. In medical research, statistical and machine learning techniques have become more popular in recent years for forecasting the course of diseases. In addition, multiple regression is applied to predict coronary heart disease and considerable research has been conducted within this field.

Multiple linear regression (MLR) analysis is a crucial method for analyzing the relationship between a dependent variable and multiple independent variables, determining the model's overall fit and predictor contribution [8]. Regression analysis is the most famous methodology for medical studies with continuous independent variables [9]. Multiple regression is a valuable tool for predicting CHD by combining risk factors into a single model, enabling the classification of diseases based on individual profiles [10].

Formally, the linear regression model is represented by equation (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon_i \quad (1)$$

The interception is represented by β_0 while the regression coefficients range from β_1 to β_n with the independent variables X_1 to X_n, and the error term ε_i [11].

Linear regression is among the most applied predictive models in both statistics and machine learning [12]. Machine learning algorithms, alongside traditional methods, have gained popularity for their ability to handle large datasets and enhance diagnostic accuracy.

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) technology have brought substantial strides in predicting and identifying health emergencies, disease populations, and disease state and immune response, amongst a few [13]. A Principal Component Analysis (PCA) in binary logistic regression was used to predict heart disease, enhancing precision to 85% [14]. An automated heart disease identification strategy using advanced logistic regression, demonstrating superior accuracy rates compared to linear regression models, was suggested in recent research [15].

The relationship between cardiovascular disease risk factors and arterial stiffness in the Mashhad stroke and heart atherosclerotic disorder was explored in a cohort study, [16]. A mathematical correlation through multiple linear regression and recognized cycling was established as the most straightforward and effective method to mitigate the risk of developing cardiovascular disease [17]. A machine learning-based classification model to identify and address heart-related issues, was formulated and developed thereby reducing the incidence of heart attacks in individuals [18]. An artificial neural network to assess and predict cardiovascular disease risk factors was used to reveal that anxiety is a prevalent risk factor for cardiovascular disease in hospitalized populations, [19]. Techniques are required to raise awareness of psychological variables that can be controlled in those who are at a higher risk of developing cardiovascular disease in the future. A diagnostic model that integrates clinical and laboratory parameters of coronary heart disease in middle-aged and elderly people was developed and validated, based on clinical data from 839 eligible patients to assess whether these factors could be integrated into the model as a support tool for the identification of cardiovascular disease in patients [20]. A logistic regression modelling to estimate the risk of coronary heart disease was applied and it has identified significant influencing factors. The findings showed the high performance of logistic regression in the domain of cardiovascular disease predictions, achieving a model prediction accuracy of 0.86 [21].

This study aims to create a predictive model for coronary heart disease using multiple regression. It integrates traditional and emerging risk factors, such as age, systolic blood pressure, cholesterol, diabetes mellitus, smoking status, physical activity and body mass index (BMI) obesity, to quantify the relative contribution of each factor to CHD risk. The model aims to provide a reliable tool for early identification of individuals at high risk of developing CHD, enabling healthcare providers to implement targeted preventive interventions. This research seeks to contribute to the reduction of CHD-related morbidity and mortality by improving risk stratification and informing evidence-based prevention strategies.

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

2. Materials & Methods

2.1 Hypothesis

The main hypothesis of this study is that a combination of modifiable and non-modifiable risk factors, when analyzed using multiple regression, will significantly predict the likelihood of an individual developing coronary heart disease (CHD). Specifically, we hypothesize that variables such as age, high blood pressure, total cholesterol levels, physical activity levels, smoking status, body mass index (BMI), presence of diabetes, will, collectively and independently, contribute to a statistically significant model that can predict the presence or absence of CHD. Furthermore, we anticipate that the strength and direction of each predictor variable's association with CHD will align with existing literature, leveraging the predictive power of multiple regression, aiming to offer a more comprehensive and nuanced understanding of CHD risk compared to assessments based on single risk factors alone.

2.2 Study design

This study uses a quantitative, retrospective design utilizing multiple regression analysis to predict the likelihood of coronary heart disease (CHD). The methodology was chosen specifically to address the research objective of identifying and quantifying the relationships between various risk factors and the presence of CHD. Multiple regression is ideally suited for this purpose as it allows for the simultaneous examination of the impact of several independent variables such as age, cholesterol levels, blood pressure, diabetes level, smoking status, physical activity level, on a single dependent variable that is the presence or absence of CHD, often measured as a continuous score representing risk or a binary outcome. This method enables us to not only determine which risk factors are statistically significant predictors of CHD, but also to assess the relative strength and direction of their influence. Furthermore, the regression coefficients generated provide a means to quantify the predicted change in CHD risk for a given change in each independent variable, aligning with the study's aims to provide actionable insights into the modifiable factors that contribute to coronary heart disease. Crucially, this detailed analysis is important in the Albanian context because it enables the development of a predictive model directly tailored to the nation's specific risk factor profile, which may differ from those identified in studies from other regions. Further, this investigation holds wider interest due to the potential to establish a robust and generalizable model that could inform preventative healthcare strategies and resource allocation in similar resource-challenged health systems. Therefore, this type of model is important to be investigated in all contexts, providing a critical basis for implementing preventive measures and enhancing public health.

2.3 Instrument for data collection

This study has employed a quantitative, retrospective approach utilizing existing datasets containing relevant health information, likely including demographic, lifestyle, and clinical measurements. This study, aimed at predicting coronary heart disease (CHD) using multiple regression analysis, was conducted within a large, urban medical research facility in Polyclinic Specialty Health Center No. 2 in Tirana city. The database was obtained from the family physicians near this center. The focus was on human participants representing the general population, encompassing a broad age range from 24 to 85 years. To ensure a diverse and representative sample, recruitment has included different age's subjects, to identify potential participants; and targeted outreach between groups. A questionnaire method was used to assess the heigh, body mass, physical activity level and smoking level of every subject, while the blood pressure, cholesterol and diabetes were measured through biochemical analysis. The study received ethical approval with protocol no. 2609-1, from the Ethical Committee of the Polytechnic University of Tirana. Approval was obtained from the subjects to consider them as study participants.

2.4 Inclusion and exclusion criteria

Inclusion criteria have stipulated participants between 24 and 85 years of age, be able to provide informed consent, and have available relevant health history data, particularly concerning cardiovascular risk factors such as blood pressure, cholesterol levels, smoking status, BMI, and family history of heart disease. Individuals with pre-existing diagnosed coronary heart disease requiring treatment, those with significant cognitive impairments that would impede their ability to participate, and those currently involved in other clinical trials impacting cardiovascular health were excluded from the study. This structured selection process aims to gather the necessary data for robust multiple regression analysis of CHD risk in a diverse population.

2.5 Population and Sample size

The present study employed a rigorous methodology to ensure that the selected sample was statistically representative of the target population, facilitating the generalizability of the findings. To sample the target population, which was defined as adults in Tirana city between the ages of 24 and 85, stratified random sampling was used. International statistical calculators were used to do power analysis to estimate the proper sample size, with an emphasis on obtaining confidence levels greater than 95%. In this computation, the needed statistical

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

power, the allowable margin of error, and the expected effect size of the main predictor variables were considered. These calculations showed that a sample size of 130 participants was required for sufficient precision and statistical power. This number minimizes the potential biases associated with an inadequate or non-representative sample and guarantees that the results produced appropriately reflect the characteristics of the larger population.

2.6 Reliability and validity

Prior to regression analysis, all continuous variables were assessed for normality using histograms and Q-Q plots, with appropriate transformations applied as necessary to mitigate skew. Categorical variables were incorporated into the regression model. The data were screened for outliers and multicollinearity (using variance inflation factors), and any identified issues were addressed prior to model estimation. The regression model was fit using ordinary least squares, and the statistical significance of each predictor variable was assessed using Fisher-tests and p-values. Model fit was evaluated using the adjusted R-squared value and residual analysis was conducted to check for violations of regression assumptions, such as homoscedasticity and normality of residuals. Furthermore, a separate dataset was used for model validation before actual implementation.

2.7 Statistical Analysis

This study has employed multiple regression analysis to model the relationship between several independent variables (predictors) and the dependent variable, the presence or absence of coronary heart disease (CHD). The dependent variable will be coded as a binary outcome (0 = no CHD, 1 = presence of CHD). The core of our analysis will involve a standard multiple logistic regression model. The independent variables include: age (continuous, measured in years), Body Mass Index (BMI, continuous, kg/m²), systolic blood pressure (continuous, mmHg), total cholesterol level (continuous, mg/dL), diabetes status (continuous, mg/dL), physical activity (PA) (categorical, potentially coded as ordinal as follows: 1=sedentary, 2=light activity, 3= moderate activity, 4=vigorous activity), and smoking status (binary, 0= non-smoker, 1 = smoker). The multiple regression has allowed us to examine the unique contribution of each predictor while controlling the effects of the others. The overall fit of the model was assessed using measures like the R-value which represents the coefficient of multiple correlations and R2 that represents the proportion of variance in the dependent variable explained by the model, for predicting heart disease. Regression coefficients (B values) for each independent variable have indicated the direction and magnitude of their association with CHD risk, and their significance has determined through the assessment of p-values associated with those coefficients. The CHD presence is associated with a one-unit change in a continuous independent variable or presence of an independent variable being present compared to its absence, while keeping all other independent variables constant. The multiple regression analysis has evaluated goodness-of-fit of the model. Additional analyses have included stepwise interaction terms to assess if the effect of one variable on CHD risk varies depending on the values of another variable. Furthermore, all p-values are assessed with a significant level of 0.05.

3. Results

The multiple regression approach is used to determine whether independent variables like age, BMI, blood pressure, cholesterol, diabetes, physical activity, and smoking can predict the dependent variable of heart disease value. To determine the best predictor in multiple regression analysis, Table 1 provides the descriptive statistics for each of the study's independent variables—age, BMI, blood pressure, cholesterol, diabetes, physical activity, and smoking—as well as the dependent variable, the heart disease value.

Table 1: Statistical descriptive for all variables of the study.

Variable	Mean ± SD	Min. value	Max. value	Range	Variance
Disease (0-1 levels)	0.76 ± 0.298	0.375	1.00	0.625	0.178
Age (Years)	53.666± 14.763	24.00	85.00	61.00	217.954
BMI (kg/m^2)	27.383 ± 5.404	20.20	47.50	27.30	29.208
Blood Pressure (mmHg)	102.30 ± 14.242	70.00	133.00	63.00	202.838
Cholesterol (mg/dL)	176.60 ± 40.39	120.00	291.00	171.00	163.135
Diabetes (mg/dL)	130.20± 51.457	45.00	300.00	255.00	264.789
PA (1- 4 levels)	2.452 ±0.255	1.00	4.00	3.00	0.26
Smoking (0-1 levels)	0.12 ±0.160	0.00	0.80	0.80	0.026

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

3.1. Results of Multiple Regression to heart disease

The overall model fit and the relative contribution of each predictor to the total variance can be assessed using multiple regression techniques. The percentage of the dependent variable's variation that can be accounted for by the independent variables, as well as the "relative contribution" of each independent variable to the explanation of variance, should be known at heart disease values.

3.2. Adjusting the Overall Model

Table 2 Table 2 is produced by using the multiple regression approach to each of the independent variables, and it gives a summary overview of the model. → R value represents the coefficient of multiple correlations. R can be a measure of predicting heart disease. The value of R = 0.671 in this case shows a good level of prediction. The value of R2 = 0.45 indicates that the independent variables included in this model can account for 45% of the variation of the dependent variable. R2 is also known as the coefficient of determination, or the portion of the variance in the dependent variable that can be explained by the independent variables.

Table 2: Summary of the overall model

Model Summary				
Model	R	R Square	Adjusted Square	RStd. Error of the Estimate
1	0.671[a]	0.450	0.275	0.50887
a. Predictors: (Constant), Smoking, BMI, Diabetes, Cholesterol, Blood Pressure, PA, Age				

R^2 value adjusted is used to report and accurately interpret the model, then R^2=0.428 value accurately explains 42.8% of the variance of heart disease value.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.661	7	0.666	2.572	0.043[b]
	Residual	5.697	122	0.259		
	Total	10.358	129			
a. Dependent Variable: Disease						
b. Predictors: (Constant), Smoking, BMI, Diabetes, Cholesterol, Blood Pressure, PA, Age						

In the Anova table, value F- Fisher tests whether the full regression model fits well with the data. For this parameter, the value of Fischer's, F (7, 122) = 2.572 and p= 0.004. p = 0.043< 0.050 means that the regression model fits well with the data.

3.3. The Model Coefficient Calculation

The regression coefficients table provides the data required to calculate the statistical model significance (based on the values of the columns t and sig column, value (p) and to forecast the dependent variable heart disease value by all independent variables. The general form of regression equations for predicting the dependent variable (Y) of heart disease value is based on beta coefficients (β) as follows:

$$y = -1.398 + 0.032\,(BMI) + 0.01(Blood\ Pressure) + 0.006\,(Cholesterol) - 0.023\,(PA) + 0.405\,(Smoking) \quad (2)$$

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.398	1.438		-0.972	0.342
	Age	0.000	0.010	0.011	0.044	0.965
	BMI	0.032	0.021	0.293	1.519	0.143
	Blood Pressure	0.010	0.008	0.228	1.239	0.228
	Cholesterol	0.006	0.003	0.387	1.968	0.062
	Diabetes	0.000	0.002	-0.014	-0.065	0.949
	PA	-0.023	0.592	-0.010	-0.038	0.970
	Smoking	0.405	0.644	0.109	0.628	0.536

a. Dependent Variable: Disease

Although the overall prediction is analyzed by this equation, the best predictors are not included. To determine if the coefficients in the population are equal to zero, the coefficient table additionally assesses the statistical significance of each independent variable. Information on which independent variables are the most significant and effective regression model predictors is obtained through exploratory research.

3.4. Finding the Best Models on heart disease

Table 3 provides a summary of exploratory models, indicating which variables are included in the model step by step.

Table 3: Summary exploratory model

Model Summary

Model	R	R Square	Adjusted Square	RStd. Error of the Estimate
1	0.529[a]	0.280	0.254	0.51619
2	0.620[b]	0.384	0.339	0.48596

a. Predictors: (Constant), Cholesterol

b. Predictors: (Constant), Cholesterol, BMI

Model (A) indicates that the best predictor by itself is cholesterol. R = 0.529, R^2 = 0.28, and the initial step calculates 28% of the independent variable's variance for this variable. F (1, 128) = 10.875, p = 0.003 <0.05, which are the F and p values that Anova reported.

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.898	1	2.898	10.875	0.003[b]
	Residual	7.461	128	0.266		
	Total	10.358	129			
2	Regression	3.982	2	1.991	8.431	0.001[c]
	Residual	6.376	127	0.236		
	Total	10.358	129			

a. Dependent Variable: Disease

b. Predictors: (Constant), Cholesterol

c. Predictors: (Constant), Cholesterol, BMI

According to model B), the best prediction models are BMI and cholesterol. BMI, which is a component of the model, is therefore the second-best predictor after cholesterol. The second stage determines 38.4% of the variance of the two variables, with R = 0.620 and R^2 = 0.384 for both variables. The corresponding values are F (2, 127) = 8.431 and p = 0.05.

Both models together have significant results, since both p<0.05.

4. Discussion

4.1. The Regression Equation for the Best Predictors on heart disease

Regression equations can be formed using the beta coefficients array (β), which provides the corresponding values for a better predictor model. According to Table 5, the β coefficients change according to which predictors are used in the model.

Table 5: Beta coefficients of heart disease prediction.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	0.138	0.430		0.322	0.750
	Cholesterol	0.008	0.002	0.529	3.298	0.003
2	(Constant)	-0.663	0.551		-1.204	0.239
	Cholesterol	0.007	0.002	0.449	2.891	0.007
	BMI	0.037	0.017	0.333	2.143	0.041

a. Dependent Variable: Disease

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

The regression equations for both models involved in exploratory study are as follows:

Equation according to model A)

$$y = 0.138 + 0.008\ (Cholesterol) \quad (3)$$

This is the weight of an equation that includes: Cholesterol as the best predictors of this model. According to this equation, for any cholesterol increase of 1mg/dl there is an increase of heart disease with a value of 0.008.

Equation according to model B)

$$y = -0.663 + 0.007\ (Cholesterol) + 0.037\ (BMI) \quad (4)$$

These are the weights for an equation that uses BMI and cholesterol as the model's top predictors. In accordance with this model, heart disease increases with a value of 0.007 for every 1 mg/dL increase in cholesterol and 0.037 for every 1 kg/m2 increase in BMI. Models A and B remove other predictors as not being particularly essential in the regression model since the inverse analysis shows extremely tiny changes of R2 between the initial step and exploratory study.

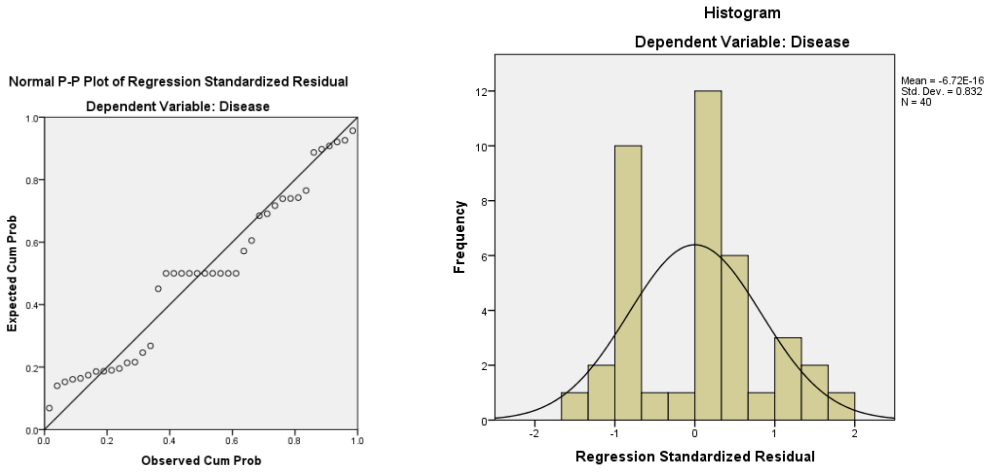

Figure 1. (*a*) The graph of the expected cumulative probability of heart disease in function of observed cumulative probability. (*b*) The Histogram of frequency distribution for heart disease based on regression standardized residual.

Figure 1 shows the graph of the cumulative probability and the histogram distribution for heart disease. By obtaining a statistically significant model that explained a large amount of the variance in CHD risk, our work effectively illustrated the feasibility of multiple regression in forecasting the probability of coronary heart disease. Our model specifically found that total cholesterol and BMI were all strongly linked to an elevated risk of coronary heart disease (CHD) ($p < 0.05$). For example, if all other factors were held constant, a hypothetical patient named "Patient A" who is 60 years old, has 240 mg/dL of total cholesterol, and a systolic blood pressure of 150 mmHg would have a predicted risk score that is significantly higher than that of "Patient B," who is 45 years old, has 180 mg/dL of total cholesterol, and has a systolic blood pressure of 120 mmHg. This demonstrates the model's prediction ability and is consistent with the epidemiological data that currently links these factors to CHD. Though they were included in our model, other factors like systolic blood pressure, diabetes, physical activity level and smoking status, were shown to have lower, but still meaningful, contributions to the prediction equation, even though the best predictors such as cholesterol and BMI, both these two variables explained a considerable percentage of the variation in CHD.

Over the past two decades, numerous prediction models have been developed, which mathematically combine multiple predictors to estimate the risk of developing CVD, such as Framingham, SCORE, and QRISK models [22]. One of the most popular methods for predicting coronary heart disease (CHD) is the Framingham Risk Score, which includes various multivariate risk scores to identify those at highest risk for developing CVD [23]. The researchers can better understand the predictors, using risk factor categories for prediction of CHD through Framingham Heart Study examination [24]. With emphasis on the relative contributions of independent variables, a study used logistic regression to examine risk factors for lifestyle diseases in the youth population [25]. Exercise therapy is crucial for elderly cardiovascular disease prevention and treatment, promoting heart health, functional independence, and overall well-being. However, inconsistent results exist compared to a control group.

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

An innovative model for predicting and preventing CHD using triglyceride-glucose index (TyGI) can be used in clinical practice, as a highly valuable index, but further studies are needed to validate the findings [26]. Even though Machine Learning-Based Predictive Models for Detection of CHD are developed, it still presents a significant global health challenge that emphasizes the critical need for developing accurate and more effective detection methods [27]. Although many prediction studies have developed, additional research is required to find if the combination of several characteristics, such as high LDL cholesterol and a sedentary lifestyle, may potentially affect the overall risk of CHD. An automated diagnostic system for heart disease prediction, utilizing a $\chi 2$ statistical model and an optimized deep neural network, achieves a prediction accuracy of 93.33% [28]. CHD mortality rates have decreased in western countries, affecting older adults. Effective prevention strategies and risk identification are needed. Variation in cardiac autonomic modulation may be due to healthy patient exercise, beta-blocker use, and coronary angioplasty. Even though different models are widely used to predict CHD, there are still issues with improving the model's generalizability and accuracy. Multiple regression, however, continues to be a fundamental component of CHD risk prediction because of its ease of use, interpretability, and capacity to offer practical insights for clinical judgement.

5. Conclusion

The study on predicting coronary heart disease using multiple regression method highlights the importance of integrating multifactorial risk assessment for early detection and prevention, highlighting the complex interplay of risk factors. The study uses multiple regression to predict coronary heart disease in Albania, a developing country with a growing cardiovascular disease burden. It analyzes risk factors like age, cholesterol, blood pressure, and lifestyle choices, aiming to identify local patterns and develop targeted interventions. Significant relationships between responder variables and predictor factors in a multiple linear function are identified using multiple regression analysis. Using this model, the results have determined the important factors influencing coronary heart disease. By applying the MLR model to the patient's data, we were able to determine the optimum model for this approach by comparing the predictors and estimating the risk of coronary heart disease. Our model discovered that a higher risk of coronary heart disease (CHD) was closely associated with both total cholesterol and BMI. The multiple regression model showed good accuracy in predicting CHD risk, but its performance may be affected by data quality, sample population biases, and linear relationships. It can be used to develop risk assessment tools, guide personalized prevention strategies, and ensure relevance in cardiovascular disease research.

Limitations of the study

The multiple regression model offers valuable insights into factors contributing to coronary heart disease, but several limitations must be acknowledged. Research on cardiovascular health in Albania is limited, highlighting a knowledge gap. Multiple regression models' predictive power depends on population characteristics, and findings from other countries may not accurately reflect Albania's situation. Studying within Albania can help develop preventative strategies, improve resource allocation, and reduce cardiovascular health burden in the country.

Impact of the study

Multiple regression methods are crucial for predicting coronary heart disease (CHD) in Albania, a developing nation with a growing burden of cardiovascular diseases. By identifying local risk factors like dietary habits, smoking prevalence, and genetic predispositions, researchers can develop targeted prevention and intervention strategies. This approach also helps identify areas where standard models may fall short, such as non-linear relationships or non-global variables. The evaluation could lead to the refinement of existing methods, including alternative statistical approaches and advanced machine learning techniques, to enhance predictive accuracy. Innovation could involve adapting regression to Albanian data, identifying novel risk factors, and developing unique algorithms.

Acknowledgments

The authors acknowledge the group of the physicians, nurses and all health personnel of the Polyclinic of Specialty Health Center No. 2 in Tirana city for their assistance and support during the data collection process.

Declarations

Author Contribution: The authors take all the responsibilities of the paper.

Funding statement

This research was unfounded. All the data are taken from the Polyclinic of Specialty Health Center No. 2 in Tirana, Albania. No author has any financial interest or received any financial benefit from this research.

Conflict of interest

The authors declare no conflict of interest.

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

References

[1] R. López-Osca, G. López-García, A. Granero-Gallegos, and M. Carrasco Poyatos, "Psychophysiological profile of ischemic patients according to their resting heart rate variability: one step closer to training individualization," *Retos*, vol. 62, pp. 196–204, 2025. https://doi.org/10.47197/retos.v62.107794

[2] C. H. Chu, H. M. Shih, S. H. Yu et al., "Risk factors for sudden cardiac arrest in patients with ST-segment elevation myocardial infarction: a retrospective cohort study," *BMC Emerg. Med.*, vol. 22, p. 169, 2022. https://doi.org/10.1186/s12873-022-00732-3

[3] J. L. Liu, N. Maniadakis, A. Gray, and M. Rayner, "The economic burden of coronary heart disease in the UK," *Heart*, vol. 88, no. 6, pp. 597–603, 2002. https://doi.org/10.1136/heart.88.6.597

[4] A. Henderson, "Coronary heart disease: Overview," *The Lancet*, vol. 348, no. Suppl 1, pp. S1-S4, 1996. DOI: 10.1016/S0140-6736(96)98001-0

[5] G. M. Blue, E. P. Kirk, G. F. Sholler, R. P. Harvey, and D. S. Winlaw, "Congenital heart disease: current knowledge about causes and inheritance," *Med. J. Aust.*, vol. 197, no. 3, pp. 155–159, 2012. https://doi.org/10.5694/mja12.10811

[6] R. Hajar, "Risk Factors for Coronary Artery Disease: Historical Perspectives," *Heart Views*, vol. 18, no. 3, pp. 109–114, 2017. https://doi.org/10.4103/HEARTVIEWS.HEARTVIEWS_106_17

[7] L. E. Chambless, C. P. Cummiskey, and G. Cui, "Several methods to assess improvement in risk prediction models: Extension to survival analysis," *Statist. Med.*, vol. 30, no. 1, pp. 22-38, 2010. https://doi.org/10.1002/sim.4026

[8] A. Bendo and F. Mara, "An Experimental Model on Multiple Regression Analysis in CMJ and SJ Jump Tests on 10-14 Years Old Players of Tirana Football Club," *Int. J. Hum. Mov. Sports Sci.*, vol. 8, no. 5, pp. 292-297, 2020. DOI: 10.13189/saj.2020.080518

[9] S. W. Lee, "Regression analysis for continuous independent variables in medical research: statistical standard and guideline of Life Cycle Committee," *Life Cycle*, vol. 2, p. e3, pp. 1-8, 2022. https://doi.org/10.54724/lc.2022.e3

[10] E. Maraj and S. Kuka, "Prediction of Coronary Heart Disease Using Fuzzy Logic: Case Study in Albania," in *2022 Int. Conf. Electr., Comput. Energy Technol. (ICECET)*, 2022, pp. 1-6. DOI: 10.1109/ICECET55527.2022.9872569

[11] A. Bendo and F. Brovina, "A statistical model using multiple regression analysis to predict equilibrium and sway index," *J. Phys. Educ. Sport*, vol. 24, no. 6, pp. 1446-1456, 2024. DOI:10.7752/jpes.2024.06164

[12] Q. Kecheng, "Research on linear regression algorithms," *MATEC Web Conf.*, vol. 395, p. 01046, 2024. https://doi.org/10.1051/matecconf/202439501046

[13] H. Habehh and S. Gohel, "Machine Learning in Healthcare," *Curr. Genomics*, vol. 22, no. 4, pp. 291–300, 2021. https://doi.org/10.2174/1389202922666210705124359

[14] M. Azhari and F. Fitriani, "Coronary Heart Disease Risk Prediction Using Binary Logistic Regression Based on Principal Component Analysis," *ENTHUSIASTIC Int. J. Stat. Data Sci.*, vol. 2, no. 1, pp. 47-55, 2022. https://doi.org/10.20885/enthusiastic.vol2.iss1.art6

[15] B. Kumar and U. Priyadarsini, "Accuracy Analysis of Heart Disease Prediction using Logistic Regression in Comparison with the Linear Regression Algorithm," *J. Pharm. Neg. Results*, vol. 13, no. 4, pp. 1666-1672, 2022. https://doi.org/10.47750/pnr.2022.13.S04.199

[16] S. Darroudi et al., "Multivariate linear regression to predict association of non-invasive arterial stiffness with cardiovascular events," *ESC Heart Fail*, pp. 1-10, 2024. https://doi.org/10.1002/ehf2.15077

[17] A. Mohanavel and J. Mathew, "Heart Disease Analysis using Multiple Linear Regression," *Int. J. Eng. Res. Technol.*, vol. 10, no. 9, pp. 718-721, 2021. DOI: 10.17577/IJERTV10IS090248

[18] V. Chavan, N. Doaj, and N. Vaswani, "Predicting Coronary Heart Disease using Various Regression Analysis," *Int. J. Innov. Sci. Res. Technol.*, vol. 9, no. 3, pp. 2872-2877, 2024. DOI: https://doi.org/10.38124/ijisrt/IJISRT24MAR2107

[19] N. Azdaki et al., "Which risk factor best predicts coronary artery disease using artificial neural network method?," *BMC Med. Inform. Decis. Mak.*, pp. 1-12, 2024. https://doi.org/10.1186/s12911-024-02442-1

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025

[20] S. Tao et al., "Development and validation of a clinical prediction model for detecting coronary heart disease in middle-aged and elderly people: a diagnostic study," *Eur. J. Med. Res.*, pp. 1-13, 2023. https://doi.org/10.1186/s40001-023-01233-0

[21] H. Wu, "Using logistic regression model to predict the future coronary heart disease," *Theor. Nat. Sci.*, vol. 50, pp. 83-90, 2024. DOI:10.54254/2753-8818/50/2024AU0147

[22] J. A. A. G. Damen et al., "Prediction models for cardiovascular disease risk in the general population: systematic review," *BMJ*, vol. 353, p. i2416, 2016. doi: https://doi.org/10.1136/bmj.i2416

[23] A. Bitton and T. A. Gaziano, "The Framingham Heart Study's impact on global risk assessment," *Prog. Cardiovasc. Dis.*, vol. 53, no. 1, pp. 68–78, 2010. https://doi.org/10.1016/j.pcad.2010.04.001

[24] P. W. Wilson et al., "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998. https://doi.org/10.1161/01.cir.97.18.1837

[25] B. Ismail and M. Anil, "Regression methods for analysing the risk factors for a lifestyle disease among the young population of India," *Indian Heart J.*, vol. 66, no. 6, pp. 587–592, 2014. https://doi.org/10.1016/j.ihj.2014.05.027

[26] S. R. Mirjalili et al., "An innovative model for predicting coronary heart disease using triglyceride-glucose index: a machine learning-based cohort study," *Cardiovasc. Diabetol*, vol. 22, p. 200, 2023. https://doi.org/10.1186/s12933-023-01939-9

[27] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics*, vol. 14, no. 2, p. 144, 2024. https://doi.org/10.3390/diagnostics14020144

[28] L. A. Ali et al., "An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network," *IEEE Access*, vol. 7, pp. 35 623-35 631, 2019. DOI:10.1109/ACCESS.2019.2904800

DOI: https://doi.org/10.54216/FPA.200210
Received: February 05, 2025 Revised: April 12, 2025 Accepted: June 05, 2025