https://doi.org/10.54361/ajmas.2581025

Original article

Utilizing Binary Logistic Regression Models for Predicting the Risk Factors of Heart Disease

Kasem Farag[1], Eman Shuaeib[2], Husam Aqel[1], Marfoua Ali[3*]

[1]*Department of Mathematical, Faculty of Science, Omar Al-Mukhtar University, EL-Beyda, Libya.*
[2]*Department of Pre- economic, Faculty of Economy, Omar Al-Mukhtar University, EL-Beyda, Libya.*
[3]*Department of Zoology, Faculty of Science, Omar AL-Mokhtar University, EL-Beyda, Libya.*
Corresponding Email. *marfouas@yahoo.com*

Abstract
With rising yearly death rates, cardiovascular disease (CVD) continues to be a major global health concern. Effective prevention and management of risk factors depend on an understanding of how they interact. Predicting the primary causes of heart disorders, examining the impact of each of these elements, and organizing them according to model choice and classification capabilities were the objectives of this work. To forecast the probability of CVD, a logistic regression model was created. The parameters impacting infection risk were characterized using descriptive analysis. Data were collected from 82 patients at Benghazi Cardiac Centre, comprising 41 infected patients and 41 control subjects selected using the Stephan Thompson Equation. Binary logistic regression analysis, conducted using SPSS 23, identified four significant predictors of cardiac infection: diabetes, BMI, cholesterol level, and blood pressure. The model demonstrated an 89% accuracy in classifying patients. Further research is recommended to investigate the impact of additional factors on cardiac infection risk in this population.

Keywords. Heart Disease, Risk Factor, Binary Logistic Regression Technique.

Introduction

A major global health burden is posed by cardiovascular disease (CVD). The World Health Organization (WHO) estimates that cardiovascular disease (CVD) caused 17.9 million deaths in 2016, accounting for almost 30% of all fatalities worldwide [1]. With treatment expenses accounting for about 4% of yearly healthcare budgets, the high mortality rate nearly 55% of those with heart disease die within three years of being diagnosed highlights the condition's substantial human and financial effect [2]. The effectiveness of good lifestyle choices in preventing and controlling heart disease has been shown in numerous research [3]. Thus, it is essential to recognize and comprehend the risk factors linked to this condition in order to create and carry out efficient treatment and preventive measures. Regarding studies on cardiac disease, the binary A particular type of regression analysis in which the outcome variable is divided into two or more categories is called logistic regression analysis. In many cases, the link between the independent variables and the response variable is not sufficiently clarified by the traditional regression method [4].

The binary logistic regression model, also known as the logit model, is the most prevalent type of analysis where the response variable has only two possible values. Nonetheless, the technique can also be expanded when there are three or more types of dependent variables, such as polychromous or multinomial. Like any other regression model, the use of logistic regression can primarily be categorized into two groups based on the research inquiry: either identifying the relationship between dependent and independent variables or developing a predictive model that aids in forecasting [5, 6]. This study sought to predict the primary causes of heart disease by employing binary logistic regression models to examine the impact of each of these factors and to group them according to the model's preference and classification and differentiation capabilities.

Methods

Study design and setting

A prospective cross-sectional study design was used in this investigation. The Benghazi Cardiac Center's medical records were examined in order to gather the data. The time frame for inclusion was January 2023–April 2023.

Data collection

The Benghazi Cardiac Center in Libya, a trustworthy source, provided the data for the 82 cases. There were 41 cases, or patients with heart disease, and 41 controls, or persons with just chest pain. There was one goal variable and five parameters in the dataset. A binary class label, with "1" denoting the presence of heart disease and "0" denoting its absence, was used to represent the target variable, which was the patient's presence of heart disease. Furthermore, the current study only included gender, diabetes, and Body mass index (BMI), Systolic Blood Pressure (SBP) and Cholesterol level among subjects.

Received*: 14-11-2024 -* ***Accepted****: 16-01-2025 -* ***Published****: 22-01-2025*

Data analysis

Two statistical analyses have been used to identify and evaluate the risk factors for heart disease. Exploratory Data Analysis (EDA) was conducted to obtain a more profound comprehension of the data and its characteristics. The EDA process involved descriptive statistics and the logistic regression analysis utilized to evaluate above parameters.

Results and discussion

Descriptive statistics

Table 1 illustrated general characteristics of the total sample. Data shows the distribution of genders in a sample of 82 instances. Females were represented by 52.4% and males 47.6% of the total sample. Distribution of cholesterol levels was classified among cases. This data suggests that a majority of individuals in the sample (69.5%) have normal cholesterol levels, while a smaller proportion (30.5%) have high cholesterol levels. Presence of diabetes of total cases was determined.

The results indicate that 44 individuals (53.7%) in the sample are diabetic, while 38 individuals (46.3%) are non-diabetic. This suggests that diabetes is a relatively common condition within this particular group. Distribution of blood pressure among cases, which shows the distribution of systolic blood pressure (SBP) in a sample of 82 individuals. The data indicates that 32 individuals (39%) have diseased systolic blood pressure, while 50 individuals (61%) have non-diseased systolic blood pressure. This suggests that a higher proportion of individuals in this sample have non-diseased SBP compared to those with diseased SBP. Another parameter that was also noticed is the distribution of body mass index (BMI) of total cases. The majority of individuals in the sample (63.4%) have a healthy weight, while a significant number (15.9%) are overweight and another 15.9% are obese. Only a small proportion (4.9%) are underweight. This suggests that a considerable portion of this population may be at risk for health problems associated with overweight and obesity.

Table 1. General characteristic of total sample.

General characteristics	Frequency	Percentage
Gender		
Female	43	52.4
Male	39	47.6
Total	82	100.0
Cholesterol levels		
Normal	57	69.5
High	25	30.5
Total	82	100.0
Presence of diabetic		
Non diabetic	38	46.3
Diabetic	44	53.7
Total	82	100.0
Systolic blood pressure (SBP)		
Diseased	32	39.0
Non diseased	50	61.0
Total	82	100.0
Body Mass Index (BMI)		
Underweight	4	4.9
Healthy weight	52	63.4
Overweight	13	15.9
Obesity	13	15.9

Logistic regression analysis results

Before performing the binary logistic regression analysis, the multicollinearity of the independent variables was investigated by the variance inflation factor (VIF) for less than 10 and tolerance value greater than 0.1. Therefore, multicollinearity was undetected among the independent variables because the VIF was less than 10 as well as the values of tolerance for the predictors were all greater than 0.1. As a result, all of our independent variables could be used to fit the binary logistic model. Tables 2 revealed the VIF and tolerance for various predictor variables.

Received*: 14-11-2024 -* ***Accepted****: 16-01-2025 -* ***Published****: 22-01-2025*

Table 2. Checking for multicollinearity problems by Variance Inflation Factor (VIF).

Model	**Collinearity Statistics**	
	Tolerance	**VIF**
Cholesterol level	.699	1.431
Diabetes	.719	1.390
Blood pressure	.791	1.265
gender	.980	1.021
BMI	.888	1.126

Table 3 presents two statistical measures used to assess the fit of a logistic regression model. the Cox & Snell R Square and the Nagelkerke R Square. The Cox & Snell R Square value of .560 indicates that the model explains 56% of the variance in the dependent variable. The Nagelkerke R Square, which adjusts the Cox & Snell R Square, has a value of .746, suggesting that the model explains a substantial proportion of the variance.

Table 3. Goodness-of-fit Coefficients

Cox & Snell R Square	**Nagelkerke R Square**
.560	.746

The calculated coefficients, their S.E., and the Wald test for the entire model were displayed in Table 4. The table's data showed that the following factors were significant: BMI (p-value = 0.010 < 0.05), diabetes (p-value = 0.009 < 0.05), cholesterol (p-value = 0.004 < 0.05), and SBP (p-value = 0.007 < 0.05). On the other hand, the variables that were statistically insignificant were gender (p-value = 0.758 > 0.05).

According to the odds ratio for cholesterol level. An odds ratio of 31.054 indicates that individuals with high cholesterol are significantly more likely to develop heart disease compared to those with normal cholesterol levels. Compared to people with normal cholesterol levels, those with high cholesterol have a much higher risk of developing heart disease, as indicated by an odds ratio of 31.054. Additionally, the odds ratio for diabetes is 7.926, indicating that those with diabetes are significantly more likely to have heart disease than people without the disease. Furthermore, SBP shows that those with high blood pressure have a much higher risk of developing heart disease than people with normal blood pressure, with an odds ratio of 10.274. Males are also less likely than girls to acquire heart disease, according to an odds ratio of 0.786. Lastly, a 3.607 odds ratio shows that people with higher BMIs had a larger chance of developing heart disease than to those with lower BMI.

Table 4. Summarized results for the full model

Variables	**B**	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Cholesterol	3.436	1.207	8.097	1	.004	31.054
Diabetes	2.070	.792	6.840	1	.009	7.926
Blood pressure	2.330	.856	7.402	1	.007	10.274
Gender	-.241	.782	.095	1	.758	.786
BMI	1.283	.497	6.664	1	.010	3.607
Constant	-8.562	2.186	15.346	1	.000	.000

Table 5 displays the performance of a classification model in predicting heart disease. The matrix compares the model's predicted outcomes (heart disease or non-diseased) against the actual observed outcomes. The diagonal elements (37 and 36) represent correct predictions, while the off-diagonal elements (4 and 5) represent incorrect predictions. The "Percentage Correct" column shows the accuracy of the model for each category and overall. With an overall accuracy of 89%, the model demonstrates a relatively high level of accuracy in predicting heart disease.

Table 5. Confusion Matrix

Observed		**Predicted**		
		Heart Disease		**Correct Percentage**
		Non- Diseased	Diseased	
Heart Disease	Non diseased	37	4	90.2
	Diseased	5	36	87.8
Overall Percentage				89.0

Discussion

The results of this study, which used binary logistic regression models to predict the risk factors for heart disease, indicate that the most significant factors in defining CVD are systolic blood pressure (SBP), cholesterol, the presence of diabetes, and BMI since the p-value is significantly less than 0.05. Backward

Received*: 14-11-2024 -* ***Accepted****: 16-01-2025 -* ***Published****: 22-01-2025*

and forward-backward selection results, however, indicate that every risk factor examined aside from gender has a substantial correlation with CVD. the potential to build an appropriate model that describes the most significant explanatory factors for the development of heart disease based on the data. Blood pressure, body mass index, diabetes, and cholesterol all have very significant effects, which is consistent with medical advice [7, 8]. There is no significant effect of the gender variable on the development of gender [9]. An odds ratio was employed to make a statistical analysis of the relationship between gender, SBP, cholesterol, presence of diabetic and BMI and cardiovascular disease.

It can be found that, if all other predictor variables are held constant, the odds of getting CVD occurring increased by increasing SBP, cholesterol, presence of diabetic and BMI. These observations were in agreement with another studied found that age, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoke, alcohol intake, and physical activity are the most important factor in determining CVD [10, 11]. In a prior work, binary logistic regression and decision tree induction were tested for predicting CVD risk. Additionally, the findings indicate that systolic blood pressure, age, weight, cholesterol, and other factors are linked to the development of CVD [12], which is quite in line with the findings of the present investigation.

BMI is useful in preventing CVD, according to current research. But because this factor is not well understood. Therefore, it is unable to gather sufficient data to recommend to patients the right amount of exercise to accomplish the goal of preventing CVD. The generalization of these findings has various drawbacks notwithstanding its benefits and contributions. First off, the suggested model leaves out important societal factors that can contribute to heart disease, like socioeconomic position and current smoking status, as well as patient medical data. To improve the quality of the data and validate the findings, more research could expand this study by include new clinical, demographic, and social variables.

Conclusion

The most significant findings from the statistical analysis and its indicators applied to the study sample are cholesterol level, diabetes, blood pressure, and body mass index variables. Results recommend continuing statistical research on the disease due to its prevalence and severity in order to better understand it and find effective solutions to reduce and treat it. It also recommends conducting statistical research on the impact of other contributing factors such as genetic factors, sudden shocks, psychological state, and other factors not addressed in this study. Raising public awareness by activating the role of the media in informing people about the risks associated with disease and the importance of adopting appropriate diets and engaging in physical activities.

Acknowledgments

Heartily thankful to staff in the Benghazi Cardiac Center for co-operation during collecting samples. Our thanks are conveyed to the patients who accept in participate in our research. Many thanks to our student Omaima Saeid for her helping to collect data.

Conflict of Interest

There are no financial, personal, or professional conflicts of interest to declare.

References

1. WHO, World Health Organization. Cardiovascular diseases. Available from: https://www.who.int/cardiovascular_diseases/en/. [accessed 21 February 2019.
2. Manji RA, Witt J, Tappia PS, Jung Y, Menkis AH, Ramjiawan B. Cost–effectiveness analysis of rheumatic heart disease prevention strategies. Expert review of pharmacoeconomics & outcomes research. 2013 Dec 1;13(6):715-724.
3. Elkheshebi A, Alakhder F, Zarti S. The Effects of Non-Pharmacological Intervention in the Management of Essential Blood Pressure. AlQalam Journal of Medical and Applied Sciences. 2021 Jul 6:143-51.
4. Lea S. Multivariate Analysis II: Manifest variables analysis. Topic 4: Logistic Regression and Discriminant Analysis. University of EXETER, Department of Psychology. Available at: www. exeter. ac. uk/~ SEGLea/multivar2/diclogi. html. 1997.
5. Stoltzfus JC. Logistic regression: a brief primer. Academic emergency medicine. 2011 Oct;18(10):1099-1104.
6. Reed P, Wu Y. Logistic regression for risk factor modelling in stuttering research. Journal of fluency disorders. 2013 Jun 1;38(2):88-101.
7. Ahmed AM, Hersi A, Mashhoud W, Arafah MR, Abreu PC, Al Rowaily MA, Al-Mallah MH. Cardiovascular risk factors burden in Saudi Arabia: the Africa Middle East cardiovascular epidemiological (ACE) study. Journal of the Saudi Heart Association. 2017 Oct 1;29(4):235-243.
8. Robinson T and Garcia B. A comparative study of logistic regression and chi-square technique for identifying risk factors of heart disease. Journal of Epidemiology and Community Health, 2015, 31(1), 129-141.
9. Patel K and Kumar S. Evaluating the performance of the person's chi-square technique for identifying risk factors of heart disease in a rural Indian population. Journal of Rural Health, 2006, 27(2), 198-211.
10. Yu Y. Statistical Analysis of CVD using Binary Logistic Regression. International Conference on Biotechnology, Life Science and Medical Engineering, 2022, BLSME. Published by CSP

Received*: 14-11-2024 -* ***Accepted****: 16-01-2025 -* ***Published****: 22-01-2025*

11. Guo Y. Research on the Influencing factors of heart disease based on Binary Logistic Regression. Science and Technology of Engineering, Chemistry and Environmental Protection. 2024 Jun 6;1(7).1-7.
12. Grabauskytė I, Tamošiūnas A, Kavaliauskas M, Radišauskas R, Bernotienė G, Janilionis V. A comparison of decision tree induction with binary logistic regression for the prediction of the risk of cardiovascular diseases in adult men. Informatica. 2018 Jan 1;29(4):675-692.

المستخلص

مع ارتفاع معدلات الوفيات السنوية، لا تزال أمراض القلب والأوعية الدموية تشكل مصدر قلق صحي عالمي رئيسي. وتعتمد الوقاية الفعالة وإدارة عوامل الخطر على فهم كيفية تفاعلها. وكان التنبؤ بالأسباب الرئيسية لاضطرابات القلب، وفحص تأثير كل من هذه العناصر، وتنظيمها وفقًا لاختيار النموذج وقدرات التصنيف، من أهداف هذا العمل. وللتنبؤ باحتمالية الإصابة بأمراض القلب والأوعية الدموية، تم إنشاء نموذج الانحدار اللوجستي. وتم وصف المعلمات التي تؤثر على خطر الإصابة بالعدوى باستخدام التحليل الوصفي. وتم جمع البيانات من 82 مريضًا في مركز بنغازي للقلب، بما في ذلك 41 مريضًا مصابًا و41 شخصًا من المجموعة الضابطة تم اختيارهم باستخدام معادلة ستيفان طومسون. وقد حدد تحليل الانحدار اللوجستي الثنائي، الذي أجري باستخدام برنامجSPSS 23 ، أربعة عوامل تنبؤ مهمة للإصابة بالعدوى القلبية: مرض السكري، ومؤشر كتلة الجسم، ومستوى الكوليسترول، وضغط الدم. وأظهر النموذج دقة بنسبة 89٪ في تصنيف المرضى. ويوصى بإجراء المزيد من البحوث للتحقيق في تأثير العوامل الإضافية على خطر الإصابة بالأمراض القلبية.

***Received**: 14-11-2024 - **Accepted**: 16-01-2025 - **Published**: 22-01-2025*