e-ISSN: 0975-1556, p-ISSN:2820-2643

Available online on www.ijpcr.com

International Journal of Pharmaceutical and Clinical Research 2022; 14 (6); 764-771

Original Research Article

Evaluation of Risk Factors for Ten-Year Coronary Heart Disease using Logistic Regression Modeling

Kshitij Darshan Shah[1], Aruna Pancharia[2], Hemendra Bamaniya[3], Anshul Sharma[4], Sheshav Somani[5], Amit Tak[6]

[1]MBBS, Orthopaedics, Medical Officer, Arthrocare Hospital, Ahmedabad, Gujarat

[2]MD, Pathology, Associate Professor, RVRS Medical College, Bhilwara, Rajasthan, India

[3]MD, ENT, Professor, Ananata Institue of Medical Sciences, Rajsamand, Rajasthan, India

[4]MD, Dept of Physiology, Assistant Professor, Government Medical College, Bharatpur, Rajasthan, India

[5]MD, Dept of Physiology, Associate Professor, RVRS Medical College, Bhilwara, Rajasthan, India

[6]MD, Dept of Physiology, Assistant Professor, RVRS Medical College, Bhilwara, Rajasthan, India

Received: 10-04-2022 / Revised: 20-05-2022 / Accepted: 25-05-2022
Corresponding author: Dr. Anshul Sharma
Conflict of interest: Nil

Abstract

Background: Machine learning techniques were used to predict Coronary heart disease (CHD) risk. The machine learning models might be good prediction models, but they have pitfalls and provide less casual insights. The present study evaluated risk factors for the ten-year risk of coronary heart disease using logistic regression analysis.

Material and Method: The present study was conducted at the Department of Physiology, RVRS Medical College, Bhilwara (Rajasthan, India), to evaluate risk factors of the ten-year CHD risk using logistic regression. The dataset (N = 4000) was publicly available from the ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The data were divided into two groups based on risk: 'CHD Risk' and 'No CHD risk.' Among 14 risk factors, those that significantly differed were selected for logistic regression analysis. The model with the lowest AIC was chosen. The Wald test was used to test the significance of logistic coefficients at 5%.

Results: Except for the heart rate variables, including age, cigarettes per day, total cholesterol, systolic and diastolic BP, BMI, and plasma glucose significantly differed between two groups. Similarly, males, smokers, patients with a history of stroke, hypertension, and diabetes were associated with higher cardiovascular risk. The stepwise logistic regression model used six regressors. The model's accuracy, the area under the ROC curve, sensitivity, and specificity was 85.58%, 73.7%, 8.6%, and 99.4%, respectively

Conclusion: The present study focused on identifying and weightage of risk factors and assessing their predictive ability and implications for disease prevention.

Keywords: coronary heart disease risk, logistic coefficients, machine learning, prediction

Introduction

Coronary Heart Disease, also known as coronary artery disease, is the leading cause of death in the United States.[1] As per World Health Organization, in 2019, 32% of all global deaths were cardiovascular deaths, and among them, 85% were due to heart attack and stroke. [2]

Atherosclerosis is the primary cause of coronary artery disease. The pathogenesis of atherosclerosis involves correlative processes such as lipid disturbances, thrombosis, inflammation, vascular smooth cell activation, remodeling, platelet activation, endothelial dysfunction, oxidative stress, altered matrix metabolism, and genetic factors. [3] The association between premature occurrence of CAD and chromosome 9p21.3 is shown in genome-wide association studies. The disease is associated with multiple risk factors, including diabetes mellitus, smoking, hyperlipidemia, obesity, homocystinuria, and psychosocial stress. [4] Cardiovascular dataset stored in large repositories, transformed by statistical and machine learning methods, helps in the prediction and prevention of cardiovascular diseases. Machine learning is a type of artificial intelligence wherein patterns in datasets train complex statistical algorithms. [5,6,7] Moreover, deep learning as a branch of machine learning combines statistics, computer science, and decision theory to take a clinical decision in medicine to assist in disease diagnosis and disease phenotyping. [8] Datasets have been used jointly by the American College of Cardiology Foundation (ACCF) and the American Heart Association to frame evidence-based strategies for treating cardiovascular diseases since 1980. [9] The machine learning models might be good prediction models, but they have pitfalls and provide less casual insights. The present study aimed to evaluate risk factors for the ten-year risk of coronary heart disease using logistic regression analysis. [10] The study emphasizes the identification and weightage of risk factors, having implications for coronary heart disease prevention.

Material and Methods:

The present study evaluated the ten-year risk of coronary heart disease using logistic regression modeling. The dataset was publicly available from the ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. [11] The data include 4000 observations and 14 attributes. The demographic attributes were gender and age (in years). The behavioral attributes include current smoking habits and the number of cigarettes per day. The medical risk factors consist of medication for blood pressure (MedsBP), history of stroke, history of hypertension, history of diabetes, total cholesterol, systolic and diastolic blood pressure (in mmHg), body mass index (kilogram per square meter), heart rate (beats per minute) and plasma glucose (mg per deciliter). The target (or predicted) variable was the ten-year risk of coronary heart disease (CHD) with binary levels: '1' for the presence of 10 year CHD risk (‘CHD Risk’) and '0' for the absence of 10 year CHD risk (‘No CHD Risk’).

Data analysis:

Researchers compared various predictors in two predicted groups based on the ten-year CHD risk: ‘CHD Risk' and 'No CHD Risk.' The predictors found significantly differed in the two groups were selected for further analysis. The selected predictors were used to run a stepwise logistic regression model. The best model was selected using Akaike Information Criterion (AIC). The best model was one with a minimum AIC value. The performance metrics of the model, including area under the receiver operator characteristic (ROC) curve, sensitivity, specificity, and accuracy, were calculated (JASP Team, 2019). The success was defined when the dependent variable took the value '1' (‘CHD Risk’). The coronary heart disease risk was calculated using the following equation:

$$p = \frac{1}{1 + e^{-\hat{\lambda}}}$$

where p is the probability of the outcome, and

$$\hat{\lambda} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k$$

for K regressors, where $\hat{\lambda}$ is the estimate of λ and and $b_0, b_1, b_2, \dots b_k$ were estimates of logistic coefficients. [12]

Statistical analysis

The continuous variables were expressed as mean (SD) or median (IQR) and compared using the student or non-parametric test based on whether the data followed assumptions of normality and equality of variances. Similarly, the categorical variables were expressed as proportions and compared using the Chi-squared test. The Wald test was used to test the significance of logistic coefficients. The best model was selected using Akaike Information Criterion and tested with the chi-squared test. The level of statistical significance was considered at 5%. Statistical analysis was performed using JASP software version 0.1.16.0 (JASP Team, 2019) [13] and MATLAB 2019a (MATLAB Team) [14]

Results

The comparison of quantitative variables between the two groups (‘CHD Risk’ and ‘No CHD Risk’) showed significantly higher age [W =473036.5; p < 0.001], cigarettes per day [W = 667949.5; p = 0.001], total Cholesterol [W =610705.5; p < 0.001], systolic BP [W =506595.5; p < 0.001], diastolic BP [W = 590483; p < 0.001], BMI [W = 645233.5; p < 0.001] and glucose level [W = 55659.5; p < 0.001] in ‘CHD Risk’ group. However, heart rate did not differ significantly between the two groups. (Table 1)

Table 1: Comparison of quantitative variables between two groups based on the ten-year CHD risk: 'CHD Risk' and 'No CHD Risk.'

Variables		N	Median	IQR	W	p
Age	No CHD Risk	2879	48	13	473036.5	< .001
	CHD Risk	511	55	13		
Cigarettes per day	No CHD Risk	2858	0	20	667949.5	0.001
	CHD Risk	510	4	20		
Total cholesterol	No CHD Risk	2848	232	57	610705.5	< .001
	CHD Risk	504	243	58		
Systolic BP	No CHD Risk	2879	127	25	506595.5	< .001
	CHD Risk	511	139	32		
Diastolic BP	No CHD Risk	2879	81	14	590483	< .001
	CHD Risk	511	85	17		
BMI	No CHD Risk	2872	25.23	4.96	645233.5	< .001
	CHD Risk	504	26.19	5.603		
Heart Rate	No CHD Risk	2879	75	15	715481	0.358
	CHD Risk	510	75	16		
Glucose	No CHD Risk	2614	78	15	556569.5	< .001
	CHD Risk	472	79.5	20		

The males showed significantly higher CHD Risk [□^{2} = 24.29; p < 0.001] compared to females. Smokers showed higher CHD risk [□^{2} = 3.95; p = 0.047]. Similarly, history of stroke [□^{2} = 15.97; p < 0.001], hypertension [□^{2} = 94.03; p < 0.001], and diabetes [□^{2} = 36.44; p < 0.001]. were associated with higher cardiovascular risk. (Table 2)

Table 2: Comparison of categorical variables between the two groups based on the ten-year CHD risk: 'CHD Risk' and 'No CHD Risk.'

Attributes	Levels	No CHD Risk Count (%)	CHD Risk Count (%)	Chi-square	p
Gender	**Female**	1684(58.493)	239(46.771)	24.29	< .001
	Male	1195(41.507)	272(53.229)		
Education	1	1135(40.463)	256(51.406)	22.16	< 0.001
	2	872(31.087)	118(23.695)		
	3	479(17.077)	70(14.056)		
	4	319(11.373)	54(10.843)		
Smoking	**No**	1467(50.955)	236(46.184)	3.952	0.047
	Yes	1412(49.045)	275(53.816)		
Medication for BP	No	2775(97.643)	471(93.452)	25.92	< 0.001
	Yes	67(2.357)	33(6.548)		
History of Stroke	No	2867(99.583)	501(98.043)	15.966	< .001
	Yes	12(0.417)	10(1.957)		
History of Hypertension	No	2065(71.726)	256(50.098)	94.029	< .001
	Yes	814(28.274)	255(49.902)		
History of Diabetes	No	2825(98.124)	478(93.542)	36.442	< .001
	Yes	54(1.876)	33(6.458)		

The best model chosen by the stepwise logistic regression has six risk factors, which is given below (Table 3)

Table 3: Showed the Logistic regression model estimates for predicting the ten-year coronary heart disease risk.

					Wald Test		
Parameter	**Estimate**	**Standard Error**	**Odds Ratio**	**z**	**Wald Statistic**	**df**	**p**
(Intercept)	-9.292	0.533	9.21E-05	-17.439	304.123	1	< .001
Age	0.066	0.007	1.069	9.177	84.212	1	< .001
Systolic BP	0.016	0.002	1.016	6.714	45.081	1	< .001
Cigarettes per day	0.022	0.005	1.023	4.758	22.643	1	< .001
Glucose	0.009	0.002	1.009	4.64	21.526	1	< .001
Gender (M)	0.489	0.12	1.63	4.06	16.48	1	< .001
Total Cholesterol	0.003	0.001	1.003	2.544	6.472	1	0.011

$$\begin{aligned}\hat{\lambda} = -9.29 + 0.066 \times \text{age} + 0.016 \times \text{Systolic BP} + 0.022\ \times \text{Cigarettes per day} \\ + 0.009 \times [\text{glucose}] + 0.0489\ \times \text{gender (M)} + 0.003 \times [\text{total cholesterol}]\end{aligned}$$

The model's accuracy as calculated from the confusion matrix was 85.58%. (Table 4).

Table 4: Confusion Matrix showed observed cases and predicted cases by the logistic regression model.

	Predicted		
Observed	No CHD Risk	CHD Risk	% Correct
No CHD Risk	2467	16	99.356
CHR Risk	406	38	8.559
Overall % Correct			85.583

The receiver operating curve showed that the area under the curve, sensitivity, and specificity were 73.7%, 8.6%, and 99.4%, respectively. (Figure 1)

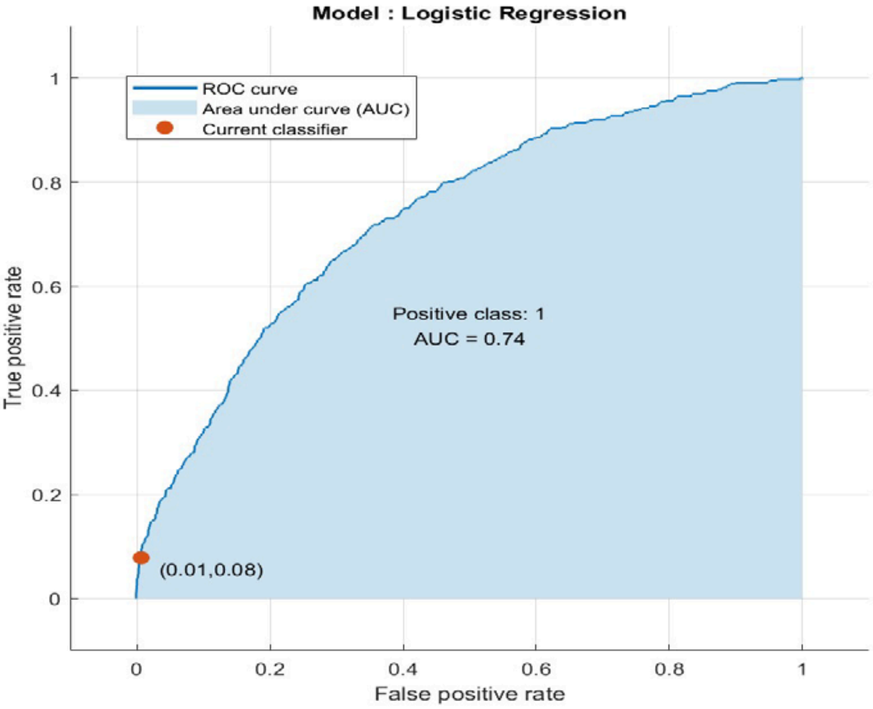

Figure 1: Area under the receiver operator characteristic (ROC) curve for logistic regression model.

Discussion

The Framingham risk score (FRS) is one of the standard tools used to predict the incidence of coronary heart disease (CHD) [15]. Four years after the Framingham Heart Study started, researchers found high cholesterol and high blood pressure levels as important CHD risk factors. The expression' risk factor' took origin in the Framingham study. Today, a risk factor is a measurable characteristic causally associated with increased disease frequency and is a significant independent predictor of an increased risk of presenting with the disease. [16] The present study developed a logistic regression model to predict the ten-year risk of CHD using the basic set of variables as regressors. All the predictors showed significant differences between the two groups except for heart rate. The regressors of the logistic regression model in descending order of weightage include age, gender, cigarettes per day, systolic blood pressure, plasma glucose, and total cholesterol. The model showed high accuracy, AUC, and specificity of 85.58%, 73.7%, and 99.4%. However, the sensitivity was very low.

With the advent of modern machine learning tools, the CHD risk algorithms showed increased accuracy. Risk scores for the prediction of coronary heart disease (CHD) have greatly improved in the past 30 years. [17] A meta-analysis by Krittinawong et al., including 344 studies, found boosting algorithms had a pooled area under the curve (AUC) of 0.88 (95% CI 0.84–0.91) prediction of coronary artery disease. [18] Johri et al. found ML-based system prediction was superior to conventional statistical methods to predict coronary artery disease using carotid plaque characteristics on 459 participants. Baseline plaque characteristics such as carotid intima-media thickness (cIMT), maximum plaque height (MPH), total plaque area (TPA), and intraplaque neovascularization (IPN) were measured. Researchers compared two ML-based algorithms - random forest (RF) and random survival forest (RSF) with (i) univariate and multivariate CAD prediction using AUC and (ii) Cox proportional hazard model for cardiovascular event prediction using the concordance index (c-index). CAD and carotid plaque characteristics were significantly associated [cIMT (odds ratio (OR) = 1.49, p = 0.03), MPH (OR = 2.44, p < 0.0001), TPA (OR = 1.61, p < 0.0001), and IPN (OR = 2.78, p < 0.0001)]. IPN alone reported significant CV event prediction (hazard ratio = 1.24, p < 0.0001).(23) However, using a basic set of variables, including age, systolic blood pressure, smoking, hypertension, exercise, body mass index, diabetes, and family history, can potentially be self-administered. Damel et al. systematically reviewed the Framingham risk models and pooled cohort equations, especially Framingham Wilson 1998, Framingham ATP III 2002, and PCE 2013, which are widely used for predicting the 10-year risk of developing coronary heart disease. Researchers found all the models overestimate the 10-year risk of CHD and CVD (pooled observed versus expected ratio ranged from 0.58 (95% CI 0.43-0.73; Wilson men) to 0.79 (95% CI 0.60-0.97; ATP III women). [19] Hippseley et al. developed QRISK cardiovascular disease risk algorithm and compared it with the Framingham score recommended by the National Institute of Health and Clinical Excellence (NICE). [20] Anderson et al. further developed the prediction equation where 5573 subjects aged 30-74 years were taken from the Framingham Heart Study. The model stressed the importance of multiple risk factors in prediction, including blood pressure, total cholesterol, high-density lipoprotein cholesterol, smoking, glucose intolerance, and left ventricular hypertrophy. Researchers developed a model that can predict outcomes for different lengths of time. [21] The SCORE project initiated by Conroy et al. developed a risk scoring system representing 2.7 million person-years of follow-up. Researchers used the Weibull model to calculate the ten-year risk of fatal cardiovascular disease. They developed two parallel estimation models based on total cholesterol and the other on total cholesterol/HDL cholesterol ratio. The areas under ROC curves ranged from 0.71 to 0.84. [22,23] The association of risk factors with CHD was further emphasized by Wilson et al. in a study consisting of 2489 men and 2856 women aged 30 to 74 years. Similar to the present study, they examined the association between blood pressure and cholesterol categories with CHD risk. They also compared the discrimination properties of the categorical and continuous variables approach. After 12 years of follow-up, researchers found a significant association of CHD with blood pressure categories, total Cholesterol, LDL cholesterol, and HDL cholesterol (all P, 0.001). The accuracy of the categorical approach was found comparable to the use of continuous variables.[24] The present study emphasizes identifying and weightage risk factors, having implications for coronary heart disease prevention. [25]

Conclusion: The present study covers multiple facets of disease prediction and describes important insights into coronary heart disease risk factors. Researchers focused on identifying and weightage of risk factors and assessing their predictive ability and implications for disease prevention.

References

1. Coronary Heart Disease | NHLBI, NIH [Internet]. [cited 2022 Mar 22]. Available from: https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease
2. Cardiovascular diseases (CVDs) [Internet]. [cited 2022 Mar 26]. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
3. Shao C, Wang J, Tian J, Tang Y da. Coronary Artery Disease: From Mechanism to Clinical Practice. Adv Exp Med Biol [Internet]. 2020 [cited 2022 Mar 22]; 1177:1–36. Available from: https://link.springer.com/chapter/10.1007/978-981-15-2517-9_1
4. Malakar AK, Choudhury D, Halder B, Paul P, Uddin A, Chakraborty S. A review on coronary artery disease, its risk factors, and therapeutics. J Cell Physiol [Internet]. 2019 Oct 1 [cited 2022 Mar 22];234(10):16812–23. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/jcp.28350
5. J Hippisley-Cox CCPB. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. BMJ (Clinical Res ed). 2017;357: j2099.
6. Ranka S, Reddy M, Noheria A. Artificial intelligence in cardiovascular medicine. Curr Opin Cardiol. 2021 Jan 1;36(1):26–35.
7. Bhandari S, Singh Shaktawat A, Tak A, Patel B, Gupta J, Gupta K, et al. Independent Role of CT Chest Scan in COVID-19 Prognosis: Evidence From the Machine Learning Classification (1) (2) (3) (4) (5) (6) (7). Scr Med. 2021;52(4):273–81.
8. C Krittanawong KJRR. Deep learning for cardiovascular medicine: A practical primer. Eur Hear J. 2019; 40:2058–73.
9. P Greenland JAGB. 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation. 2010;122: e584-636.
10. Bhandari S, Shaktawat AS, Tak A, Patel B, Shukla J, Singhal S, et al. Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters. Ibnosina J Med Biomed Sci [Internet]. 2020 [cited 2022 Mar 26];12(2):123. Available from: http://www.ijmbs.org/article.asp?issn=1947-489X;year=2020;volume=12;issue=2;spage=123;epage=129;aulast=Bhandari
11. Cardiovascular Study Dataset | Kaggle [Internet]. [cited 2022 Mar 26]. Available from: https://www.kaggle.com/datasets/christofel04/cardiovascular-study-dataset-predict-heart-disea
12. Abhaya I, Malhotra RK. Medical Statistics [Internet]. Fourth. CRC Press, Taylor & Francis Group; 2018. 470–71 p. Available from: www.taylorandfrancis.com
13. JASP Team. JASP (Version 0.16.1) [Computer software] [Internet]. 2022. Available from: https://jasp-stats.org/
14. Team MATLAB. Statistics and Machine Learning Toolbox. Natick, Massachusetts: The Mathworks Inc; 2019.
15. Nishimura K, Okamura T, Watanabe M, Nakai M, Takegami M,

Higashiyama A, et al. Predicting coronary heart disease using risk factor categories for a Japanese urban population, and comparison with the framingham risk score: the suita study. J Atheroscler Thromb [Internet]. 2014 [cited 2022 Mar 26];21(8):784–98. Available from: https://pubmed.ncbi.nlm.nih.gov/24671110/

16. O'Donnell CJ, Elosua R. [Cardiovascular risk factors. Insights from Framingham Heart Study]. Rev Esp Cardiol [Internet]. 2008 [cited 2022 Mar 26];61(3):299–310. Available from: https://pubmed.ncbi.nlm.nih.gov/18361904/
17. Wilson PWF. Risk scores for prediction of coronary heart disease: an update. Endocrinol Metab Clin North Am [Internet]. 2009 Mar [cited 2022 Mar 26];38(1):33–44. Available from: https://pubmed.ncbi.nlm.nih.gov/19217511/
18. Machine learning prediction in cardiovascular diseases: a meta-analysis | Scientific Reports. [cited 2022 Mar 22]; Available from: https://www.nature.com/articles/s41598-020-72685-1
19. JA Damen RPPH. Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: A systematic review and meta-analysis. BMC Med. 2019; 17:109.
20. J Hippisley-Cox CCYV. Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. BMJ (Clinical Res ed). 2008; 336:1475–82.
21. Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. Am Heart J. 1991;121(1 PART 2):293–8.
22. RM Conroy KPAF. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. Eur Hear J. 2003; 24:987–1003.
23. Johri AM, Mantella LE, Jamthikar AD, Saba L, Laird JR, Suri JS. Role of artificial intelligence in cardiovascular risk prediction and outcomes: comparison of machine-learning and conventional statistical approaches for the analysis of carotid ultrasound features and intra-plaque neovascularization. Int J Cardiovasc Imaging. 2021 Nov 1;37(11):3145–56.
24. PW Wilson RDDLABHSWK. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97:1837–47.
25. Margute, T. G., Ferreira, P. C., Almeida, I. M. M., Denardin, C., Silva, T. Q. M. da., Margute , T. G., Maione, M. S., Rossato, A. R., & Santos, I. F. dos. Use of tricyclic antidepressants in trigeminal neuralgia. Journal of Medical Research and Health Sciences, 2022:5(5), 2008–2012.