الهيئة الليبية للبحث العلمي
Libyan Authority For Scientific Research

Libyan International Journal of Natural Sciences
2025
LIJNS
Peer-Reviewed Journal

(LIJNS)

Libyan International Journal of Natural Sciences

VOLUME 1 ISSUE 2 November 2025

Bi-Annual, Peer- Reviewed, and Open Accessed e-Journal

turnitin Google Scholar

Puplished by

MILJS@aonsrt.ly

Libyan International Journal of Natural Sciences (LIJNS)

Vol.1 Issue No 2 November 2025

الهيئة الليبية للبحث العلمي
Libyan Authority for Scientific Research

Application of the Binary Logistic Regression Model in Studying Coronary Artery Disease Among Patients Undergoing Catheterization

[1] *Wafa Al-Shikhe* , [1]*Faiza Farag El- khafifi* , [1*]*Intesar El-Saeiti* .
[1] *Statistics Department, Faculty of Science, University of Benghazi* .

ARTICLE INFORMATION

Article history:
Received 24 May 2025
Accepted 15 September 2025
Published 22 September 2025

Keywords:

Coronary Artery Disease (CAD); Binary Logistic Regression; Risk Factors; Predictive Modeling; ROC Curve Analysis.

ABSTRACT

Coronary artery disease (CAD) is a leading cause of morbidity and mortality worldwide, underscoring the need for reliable methods to identify key risk factors. This study applied binary logistic regression to assess the most significant predictors of CAD among 856 patients who underwent cardiac catheterization at the Benghazi Heart Center between 2020 and 2022. The independent variables examined were age, gender, diabetes, blood clots, smoking, and hypertension. Results from contingency tables and logistic regression analysis revealed that age, male gender, diabetes, and blood clots were statistically significant predictors of CAD ($p < 0.05$), with diabetes and blood clots having the highest odds ratios (OR = 2.851 and OR = 2.941, respectively). All regression methods (Enter, Forward Stepwise, Backward Stepwise) demonstrated consistent findings, with a correct classification rate of 78.2% and AUC values around 0.727, indicating good model performance. These findings confirm that binary logistic regression is an effective tool for identifying high-risk groups and can support clinical decision-making and preventive healthcare strategies.

Introduction

Researchers depend on statistical analysis to study, make sense of, and explain information or changes seen in scientific experiments. It is understood that binary logistic regression is one of the best approaches to study categorical dependent variables. This investigation applies binary logistic regression to determine what key features increase the risk of CAD among patients with cardiac catheterization. Logistic regression is a valuable tool for analyzing binary data because it can process both types of variables. Because it is used widely in healthcare, it offers a reliable way to study how explanatory factors affect whether something occurs or not. Cardiovascular diseases, including CAD, are one of the most significant worldwide health problems. In 2019, CAD caused one-third of all global deaths. The number of people who die each year from cardiovascular diseases has gone up, from 12.1 million to 18.6 million, with most of these fatalities taking place in low- and middle-income countries. Mortality is reduced if detection and intervention take place

* Corresponding author: *E*-mail addresses: entesar.el-saeiti@uob.edu.ly

early. Yet, only in about 67% of cases can medical professionals accurately predict heart disease, so better predictive methods and tools are needed urgently [1].

The coronary arteries, which carry oxygen-rich blood to the heart, are necessary for maintaining normal heart function. Channel arteries become narrow because they fill with cholesterol and other materials, causing the vessels to constrict or become blocked and threatening serious events such as heart attacks. You can lower the risk of CAD if you do not smoke, keep your blood sugar and pressure down, maintain a healthy weight, and have no family history of the disease. Figure out how these elements lead to CAD is necessary to form better ways to prevent and treat it.

This study aims to identify important factors for CAD in patients by applying binary logistic regression to their data. The goal of creating a strong predictive model is to equip healthcare workers to detect and treat CAD and help lessen its effects. The research stresses that medical research on complex conditions like CAD depends heavily on statistical methods. Past research shows that binary logistic regression can discover risk factors connected to CAD. A study in Sudan reported that logistic regression identified through logistic regression that heart disease is most strongly predicted by age, gender, and diabetes [2]. Similarly, CAD risk was estimated using logistic regression, while considering how blood pressure and smoking affect the risk. However, these studies were done under different conditions, so their results may not be easily used for Libya, particularly during the COVID-19 pandemic, which added new challenges and factors [3].

Another study [4] looked at logistic regression to see whether the presence of satellite males could be predicted near female horseshoe crabs, based on their physical characteristics. The researchers found that logistic regression helps in studying binary results, and weight was revealed to be essential for predicting outcomes. Using the Hosmer-Lemeshow test, the model was shown to fit the data well, and its predictive ability was assessed with ROC curve analysis, leading to an AUC of 0.7379. They prove that logistic regression has a wide range of uses in many research areas, including ecology and medicine, and allows researchers to analyze outcomes that have only two values.

Based on previous findings, logistic regression was further examined in [5] looked at how well logistic regression models with various parameter settings can distinguish two outcome classes. By leveraging SAS, the analysis examined sensitivity, specificity, and accuracy for predictive purposes. A concordance index of 0.738 was calculated, and this, along with the area under the ROC curve, determined that the model was able to provide reasonable classifications The study found that both the Hosmer-Lemeshow and Pearson tests indicate that deviations from the model are not significant. The model was developed by using backward elimination, suggesting only the significant main effects and interaction variables to maintain both accuracy and clarity. Identifying smoking and alcohol use as substantial predictors, it appeared that they were strongly linked to binary health outcomes. The work analyzed how minor differences in single observations could influence the outcomes of the model. The findings highlighted that selecting the right model and checking its fit is key in logistic regression, which gives essential tips for improving categorical data classification models. Within Libya, the situation is particularly critical, as cardiovascular diseases account for a staggering 37% of all deaths, making them the foremost cause of mortality in the country. A study conducted in Tripoli revealed that risk factors such as diabetes, hypertension, and smoking are major contributors to disease progression. Early detection and intervention can significantly reduce mortality. However, medical professionals can accurately predict heart disease in only about 67% of cases, highlighting the urgent need for better predictive methods and tools [1]. Research has shown that COVID-19 not only exacerbates pre-existing heart conditions but can also cause cardiovascular complications such as myocardial injury and inflammation. Patients with a history of cardiovascular diseases are more susceptible to severe outcomes when infected with COVID-19 [2].

Many international studies, and the one by [6] is an example, have also focused on how age, cholesterol, and diabetes help predict CAD. The studies [7] and [8], as well as more recent work, have confirmed that logistic regression can be used effectively to predict CAD, particularly when working on making the results as precise and distinct as possible. All of the studies emphasize the ability of statistics to aid doctors and improve methods for disease prevention. Using these key points as a base, this study hopes to add to what is known about CAD risk factors, with a special focus on COVID-19. The study's results should assist both medical experts and decision-makers in providing better care and lead to improved patient care.

Materials and Methods

Six main variables are included in this study and will be discussed in depth further on. Initially, the sample distribution was looked at by categorizing people as having or not having coronary artery disease (CAD). In addition, information about smoking, age, gender, blood clots, blood pressure, and diabetes was used to analyze the sample distribution. The variables were organized and examined using SPSS software, version 26.

Logistic Regression Model

When performing logistic regression, the dependent variable Y is a random binary value, where it has a probability p of being 1 and a probability (1 - p) of being 0. As noted by [9] and [10], the Bernoulli distribution is a discrete random variable. The logistic regression model is used to predict the likelihood of having CAD (since it is a yes or no situation) based on different independent variables. The authors chose this strategy to study binary information, which is helpful when analyzing what causes CAD in this sample.

$p_r(Y = y, p) = p^y(1 - p)^{1-y};\ y = 0{,}1$ (1)

Suppose there are n people and k different independent variables such as$(x_1, x_2, \dots x_k)$. Sometimes, independent variables are measured using numbers, categories, or features of both [11].

For individual i, it can be written as

$p_i = p(y_i = 1)$ and $1 - p_i = p(y_i = 0)$ (2)

The logistic regression technique assumes that the logistic function can explain the connection between the binary variable and the other variables.

$$p_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} \qquad (3)$$

$\beta_{(k+1)(1)}$ It is the vector of the model's parameters and the vector of the model's independent variables. The odds ratio is the ratio of the parameter pi to 1 minus pi (see [12]). We get it from:

$$odds = \frac{p_i}{1-p_i} = e^{x_i\beta} \qquad (4)$$

Probability is the ratio between the chance of an event happening and the opportunity that it does not occur. With logistic analysis, researchers can directly estimate the chance of an event taking place (Refer to [13]).

The logarithm or the principle of odds

$$ln(odds) = \ln\left(\frac{p_i}{1-p_i}\right) = e^{x_i\beta} \qquad (5)$$

Researchers often call it the logit value of p [11]. This means we use ordinary multiple regression, but with logit as the dependent variable. This formula is applied when we want to study the relationships between k different explanatory factors and the dependent variable. Probability estimation is possible because this form ensures the predicted value of the dependent variable always remains within the range 0 and 1, regardless of the amount taken from the explanatory variables [14]. Often, the logit of p is termed multiple logistic regression [15], and this is an effective way to predict binary outcomes from several explanatory factors.

Source and description of data

People with coronary artery disease (CAD) being seen at the Benghazi Heart Center from 2020 to 2022 were the subjects of the study. In total, 6,777 patients were treated during this period. 1,355 people were included in the study, as they formed a random 20%, and their sample size was approved for statistical use. The reason for this approach was that it works best for groups with slight variation among their members. The total population (6,777) was divided by the sample size we wanted (1,355), and this gave us a sampling interval of 5. We chose the first random date to start with, and every fifth case was selected in the sample. When we removed the incomplete records, the total number of participants reduced to 856, between the ages of 26 and 86.

We excluded the following from the study:
1. People who had surgery on their hearts or received pacemakers.
2. When equipment breaks down or is delayed for some reason, the performance of catheterization is delayed.
3. Persons who choose not to have catheterization performed.
4. Performing regular check-ups and echocardiograms in people with no CAD diagnosis.

Variable Description

At this point, you will mention all the variables under study (such as age, gender, smoking, diabetes, blood pressure, and blood clots).

Six variables were studied: one numerical variable, age (divided into three groups), and five qualitative variables, gender, smoking, blood pressure, diabetes, and blood clots. Each variable was represented by putting 1 for its presence and 0 for its absence. A logistic regression technique was used to assess how the independent factors play a role in CAD.

There were six key variables studied during the study:
1. For this study, age was clustered into three age groups
2. The gender variable is recorded as a dummy variable.
- ➢ 1 is Male
- ➢ 0 is assigned to Female.

3. Smoking is considered a qualitative variable that is coded as a dummy variable.
- ➢ 1 refers to being a smoker
- ➢ 0 means you do not smoke.

4. Blood Pressure (Qualitative Variable): Assigned a dummy variable:
- ➢ A diagnosis code of 1 indicates Hypertension (high blood pressure).
- ➢ Normal blood pressure is a reading of 0.

5. Diabetes was recorded as a dummy variable, with two levels:
- ➢ When there is one, the condition is called diabetes.
- ➢ When blood sugar is 0, a person is not considered to have diabetes.

6. Blood Clots is considered a dummy variable that uses binary coding.
- ➢ 1 means blood clots are present
- ➢ When there is no sign of blood clots, it is found in 0.

The variables were examined systematically using binary logistic regression. The investigators used logistic regression to determine whether these independent variables affect the likelihood of someone having coronary artery disease (CAD). This approach allows us to see which risk factors are most important for CAD and how they affect disease risk.

Ethical Approval

This study was conducted in agreement with the ethical values of the organized and national research commission and with the 1964 Helsinki Declaration and its later amendments. Ethical approval for this research was obtained from the Research Ethics Committee of the University of Benghazi, Faculty of Science. All data used in this study were collected retrospectively from medical records at the Benghazi Heart Center between 2020 and 2022. The patients' personal identifiers were removed prior to analysis to ensure anonymity and confidentiality. Since the study used anonymized secondary data, the ethics committee waived the requirement for informed consent.

Results

Relationships between categorical variables were studied using contingency tables, and frequencies and percentages revealed which associations and trends were significant.

The Hypothesis of Chi-Square Test

Null hypothesis (H_0): there is no statistically significant association between age, gender, diabetes, hypertension, and blood clot and coronary artery disease (DAC). The main points of the analyses are listed as follows.

- CAD affected men more often than it affected women; males had a prevalence of 57.4%, while females had 18.5%. Using a chi-square test, we found that males are at increased risk for CAD, compared to females (p- value < 0.05).
- CAD was more widespread among adults aged 45 to 64 than in any other age group (48.0%). Chi-square testing showed that age and CAD are strongly linked (p- value < 0.05), demonstrating that age significantly affects the risk of CAD.
- No evidence that being hypertensive significantly predicts CAD in those studied (p- value > 0.05).
- There is a meaningful relationship between diabetes and CAD (p- value < 0.05). This study points out that diabetes increases the chance of CAD.
- Those with blood clots had a greater chance of CAD compared to others (p- value < 0.05).
- It was observed that smoking is a significant cause of CAD in the population, since their association was substantial (p- value < 0.05).

With this research, we see that age, gender, diabetes, blood clots, and smoking are significant factors increasing the risk of coronary artery disease, which is very helpful when choosing proper treatment plans and treatment programs.

Binary logistic regression was used to identify the most significant predictors of CAD. The analysis was conducted using three methods: The Enter method, the Forward Stepwise method, and the Backward Stepwise method.

Significant Predictors of CAD

When using Enter, Forward Stepwise, and Backward Stepwise, the same significant predictors of CAD were found in each case:

1. The model shows that age (OR = 1.715 for Enter, 1.702 for Forward and Backward Stepwise).

2. Gender proved to be important for both the complete model and the models built by stepwise regression.

3. The model found that diabetes has a relative risk of 2.935, which decreased to 2.851 using the forward and backward stepwise method.
4. Blood Clots (OR = 2.822 for Enter on Standard Cox Regression and 2.941 on Forward and Backward Stepwise Cox Regression)
All of these predictors were significant, with p-values under 0.05. This means these factors are highly linked to the risk of CAD

Variable	Coefficient (B)	Standard Error (SE)	Wald Value	p-value	Odds Ratio (OR)
Age	0.539	0.112	23.14	< 0.001	1.715
Gender	1.350	0.245	30.36	< 0.001	3.858
Diabetes	1.077	0.231	21.68	< 0.001	2.935
Blood Clots	1.037	0.227	20.85	< 0.001	2.822
Constant	-2.153	0.356	36.57	< 0.001	0.116

Comment: table summarizes the results of a binary logistic regression model identifying key factors associated with CAD. All listed variables (Age, Gender, Diabetes, and Blood Clots) show a statistically significant association with CAD (p-value < 0.001). The Odds Ratio (OR) indicates the increased likelihood of CAD for each factor. For instance, being male increases the odds of CAD by 3.858 times compared to females, while diabetes increases the odds by 2.935 times. These findings highlight the importance of these factors in assessing CAD risk.

Model Significance

H_0: The proposed regression model is not statistically significant.

H_1: The proposed regression model is statistically significant

All models showed a low p-value, less than 0.05, with a χ^2 value of 114.969 for Enter, so we reject the null hypothesis) ,111.976 for Forward Stepwise, and 112.988 for Backward Stepwise. The results show that the models fit the data well and that the predictors found matter.

Goodness of Fit (Hosmer and Lemeshow Test)

With p-values higher than 0.05, the Hosmer and Lemeshow test revealed that our model adequately represented the data.

Predictive Accuracy

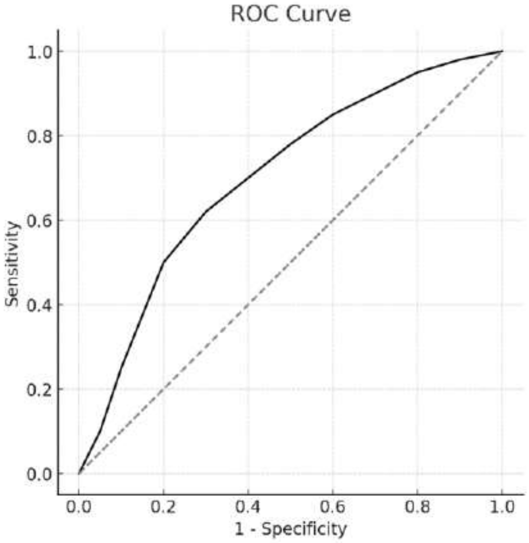
The models were able to make accurate predictions 78.2% of the time on average. Yet, it was better at predicting CAD cases where there was heart disease (95.8%) than cases where there was not (22.7%). As a result, the models can accurately detect CAD better than they can find those who do not have it.

ROC Curve Analysis

The ROC AUC values were.

The AUROC obtained was 0.727 when using the entry-to-entry method.
- Forward Stepwise Method: 0.725
- The backward stepwise method led to an AUC of 0.725.

A score of 0.727 proves that the models could tell apart CAD-positive and CAD-negative cases, as shown in Fig. 1 below:



Discussion

The study found that binary logistic regression effectively finds the main factors that predict CAD. All methods—Enter, Forward Stepwise, and Backward Stepwise—had a correct classification rate of 78.2%. This shows that the model can be trusted to identify whether CAD is present.

Although the model performed well overall, its ability to correctly identify patients who do not have coronary artery disease (CAD) was low, with a specificity of only 22.7%. This means the model often predicted that patients had CAD even when they did not. While this may seem safer in a hospital setting, it could lead to unnecessary tests, stress for patients, and extra use of resources. The discussion of this issue was limited, and more explanation is needed.

One possible reason for this result is that the data may have included more CAD-positive cases than negative ones, causing the model to focus on predicting the positive group. Another reason could be that important factors such as cholesterol levels, family history, or medications were not included in the analysis. Also, some patients without clear symptoms might have been misclassified, making it harder for the model to learn how to recognize true negative cases.
To improve the model in future studie, researchers could use balanced samples, add more medical and lifestyle variables, or try other models like machine learning methods. These steps might help the model become more accurate in detecting both CAD and non-CAD cases.

The most significant discoveries from the study are located here:

1. Gender, age, diabetes, and blood clots were the impacted variables for CAD. All three analyses found the same factors to be strongly related to CAD.
2. The likelihood of having CAD grew as age increased, with an odds ratio of 1.702. According to these results, older people tend to get coronary artery disease, in agreement with other studies on aging and heart disease.

3. Men had a higher risk of CAD than women did, so males were 3.858 times as likely to get it. The same finding has been made in other studies, which link gender with a greater risk of CAD in younger people.
4. Among patients, those with diabetes were more likely to develop CAD by nearly three times, and those with blood clots were almost three times as likely, according to the results. The results suggest that managing diabetes and monitoring blood clots help to prevent CAD.
5. Confirming Model Fit: Using the Hosmer and Lemeshow test, there was no significant gap between the actual values and the values predicted from the model. Positive cases were identified accurately (95.8%), though the model missed most of the time (77.3%) when trying to determine those who were negative. Therefore, we should look for ways to improve the model so it predicts even better.
6. According to the ROC curve analysis, the model had an area under the curve (AUC) of 0.725 and could identify positive cases about 72.5% of the time. Consequently, the model appears to be effective at forecasting CAD.
7. Same Outcome: The similar results using the Enter, Forward Stepwise, and Backward Stepwise approaches prove that the model is strong. The fact that changing the order of entry does not affect the results supports a strong statistical relationship between the independent variables and the outcome.

Conclusion

This study aimed to find the main things that cause CAD using a simple statistical model (binary logistic regression). The study successfully showed that things like being older, being male, having diabetes, and having blood clots are strongly linked to a higher chance of getting CAD. Specifically, men were 3.8 times more likely to get the disease. The risk also doubled for older people and those with diabetes or blood clots. This highlights how important these factors are for doctors when checking patients.

The model was good at correctly identifying cases overall (78.2% accuracy) and was reasonably good at telling the difference between sick and healthy people (AUC = 0.725). This means it can be a useful tool for predicting if someone has the disease. However, the study found that the model was very bad at correctly identifying people who don't have the disease (Specificity = 22.7%). This means it often says someone has CAD even when they don't. This problem could lead to unnecessary tests and cause stress and financial burden for patients and the healthcare system.

Because of these findings, the study suggests that we should focus on finding CAD early and managing risk factors in people who are more likely to get it, especially older men with long-term illnesses like diabetes. To make the model better in the future, it's suggested to use more balanced data, add more information (like cholesterol levels and family history), and use newer machine learning methods that might be better at handling complex data.

In conclusion, this study gives important insights for CAD research. It confirms that the logistic regression model is a strong starting point. Fixing the current problems will help create better and more reliable prediction tools, which will improve how we treat patients.

Conflict of interest: The authors certify that there are no conflicts of interest.

References

[1] A. Mustafa, R. Hassan, N. Omar, et al., "Factors influencing heart disease in Sudan: a logistic regression approach," *J. Health Stat.*, vol. 12, no. 3, pp. 45–60, 2015.
[2] A. Pannu and I. N. El-Saeiti, "Evaluating predictive accuracy and model selection in logistic regression: a statistical approach using sensitivity, specificity, and ROC analysis," *Int. J. Manag.*, vol. 16, no. 1, pp. 16–23, 2025, doi: 10.34218/IJM_16_01_002.
[3] A. Zulkiflee, H. Rahman, and P. Singh, "Heart disease prediction using binary logistic regression: a comparative study," *Int. J. Med. Inform.*, vol. 145, pp. 104–112, 2021.
[4] A. G. S. Alzwawy and R. F. Aldarnawi, *Hyperferritinemia as a Potential Predictor of COVID-19 Severity and Mortality among a Sample of Quarantine Hospital's Patients in Ajdabiya /Libya*, 2025. [Online]. Available: http://www.doi.org/10.62341/arsh1727
[5] B. Douma, J. Smith, K. Lee, et al., "Predicting heart disease using logistic regression: a case study," *Int. J. Epidemiol.*, vol. 47, no. 2, pp. 123–135, 2018.
[6] D. G. Kleinbaum, L. L. Kupper, and K. E. Muller, *Applied regression analysis and other multivariable methods*, Boston: PWS-Kent, 1987.
[7] D. W. Hosmer and S. Lemeshow, *Applied logistic regression*, 2nd ed., New York: Wiley, 2000.
[8] I. N. El-Saeiti and A. Pannu, "Evaluating the predictive power of logistic regression models in classifying binary outcomes," *Int. J. Educ.*, vol. 5, no. 2, pp. 93–103, 2024, doi: 10.34218/IJE_05_02_007.
[9] M. D. Intriligator, *Econometric models, techniques, and*

applications, Englewood Cliffs: Prentice-Hall, 1978.

[10] M. J. Norusis, *SPSS for Windows: base system user's guide*, Chicago: SPSS Inc., 1993.

[11] R. D. Retherford and M. K. Choe, *Statistical models for causal analysis*, New York: Wiley, 1993.

[12] S. Chand, R. Patel, J. Kim, et al., "Risk factors for coronary artery disease: a logistic regression analysis," *Am. J. Cardiol.*, vol. 95, no. 6, pp. 789–795, 2005.

[13] S. Sharma, *Applied multivariate techniques*, New York: Wiley, 1996.

[14] S. Swain, L. Thompson, M. Yang, et al., "Predictive modeling of coronary artery disease using logistic regression," *J. Cardiovasc. Res.*, vol. 58, no. 4, pp. 456–470, 2019.

[15] T. Chap, *Applied logistic regression analysis*, New York: Wiley, 1998.

[16] World Health Organization, "cardiovascular diseases (CVDs): key facts," Geneva: WHO, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed: Jan. 30, 2025].