IMPACT: International Journal of Research in Applied, Natural and Social Sciences (IMPACT: IJRANSS)
ISSN(P): 2347-4580; ISSN(E): 2321-8851
Vol. 5, Issue 3, Mar 2017, 43-58

BINARY RESPONSE MODELS FOR OUTCOME OF PROSTATE CANCER SURGICAL OPERATIONS AMONG MALE NIGERIANS

ADEKANMBI, D. B

Department of Statistics, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

ABSTRACT

The goal of this study is to predict how the death of male prostate cancer patients subjected to surgical operation is associated with their prostate specific antigen (*PSA*), *age*, *grade* and *stage* of prostate cancer, and thereby determine the relationship between the likelihood of survival from the disease and other risk factors, using the most parsimonious model. Three logistic regression models were fitted to the outcome of prostate cancer surgical operation data. Model 1 fits all the four predictors, model 2 is a reduced model involving *PSA*, *age* and *grade* as predictors, and is nested in model 1, while model 3 involves *PSA*, *age*, *grade* and interaction term of age*PSA. Model1 gives a deviance value of 144.02 on 116 d.f (p=0.0399), model 2 yields a deviance of 145.27 on 119 d.f (p= 0.0511), while model 3 gives deviance of 0.7293 on 115 d.f (p=0.7293). Model 3 appeared to have fitted the data well but none of the predictors including the interaction term is significant in predicting the status patients to undergo surgical operation. The focus is therefore on model 2 for interpretation, which is moderately reasonable and revealed some of the predictors as significant. In fact, the AIC value for model 2 clearly revealed the model as the most parsimonious, compared to the other two models. The results of model 2 revealed that *PSA* and *age* were the two significantly important predictors to the model specification. Surprisingly *grade* made no significant contribution to the model in the presence of other predictors. The results also showed that for a unit increase in PSA of patients, the odds of dying from the surgical operation increases by 0.4783 with other variables fixed. Also, the older prostate cancer male patients have higher possibility of dying from the surgical operation of prostate cancer removal compared with the younger males. The odds of dying from prostate cancer surgical operation therefore increased with *PSA* and *age* of patients, so that the two predictors are critical to the survival of patients subjected to surgical operation. Few observations were identified as outliers from the residual plots, but they did not cause much perturbation in the model parameters on omission. Diagnostic evaluation of the model therefore revealed no major problem in the model. The area under the ROC for the three models ranges between 0.77 and 0.79, giving acceptable discrimination of the models. There was an indication of slight overdispersion in the data but does not call for concern. The proposed logistic models are useful in predicting the outcome of surgical operation of male prostate cancer patients; and could be used to generalize for other male Nigerians since genetics and environment have effect on the disease.

KEYWORDS: Prostate-Cancer, Logistic Regression Model, Effect Modifier, Residual Deviance; Receiver Operating Characteristic (ROC) Curve, Akaike Information Criterion

1.0 INTRODUCTION

Prostate is part of male reproductive system, located immediately below bladder, just in front of the rectum. It is about the size of walnut and surrounds part of the urethra, which is the tube that empties urine from the male bladder.

When cancer is formed in tissue of the prostate, it is referred to as prostate cancer. Prostate cancer occurs when one of the cell of the prostate reproduce rapidly than expected, which result into swelling or tumour, [4, 6]. Prostate cancer is becoming a major health concern in most developing nations, and many other parts of the world. It has been established that almost one man in eleven will develop prostate cancer during his life time, [21, 22,]. It was reported that between 1987 and 1992, the incidence rate of prostate cancer increased by 84% followed by a decline of 46% between 1992 and 1994, [22]. The disease is the leading cause of death among men aged 60 to 79. As reported, Nigeria ranked as the third highest among the countries of the world with significant prostate cancer burden, [16, 22].

Some of the symptoms of prostate cancer that can be identified in a patient having the disease include; frequent waking up at night to urinate, difficulties in starting to urinate, sudden needs to urinate, slow flow of urine and difficulty in stopping, painful ejaculation, blood in the urine or semen, decrease in libido, pain at back, hips, pelvis, shortness of breath and dizziness, [12, 16, 4, 6]. Various medical treatment have been developed, but the choice of treatment will differ for each individual based on a person's age, general health condition of the patient, grade and disease stage of the cancer, symptoms, lifestyle and personal choice, [12, 4, 21]. Medically, the causes of prostate cancer are not known yet, but researchers have identified risk factors of developing the disease. Some of the risk factors of developing prostate cancer are smoking, ageing, family history, and genes, high consumption of fat and red meats, obesity, use of sex hormones, sexually transmitted infections, and vasectomy, [22, 23, 12].

Common medical diagnosis tests for screening prostate cancer are digital rectal examination (DRE), prostate specific antigen (PSA), and Biopsy, [4, 6, 9, and 12]. DRE involves examination of a patient rectum trough insertion of a lubricated glove finger by a doctor into the rectum to feel the prostate through the rectal wall for lumps or abnormal enlargement. PSA is a substance made by the prostate that may be found in an increased amount in the blood of men who have prostate cancer. PSA test therefore measures the level of protein produced by the prostatic epithelium that can be detected in the blood. Biopsy involves removal of cells or tissues so that they can be viewed under a microscope by a pathologist, [4]. The pathologist will examine the tissue sample to see if there are cancer cells and find the *gleason score, (GS).*

The most important aspect of evaluating prostate cancer is to determine the disease stage to know how far the cancer has spread in a human body. This helps in defining prognosis which is useful when selecting therapies. The prognosis of a patient diagnosed with prostate cancer is the chance that the disease will be treated successfully and that the patient will recover, [12, 9]. The prognosis involves grading and staging of the prostate cancer of a patient. The system used to grade prostate cancer is known as the *Gleason score*, (GS), [12, 9, and 23]. The GS ranges from 2-10, describing how likely the tumour will spread, so that the higher the GS, the more aggressive the tumour is likely to be and the greater the chance that it has spread within the body. The system usually employed to stage prostate cancer is TNM, where T refers to the extent of the tumour, N refers to whether the lymph nodes are involved, while M refers to whether cells have metastasized or spread. After determining the TNM category of a patient, the information is combined with the gleason score and PSA, in a process called *disease stage grouping*. The overall disease stage is expressed in Roman numerals I for the least advanced to IV, the most advanced. This process assists in determining treatment options and the outlook for survival or cure for a prostate cancer patient, [9, 23].

The classical linear model cannot handle non-normal responses, such as counts or proportions, [11]. Generalized

Linear Model (GLM) extends the ideas underlying Linear models to situations when the response has binomial, Poisson, gamma and any other distribution that belong to the exponential family of distributions, [1, 7]. Link function is the main central ideas of GLMs in that it is used to link the linear predictors to the mean of the response. The choice of link function in GLM is based on assumptions derived from physical knowledge or convenience, [11]. If the predictors are discrete and the outcome variable is independent, then binomial distribution can be used for grouped data consisting of counts of successes in each group. Logistic regression analysis, which is a special case of GLM is a tool for modeling the effect of one or more risk factors on a binary (dichotomous) response, with one or more predictor that can be binary, categorical or continous, [8, 10, 13]. The data on prostate cancer can therefore be analysed by fitting logistic regression model to the binary responses. Logistic regression model is a generalized linear model with binomial response and link logit.

The basic focus of this study is to determine the significant factors among the possible risk factors that could lead to the survival or cure of prostate cancer patients, subjected to surgical operation. The risk factors that will be considered in this study are age of patients, PSA level, grade and disease stage of prostate cancer. The response variable is the status of the patients subjected to surgical operation which is dichotomous; either dead or alive. The description of the data used in this study is given in section 2.0. Section 3.0 focuses on theoretical model formulation of logistic model, while explanation on fitting logistic model and model selection are discussed in sections 4 and 5 respectively. Effect modification, goodness-of-fit test assessing contributions of predictors, model validation and measures to determine influential observations are discussed in sections 6.0, 7.0, 8.0, 9.0 and 9.1 respectively. Explanation on receiver operating characteristic curve is given in section 10.0. The results of the analysis and the interpretations are given in section 11.0. Important issues arising from the study are discussed in section 12.0.

2.0 DATA

The data employed in this study are secondary data based on the outcome of surgical operations of prostate cancer of male patients; as extracted from the records of Lagos State University Teaching Hospital (LASUTH) in Idi-Araba, Mushin Lagos, Nigeria. The data consist of frequencies of males diagnosed of prostate cancer and subjected to surgical operation in the teaching hospital. Clinical information such as age of the patients, the disease stage of the prostate cancer, and their Prostate Specific Antigen value (PSA) measured in mg/ml; were extracted from the patients' medical records. The data were recorded on monthly basis, spanning the period of four consecutive years, 2010 to 2013. Data were collected on a total number of 127 male patients in the teaching hospital, out of which 55 of them died after surgical operation.

3.0 MODEL FORMULATION

Logistic regression model is suitable for modelling discrete response variable having binary or dichotomous categories, [13, and 19]. The model is part of a category of statistical models called generalized linear models, and is simply referred to as model for binary responses, [19, 25, and 18]. With two categories of prostate cancer patients: either dead or alive, logistic model can be used to predict which of the two statuses a prostate cancer patient is likely to have before surgical operation, given certain other information.

Given that n responses of prostate cancer patients Y_i for $i = 1,.....n_i$ are independent and are binomially distributed with $B(n_i, p_i)$, then the binary random variable is defined as:

$$Y_i = \begin{cases} 1 \text{ if the prostate cancer patient is alive} \\ 0 \text{ if the prostate cancer patient is deceased} \end{cases}$$

The observed values of the response variable are linked to the model by the Binomial distribution, so that in cell i, if y_i be the number of prostate cancer patients dead from surgical operation observed in group *i*, then it could be assumed that y_i are distributed binomially with probability p_i. The count variable n_i is the total number of males in group *i* and p_i is the probability of observing prostate cancer in any male in group *i*. Then, Y_i's are independent binomial random variables with parameters n_i, p_i. The probability distribution function of Y_i is therefore given by:

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \quad \text{For } y_i = 0, 1, 2 \ldots n_i \tag{1}$$

With $Y_i \sim B(n_i, p_i)$ under the assumption that p_i is constant, it follows that $\mu_i = E(Y_i) = n_i p_i$ so that $p_i = \frac{\mu_i}{n_i}$, and $Var(Y_i) = n_i p_i (1-p_i)$.

That linear regression can be written as

$$\text{logit}(p_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots\ldots + \beta_k x_{ik} \tag{2}$$

In matrix form,

$$\eta_i = x_i' \beta$$

Where

$$x_i' \beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots\ldots + \beta_k x_{ik} .$$

x : is a vector of covariates, so that $x_{i1}, x_{i2}, \ldots\ldots\ldots, x_{ik}$ are the predictors.

β : is a vector of regression coefficients. They β's are the regression coefficients associated with the *k* exposure variables.

i: indicates individual observations.

The only continuous predictor considered in this study is *PSA* of prostate cancer patients. The other predictors are categorical in nature. The first logistic model referred to as model 1 then has linear predictor:

$$\begin{aligned} \eta_i = {} & \beta_0 + \beta_1 X_{i1} + \beta_2^{(1)} X_{i2}^{(1)} + \beta_2^{(2)} X_{i2}^{(2)} + \beta_2^{(3)} X_{i2}^{(3)} + \beta_2^{(4)} X_{i2}^{(4)} + \beta_3^{(1)} X_{i3}^{(1)} + \beta_3^{(2)} X_{i3}^{(2)} + \beta_3^{(3)} X_{i3}^{(3)} \\ & + \beta_4^{(1)} X_{i4}^{(1)} + \beta_4^{(2)} X_{i4}^{(2)} \end{aligned} \tag{3}$$

Where

X_{i1} : is the effect of *PSA* of prostate cancer patient.

X_{i2} : is the effect of *age* of prostate cancer patients, fitted as a categorical variable, with 4 dummy variables for the 5 levels of age.

X_{i3} : is the effect of Disease *stage* of prostate cancer, fitted as a categorical variable, with three dummy variables for the 4 levels of disease stage.

X_{i4} : is the effect of *Grade* of prostate cancer, also fitted as a categorical variable with two dummy variables for the 3 levels of grade.

4.0 FITTING LOGISTIC REGRESSION MODEL

Fitting a logistic regression model to a data set requires that the unknown parameters in the model should be estimated, [15]. In order to model the dependence of n binomial observations, $y_1, y_2, \ldots\ldots\ldots y_n$ on *k* predictors $x_1, x_2, \ldots\ldots\ldots x_n$, with an assumption that $Y_i \sim B(n_i, p_i)$. For polychotomous risk factors, the multiple logistic models, which is the logit of the underlying probability p_i is a linear function of the predictors, such that:

$$\text{logit}(\hat{p}_i) = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \eta_i = \beta_1 x_{i1} + \ldots\ldots + \beta_k x_{ik} \tag{4}$$

So that

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \sum \beta_k x_{ki} \tag{5}$$

The regression coefficients are estimated by the maximum likelihood method which is designed to maximize the likelihood of producing the data given the parameter estimates. The link function is $g(p_i) = \eta_i = \log\left(\frac{p_i}{1-p_i}\right)$.

The coefficients β's are the log-odds ratio. The sign of log-odds indicates the direction of its relationship, so that negative values indicate a negative relationship between the probability of 'success' and a predictor, while positive value indicates a positive relationship. Once the parameters of the model have been estimated, a back-transformation can be done to obtain estimates for p using

$$\hat{p}_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} = \frac{\exp(\beta_1 x_{i1} + \ldots\ldots + \beta_k x_{ik})}{1+\exp(\beta_1 x_{i1} + \ldots\ldots + \beta_k x_{ik})} \tag{6}$$

As $\eta_i \to \infty$, $p_i \to 1$ and as $\eta_i \to -\infty$, $p_i \to 0$. Therefore η_i is the logistic transform of p_i.

This is the linear logistic model for binary data. The logistic transformation is used for stretching the scale of the probabilities (p) from $(0,1)$ to $(-\infty, +\infty)$ thus ensuring that $0 < p < 1$, [17].

5.0 MODEL SELECTION

In order to determine the best subsets for logistic regression, the number of parameters may be taken into account using Akaike Information Criterion (AIC). It is a measure of relative goodness of fit of a statistical model, [19]. AIC is a measure of fit that penalizes for the number of parameters.

$$\begin{aligned} AIC &= D + 2p \\ &= -2l_{mod} + 2p \end{aligned} \tag{7}$$

Where

D: deviance statistic

P: number of parameters in the linear predictor of the model under consideration.

l_{mod} : Log-likelihood of the fitted model.

Smaller values of AIC indicate better fit, and thus the AIC can be used to compare models, whether nested or not, [1, 18].

6.0 EFFECT MODIFICATION

When degree of association between a disease and an exposure is different for each level of another variable, then there is an interaction between the exposure factor and the variable. The variable is said to modify the effect of the exposure factor on the disease and is referred to as an *effect modifier*. It is actually the interaction between exposure variables. Interaction could occur between two categorical variables, between a quantitative and a categorical variable and between two quantitative variables, [25, 8].

For effect modifier, the interest is to compute different odds ratios and relative risks for each level of the effect modifier. In order to determine whether a variable is an effect modifier or not, an interaction term between the variable and the risk factor of interest can be included in the logistic model. [25]. If the interaction term is both meaningful and statistically significant, then the variable is said to be an effect modifier, [15, 7].

7.0 GOODNESS-OF-FIT OF LOGISTIC REGRESSION MODEL

Deviance is a measure that can be used to determine how well a proposed logistic regression fits a set of data. It provides a measure of discrepancy between the current and full models, [17, and 11]. In order to assess the overall fit of a logistic regression, deviance can be employed which gives the extent to which the current model adequately represents the data. To compare L_f and L_c it is convenient to use deviance which is:

$$D = -2\log\frac{L_c}{L_f} = -2\{\log L_c - \log L_f\} \tag{8}$$

8.0 ASSESSING THE CONTRIBUTION OF PREDICTORS: WALD STATISTIC

The Wald statistic is usually used to determine whether a predictor is a significant predictor of the response. The statistic is used to test the significance of individual logistic regression coefficient for each predictor, [15]. The Wald test is

obtained by comparing the maximum likelihood estimate of each of β_j to their respective estimates of their standard errors, and Wald statistic (W) can be computed from the formula:

$$W = \frac{\hat{\beta}_j}{se\left(\hat{\beta}_j\right)} \tag{9}$$

The statistic should be used cautiously because when regression coefficient is large, the standard error becomes inflated, so that the Wald statistic is underestimated. Also the statistic is sensitive to changes in parameterization, [20].

9.0 MODEL VALIDATION

This refers to verification of the underlying assumptions of a logistic model to determine if they are well satisfied or not, and to determine the adequacy of the link function, [15, 8, 7, 11, 10]. Diagnostic methods can be graphical or numerical. Another aspect of model checking or model validation is checking for outliers or influential observations. These checks are usually based on graphical analysis of residuals or a transformation of these residuals, [19, 24].

Diagnostic measures in logistic regression model are based on residuals. There are several forms of residual in the binomial case.

Given that $\hat{e}_i = y_i - n_i\hat{p}_i$, then the Pearson Residuals is

$$\chi_i = \frac{\hat{e}_i}{\sqrt{n_i\hat{p}_i\left(1-\hat{p}_i\right)}} \tag{10}$$

So that the Chi-squared statistic $\chi^2 = \sum_{i=1}^{N} X_i^2$

The standardized Pearson residual is therefore:

$$r_{p_i} = \frac{X_i}{\sqrt{1-h_{ii}}} \tag{11}$$

Where h_{ii} the leverage for the i^{th} observation, is the i^{th} diagonal element of the hat matrix which for a GLM is:

$$H = W^{\frac{1}{2}}X\left(X^T W X\right)^{-1}X^T W^{\frac{1}{2}}$$

The Devaince residuals is

$$d_i = \pm\left[2\left\{y_i In\left(\frac{y_i}{n_i p_i}\right) + \left(n_i - y_i\right)In\left(\frac{n_i - y_i}{n_i\left(1-p_i\right)}\right)\right\}\right]^{1/2} \tag{12}$$

Either of the two residuals can be plotted versus the observation number to check for the form of the linear predictor. A systematic pattern in this plot is an indication that the model is incorrect. For a binary response, the logit

transformation is always appropriate and does not need checking, [11].

9.1 INFLUENTIAL OBSERVATIONS

Generally, outliers can be identified by standardized residuals greater than +2 or smaller than -2. Other diagnostics for identifying influential observations are:

- **Leverage** (h_i)**:** Is a measure of the leverage of covariate pattern 'i' to determine how much of an effect does an observation have on the estimated model. A point with high leverage has the potential to be influential. The diagonal elements of the 'hat matrix' are a measure of leverage of covariate pattern 'i'. A threshold used is that values greater than $2p/n$, indicate large leverage, where p is the number of variables in the model and n is the number of covariate patterns, [19, 15].
- **Cook's Statistic**

It is a popular influence diagnostic. Cook's statistic is a measure of the distance between the fitted logistic regression coefficients with and without each observation.

$$D_i = \frac{1}{p}\left(\hat{\beta} - \hat{\beta}_{(i)}\right)^T X^T W X \left(\hat{\beta} - \hat{\beta}_{(i)}\right) \tag{13}$$

D_i Can be computed using the formula

$$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_{p_i}^2 \tag{14}$$

The thresholds of 0.33 0r 1 signaled unusual observation, [8].

10.0 RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

Sensitivity and specificity are measures of model performance that are useful in determining the overall tests of goodness of fit. ROC curve is the plot of sensitivity versus 1-specificity over all possible cutpoints. The area under the curve provides a measure of discrimination, [2, 3, 2]. It is a visual index to compare competing logistic models. It plots the probability of detecting true signal (sensitivity) and false signal (1-specificity) over all possible cutpoints, [15, 5]. The curve generated by these points is called ROC curve. ROC curve provides a useful tool to evaluate the performance of classification schemes that categorise cases into one of two groups, [2, and 26]. The general rule for ROC curve is given in table 1, [15].

Table 1: General Rule of ROC Curve

ROC Area	Decision
ROC=0.5	No discrimination
$0.7 \le ROC < 0.8$	Acceptable discrimination
$0.8 \le ROC < 0.9$	Excellent discrimination
$ROC > 0.9$	Outstanding discrimination

11.0 RESULTS AND MODELS' INTERPRETATIONS

The data on outcome of surgical operation of prostate cancer patients was first subjected to data exploration, which is valuable in gaining understanding of the data. Figure 1 is the scatterplot of the pair of predictor versus predictor and response versus predictors, arranged in matrix form. It appears logistic regression model is appropriate in fitting the data. Table 1 shows the data expressed as counts, with values in parentheses as percentage. The age of the patients was categorized into five age groups with age-group 40-50 as the reference. The PSA of the patients ranges from 1.23 to 6.7 with a mean of 4.22. The variable disease stage and grade are categorized, having 4 and 3 categories respectively. The response variable is the status of prostate cancer patient: either dead or alive.

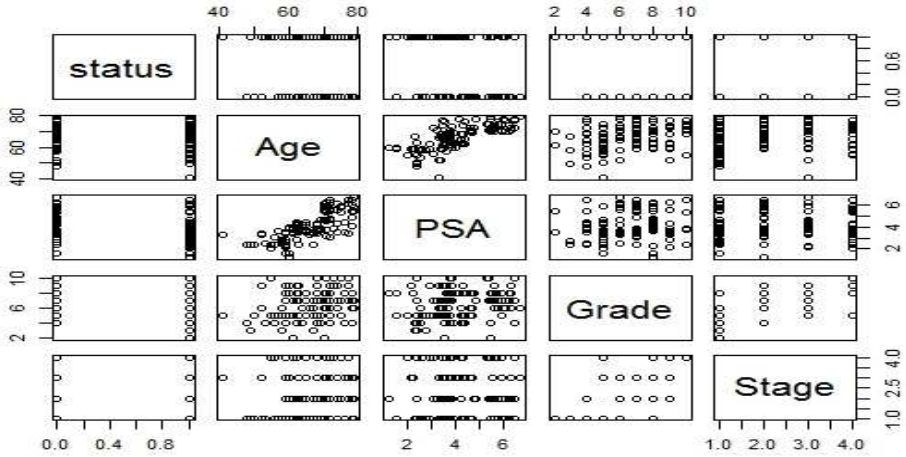

Figure 1: Scatterplot Matrix of Prostate Cancer Data

Table 2: Descriptive Analysis of the Prostate Cancer Data

Variables	Status		
	Alive(n=73)	Dead (n=54)	Total (n=127)
PSA	**73(57.5)**	**54(42.5)**	**127(100.0)**
Age 40-50	1(1.4)	2(3.7)	3(2.4)
51-60	18(24.7)	5(9.3)	23(18.1)
61-70	30(41.1)	18(33.3)	48(37.8)
71-80	22(30.1)	23(42.6)	45(35.4)
81-90	2(2.7)	6(11.1)	8(6.3)
Disease stage 1	29(39.7)	16(29.6)	45(35.4)
2	14(19.2)	16(29.6)	30(23.6)
3	16(21.9)	14(25.9)	30(23.6)
4	14(19.2)	8(14.8)	22(17.3)
Grade 2-4	15(20.5)	3(5.6)	18(14.2)
5-7	32(43.8)	31(57.4)	63(49.6)
8-10	26(35.6)	20(37.0)	46(36.2)

Table 3: Logistic Regression Output, Showing Parameter Estimates, Associated Standard Errors and Inferences for the Parameters in the Model

Model	Parameter	Variable Name	Estimate	Std. Error	Z -Value	$p > \|z\|$	Null Deviance (df)	Residual Deviance
Model 1	β_0		**1.6493**	**1.4579**	**1.131**	**0.2579**	**173.21 (126)**	**114.02 (116)**
	β_1	PSA	-0.7652	0.2557	-2.993	0.0028**		
	$\beta_2^{(1)}$	Age$^{(1)}$	3.5247	1.5777	2.234	0.0255*		
	$\beta_2^{(2)}$	Age$^{(2)}$	3.4310	1.5478	2.217	0.0267*		
	$\beta_2^{(3)}$	Age$^{(3)}$	4.0478	1.7098	2.367	0.0179*		
	$\beta_2^{(4)}$	Age$^{(4)}$	3.0774	1.9522	1.576	0.1149		
	$\beta_3^{(1)}$	Disease stage$^{(1)}$	-0.1554	0.5808	-0.268	0.7890		
	$\beta_3^{(2)}$	Disease stage$^{(2)}$	0.0971	0.7682	0.126	0.8994		
	$\beta_3^{(3)}$	Disease stage$^{(3)}$	0.5888	0.7416	0.794	0.4272		
	$\beta_4^{(1)}$	Grade$^{(1)}$	-2.0212	0.8668	-2.332	0.0197*		
	$\beta_4^{(2)}$	Grade$^{(2)}$	-1.8262	1.0473	-1.744	0.0812		
Model 2	β_0		1.5776	1.4525	1.086	0.2774	173.21(126)	145.27(119)
	β_1	PSA	-0.7375	0.2526	-2.920	0.0035**		
	$\beta_2^{(1)}$	Age$^{(1)}$	3.5598	1.5704	2.267	0.0234*		
	$\beta_2^{(2)}$	Age$^{(2)}$	3.3663	1.5399	2.186	0.0288*		
	$\beta_2^{(3)}$	Age$^{(3)}$	3.9369	1.6928	2.326	0.0200*		
	$\beta_2^{(4)}$	Age$^{(4)}$	2.9711	1.9363	1.534	0.1249		
	$\beta_3^{(1)}$	Grade$^{(1)}$	-1.9900	0.8322	-2.391	0.0168		
	$\beta_3^{(2)}$	Grade$^{(2)}$	-1.5709	0.8695	-1.807	0.0708		
Model 3	β_0		-2.604e+03	2.289e+05	-0.011	0.9909	173.21(126)	137.75(115)
	β_1	PSA	1.078e+03	9.501e+04	0.011	0.09909		
	$\beta_2^{(1)}$	Age$^{(1)}$	2.610e+03	2.289e+05	0.011	0.09909		
	$\beta_2^{(2)}$	Age$^{(2)}$	2.612e+03	2.289e+05	0.011	0.09909		
	$\beta_2^{(3)}$	Age$^{(3)}$	2.608e+03	2.289+05	0.011	0.09909		
	$\beta_2^{(4)}$	Age$^{(4)}$	2.602e+03	2.289+05	0.011	0.09909		
	$\beta_3^{(1)}$	Grade$^{(1)}$	-2.197	9.162e-01	-2.398	0.0165		
	$\beta_3^{(2)}$	Grade$^{(2)}$	-1.699	9.443e-01	-1.799	0.0720		
	$\beta_4^{(1)}$	Age$^{(1)}$:PSA	1.079e+03	9.501e+04	-0.011	0.9909		
	$\beta_4^{(2)}$	Age$^{(2)}$:PSA	1.080e+03	9.501e+04	-0.011	0.9909		
	$\beta_4^{(3)}$	Age$^{(3)}$:PSA	1.079e+03	9.501e+04	-0.011	0.9909		
	$\beta_4^{(4)}$	Age$^{(4)}$:PSA	1.078e+03	9.501e+04	-0.011	0.9909		

Three logistic models were fitted to the data. The estimates of the parameters of the logistic model of prostate cancer, the associated standard errors and inferences for the parameters in the models are shown in table 3. Model 1 contains all the four predictors, while model 2 is a reduced model involving three predictors only. The effect of *stage* variable is not significant in model 1 and was consequently removed to achieve reduced model 2. Model 3 contains three significant predictors, including interaction term of age and PSA. The reference levels of all the categorical variables are suppressed in the regression equations of the three models.

The estimated logistic regression equations for model 1 and model 2 are therefore:

Model 1

$$\begin{aligned}\text{logit} = {} & 1.6493 - 0.7652\text{PSA} + 3.5247\text{Age}^{(1)} + 3.4310\text{Age}^{(2)} + 4.0478\text{Age}^{(3)} + 3.0774\text{Age}^{(4)} \\ & -0.1554\text{Stage}^{(1)} + 0.0971\text{Stage}^{(2)} + 0.5889\text{Stage}^{(3)} - 2.0212\text{Grade}^{(1)} - 1.8262\text{Grade}^{(2)}\end{aligned}$$

Model 2

$$\begin{aligned}\text{logit} = {} & 1.5776 - 0.7375\text{PSA} + 3.5598\text{Age}^{(1)} + 3.3663\text{Age}^{(2)} + 3.9369\text{Age}^{(3)} + 2.9711\text{Age}^{(4)} \\ & -1.9900\text{Grade}^{(1)} - 1.5709\text{Grade}^{(2)}\end{aligned}$$

The residual deviance of model 1 is 114.02 on 116 d.f yields p-value of 0.0399, while the residual deviance of model 2 is 145.3 on 119 d.f yields a p-value of 0.0511; and the residual deviance for model3 is 137.75 on 115 d.f yields a p-value of 0.7293. The model with the highest p-value for residual deviance has a better prediction. This is an indication that model 1 is not adequate, while model 2 is moderately adequate and model 3 is more reasonable in interpreting the data, but none of the predictors including the interaction term in model 3 is significant in predicting the status of patient subjected to surgical operation. The focus is therefore on model 2 for interpretation, which is moderately reasonable and revealed some of the predictors as significant. Model 2 is clearly nested within model1; the difference between them is the variable for the risk factor disease stage.

According to model 2, the log of odds of a prostate cancer patient surviving is negatively related to his *PSA* and *grade* variables, but is positively related to his age. The model indicates that the intercept is not significant, so that it can be ignored. The association between *PSA* and *status* of patients is statistically significant, (p= 0.0035), and *status* is also significant with the first three dummy variables of age (p=0.0234, 0.0288, 0.0200). Disease *stage* is not significant in the presence of other three variables, indicating that the predictor appears to be redundant in the reduced model. The coefficient for PSA is -0.7375, which implies that for a unit increase in patients PSA, a patient subjected to surgical operation has a log-odds of surviving of -0.7375, which translates into an odds of 0.4783, with other variables fixed. This also translates into a probability of dying of 0.3236. The odds of dying from surgical operation for males in age interval 51-60 are 35.16 times higher than males in the age interval 40-50, controlling for other variables. Also the odds of dying from surgical operation for a male patient in the age interval 61-70 are 28.97 times higher than those in the age interval 40-50, with other variables fixed. The odds of a patient in the age interval 71-80 dying from surgical operation are 51.23 times higher than the odds for patients in the age interval 40-50, when other variables are controlled. The odds of dying from prostate cancer surgical operation therefore increase with PSA and age of patients. Older prostate males have higher possibility of dying from the surgical operation of prostate cancer removal compared with the younger males, so that age is

positively and significantly associated with the response variable, *status*.

The AIC value for model1 is 166.02, for model 2 are 161.27 and for model 3 is 161.75. Clearly model 2 is the most parsimonious based on the AIC values.

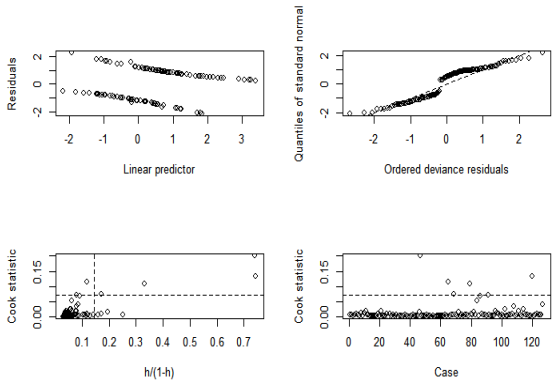

Figure 2: Residual Diagnostic Plots for Model 2 of Prostate Cancer

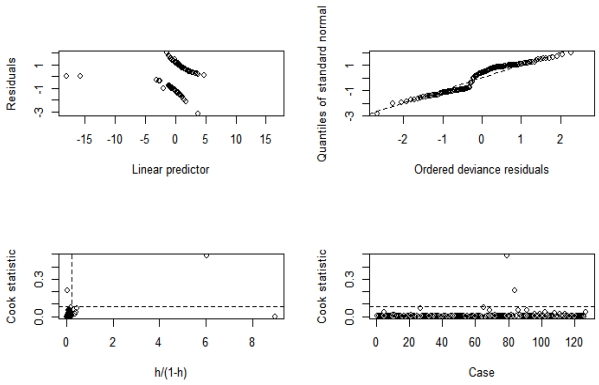

Figure 3: Residual Diagnostic Plots for Model 3 of Prostate Cancer

The plots of the residuals for model 2 are shown in figure 2, and the plots of residuals for model 3 are shown figure 3. The plots of jackknife deviance residuals against linear predictor for the two models displayed in the upper left panel of the figures, reveals dichotomous nature of logistic residuals which makes it almost impossible to discern any pattern in the plot. The normal Q-Q plots in the upper part of second panel of both figures 3 and 4; indicate that case 65 could be an outlier. The plots of the leverage statistic are shown in the lower part of the second panel of figures 2 and 3. From the plots, there are indications that observations 65, 47 and 121 are outliers, but not influential, since they do not cause much perturbation in the model parameters of both models 2 and 3 on omission. Generally, the residuals do not indicate any major problem with the modeling assumptions.

The estimates of overdispersion (φ) which is the ratio of residual deviance/d.f for the three models are $(\varphi = 1.24, 1.22 \text{ and } 1.20)$ respectively. There is therefore slight indication of overdispersion in the data, since for all the models the ratios are slightly >1. Model 3 shows the lowest overdispersion, which could be due to the influence of the interaction term in the model.

Figure 4 shows the Receiver Operating Characteristic (ROC) curve for the three logistic models which provides

index accuracy by showing the ability of the models to discriminate between the statuses of prostate cancer patients. The diagonal line is the reference line, and the further the curves are above the reference line, the more accurate the model. Based on their distances from the reference line, models 2 and 3 are good and are nearly indistinguishable, while model 1 is the worst. As shown in table 3, the area under the ROC curve for model1 is 0.766, with 95% confidence interval 0.683-0.848. For model 2, the area under the ROC curve is 0.768 with a 95% CI of 0.684-0.851, while the area under the ROC curve for model3 yields 0.797 with a 95% CI of 0.718-0.876. The p-value of each of the three models is less than 0.05. Model 3 has the largest area under the curve, though with a slight difference from that of model 2.

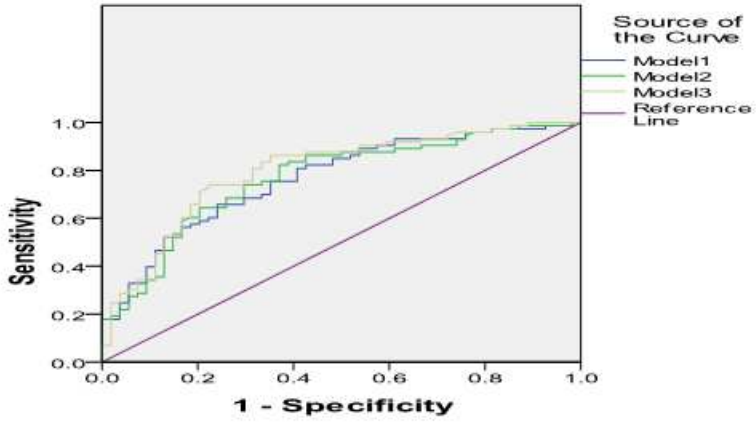

Figure 4: Receiver Operating Characteristic Curve of the Prostate Cancer Logistic Model

Table 4: Area of ROC Curve for the Prostate Cancer Logistic Model

Test Result	Area	Std. Error	Asymptotic Sig	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Model1	0.766	.042	.000	.683	.848
Model2	0.768	.043	.000	.684	.851
Model3	0.797	.040	.000	.718	.876

12.0 DISCUSSIONS

The focus of this study was to determine the risk factors that have significant effect on the status of prostate cancer patients that were subjected to surgical operation. The interest was to determine the effect of the following risk factors: *PSA, age, grade and disease stage* on whether a prostate cancer patient will survive surgical operation or die in the process. The logistic regression models of prostate cancer are based on the sample of prostate cancer patients in Lagos State University Teaching Hospital (LASUTH), which may not be representative of the Nigeria as a whole. The logistic models are based on the average outcome of the prostate cancer patients in LASUTH subjected to surgical operation, and may overestimate or underestimate individual risk due to differences in exposure or genetic susceptibility.

Risk factors of dying from prostate cancer though yet to be determined medically, but this study has identified PSA and age as the most statistically significant risk factors from dying from the surgical operation among male patients. This is an indication that the risk of dying from the surgical operation increases with PSA and age of the patients. The interaction term (age*PSA) though meaningful, but was not statistically significant to be regarded as an effect modifier.

CONCLUSIONS

A common problem in logistic modeling is over-dispersion also referred to as extra-binomial variation. There was an indication of slight over dispersion in the data as evidenced by the estimate of the dispersion parameter of the models, but the slight over dispersion does not call for concern. There are also unusual surprises noticed in the result of the analysis, which should be interpreted with caution, such as predictor *stage* and *grade* not being significant in the models. This could be as a result of correlation between the predictors. On its own, disease stage and grade could be significant, but in the presence of many other correlated predictors, they were no longer needed. Despite the stated limitations, the models provide an indication to determine the status of prostate cancer patients to be subjected to surgical operation.

REFERENCES

1. Agresti, A. Categorical data analysis. 2nd. Ed. New-York: John Wiley, 2002.
2. Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology*, 1975; 12: 387- 415.
3. Beck, R. J., and Shultz, E. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Arch. Pathol. Lab. Med.*, 1986; 110: 13-20.
4. Catalona W. J., Smith D. S., Ratliff T. L., and Basier JW. Detection of organ-confined prostate cancer is increased through prostate-specific antigen-based screening. JAMA 1993; 270: 948-954.
5. Centor, R. M., and J. S. Schwartz. An evaluation of methods for estimating the area under the receiver operating statistic (ROC) curve. *Med. Decis. Making*, 1985; 5: 149-156.
6. Chodak G. W., Keller P., and Schoenberg H. W. Assessment of screening for prostate cancer using the digital rectal examination. J Urol 1989; 141: 1136-1138.
7. Collet, D. Modelling binary data. (2nd Ed.). Chapman & Hall, 2002.
8. Dobson, AJ. An intodction to generalised linear models. CRC Press, 2010.
9. Dyke, C. H., Toi A., and Sweet, J. M. Value of random US-guided transrectal prostate biopsy. Radiology 1990; 176: 345-349.
10. Everitt, B. S and Hothorn, T. A handbook of statistical analyses using R. CRC Press, 2005.
11. Faraway, J. J. Extending the linear model with R: generalised linear, mixed effects and non-parametric regression models. USA: Chapman and Hall/CRC, 2006.
12. Friedman G. D., Hiatt, R. A., Quesenberry Jr, C. P., and Selby, J. V. Case-control study of screening for prostatic cancer by digital rectal examinations. Lancet 1991; 337:1526-1529.
13. Green, P. J., and Silverman, B. W. Non-parametric regression and generalized linear models: a roughness penalty approach. London: Chapman & Hall, 1993.
14. Hanley, J. A., and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 1982; 143:29-36.

15. Hosmer, D. W., and Lemeshow, S. Applied logistic regression. 2nd Ed. New-York: Wiley and Sons, Inc, 2000.
16. Jemal A., Siegel R., Ward E., Murray T., Xu, J., and Thun, M. J. Cancer statistics, 2007. CA Cancer J Clin 2007; 57: 43-66.
17. Krzanowski, WJ. An introduction to statistical modelling. London: Arnold, 1998.
18. Lawal, H. B. Categorical data analysis with SAS and SPSS applications. New-Jersey: Lawrence Erlbaum Associates, Inc. 2003.
19. McCullagh, P., and Nelder, J. A. Generalised linear models, 2nd Ed. London: Chapman and Hall, 1989.
20. Molenberghs, G., and Verbeke, G. Models for discrete longitudinal data, 2006.
21. Parkin D. M., Bray, F., Ferlay, J., and Pisani, P. Global cancer statistics, 2002. CA cancer J Clin 2005; 55: 74-108.
22. Reis, L. A. G., Eisner, M. P., Kosary, C. L. SEER Cancer Statistics Review, 1975-2001. Bethesda, MD: National Cancer Institute. Available at http://seer.cancer.gov/csr/1975 2001/Accessed March 20, 2015.
23. Smith, D. S., Catalona, W. J. Interexaminer variability of digital rectal examination in detecting prostate cancer. Urology 1995; 45: 70-74.
24. Weisberg, S. Applied linear regression. New-York: Wiley, 1985.
25. Woodward, M. Epidemiology: study design and data analysis 2nd Ed. Florida: Chapman and Hall/CRC, 2005.
26. Zweig, M. H., and Campbell, G. Receiver Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. Clinical Chemistry, 1993; 39:4, 561-577.