IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 13, Issue 2 Ver. IV (Mar. - Apr. 2017), PP 42-46
www.iosrjournals.org

Predictive Risk Factors of Prostate Cancer Incidence in Khartoum State, Sudan

Rayyan Ibrahim Mukhtar Ahmed [*1], Ahmed Mohamed Abdalla Hamdi[2], Altaiyb Omer Ahmed Mohmmed [3]

[1]*(Department of Statistics, Faculty of Science/ Sudan University of Science & Technology, Sudan)*
[2]*(Department of Statistics, Faculty of Science/ Sudan University of Science & Technology, Sudan)*
[3]*(Department of Statistics, Faculty of Science/ Sudan University of Science & Technology, Sudan), P.O BOX 407*

Abstract: *Prostate cancer is the most common men's cancer in the world. This study aimed to identify the predictive risk factors of prostate cancer incidence in order to set priorities for public heath interventions and to reduce the incidence of the disease. This study included patients with prostate cancer who were being treated at the National Center for Radiotherapy and Nuclear Medicine in Khartoum State, Sudan. 250 patients were chosen by interviews and from their medical history. The risk factors that increase the incidence of the disease were identified by using multiple logistic regression models. The odds ratio for the prostate-specific antigen (PSA) was 101, and for age was, 1.2, which significantly increased the risk of the incidence of prostate cancer. The odds ratio for states of (the former Central Region in Sudan was greater 77.9 times compared to Khartoum State. These potentially modifiable risk factors could be taken into account in making preventive interventions for prostate cancer patients.*

Keywords: *Incidence, PSA, Risk factor, Odds Ratio, Logistic.*

I. Introduction

Prostate cancer is a major public health problem in countries with aging population, and in places where people do not follow proper food habits [1]. Regarding cancer types, prostate cancer is the second most common cancer in men. An estimated 1.1 million men worldwide were diagnosed with prostate cancer in 2012, with an estimated 300,000 deaths in 2012. It is the fifth leading cause of death from cancer in men. In the GLOBOCAN 2012 report, prostate cancer incidence and mortality rate in Africa were reported to be 23.2 and 17.0 per 100,000 respectively [3]. Mortality rate is generally high in black population .Prostate cancer is considered the second among cancers in Sudan, with a high mortality rate [2]. This high mortality rate may be the result of late detection, since studies show that, by early diagnosis of the disease, 87 percent of men would be able to survive up to five years. Prostate Cancer should not be linked with benign prostate hypertrophy (BPH). BPH is the slow enlargement of the prostate gland that occurs in more than half of the men above 45, and while it is not a malignant condition, prostate cancer is prevalent in about 38 percent of men who undergo surgery to ease the symptoms caused by an enlarged prostate [5].

The research problem lies in how to build an accurate statistical model that describes the relationship between the incidence of prostate cancer and the risk factors, which helps in the diagnosis of the disease and makes the early treatment possible. The objective of this study is to identify the risk factors that increase the incidence of prostate cancer, and the necessity to classify the individuals in their appropriate groups (cases or control). This study included patients with prostate cancer who are being treated at the National Center for Radiotherapy and Nuclear Medicine in Khartoum State, Sudan. In this study, binary logistic regression is used in data analysis and to conduct an accurate model to describe the relationship between the risk of prostate cancer incidence and risk factors, and to classify new individuals in the appropriate group.

II. Logistic Regression Model

Logistic regression analysis is a statistical technique often used in different fields of research such as medical and social sciences, marketing, finance, etc. It was introduced in the late 1960s, as an alternative to ordinary least squares (OLS) regression. It established a wide application in statistical software programs during the 1980s, [6]. Its main goal is to find the best fitting model that best describes the relationship between an outcome and the set of independent variables [7]. The main mathematical concept under the logistic regression is the logit or the natural logarithm of an odds ratio. Logistic regression as a statistical method is suitable and usually used for testing hypothesis of the relationship between a categorical dependent or an outcome variable and one or more categorical or continuous predictors or independent variables. The dependent variable in logistic regression is binary or dichotomous. Logistic regression predicts the logit of Y to X. Since the logit is the natural logarithm (ln) of odds of Y, and the odds are the ratios of probabilities (π) of Y happening to probabilities (1–π) of Y not happening. The dependent variable in logistic regression can be presented as follows:

$$y = \begin{cases} 1 \\ 0 \end{cases}$$

Let us denote the p independent variables by the vector $X = \left(x_1, x_2, x_3, \dots x_p\right)$. If the conditional probability of the outcome $\pi(y = 1|x)$ is$\pi(x)$, the logistic regression model is given by the following equation:

$$\pi(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

The logistic regression has the following form:

$$logit\ (y) = natural\_log(odds) = \ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x$$

$$\pi = probability(Y = outcome\ of\ interest|X = x, specific\ value\ of\ x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Where (π) is probability of the outcome of interest, β_0 intercept of Y and β_1 regression coefficient. The logistic regression can also be stated in the form of odds:

$$\frac{\pi_i}{1 - \pi_i} = exp(\beta_0 + \beta_1 x_i)$$

It can also be presented in terms of probability:

$$\pi_i = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$$

In logistic regression the value of the coefficient β_1 determines the direction of the relationship between X and the logit of Y, while β_0 and β_1 are usually determined by maximum likelihood method (ML). The data are entered into the analysis as 0 and 1 coding for the dichotomous outcome. Usually, the null hypothesis of the logit model states that all β, s are equal to zero. If there exists at least one β different from zero the null hypothesis is rejected, and the logistic regression model predicts the probability of an outcome better than only the mean of the dependent variable marked as Y. The probability of X or $\pi(x)$ does not have a linear relation to coefficients in the logistic function, and the maximum likelihood is used. The maximization of the likelihood function expresses the probability of the data set as a function of the unknown parameters:

$$l(\beta) = \prod_{i=1}^{n} \pi_i(x_i)^{y_i}[1 - \pi_i(x_i)]^{y_i}$$

The use of logistic regression requires certain data pre-processing and model building methodology. In data pre-processing phase a question of variable selection is raised. The dependent variable is dichotomous, while the dependent variables can be usually dichotomous or continuous. It is less sensitive to statistical assumptions than the other statistical techniques. Many different model-building methodologies exist in logistic regression such as stepwise procedure (forward and backward). The statistical significance of individual regression coefficients is tested using the Wald chi-square statistics. Goodness-of-fit statistics assesses the fit of a logistic model against actual outcomes; the Hosmer–Lemeshow (H–L) test statistics (Pearson chi-square statistic) is used to test classification power for the logistic regression model [6].

III. Methodology

The maximum likelihood method, which yields values for the unknown parameters, is used for estimating the least squares function. Logistic regression solves such problems by applying the logit transformation. The case–control study was carried out in the national center for radiotherapy and nuclear medicine in Khartoum state, Sudan. Study subjects consisted of patients who are treated for prostate cancer during one year (2015 – 2016). 250 patients were collected through patient's interviews and from their medical history. Potential risk factors for the prostate cancer incidence were estimated using multiple logistic regression models using NCSS11. By using the forward selection method, the risk factors which significantly associated with the outcome were identified. (PSA, age, state=3; the states of (the former Central Region in Sudan)) were considered in the final model.

IV. Study Design And Data Collection:

The data were collected based on the logistic regression and the results were analyzed. The data consist of 16 independent variable: age, occupation, the state, marital status, age at marriage, family history, eating red meats and animal fats regularly, eating green vegetables and fruits regularly, suffering from overweight, high cholesterol, high blood pressure, ingestion of prostate medication, alcohol, smoking, developing one or more of these diseases: “syphilis, gonorrhea, Chronic prostatitis, enlarged prostate” and prostate specific antigen (PSA). The sample size was 250 individuals; 150 cases (with prostate cancer) and 100 were control (without prostate cancer). The outcome variable is (Diagnostic) represented the incidence of prostate cancer (1: Yes and 0: No)**.**

V. Results And Discussion

More variables that increase the incidence of the prostate cancer have been chosen from 16 variables, by using forward selection, the following tables illustrate the results:-

Table (1): Run summary

Item	**Value**	**Item**	**Value**
Y Variable	diagnostic	Rows Processed	250
Reference Value	without disease	Rows Used	250
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	0
Numeric X Variables	1	Rows Freq Miss. or 0	0
Categorical X Variables	15	Rows Prediction Only	0
Final Log Likelihood	-20.05139		
Model R²	0.88083	Sum of Frequencies	250
Actual Convergence	7.453633E-09	Likelihood Iterations	9
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	7	Completion Status	Normal Completion
Priors	Ni/N		
Subset Selection Method	Forward Selection		

This table specifies the independent variable’s name (diagnostic), and reference value is (without disease); this option specifies a reference value for the dependent variable, it is the outcome for which no regression equation is generated. This value could be text or numeric. Forward selection method was used to select the best subset from independent variables (X’s) with maximum iteration 20. The selection stops at five steps. The final log likelihood is equal (-20.05). The R^2 values tell us approximately how much variation in the outcome is explained by the model (88%), this implies that 88% percent of variation in Y caused by the independent variables. The Target Convergence is the amount that is used to stop the iterative fitting of the maximum likelihood algorithm (0.000001). If the Actual Convergence amount is larger than the Target amount, the algorithm ended before converging, and care must be taken in using any of the results. We used the choice that: Ni/N (Y-Value Proportions) for the prior probabilities as estimated by the Y-value proportions of the data. The Likelihood Iterations are the number of iterations necessary to solve the likelihood equations. In this case nine iterations are necessary.

Table (2): Y Variable Summary

Y Diagnostic	count	Unique rows (Y and X’s)	Y proportion	Y prior	R² (Y vs. Pred. probability)	Percent correctly classified
Without disease	100	96	0.40	0.40	0.92602	97.00
With disease	150	150	0.60	0.60	0.92602	99.33
Total	250	246				98.40

Variable summary describes number of individuals with and without disease, 250 individuals were collected which 100 without disease and 150 with disease, with portion 40% and 60% respectively. And the R^2 for the Y vs. predicted probability was 0.93, and 97% correctly classified the individuals without disease whoever, 99.3% correctly classified as with disease, with percentage of total for all 98.4%.

Table (3): Subset Selection Summary

Subset Selection Method = Forward Selection

No. Term	No. X’s	Log Likelihood	R² value	R² change	Entered
1	1	-168.25292	0.00000	0.00000	Intercept
2	2	-38.69427	0.77002	0.77002	PSA
3	3	-24.90294	0.85199	0.08197	Age
4	6	-22.15279	0.86834	0.01635	State
5	7	-20.05139	0.88083	0.01249	Alcohol

The Forward selection method was used to choice the best subset variables from the independent variables (X's). A five steps criterion conducted to get the best variables with the value of log likelihood, first step for intercept so the R^2 (usually use R^2 to determinant the important variable) is zero, so, there is no variables, with the (-168.25292) log likelihood, in the second step, (PSA) entered to the model with $R^2 = 0.77$ and log likelihood (-38.69427), in third step age entered so, R^2 was changed by 0.08 ($R^2 = 0.85$) and the log likelihood decreased to (-24.90294), fourth step state was entered also, R^2 was changed by 0.016 ($R^2 = 0.868$) also log likelihood decreased to (-22.15279), the last step (step five) the variable alcohol consumptions was entered, also R^2 was changed by 0.012 ($R^2 = 0.88$) also log likelihood increased to (-20.05139), so, the ranking of the important variables as in above.

Table (5): Coefficient Significance Tests

Independent Variable X	Regression Coefficients b(i)	Standard Error sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept	-10.71275	2.81841	-3.801	0.00014	0.00002
Age	0.16762	0.04580	3.660	0.00025	1.18249
(State=1)	1.29516	1.12620	1.150	0.25013	3.65158
(State=2)	1.74882	1.43864	1.216	0.22413	5.74784
(State=3)	4.35512	2.05720	2.117	0.03426	77.87620
(PSA=1)	4.62114	1.28241	3.603	0.00031	101.60999
(Alcohol =1)	-2.28197	1.28458	-1.776	0.07566	0.10208

In this study alcohol consumptions, state=2 (the states of northern and eastern Sudan) and state=1 (Darfur and kurdufan states) are insignificant, so they have no effect in the model. Also, this study showed the significant variables are age, state=3 and PSA with p-value 0.00025, 0.03426 and 0.00031 respectively. The prostate cancer incidence increased in men age over 50 years with ($\beta = 0.16762$) and odds ratio ($OR = 1.2$), which means that log of odds for incidence of prostate cancer was greater in men over 50 years (1.2) times than men less than 50 years, there are some studies that confirm validity of this study. Carter and Colleagues showed that 50% of men between 70 and 80 years of age showed histological evidence of malignancy. At that time risk of 42% for developing histological evidence of prostate cancer in 50-year-old men had been calculated. In men at this age, however, the risk of developing clinically significant disease is only 9.5%, and the risk of dying from prostate cancer was only 2.9% (8). Abnormal PSA increases the risk of the disease with very large odds ratio (101) and ($\beta = 4.62114$), that means PSA is most important variable in this study. Other studies have reached similar results, Ernesto P. Esteban et al (9); conducted the analytical study of 218 Japanese patients. They had first developed a theoretical framework to study PSA dynamics for BPH and prostate cancer patients. This analytical study then was applied to obtain monograms for a better understanding of the relationship among PSA and tumor volume in Japanese men with proven BPH or proven prostate cancer. This novel approach which does not neglect PSA contribution due to BPH may provide new information useful for a better diagnostic and prognosis of prostatic diseases or localized prostate cancer. They provided a relationship among PSA, age, and tumor volume. Another study provided by, Swanson KR, True LD et al (10), developed a mathematical model for the dynamics of serum levels of PSA as a function of the tumor volume. Their model results show good agreement with experimental observations and provide an explanation for the existence of significant prostatic tumor mass despite a low-serum PSA. This result can be very useful in enhancing the use of serum PSA levels as a marker for cancer growth. The state variable also has important role in this study, this variable consists of 4 categories (Khartoum state is the reference group). States of (formerly central region) "State=3" has large odds ratio equal to (77.9), that means men in States of (the former central region) have greater incidence rate (77.9) times than men live in Khartoum state, and (β =4.35512).

Table (6): Classification Table

		Estimated		Total
		Without disease	With disease	
Actual	Without disease	97	3	100
	With disease	1	149	150
	Total	98	152	250

To know the difference between the actual and estimated values were conducted by the model. There were 100 individuals actually without disease, while the number of the individuals without disease estimated by the model was (98). On the other hand, the model estimated (152) people to be diagnosed with the disease. There were (150) diagnosed with disease. So, the classification percentage of the model was 98.4%.

Estimated Logistic Regression Model(s) in Reading Form

Model for Logit(diagnostic) = XB when diagnostic = with disease

-10.71 + 0.17 * age + 1.30 * (state=1) + 1.75 * (state=2) + 4.36 * (state=3) + 4.62 * (PSA=1) - 2.28 * (alcohol consumptions=1)

Each model estimates XB (where Logit(Y) = XB) for a specific Y outcome. To calculate the Y-value probabilities when there are only 2 outcomes, transformation of the logit can be used as:

Prob(Y = outcome) = 1/ (1+Exp (-XB)) or Prob(Y ≠ outcome) = Exp (-XB)/ (1+Exp (-XB)).

VI. Conclusion

It has been found that, the most important predictive risk factors for prostate cancer incidence were age, PSA and state=3; States of (The former Central Region in Sudan), Khartoum state as reference group. By comparing p-value with ($\alpha = 0.05$), there is no strong evidence to reject the null hypothesis. So, alcohol, state=2(the states of northern and eastern Sudan) and state=1(Darfur and kurdufan states) were insignificant. The percentage correctly classified was 98.4%., so the resulting model was appropriate and accurate.

VII. Recommendations

Based on the research findings, the following points are to be recommended:

1- There should be an optimum use of the Multiple Logistic Regression in designing statistical classification models or in group separation, especially when there is a mixture of variables between the continuous and discrete variables, or when the variables or do not follow a normal distribution.

2- Further studies need to be conducted about prostate cancer to investigate the most dangerous risk factors that increase the prevalence of the disease, because it affects a large group of people in the society.

3- More research and studies should be carried out in the former Central Region in Sudan because there is a high prevalence of prostate cancer in this area.

References

[1]. https://www.pcrm.org/health/cancer-resources/diet-cancer/type/nutrition-and-prostate-health

[2]. http://globocan.iarc.fr/old/Factsheets/cancers/prostate-new.asp

[3]. F. Jacques, S. Isabelle, D. Rajesh, E. Sultan, M. Colin, R. Marise, P. Donald Maxwell, F. David, B. Freddie, Cancer incidence and mortality worldwide, *Cancer International Journal of Cancer, 136(5)*, 2012, E359-E386.

[4]. Report of the Sudanese Federal Ministry of Health for 2016

[5]. http://training.seer.cancer.gov/prostate/intro/ [National cancer institute & SEER Training Modules]

[6]. Y. Chao, P. Joanne, L.L. Kuk & I.M. Gary, An Introduction to Logistic Regression Analysis and Reporting, *The Journal of Educational Research, 96(1)*, 2002, 3-14

[7]. H.W. David, L. Stanley, *Applied Statistics Regression*(Wiley, New York, 2002)

[8]. Carter HB, Piantadosi S, Isaacs JT, Clinical evidence for and implications of the multistep development of prostate cancer, *J Urol, 143(4)*, (1990), 742-6.

[9]. E.P. Ernesto, D. Giovanni, R. Jaileen and L.M. Stephanie, An analytical study of prostate-specific antigen dynamics, *computational and mathematical methods in medicine, Volume 2016 (2016), Article ID 3929163*, 6 pages

[10]. Kristin R. Swanson, Lawrence D. True, Daniel W. Lin, Kent R. Buhler, Robert Vessella, and James D. Murray, A Quantitative Model for the Dynamics of Serum Prostate-Specific Antigen as a Marker for Cancerous Growth, *Am J Pathol, 158(6)*, (2001), 2195–2199.