International Journal of Innovative Scientific & Engineering Technologies Research 5(3):1-5, July-September, 2017

© SEAHI PUBLICATIONS, 2017 **www.seahipaj.org** **ISSN: 2360-896X**

Logistic Regression Analysis of Risk Factors for Hepatitis B Virus Infection in Jos – Plateau, Nigeria

[1]Datong, Godwin Monday; [2]Buba, Audu & [3]Ujah, Hilary, A.

[1]Department of Mathematics
University of Jos
Jos, Nigeria.
Email of Corresponding Author: mallong2007@yahoo.com

[2]Department of Mathematics,
University of Jos,
Jos, Nigeria.

[3]Department of Mathematics,
University of Jos
Jos, Nigeria
Email: hills4manutd@gmail.com

ABSTRACT
A logistic regression model was used for the analysis of risk factors for Hepatitis B Virus infection. The risk factors used are Age, History of Blood transfusion and Alcohol consumption from a sample of 693 patients of Plateau State Specialist Hospital, Jos, Nigeria. The results of the analysis showed that there is a moderate relationship of 12.8% between infection and the risk factors of Hepatitis B Virus. A Wald test statistic criterion demonstrated that all predictor variables made significant contributions for the analysis of HBV. The Hosmer and Lesmeshow (H-L) test statistic of 0.457, indicates that the model's estimates fit the data at an acceptable level of significance making the logistic regression model a good model for analyzing hepatitis B virus infection.
Keywords: Logistic regression model, Hepatitis B Virus, Odd ratio and Logit models.

INTRODUCTION

Logistic regression is a nonlinear regression model that belongs to a family of generalized linear models where the response variables are discrete and the error terms are not normally distributed. Logistic regression analysis examines the influence of various factors on a dichotomous outcome by estimating the probability of the events' occurrence in order to measure the relationship between the response variable and the covariates (Shakesha, 2001). The goal is to predict the category of outcome for individual cases using most parsimonious model.
The generalized logistic model (GLM) is given as:

$$Logit\,[\pi(x_i)] = ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad 1.1$$

This model explains the effects of the predictor variables on the response.

Logistic regression is a promising statistical technique that is widely used in any situation where the response variable tends to take on binary form. For example, in most health research studies, it is common for the response variable to have the form "yes or no", "responded to medication versus did not respond to medication", "reactive or non-reactive", "dead or alive, "infected or not infected", and situations in which logistic regression would be employed (Chatterjee et al., 2000). Logistic regression analysis has found widespread usage in the epidemiological literature, where often the dependent variable is the presence or absence of a disease state (Chuang, 1997). One of the aims of epidemiology is to study those factors which at a given moment affect the existence of a health problem, and to control the dimension of the latter, as well as to construct models with predictive capacity that assess the mentioned health problem (Dominguez et al., 2011). The phenomenon of infection is discrete or qualitative in nature, i.e. either infection occur or it does not. This binary discrete phenomenon usually takes the form of a dichotomous indicator variable. Logistic regression analysis examines the influence of various factors on a dichotomous outcome by estimating the probability of the event's occurrence. It does this by examining the relationship between one or more independent variables and the log odds of the dependent as opposed to the dependent variable itself. The log odds ratio is the ratio of two odds and it is a summary measure of the relationship between two variables. The use of the log odds ratio in logistic regression provides a more simplistic description of the probabilistic relationship of the variables and the outcome in comparison to a linear regression by which linear relationships and more rich information can be drawn (Hosmer et al., 2000).

1.2 Hepatitis B Virus

Hepatitis B is a virus, or infection, that causes liver disease and inflammation of the liver cells. HBV belong to the family of hepadnaviride and is the only hepadna virus causing infection in humans (Finlayson et al., 1999).

Hepatitis B virus (HBV) infection remains a significant health problem and a major cause of liver infection that can be life threatening and often leads to chronic disease, liver cirrhosis and liver cancer throughout the world. Despite the continuing scientific advances geared towards the treatment of this infectious disease, early diagnosis is essential to limit the extent at which the disease is spread and increases the potential for success of any definitive therapy provided. In a developing country like Nigeria, the present trend indicates that there is an increase in this disease and it plays an important role in deciding the health status of an individual.

Hepatitis B virus is a major cause of liver cancer and Hepatocellular carcinoma (HCC) is rated as the 9^{th} cause of death worldwide (Sule, et al., 2011). The risk assessment has become increasingly important in the prevention of any viral disease. However, the change in our knowledge of susceptibility factors as they affect initiation and progression, have led to intense study of risk factors of the disease.

To investigate the risk factors associated with HBV infection, the logistic regression methods have become an integral component of any data analysis concerned with the explanation of relationship between a dichotomous response variable and one or more explanatory variables called the factors. Many different types of linear and non-linear models have been seen in the literature and its use in many areas including hepatitis epidemiology.

2.0 LITERATURE REVIEW

In the early 1960s, Cornfield et al introduced the usage of Logistic regression, and in 1967, Walter and Duncan used this methodology to estimate the probability of occurrence of a process as a function of other variables.

The use of logistic regression increased during the 1980s, and it constitutes one of the widely used methods in research in the health sciences, most especially in epidemiology and prediction of risk factors of HBV infectious diseases (Dominguez et al., 2011). Sabria (2011) used logistic regression to predict risk of HBV infections among antenatal clinic attendance. Logistic regression was used in 2012, to study the prevalence of HBV infection among HIV patients in South Africa (Bibi, 2012). Similarly,

Shivalingappa and Parameshwar (2012) used logistic regression model to predict the risk factors of oral health diseases. James and Kim, (1991) developed a logistic regression model for describing the use of child safety seats for children involved in crashes in Hawaii. They concluded that drivers and children using seat belts are less likely to suffer serious injuries from crashes. Logistic regression was also used by Lian, 2012 to model DIF dictation.

2.1 Risks factors of Hepatitis B virus infection.

Hepatitis B virus lives in the blood and other body fluids such as semen, vaginal secretion, saliva, and menstrual blood and to a lesser extent, perspiration, breast milk, tears and urine of chronic infected persons (Yagua, 2011). Transmission of the virus occurs when blood or body fluid of an infected person comes in contact with that of a person who is not immuned (CDC, 2003).

Most studies in Nigeria found a low prevalence in infancy and an increasing rate with age (Amazigo, 1990). A figure of about 2.8% has been documented as the rate of HBV transmission from Nigerian females to their offspring (Emechebe et al., 2009).Most infections in Nigeria occur through horizontal transmission (Amazigo, 1990). Various studies in Nigeria showed that blood transmission is an important source of HBV transmission (Obiaya and Ebohom, 1982 and Emechebe et al., 2009). Although Centre for Disease Control and prevention (CDC) publications in July 2003 and a study in South Africa linked HBV transmission to tattoos and body cutting/piercing, most studies in Nigeria found no link between traditional practices like, scarification, circumcision, ear piercing and HBV infection (Chukwuka et al., 2003). Studies from north-central Nigeria indicate that unprotected sex is implicated in the transmission of HBV (Mustapha and Jibril, 2004).

The risk factors of HBV infection chosen for this study are; age, history of blood transfusion and alcohol consumption.

According to Macmillan dictionary, age is the number of years one has lived. Studies have shown that chronic infection with HBv occurs in 90% of infants at birth, 30% of children infected at 1 – 5 years and 6% of persons infected above 5 years (CDC, 2003). Thus there is inverse relationship between chronic infection and age due to the maturity of immune system, Agumadu and colleagues (2002) studying 213 children with sickle cell anaemia, showed that markers of HBV infection (HBsAg and anti HBc) increased with age.

Blood transfusion is a medical treatment that replaces blood lost through injury, surgery or illness. Various studies in Nigeria have shown that blood transfusion is an important source of HBV transmission (Obiaya and Ebohom, 1982 and Emechebe et al., 2009). Many people infected with the hepatitis B virus (HBV) wonder if their blood borne infection poses a health risk to others or not. In Benin, Obiaya et at., (1982) in their study noted that blood transfusion was hazardous in view of the high prevalence of HBsAg in donor blood. Multimer et al., (1994) found that blood transfusion clearly increased the risk of HBV infection as shown by significantly higher markers of HBV infection in subjects who were transfused. In Ibadan, Olubyide et al., (1997) found that a high (39%) prevalence of HBsAg was associated with surgeons and dentists, with high potential of transmissibility. They speculated that it was due to lack of vaccination and infrequent application of universal precautions.

Alcohol consumption is the drinking of beverages that contains ethanol. Excessive alcohol consumption increases the risk of liver injury, disease and death. Ndako et al., (2009) studied the prevalence of HBV infection amongst alcohol consumers and non-consumers in Bass local government area of Plateau state and recorded 83.61% prevalence rate amongst alcohol consumers.

3.0 THEORETICAL FRAMEWORK

Consider a simple linear regression model,

$$Y_i = \beta_0 + \beta_1 x_1 + e_i, \qquad (1)$$

where the response variable Y_i is binary, taking value of either 0 or 1. The expected response (Y_i) has a special meaning in this case, since $E(e_i) = 0$

$$E(Y_i) = \beta_0 + \beta_1 x_1 \qquad (2)$$

Now if we consider Y_i being a Bernoulli random variable for which we state the probability distribution as follows:

$$Y_i = \begin{cases} 1 & \text{if the event occurs} \\ 0 & \text{if the event does not occur} \end{cases}$$

Thus;

Y_i	Probability
1	P($Y_i = 1$) = π_i
0	P($Y_i = 0$) = 1 - π_i

Hence,

$$E(Y_i) = 1P(Y_i = 1) + 0\ P(Y_i = 0)$$
$$= 1(\pi_i) + 0(1 - \pi_i)$$
$$= \pi_i \qquad (3)$$

Equating equations (2) and (3) we have

$E(Y_i) = E(Y_i) = \beta_0 + \beta_1 x_1 = \pi_i$

Hence, the mean response $E(Y_i) = \beta_0 + \beta_1 x_1$ is simply the probability that $Y_i = 1$ when the level of the predictor variable is x_i.

To relate HBV infection with the risk factors, the logistic regression model will take on the general form;

$$Y_i = \pi(x_i) + e_i$$

where e_i is a random error that depends on the Bernoulli distribution of the response variable Y_i.

$$\pi(x_i) = E[Y \mid X] = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}$$

$\beta_i'^s$ are estimates of the model parameters while x_i is the explanatory variable. i = 0,1, 2 and 3.

Because $\pi(x_i)$ represents the probability that Y_i = 1 (i.e. not infected with HBV) given the value of x_i (i.e. expose to risk factor), 1 - $\pi(x_i)$ is the probability that Y_i = 0 (i.e. not infected with HBV) given the value of the risk factor, x_i.

The ratio $\frac{\pi(x_i)}{1 - \pi(x_i)}$ gives the odds of Y occurring, i.e. the chances of a person being infected with HBV.

The equation for the odds ratio gives a complex function to fit, the natural log of the odds, denoted as Logit $\{\pi(x_i)\}$, define thus;

Logit $\{\pi(x_i)\} = \ln\left\{\frac{\pi(x_i)}{1 - \pi(x_i)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2\, \beta_3 x_3$, which is of the linear form, provides an easier model to fit.

The parameters of the logit model are estimated by computer-intensive search procedures to find the maximum likelihood estimates b_0, b_1, b_2, b_3; with b_0 representing the slope while b_1, b_2 and b_3 representing the change in $\pi(x_i)$ for a unit change in x.

Thus;

$$b_i = \pi(x_i + 1) - \pi(x_i)$$

$$= \ln\left[\frac{\pi(x_i)}{1 - \pi(x_i + 1)}\right] - \ln\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right]$$

$= \ln(Odds_2) - \ln(Odds_1)$

$= \ln\left(\frac{Odds_2}{Odds_1}\right)$

$\rightarrow e^{b_i} = \frac{Odds_2}{Odds_1}$, which represents the percentage increase in the probability Y = 1 for each increase in X.

The significance of the model parameters are tested individually or in groups using the Wald test statistic, while the Hosmer and Lesmeshow goodness-of-fit test statistic is used to check the appropriateness of the logistic regression model. Meanwhile the Nagelkerke R^2 is used to determine the relationship between infection and the risk factors of HBV.

4.0 DATA ANALYSIS AND DISCUSSION OF RESULTS

Data collected by the Hematology department of Plateau Specialist Hospital, Jos was used for the analysis. The response variable Y is the Hepatitis B surface antigen (HBsAg), while the explanatory variables (risk factors) are Age, History of Blood transfusion and Alcohol consumption.

SPSS version 20.0 was used for the analysis of the data, yielding the following results:

Model parameters;

$$\beta_0 = 1.072,\ \beta_1 = -0.033,\ \beta_2 = -0.917,\ \beta_3 = -0.729.$$

Hence, the estimated Logit $[\pi(x)]$ function is

$$Logit[\pi(x)] = 1.072 - 0.033x_1 - 0.917x_2 - 0.729x_3.$$

To compute the predicted probabilities, its equation is:

$$\hat{\pi}(x) = \frac{e^{1.072-0.033x_1-0.917x_2-0.729x_3}}{1+e^{1.072-0.033x_1-0.917x_2-0.729x_3}}.$$

From the model summary output, the Wald test statistic values for the various variables were: Age = 18.718, HBT (1) = 22.86, Alcohol (1) = 13.825 and Constant term = 12.723; indicating that all the risk factors made significant contribution.

the Nagelkerke R^2 = 0.128, indicating a moderate relationship of 12.8% between infection and the risk factors of Hepatitis B.

For the goodness-of-fit, the Hosmer and Lesmeshow coefficient is 0.457, which is greater than 0.05; hence we conclude at 5% level of significance that the model fits the data.

Conclusively, our analysis indicates that the logistic regression model is a good model for the analysis of risk factors of Hepatitis B Virus infection.

However, further studies should be carried out on the role of socioeconomic risk factors in transmission of HBV, along with the predictive power of the logistic regression modeling of risk factors of HBV infection.

REFERENCES

Ado, A. *et al.* (2010). Sero-prevalence of Hepatitis B Surface Antigen among Blood donors attending Ahmadu Bello University Teaching Hospital, Zaria, Nigeria. *Bayero Journal of Pure& Applied Sciences, 3(1): 20 – 22.*

Amazigo, U.O. & Chime, A.B. (1990). Hepatitis B Virus infection in rural/urban population of Eastern Nigeria; Prevalence of serological markers. *East African Med J, 67(8): 539 – 544.*

Chatterjee, S., Ali, S. H and Bertram, P. (2000). Regression by Example. New York; John Wiley & Sons

Dominquez, S. A. (2011). *Logistic Regression Models*. SEIP; Elsevier Espana, S.L.

Emechebe, G. O. *et al.* (2009). Hepatis B virus infection in Nigeria – *A review Art.* 50 (1): 18-22

Hosmer, D.W. & Lesmeshow, S. *et al* (2000). *Applied Logistic Regression,* 2nd Edition, New York: John Wiley & Sons.