MODELLING THE ASSOCIATION AND PREDICTION OF CARDIOVASCULAR DISEASE USING LOG-LINEAR AND BINARY LOGISTIC REGRESSION APPROACHES

OTORI ILYASU ALIYU & OLAYIWOLA MATTHEW ADETUTU
Department of Statistics, Federal University of Technology Minna, Nigeria.
Email: ola.adetutu@futminna.edu.ng **Phone No:** +2348030737153

Abstract
Cardiovascular diseases are group of disorders of the heart and blood vessels that constitute the leading cause of death globally. This research focused on determining the factors associated with the likelihood of developing cardiovascular disease, and examining the pattern of association and interaction among categorical variables responsible for cardiovascular disease in Northwestern Nigeria. Data on cardiovascular risk factors and symptoms from a cohort comprising healthcare personnel and patients over a period extending from August 15, 2024 to September 15, 2024 were collected. Risk factors admitted by respondents including High Blood Pressure (Hypertension) acknowledged by 21.1% of respondents, High Cholesterol recognized by 15.2% of participants, Smoking practiced by 12.4% of the cohort, Alcohol Consumption reported by 8.2% of respondents, Unhealthy Diet admitted by 11.8% of participants, Physical Inactivity identified by 12.1% of the study population, Obesity and Overweight prevalent in 10.3% of respondents, and Diabetes recognized by 8.8% of the cohort which provided valuable intuitions into the prevalence of risk factors in the Northwestern Nigerian context. Findings from logistic regression affirmed age, gender, diagnosis, symptoms, risk factors, and family history as significant predictors of cardiovascular diseases with p-values 0.000, 0.000, 0.004, 0.000, 0.000, 0.002, and 0.001 respectively. Log linear modelling revealed positive partial association between age and gender, gender and residing site, age, gender, residing site with cardiovascular disease with p-values 0.0000, 0.0000, 0.0000, 0.0000, 0.0000 and 0.0000 respectively, while age and residing site with 0.0640 was not impactful.

Keywords: cardiovascular, demographic, interaction, prevalence, risk

Introduction

According to World Health Organization (WHO), cardiovascular disease CVD remains the leading cause of morbidity and mortality worldwide, accounting for an estimated 17.9 million deaths annually (WHO,2023). The growing burden of CVD is strongly linked to multiple factors, including biological predispositions, behavioral patterns, and socio-demographic characteristics such as age, gender, and place of residence. Understanding the prevalence of CVD and its risk factors is therefore critical to designing effective prevention, intervention, and policy strategies (Adebiyi, 2017; Joynt-Maddox *et al.,* 2024). Epidemiological studies often rely on statistical modeling to investigate the complex interplay of risk factors and demographic variables. Logistic regression has been extensively applied in cardiovascular research for estimating the probability of disease occurrence based on independent predictors (Ogah *et al*., 2012; Siaga *et al*., 2024). This method is well-suited for binary outcomes such as the presence or absence of CVD and allows for the quantification of associations between risk factors (for examples, hypertension, diabetes, smoking, and obesity) and disease occurrence through odds ratios. Moreover, logistic regression enables adjustment for confounding variables, making it a powerful tool for identifying independent risk factors.

However, the epidemiology of CVD often involves more intricate relationships, such as higher-order associations and interdependencies among demographic characteristics and health behaviours. Papathomas and Richardson (2016) affirmed that Log-linear modeling provides an alternative and complementary approach for examining these multidimensional interactions. Unlike logistic regression, which focuses on predicting outcomes, log-linear models analyze the

structure of contingency tables to detect associations and higher-order interactions among categorical variables (Cheng *et al.,* 2018). This makes it especially useful in exploring how demographic variables (such as age categories, gender, and residence) interact with symptoms and risk factors in shaping the distribution of CVD prevalence across populations.

Together, logistic regression and log-linear analysis offer a robust methodological technique. Logistic regression identifies and quantifies independent predictors of disease risk, while log-linear models uncover interaction patterns and higher-order dependencies that may otherwise be overlooked. Applying both approaches in cardiovascular research enhances the depth of epidemiological insight, providing a comprehensive understanding of how prevalence is shaped by individual risk factors and demographic interactions. Such combined analyses not only improve the statistical rigor of CVD studies but also provide evidence for targeted interventions and tailored public health policies. But symptoms are clinical manifestations that arise as a result of established disease and alert to its presence (Knoke and Burke, 1980, Adhikary *et al.*, 2022) while risk factors are characteristics, behaviours, or exposures that increase the probability of developing cardiovascular diseases (Teo, 2021).

Aim and Objectives.

The main focus of the research is to analysis the prevalence, risk factors, and demographic factors' interactions in cardiovascular disease in northwest Nigeria using log linear and binary logistic regression techniques.

Material and Methods

Methods

In quest to achieve the aim of the study, two statistical techniques were employed which includes:

Binary Logistic Regression

This modelling is a statistical technique used to develop a model useful in predicting the likelihood of cardiovascular disease based on risk factors and symptoms acknowledged. It is a type of regression model that is particularly well-suited for situations where the dependent variable represents a categorical outcome with only two possible categories or levels. Probability of the occurrence of interest: the probability of the outcome variable such as a patient test positive to cardiovascular is given as

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+ \ldots+ \beta_8 x_8}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+ \ldots+ \beta_8 x_8}} = \frac{1}{1+e^{-\beta_0+\beta_1 x_1+\beta_2 x_2+ \ldots+\beta_8 x_8}} \quad (1)$$

The probability that patient has no cardiovascular disease is thus given by:

$$1-\pi = 1-\frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \ldots..+ \beta_8 X_8}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots.+\beta_8 X_8}}$$

$$=\frac{1}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_8 X_8}} \quad (2)$$

x_1 : The binary predictor variable for hypertension
X_2: The binary predictor variable for high cholesterol
X_3:The binary predictor variable for smoking
X_4: The binary predictor variable for alcohol consumption
X_5 :The binary predictor variable for unhealthy diet
X_6:The binary predictor variable for physical inactivity
X_7:The binary predictor variable for obesity
X_8: The binary predictor variable for diabetes

β_0 : The Y intercept, β_1 , β_2 ,…, β_8 are the regression coefficients for the predictor variables X_1 ,X_2, …, X_8.

The model assumptions includes: binary outcome, independence observations, linearity of the log-odds, no multicollinearity while several test statistic used to evaluate the model are: Hosmer-Lemeshow test for goodness of fit, Wald test to test whether an individual predictor variable significantly contributes to the model, and Cox and Snell R square to explains variations in the outcome.

Log linear Modelling

A log linear model is a statistical technique available to analyze structures of cross classified categorical data in contingency table therefore treats all variables as categorical factors and studied their joint distribution rather than having one as dependent and others as independent variables as in logistic regression. The three demographic factors interested are age, gender, and residing site, the log linear model for a cell:

F_{ijk} (Age = i, Gender = j, Residence = k, such that i = 0, 1, 2, 3; j = 0, 1; k = 0, 1)

$$\ln(F_{ijk}) = \lambda + \lambda_i^A + \lambda_j^G + \lambda_k^R + \lambda_{ij}^{AG} + \lambda_{ik}^{AR} + \lambda_{jk}^{GR} + \lambda_{ijk}^{AGR} \quad (3)$$

Where:

λ : Grand mean (baseline)

$\lambda_i^A, \lambda_j^G, \lambda_k^R$: Main effects (distribution of Age, Gender, and Residence respectively)

λ_{ij}^{AG} λ_{jk}^{AR} λ_{jk}^{GR} : Pairwise association among the two demographic factors

λ_{ijk}^{AGR} : Association among the three demographic factors

The basic assumptions to be considered when using log linear model includes: observations should be independent from each other, all observations are identically distributed, and the relationships between variables are multiplicative.

Results and Discussions

Descriptive Analysis

Descriptive analysis of responses received from respondents which provided a preliminary insight into the study are presented in Tables 1.

Table 1: Symptoms and Risk factors Acknowledged by the Respondents

Symptoms	Freq	%	Risk Factors	Freq	%
Irregular Heartbeat (1)	364	22.3	Blood Pressure(1)	345	21.1
Shortness of Breath (2)	249	15.3	Cholesterol(2)	248	15.2
Fatigue (3)	167	10.2	Smoking(3)	203	12.4
Chest Pain (4)	395	24.2	Alcohol(4)	134	8.2
Palpitations (5)	117	7.2	Diet(5)	192	11.8
Pain (6)	122	7.5	Inactivity(6)	198	12.1
Swelling (7)	68	4.2	Obesity(7)	168	10.3
Dizziness(8)	150	9.2	Diabetes(8)	144	8.8
Total	1632	100	Total	1632	100

Eight symptoms and risk factors admitted by the cohort were summarized respectively. Chest pain (24.2%) followed by irregular heartbeat (22.3%) were the most frequent, and swelling (4.2%) preceded by palpitations (7.2%) were rare symptoms admitted by the respondents. Similarly, of all the risk factors admitted and summarized in the same Table 1, blood pressure (21%) follow by high cholesterol (15%) as the most frequent, while alcohol consumption (8.2%) and diabetes (8.8%) were descriptively least identified risk factors.

Binary Logistics Regression

Table 2: Binary Logistic Regression Equations Estimates

Variables	B	Std Error	Wald	df	Sig	Odd Ratio
Age	0.219	0.050	19.000	1	0.000	0.804
Gender	-0.985	0.163	36.693	1	0.000	0.374
State	0.243	0.039	37.761	1	0.600	1.275
Diagnose (1)	-0.447	0.794	0.316	1	0.004	0.064
Believe Cure (1)	4.542	0.814	31.151	1	0.000	93.913
Symptoms	0.134	0.034	15.756	1	0.000	0.874
Risk Factors	0.055	0.031	3.246	1	0.002	1.057
Family History	0.029	0.145	0.040	1	0.001	1.030
Constant	-0.145	0.302	23.028	1	0.000	0.234

Estimates of the predictors of CVD are presented in Table 2. The variable column depicts various variable which determined CVD status according to the logistic regression. The coefficient of *age* is 0.219, highly statistically significant ($p < 0.001$) suggested each one-unit increase in age, the odds of possibly CVDs increases by a factor of approximately 0.804, holding other variables constant. *Gender* is another explanatory variable with coefficient -0.985 is highly statistically significant ($p < 0.001$) indicating a decrease odds of having CVD by approximately 0.374 compared to the reference category (Male). *State* with coefficient 0.243, p-value 0.600 is statistically insignificant in predicting the CVD outcome.

The coefficient of *diagnose* (1) is -0.447, it is statistically significant with a p-value 0.004 suggests a decrease in the odds of experiencing cardiovascular diseases by a factor of approximately 0.640 compared to the reference category. This variable is statistically significant .The coefficient for Believe for cure (1) is 4.542, statistically significant ($p < 0.001$) meaning an increase in this variable is associated with a substantial increase in the odds of experiencing cardiovascular diseases, with an odds ratio of approximately 93.913. The coefficient for Symptoms is 0.134, is highly statistically significant ($p < 0.001$), meaning an increase in the Symptoms variable is associated with an increase in the odds of experiencing cardiovascular diseases by a factor of approximately 1.057. The coefficient for Risk Factors is 0.055, and it is statistically significant with a p-value of 0.002. An increase in Risk Factors is associated with an increase in the odds of experiencing cardiovascular diseases by a factor of approximately 1.057.

Moreover, the coefficient for Family History (1) is 0.029, and it is statistically significant with a p-value of 0.001. This variable is associated with a small increase in the odds of experiencing cardiovascular diseases by a factor of approximately 1.030. The constant term represents the baseline log-odds of the event happening when all other independent variables are set to zero or their reference levels. The coefficients and p-values provide intuitions into the direction and strength of these associations.

Log linear Modelling

Table 3: Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
Age*Gender	3	54.283	0.0000	2
Age*Residing Site	3	10.404	0.0640	2
Gender*Residing Site	1	26.829	0.0010	2
Age	3	779.67	0.0010	2
Gender	1	76.521	0.0010	2
Residing Site	1	184.85	0.0010	2

The Table 3 provided the results of partial association tests for various effects involving categorical variables. These tests assessed the significance of associations between pairs of categorical variables while controlling for the effects of other variables, this test assessed the interaction between age and gender while controlling for other variables. The p-value is highly significant ($p < 0.001$), indicating a significant association between age and gender when other variables are taken into account, the association between Age and residing site with p-value 0.064 is not impactful (statistically significant) when controlling for other variables. The interaction between gender and residing site while controlling for other variables is highly significant ($p < 0.001$), indicating a significant association between gender and residing sites when other variables are considered. Also, association between age (independent of other variables) is highly significant ($p < 0.001$) indicating a significant association between age and the outcome the association between gender (independent of other variables) and the association between gender and the outcome, residing site and outcome ($p < 0.001$) are also significant.

Table 4: Parameters Estimates of the log-linear model

Effect	Parameter	Estimate	Standard Error	z	Significance	95% Confidence Interval	
						Lower Bound	Upper Bound
Age*Gender*Residing Site	1	0.157	0.129	-1.217	0.224	-0.410	0.096
	2	0.079	0.121	0.656	0.512	-0.157	0.316
	3	0.165	0.120	1.374	0.170	-0.071	0.401
Age*Gender	1	0.404	0.129	3.132	0.002	0.151	0.656
	2	0.374	0.121	3.095	0.002	0.137	0.610
	3	0.313	0.120	2.603	0.009	0.077	0.549
Age*Residing Site	1	-0.024	0.129	-0.186	0.853	-0.277	0.229
	2	-0.060	0.121	-0.499	0.617	-0.297	0.176
	3	-0.071	0.120	0.587	0.557	-0.306	0.165
Gender*Residing Site	1	-0.241	0.112	-2.151	0.031	-0.460	-0.021
Age	1	0.292	0.129	2.269	0.023	0.040	0.545
	2	0.857	0.121	7.097	0.000	0.620	1.093
	3	0.860	0.120	7.151	0.000	0.624	1.096
Gender	1	0.056	0.112	0.498	0.619	-0.164	0.275

These parameter estimates provided information about the strength and direction of the relationships between the variables included in the model. This column specifies the combination of variables for which the parameter estimates are provided, parameter column indicated the specific parameter or combination of parameters being estimated within each effect. Estimates column provided the estimated value of the parameter, standard error column shows the standard error associated with the parameter estimate, which represents the uncertainty or variability of the estimate. The Z-statistic measures how many standard errors the estimate is away from the null hypothesis value (usually zero). Significance column provided the p-value associated with the Z-statistic indicates whether the parameter estimate is statistically significant. If the p-value is below a chosen significance level, the estimate is considered statistically significant. The confidence interval provided a range of values within which the true parameter value is likely to fall with a certain level of confidence (usually 95%). The significant ones are: age and gender, gender and residing sites, age, residing sites, while others are not significant as displayed in Table 4.

However, these estimates provided intuitions into the relationships between variables in the model. Parameters with statistically significant estimates ($p < 0.05$) are considered to have a significant impact on the outcome, while non-significant parameters may not contribute significantly to the model. Additionally, the signs of the estimates (+ or -) indicate the direction of the relationship between variables. The interaction of Age and residing either in city or village does not show any statistical impact to the model construction.

Conclusion

In logistic regression analysis, Age, Gender, Symptoms, Risk Factors, and Family History emerged as statistically significant predictors of cardiovascular diseases. These factors collectively contribute to the complex landscape of CVDs within the region. However, the variable "STATE" did not prove to be statistically significant, suggesting that geographical location may not significantly impact the likelihood of cardiovascular diseases when controlling for other variables, logistic regression model demonstrated strong predictive accuracy in determining residing environments, achieving an overall accuracy of 83.9%. The model's effectiveness underscores its potential utility in public health planning and intervention strategies.

In log linear modelling, we observed a notable interaction between age and gender, indicating that these variables jointly influence the risk of cardiovascular diseases. However, the interaction between age and residing site did not reach statistical significance, suggesting that geographical location may not have a substantial impact on age-related CVD risk.

References

Adebiyi, A. (2017). Measuring the incidence and prevalence of cardiovascular diseases in Nigeria. *Walden Dissertations and Doctoral Studies*. https://scholarworks.waldenu.edu/dissertations/3269

Adhikary, D., Barman, S., Ranjan, R., & Stone, H. (2022). *A systematic review of major cardiovascular risk factors: A growing global health concern*. *Cureus, 14*(10), e30119. https://doi.org/10.7759/cureus.30119

Cheng, P. E., Liou, J. W., Kao, H. W., & Liou, M. (2018). *A constructive procedure for modelling categorical variables: Log-linear and logit models*. *arXiv preprint arXiv:1801.01278*.

Joynt-Maddox, K. E., Elkind, M. S. V., Aparicio, H. J., Commodore-Mensah, Y., de Ferranti,

S. D., Dowd, W. N., Hernandez, A. F., Khavjou, O., Michos, E. D., Palaniappan, L., Penko, J., Poudel, R., Roger, V. L., Kazi, D. S., & Moran, A. E. (2024). *Forecasting the burden of cardiovascular disease and stroke in the United States through 2050: Prevalence of risk factors and disease: A Presidential Advisory from the American Heart Association. Circulation, 150*(4), e65–e88. https://doi.org/10.1161/CIR.0000000000001256

Knoke, D., & Burke, P.J. (1980). *Log-linear models.* Sage Publications, Inc., NewJersey, 8-17.

Lind, L., Ingelsson, M., & Sundstrom, J., (2021). Impact of risk factors for major cardiovascular diseases: A comparison of life-time observational and mendelian randomization findings Open Heart ;8:e001735. doi: 10.1136/openhrt-2021-001735

Ogah, O. S., Okpechi, I., & Chukwuonye, I. I., (2012). *Blood pressure, prevalence of hypertension and hypertension related complications in Nigerian Africans: A review*. *World Journal of Cardiology*, 4(12), 327-340.

Sianga, B. E., Mbago, M. C., & Msengwa, A. S. (2024). *Bayesian spatial-temporal analysis and determinants of cardiovascular diseases in Tanzania mainland*. *BMC Medical Research Methodology, 24*(225). https://doi.org/10.1186/s12874-024-02348-6

Papathomas, M., & Richardson, S. (2016). Exploring dependence between categorical variables: Benefits and limitations of using variable selection within Bayesian clustering in relation to log-linear modelling with interaction terms. *Journal of Statistical Planning and Inference,* 173, 47-63.

Teo, K. K. (2021). Cardiovascular risk factors and prevention: A Perspective from Developing Countries. *Canadian Journal of Cardiology,* 37(5), 733-743. https://doi.org/10.1016/j.cjca.2021.02.009

World Health Organization (WHO, 2023). *Cardiovascular diseases (CVDs)*. https://www.who.int/health-topics/cardiovascular-diseases World Health Organization