Proceedings of the 1st International Conference on Statistical Sciences and Data Analytics

LOGISTIC REGRESSION MODEL FOR CORONARY HEART DISEASE RISK PREDICTION: (A MURTALA MUHAMMAD SPECIALIST HOSPITAL KANO AS CASE STUDY)

Shazali Umar Madaki*[1], Oboyi Godwin Antony[1] and Terna Simon Terson[1]

[1]Department of Mathematics and Statistics Kaduna Polytechnic

*Corresponding Author: shazaliumar6@gmail.com

ABSTRACT

Coronary heart disease (CHD) is a major public health concern. This study utilizes logistic regression to predict coronary heart disease risk and identify significant predictor. The aim of this study is to develop a logistic regression model for predicting CHD risk among the patients and its objectives are to identify significant predictors of CHD risk, to evaluate the performance of the logistic regression model and to assess the overall significance of the model. This study analyzed a dataset of 1300 infant obtained from cardiovascular outpatient unit at Murtala Muhammad Specialist Hospital Kano. The logistic regression model achieved an accuracy of 62.31% and AUC of 0.5 in predicting CHD risk which indicated that the model predictive ability is limited, and additional features or modelling approaches may be needed to improve performance. The results indicate that weight and cholesterol levels are significant predictor's of CHD risk, with increased weight and lower cholesterol levels associated with higher CHD risk. This study develops a logistic regression model for predicting coronary heart disease risk among patients at Murtala Muhammad Specialist Hospital Kano, contributing to early detection, personalized medicine, and public health policy, while informing research and development efforts to reduce the burden of CHD in Nigeria.

KEYWORDS: Coronary Heart Disease, Weight, Cholesteral, Logistic Regression, Discriminant Analysis.

1. INTRODUCTION

Heart disease, a heart- related disorder, is the leading cause of dealth worldwide. It's the most common non-communicable disease, making early treatment difficult. The World Health Organization predicts an increase in heart deaths in the coming years, with the average increasing annually (istanto *et al*., 2023). The burden of CHD is increasing in low and middle income countries, including Nigeria, where cardiovascular disease are becoming a major public health concern (Mensah & Brown, 2007). Every year, the number of people suffering from cardiovascular disease rises as well. Numerous factors, including age, blood pressure, cholesterol, diabetes, hypertension, heredity, obesity, and bad lifestyle choices, contribute to the development of this disease. Physical indications such as dizziness, shortness of breath, exhaustion, and chest pain can be used to identify a variety of symptoms. Statistical modelling approach have been widely used to predict CHD risk, enable early detection approach and target interventions (D'Agostino *et al*., 2008). However, the existing models often and may not perform well in diverse populations (Damel *et al*., 2016). Recent studies have highlighted the potential of machine learning and logistic regression models in improving CHD risk prediction (Li *et al*., 2020). Despite this

progress, there is scarcity of research on CHD risk prediction models specifically tailored to Nigerian populations (Ogah *et al*., 2019). This study aims to address this gap by developing a regression model for predicting CHD risk among patients at Murtala Muhammad Specialist Hospital Kano. By identifying significant predictors and evaluating the model performance, this study seek to contribute to the development of effective CHD prevention and management strategies in Nigeria.

Use of logistic regression for chd prediction

Logistic regression has been widely used to predict Coronary Heart Disease (CHD) due to its ability to model binary outcomes and provide probabilities. Studies have shown that logistic regression models can accurately predict CHD risk using various predictors such as age, cholesterol levels, blood pressure, and smoking status.

A study by Wu (2024) employed logistic regression modeling to predict CHD risk and identified key influencing factors, including prevalent stroke, prevalent hypertension, and blood pressure medications. The model achieved an accuracy of (86%) and an area under the ROC curve (AUC) of (73%) (Wu, 2024). Logistic regression has been used to identify significant risk factors for CHD, such as age, sex, and cholesterol levels (Ganesh *et al*., 2022).

(Priyadarshini *et al*. (2021), used binary logistic regression analysis to examine the relationship between the dependent variable (Y) and independent factors (X) of the logit function, with Y as a dependent variable and X as a continuous predictor variable. The evaluation uses logistic regression to model factors influencing coronary heart disease, evaluating its performance using MAPE, RMSE, and SSE. The Major risk factors include of age, gender, obesity, blood stress, ex-smokers, BMI, dyspnea, chest pain, and stenosis (Ciu *et al*., 2020), discovered that the logistic regression method was categorized as an efficient and successful algorithm in predicting the primary cause of cardiovascular disease, which was the issue addressed in the investigation. Fourteen cardiovascular performance related characteristics were used as variable to create a logistic regression model. It is discovered that there is a substantial correlation between the variables. As a result, there is often less chance of multicollinearity in the study.

Comparative studies between logistic regression and other methods

Several studies have compared the performance of logistic regression with other methods in predicting CHD. (Isnanto *et al*. (2023) compared the performance of logistic regression (LR) and predictive discriminant analysis (PDA) for the two-group classification problem examined in the Monte Carlo study. The classification process employed prior probability with the assumption that the cost of misclassification would be the same. Three factors were included in the study's fully crossover experimental design: sample size, prior probabilities, and equal/unequal covariance matrices. There were 200 replications in each cell. To give the investigation a replication mechanism, two data patterns were simulated. The principal conclusions are: When two groups have equal prior probability, PDA and LR perform similarly; when two groups have different prior probabilities, LR minimizes the error rate for the smaller group and PDA minimizes the error rate for both the bigger group and the entire sample. (Munmun *et al*., 2025) compared the performance of logistic regression, random forest, and support vector machine across most evaluation metrics. (Anderson, 2024) compared the performance of logistic regression and gradient boosting for heart disease detection. The results demonstrated that gradient boosting outperformed logistic regression in terms of accuracy, precision, recall and F1- score.

Abdulqader (2015), discovered that classification method to classify datasets using linear discriminant analysis. The dataset used was divided into 25 percent tests and 75 percent training. This classification is carried out fewer than two conditions. The first condition is the number of outputs consisting of 5 labels, and the second condition is only the number of labels with 2 outputs. The classification of performance measurement based on accuracy, precision, repeatability, and F1 value shows the results of the performance of the LDA algorithm in classifying heart disease using the two labels used as targets or results. Based on the results, the precision value is 0.82, the repetition value is 0.81, the F1 value is 0.81, and with an accuracy of 81.22 percent, and the confusion matrix that is found in classic heart disease with LDA at 2 targets or outputs. Kiyoshige *et al*. (2023), used Bayesian age-period-cohort (BAPC) models to estimate future CHD and stroke mortality projections in Japan, focusing on population estimates until 2040.

2. MATERIALS AND METHODS

Study Design

This study employs a quantitative research design to develop a logistic regression model for predicting Coronary Heart Disease (CHD) risk. The dataset used for this study consists of 1300 observations and 12 attributes related to CHD patients.

Data collection

The data used in this research is a secondary data obtained from 1300 patients at cardiovascular outpatient ward Murtala Muhammad Specialist Hospital Kano.

Method of Data Analysis

The study employed the following statistical methods:

1. Logistic Regression: To predict CHD risk and identify significant predictors.
2. Performance Metrics: To evaluate the accuracy, specificity, sensitivity, precision, and F1-score of the logistic regression model.
3. Omnibus Chi-square: To assess the overall significance of the logistic regression model

Data Descriptions

The dataset used for this study purpose contains 1300 observations and 12 attributes on the coronary heart disease patients. The attribute's information is as follow:

Dependent Variable

Coronary Heart Disease – Binary (Yes/No)

Independent Variable

Diabetes – (Binary Yes/No)

Age – (Numeric)

Gender – (Male/Female)

Hypertension – (Binary yes/no)

Weight – (Numeric)

Cholesterol – (Numeric)

Family History – (Binary Yes/No)

Smoking – (Binary Yes/No)

Marital Status – (Categorical: Marriage, Single, Widowed, Divorced)

Employment – (Categorical: Employed, Unemployed, and Retired)

Ethnicity – (Categorical: Hausa, Fulani, Igbo, Yoruba, Igala)

The study included patients with a confirmed diagnosis of CHD, aged from one month to above years, and receiving treatment at a healthcare facility. Patients with incomplete medical records or those who refused to participate in the study were excluded.

Ethical Approval

This study was approved by Kano State Ministry of Health, ensuring adherence to ethical standard for research involving human subjects.

Statistical Analysis

The logistic regression analysis was used to develop a model that predicts the risk of CHD based on various factors. The performance metrics were calculated to evaluate the model`s accuracy. The omnibus chi-square test was used to determine the significance of the relationships and the overall model fit.

Logistic Regression

Logistic regression models are suitable for scenarios in which the dependent variable is binary, meaning that it can have just two possible outcomes, such as "yes/no," "success/failure," "normal/abnormal," or "sick/healthy." Disease status is another example of a variable that only accepts two possible values: it can either be present or absent. These binary results are assumed to be coded as 1 and 0(Usman, 2012).

When studying linear regression, we attempted to estimate a population regression equation;

$$Y = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots .. \beta_k X_{ki} + e_i \quad (1)$$

By fitting the following estimated model:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots .. \hat{\beta}_K X_k + e_i \quad (2)$$

The response Y was continuous, and will assume to follow a normal distribution. We are concern with predict or estimate of the mean value of the response corresponding to a given set of values for explanatory variable. Another property of logistic regression relates to situations in which explanatory variable and X is the rather than the response variable Y is random. In this case we shall consider only a simple logistic regression that is, logistic regression models with one explanatory variable. Our first strategy might be fit a model of the form.

$$P = \alpha + \beta x \quad (3)$$

This is simple the standard linear regression model in which x represents the explanatory variable and y- the outcome of a continuous, normally distributed random variable – has been replaced by (Menad, 1995). As before, α is the intercept and β its slope. On inspection, however, this model is not feasible. Since p is a probability, it is restricted to taking value between 0 and 1. The term in contrast could easily yield a value that lies outside this range. Instead we might try to solve this problem by fitting the model.

This equation guarantees that the estimate of p is positive. We would soon realize, however, that the model is also unsuitable. Although the term cannot produce a negative estimate of p, it can result in a value that is greater than 1. To accommodation this final constraint, we fit a model of the form;

$$P = e^{\alpha+\beta x} \quad (4)$$

This expression on the right, called a logistic function (logit model), cannot yield a value that is either negative or greater than 1; consequently, it restricts the estimated value of p to the required range.

$$P = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \quad (5)$$

$\frac{P}{1-P}$ to 1 (6)

Thus, if a success occurs with probability

$$P = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \quad (7)$$

The odds in favour of success are

$$\frac{P}{1-P} = \frac{e^{\alpha+\beta x}/1+e^{\alpha+\beta x}}{1/e^{\alpha+\beta x}} = e^{\alpha+\beta x} \quad (8)$$

Taking the natural logarithms of each side of this equation

$$\text{In}\left(\frac{P}{1-P}\right) = \text{In}e^{\alpha+\beta x} = \alpha+\beta x \quad (9)$$

Where; the link function is;

$$\text{In}\left(\frac{P}{1-P}\right) \quad (10)$$

Thus, modeling the probability P with a logistic regression function is equivalent to fitting a linear regression model in which continuous response y has been replaced by the logarithms of the odds of success for a dichotomous random variable. Instead of assuming that a relationship between p and x is linear, we assume that the relationship between x and the following link function is linear.

Recall that the technique of fitting model of this form is known as logistic regression. The estimated relationship the explanatory variable and the odd in favor of success is represented as follow:

$$In\left(\frac{\hat{P}}{1-\hat{P}}\right)=\hat{\alpha}+\hat{\beta} \quad (11)$$

However, we cannot apply the method of least squares to fit logistic regression model; instead we used method of maximum likelihood. Because the method of least square assumes that response is continuous and normally distributed. Let us see how this method is applied in the case of simple logit model. We start by assuming Bernoulli random variable associated with to each observation y_i there-fore, the joint distribution of the n observations is represented as follows:

$$P(y_{i,\ldots}y_{n)}=\prod_{i=1}^{n} P_i^{yi}\,(1-P)^{1-yi} \quad (12)$$

Taking the natural logarithms of this likelihood function, we obtain the following:

$$In\,P \;=\; \sum_i yi\;In\left(\frac{P_i}{1-Pi}\right)+\sum_i In(1-P_i) \quad (13)$$

Using the following fact:

$$P_{i}=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \quad (14)$$

Also written as follow:

$$\text{In}\left(\frac{P}{1-P}\right) = \alpha+\beta x \quad (15)$$

The logarithms of this likelihood function (log-likelihood), which is a function of the coefficients $L(\beta)$ can be expressed as follows:

$$\text{L}(\beta) =\sum_i yi(\alpha+\beta)+\sum_i In[1+\exp(\alpha+\beta x)] \quad (16)$$

Performance evaluation metrics

Sensitivity (True Positive Rate): The proportion of actual coronary heart disease patients correctly identified by the model.

$$sensitivity=\frac{TP}{TN+FN}\times 100 \quad (17)$$

Specificity (True Negative Rate): The proportion of non- coronary heart disease patients correctly identified by the model.

$$specificity = \frac{TN}{TN+FP} \times 100 \quad (18)$$

Precision: This measure the proportion of patients correctly predicted to have CHD among all patients predicted to have CHD.

$$Precision = \frac{TP}{(TP+FP)} \quad (19)$$

Accuracy: The overall proportion of correct predictions (both CHD and non-CHD patients):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (20)$$

F1-score :(Harmonic mean of sensitivity and specificity): A balanced measure of sensitivity and specificity.

$$F1 - Score = \frac{(SE \times SP)}{(SE+SP)} \times 100 \quad (21)$$

Omnibus chi-square

In this study, we employed the omnibus chi-square test to examine the association between independent variable and CHD risk. The omnibus chi-square test is a statistical method used to determine if there is a significant association between a categorical independent variable and a binary dependent variable CHD risk. This test is particularly useful when dealing with multiple categories in the independent variable.

Test of hypothesis
H_0: There is no significant association between the independent variables and CHD risk
H_1: There is significant association between the independent variables and CVD risk

Test Statistic

$$x^2 = 2[\sum_{i=1}^{r}\sum_{j=1}^{c} (o_{ij}\ In\ \frac{o}{eij}] \quad (22)$$

Where;
χ^2_0--- Chi-square calculated
χ^2_v---- Chi-square value with v df

Decision Criterion
Reject H_0 if P<0.05 otherwise accept H_0 at the 5% level of significance.

3. RESULT AND ANALYSIS

Adequacy checking of logistic regression model

Table 1: Adequacy checking of regression model

Variable	GVIF	Df	$_{GVIF}(1/(2*Df))$
DB	1.076122	1	1.037363
HPT	1.029860	1	1.014820
FMH	1.160945	1	1.077472
SMK	1.138356	1	1.066937
GND	1.305327	2	1.068882
MS	1.560913	10	1.022513
ETH	1.046238	7	1.003234
EDQ	1.209427	3	1.032199
EMP	1.330893	6	1.024107
AGE	1.407900	1	1.186550
WGT	1.055945	1	1.027592
CHL	1.048670	1	1.024046

From the results obtained in Table 1 it is observed that the all variance inflation factors of the independent are less than 2, thus, the variables in the model have no high correlation.

Logistic regression coefficient models

By using the given logistic regression model in equation (14), R package was used to analyze the data and the following table of model coefficients was generated.

Table 2: Estimated Coefficients Model

Variables	Coefficients(β)	Standard error	Z- value	P- value
Intercept	-3.406e-02	6.151e-01	-0.055	0.95584
BD (Yes)	1.418e-01	1.261e-01	1.124	0.26084
HPT (Yes)	-2.547e-02	1.576e-01	-0.162	0.87158
FMH (Yes)	-3.247e-01	3.068e-01	-1.059	0.28979
SMK (Yes)	-5.369e-02	1.843e-01	-0.291	0.77085
Male	1.419e+01	1.455e+03	0.010	0.99222
Married	-3.509e-01	3.537e-01	-0.992	0.32125
Single	2.278e-02	4.281e-01	0.053	0.95757
Widowed	1.533e-01	4.963e-01	0.309	0.75737
Hausa	2.302e-01	1.394e-01	1.652	0.09862
Igala	6.078e-01	4.944e-01	1.230	0.21887
Igbo	-5.174e-02	2.193e-01	-0.236	0.81347
Yoruba	1.904e-01	1.757e-01	1.084	0.27850
Secondary	-1.005e-01	2.717e-01	-0.370	0.71155
Tertiary	-8.124e-02	2.686e-01	-0.302	0.76233
Unemployment	-5.717e-02	3.818e-01	-0.150	0.88095
Retired	3.492e-03	2.636e-01	0.013	0.98943
Self employed	1.092e-01	1.829e-01	0.597	0.55048
Unemployment	-1.601e-01	1.731e-01	-0.925	0.35511
AGG	5.633e-04	4.075e-03	0.138	0.89004
WGT	7.854e-03	3.924e-03	2.001	0.04534
CHL	-2.142e-03	7.053e-04	-3.037	0.00239

The two variables weight and cholesterol, showed statistically significant associations. Weight had a positive coefficient (β = 0.007854, p = 0.04534), Indicating that increased weight is associated with higher log-odds of CHD. Conversely, cholesterol had a negative coefficient (β = -0.002142, p = 0.00239), suggesting that higher cholesterol levels are associated with lower log-odds of CHD in this model. Other variables showed non-significant associations, with some having positive coefficients (potential risk factors) and others having negative coefficients (potential protective factors). Variables with positive coefficients included BD, male gender, single, widowed, Yoruba ethnicity, and self-employment. Variables with negative coefficients included hypertension, family history of heart disease, smoking, being married, and having secondary or tertiary education. Although these variables did not reach statistical significance, they may still be worth exploring further in future research

Table 3: Odd Ratio

Coefficient	Odd ratio	(2.5% 97.5 %)
Intercept	9.665102e-01	(2.896661e-01 3.241152e+00)
BD (Yes)	1.152310e+00	(8.999394e-01 1.475577e+00)
HPT (Yes)	9.748487e-01	(7.152288e-01 1.327345e+00)
FMH (Yes)	7.227269e-01	(3.939332e-01 1.317683e+00)
SMK (Yes)	9.477225e-01	(6.606454e-01 1.362306e+00)
Male	1.451674e+06	(6.902061e-01 1.162413e+00)
Married	7.040757e-01	(3.480957e-01 1.408576e+00)
Single	1.023038e+00	(4.392596e-01 2.369132e+00)
Widowed	1.737635e+00	(1.484899e-01 4.011950e+01)
Hausa	1.258806e+00	(9.580052e-01 1.654692e+00)
Igala	1.836471e+00	(7.130984e-01 5.102384e+00)
Igbo	9.495800e-01	(6.171941e-01 1.459706e+00)
Yoruba	1.209737e+00	(8.575273e-01 1.708331e+00)
Secondary	9.043990e-01	(5.284146e-01 1.538137e+00)
Tertiary	9.219718e-01	(5.418563e-01 1.558309e+00)
Uneducated	9.444303e-01	(4.463768e-01 2.001177e+00)
Retired	1.003498e+00	(5.986712e-01 1.685295e+00)
Self employed	1.115417e+00	(7.790698e-01 1.596839e+00)
Unemployment	8.520822e-01	(6.063128e-01 1.195737e+00)
AGG	1.000563e+00	(9.925972e-01 1.008591e+00)
WGT	1.007885e+00	(1.000191e+0 1.015709e+0)
CHL	9.978600e-01	(9.964658e-01 9.992295e-01)

From the results obtained in Table 3 it is observed that the odds of CHD (Yes) are approximately 1.15 times higher for individuals with Diabetes (Yes) compared to those without Diabetes (No). The odds of outcome variable are approximately 0.9749 times higher for individuals with hypertension (No) compared to those without hypertension (Yes). The odds of CHD are approximately 0.7227 times higher for individuals with family history compared to those without. The odds of CHD are approximately 0.9477 times higher for individuals not doing smoking compared to those that are doing. The odds for CHD are approximately 1.4518 times higher for male compared to female. The odds for CHD are approximately 0.7040 times lower for married individuals compared to those that are divorced. The odds for CHD are approximately 1.0230 times higher for single individuals compared to those that are divorced. The odds for CHD are approximately 1.7376 times higher for widowed individuals compared to those that are divorced.

Evaluation metric measures

Table 4: Confusion matrix

Prediction	0	1
0	394	214
1	276	416

The confusion matrix reveals that the model correctly predicted 394 true negatives (cases without CHD) and 416 true positives (cases with CHD).However, it also incorrectly predicted 214 false positives (cases without CHD predicted as having CHD) and 276 false negatives (missed cases with CHD). This matrix can be used to calculate key performance metrics, such as accuracy, specificity, sensitivity, precision, and F1-score, which showed moderate performance. The numbers in the matrix align with the observed metrics, highlighting the model's strengths and areas for improvement in predicting coronary heart disease.

Table 5: Accuracy Classification Measures

Accuracy	Specificity	Sensitivity	Precision	F1-Score
0.6231	0.6603	0.5881	0.6480	0.6166

The model's performance metrics suggest moderate accuracy (62.31%) in predicting coronary heart disease. The specificity (66.03%) and precision (64.80%) indicate reasonable performance in identifying true negatives and actual positives. However, the sensitivity (58.81%) suggests room for improvement in detecting true positives. The F1-score (0.6166) reflects a balance between precision and sensitivity.

ROC Curve of Logistic Regression Model

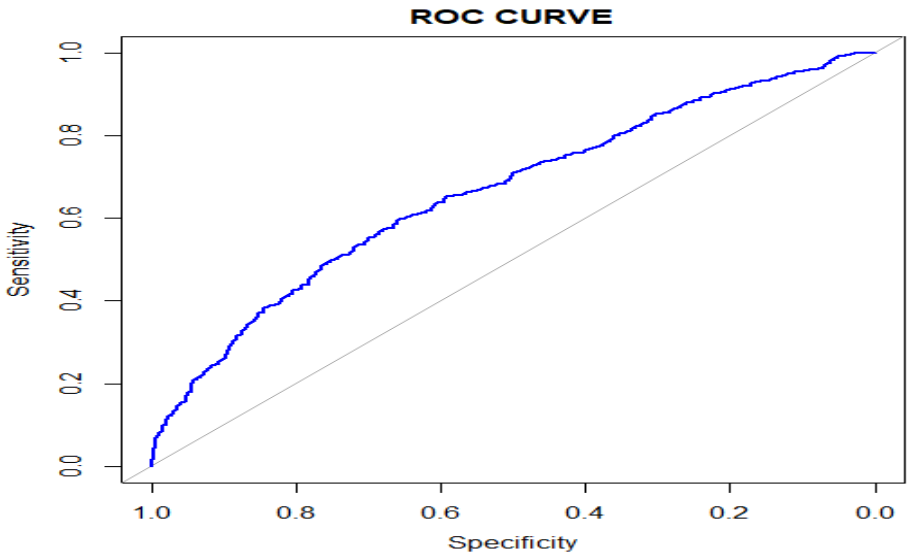

Figure 1: Reciever Operating Curve

From the result obtained in Figure 1, it is observed that the Area under the Curve (AUC) is 0.6622, which indicates that the model has a moderate level of predictive accuracy. In addition, the model is better than the random chance (AUC = 0.5).

Omnibus Chi-square Test Result

Table 5: Omnibus Chi-square

Residual Df	Residual Dev	Df	Deviance	Pr(>Chi)
1264	295.69			
1299	323.64	-35	-27.958	3.649e-11

The omnibus chi-square test evaluates the overall fit of the model. The results indicate a significant improvement in model fit (χ^2 = 27.958, df = 35, p = $3.649e^{-11}$), suggesting that the predictors collectively contribute to explaining the variation in CHD risk. This implies that the model is effective in capturing the relationships between the predictor variables and CHD risk. The significant omnibus test statistic indicates that at least one of the predictors has a significant effect on CHD risk, supporting the use of this model for predicting CHD outcomes.

4. CONCLUSION

The logistic regression analysis revealed that only two variables, weight and cholesterol were significant predictors of coronary heart disease risk. The positive coefficient for weight ($\beta = 7.854\, e^{-03}$, $p = 0.04534$) suggest that increasing in weight is associated with the higher risk of CHD. This finding is consistent with previous studies that have shown a strong link between obesity and cardiovascular disease. On the other hand , the negative coefficient for cholesterol ($\beta = -2.142e^{-03}$, $p = 0.00239$) suggests that higher cholesterol levels are associated with lower risk of CHD , which is counterintuitive and may be due to the complex relationship between cholesterol and CHD. The remaining variables, including diabetes, hypertension, family history, smoking, and demographic factors, were not significant predictors of CHD risk in this study. The model performance was evaluated using accuracy, specificity, sensitivity, precision and F1-score. The results showed that the model had a moderate accuracy of 62.31% with a specificity of 66.03% and a sensitivity of 58.81%. The precision was 64.80%, and F1-score was 61.66%. These results suggest that the model the model has some predictive value but may not be sufficient for clinical decision marking.

Limitations and future work

This study has several limitations:

i. The model was developed using data from a single hospital, which may limit its generalizability
ii. Some variables may not have been included in the model, potentially affecting its accuracy

Future work could:

i. Validate the model using data from multiple hospitals or regions
ii. Incorporate additional variables, such as genetic markers or life style factors to improve the models predictive power

iii. Explore alternative modeling techniques, such as machine learning algorithms, to compare performance and identify the most effective approach

REFERENCES

Abdulqader, Q.M. (2015). Comparison of discriminant analysis and logistic regression analysis: An application on caesarean births and natural births data. *Yuzuncu Yil Universitesi Fen Bilimleri Enstitusu Dergisi*, 20(1-2):34-46.

Anderson, K. (2024). Heart disease detection using logistic regression and gradient boosting. *Journal of Cardiovascular Medicine*, 25(12), 1-8.

Ciu, T. and Oetama, R.S. (2020). Logistic regression prediction model for cardiova Scular disease. IJNMT (*International journal of new media Technology*), 7(1):33-38.

Damen, J.A., Hooft, L., Schuit, E., Debray, T. P., Collins, G. S., Tzoulaki, I., & Moons, K. G. (2006). Prediction models for cardio vascular disease risk in the general population: systematic review. *BMJ*, 353, i2416.

D′Agostino, R.B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for used in primary care: The Framingham heart study. *Circulation*, 117(6), 743-753.

Ganesh, R., *et al*. (2022). Prediction of cardiovascular disease using logistic regression. *Journal of Medical Systems*, 46(10), 2105.

Isnanto, R. R., Rashad, I., and Widodo, C.E. (2023). Classisication of heart disease Using linear discriminant analysis algorithm. In *E3S Web of Conferences*, Volume 448, page 02053. EDP Sciences.

Kiyoshige, E, Ogata S, O'Flaherty M, Capewell, S, Takegami M, Jihara K, Kypridemos, C, And Nishimura K (2023), Projections of future coronary heart disease and Stroke mortality in japan until 2040: a Bayesian 42 age-period cohort Analysis. The Lancet Regional Health-Westem Pacific, 3

Li, K., Liu, X., & Li, Z. (2020). A logistic regression model for predicting cardiovascular disease risk in a chineese population. *Journal of Cardiovascular Medicine*, 21(12), 641-648.

Mensah, G. A., & Brown, D. W. (2007). An overview of cardiovascular disease burden in the United States. *Health Affairs*, 26(1), 38-48.

Ogah, O. S., Okpechi, I. G., & Chukwuonye, I.I. (2019). Cardiovascular disease risk prediction models in Nigeria: A systematic review. *Nigerian Journal of Cardiology*, 16(2), 53- 62.

Priyadarshini, E., Gayathri, G. R., Chakravarthy, S. E., Vidhya, M., and Memala, W. A. (2021). Analysis of heart disease using statistical techniques. In *Journal of physics: Conference series*, volume 1770, page 012105. IOP Publishing.

Rajarathinam, A. (2024). Discriminant analysis for heart disease attributes. International Journal of Statistics and Applied Mathematics, 9(2), pp. 180-188.

Usman, A. (2012). Statistical methods for biometric and medical research. *Mil-Lennium Printing ad Publishing Company Limited, Kaduna, Nigeria, pages* 486-495.