Transactions of the Nigerian Association of Mathematical Physics
Volume 2, (November, 2016), **pp 163 – 170**

Logistic Regression for Cancer Data

G.O. Korter[1] and J.O. Omolehin[2]

[1]Department of Statistics, University of Ibadan, Nigeria.
[2]Department of Mathematical Sciences, Federal University Lokoja, Nigeria.

Abstract

More than 60% of the world's total new annual cancer cases occur in Africa, Asia and Central and South America. The aim of this study was to build a predictive model for the two possible outcomes of cancer patients and to examine which of the several types of cancer was more deadly. Secondary data of 335 patients aged 11 to 90 years who received treatment for liver, lung, colon, colorectal, prostate, breast or skin cancer at LAUTECH teaching hospital between 2004 and 2010 was used for this analysis. Logistic regression analysis was conducted. The Hosmer and Lemeshow and Likelihood Ratio tests were used to determine the fit and significance of parameters of the model. Only the type of cancer suffered by patients contributed significantly to the prediction model. The odds of dying for patients with lung cancer were about 4 times that of other types of cancer. However, the incidence of liver, lung, colon, colorectal, prostate, breast and skin cancer was prevalent across patients aged 11 and 90 years, irrespective of sex. Lung cancer was found to be more deadly than other types of cancer observed in the sample.

Keywords: Cancer, deaths, developing countries, logistic regression, prevention

1.0 Introduction

Cancer is one of the leading causes of death worldwide, with approximately 14.1 million new cases, 8.2 million cancer related deaths and 32.6 million people living with cancer - within 5 years of diagnosis in 2012 [1]. More than 60% of the world's total new annual cases occur in Africa, Asia and Central and South America. These regions account for 70% of the world's cancer deaths. It is expected that annual cancer cases will rise from 14 to 22 million in 2012 within the next 2 decades [2]. It has been established that there is a link between lifestyle and cancer [1, 3 – 7].

57% (8 million) of new cancer cases, 65% (5.3 million) of the cancer deaths and 48% (15.6 million) of the 5-year prevalent cancer cases occurred in the less developed regions. The overall age standardized cancer incidence rate is almost 25% higher in men than in women, with rates of 205 and 165 per 100,000, respectively. Male incidence rates vary almost five-fold across the different regions of the world, with rates ranging from 79 per 100,000 in Western Africa to 365 per 100,000 in Australia/New Zealand (with high rates of prostate cancer representing a significant driver of the latter). There is less variation in female incidence rates (almost three-fold) with rates ranging from 103 per 100,000 in South-Central Asia to 295 per 100,000 in Northern America. In terms of mortality, there is less regional variability than for incidence, the rates being 15% higher in more developed than in less developed regions in men, and 8% higher in women. In men, the rate is highest in Central and Eastern Europe (173 per 100,000) and lowest in Western Africa (69). In contrast, the highest rates in women are in Melanesia (119) and Eastern Africa (111), and the lowest in Central America (72) and South-Central (65) Asia [1].

The causes of cancer remain a global burden of concern. The relationship between asbestosis and bronchial cancer in South Africa was examined in [8]. In a necropsy series of 339 amphibole asbestos miners, heavy smoking, age and the presence of asbestosis were significantly associated with the presence of bronchial cancer. Also, in a case - controlled study among incident African patients with cancer in the Northern province of Southern Africa. Olga et al [9] measured the association between lung cancer and exposure to tobacco, occupational pollution and environmental pollution. Similar to [8], results suggested that tobacco smoking was the most important risk factor for the development of lung cancer. Risks for lung cancer were

Corresponding author: G.O. Korter, E-mail: kortergrace@gmail.com, Tel.: +2348033578643(JOO)

reminiscent of those observed in Western countries in the 1960s and 1970s. However, environmental exposure to asbestos, a dusty occupation (in men), and perhaps indoor air pollution could have contributed to the development of lung cancer in the province.

Alexis Elbaz et al [10] conducted a population base case-controlled study to investigate the association of Parkinson's disease (PD) with preceding nonfatal cancer in Olmsted county. The authors did not find a strong association between PD and preceding nonfatal cancer, but, there were suggestive trends in analyses stratified by sex and age at the onset of PD and for specific cancers related to smoking or hormonal factors. Breast cancer is another major concern thought to be linked to smoking. For example, data was obtained from Massachusetts vital statistics registry (United States). After adjusting for potential confounders, women who smoked during pregnancy were found not to have an increased risk of breast cancer compared to women who did not smoke during pregnancy. Also, Aliza and Timothy [11] investigated an earlier report that there is a five – fold increase in breast cancer risk among women who smoked during pregnancy. Similar to [10] an increased risk of breast cancer in women who smoked during pregnancy was not observed.

Apart from tobacco smoking, family cancer history and previous lung disease are risk factors that could be associated with lung cancer. Wang et al [12] evaluated the associations of previous lung disease and family cancer history with the occurrence of lung cancer among Chinese women in Hong Kong. It was found that all previous lung diseases, except chronic bronchitis were related to an elevated risk for lung cancer. Also, the association with asthma was found to be significant. Those that had more than one previous lung diseases tend to be at higher risk than those with only one of them. Positive family history of any cancer was associated with over two fold risk than negative family history. The joint effect of positive history of previous pulmonary diseases and positive family cancer history appeared to be additive, indicating the two factors acted independently. The results supported an etiological link of pre-existing lung disease and family cancer history to the risk of lung cancer.

Invariably, tobacco use is a major cause of lung cancer in some parts of Africa and worldwide [8, 9, 13, 14]. Other risk factors include obesity, a poor diet, lack of physical activity, alcohol consumption, exposure to ionizing radiation, environmental pollutants and infections such as hepatitis B, hepatitis C, and human papillomavirus. These factors act, at least partly, by changing the genes of a cell. Typically, many such genetic changes are required before cancer develops. Cancer can be detected by certain signs and symptoms or screening tests. It could then be further investigated by medical imaging and confirmed by biopsy [5, 13 – 16].

Research have shown that many cancers can be prevented by not smoking, maintaining a healthy weight, low alcohol consumption, eating of vegetables, fruits and whole grains, being vaccinated against certain infectious diseases, not eating too much red meat, avoiding excessive exposure to sunlight, early detection and treatment, human papillomavirus (HPV) vaccination and regular checkups [3- 7, 13, 14, 17]. Also, early detection through screening is useful in mitigating the effect and magnitude of cancer. Cancer is often treated with some combination of radiation therapy, surgery, chemotherapy and targeted therapy. Pain and symptom management are an important part of care. Palliative care is particularly important in those with advanced disease. The type of cancer and extent of disease at the start of treatment determines the chances of survival [2, 3, 13, 14, 16].

To achieve maximum completeness in case finding procedures, cancer registries should ensure the calculated incidence rates (and survival statistics) are as close to the true value as possible. Maxwell et al [18] used a large case series, collected independently of the cancer registry case finding mechanisms, as part of a study of the influence of HIV infection on cancer risk to evaluate the completeness of the registry in Kampala, Uganda for the 1994 -1996 period. Results showed that completeness of registration of diagnosed cancer cases was 89.6% for adults aged 15 or more. It varied with age (better ascertainment of younger cases, aged under 30 and cancer site (with Kaposi Sarcoma cases significantly better identified). Cases with a histology report were more likely to be registered than those without (though the difference was not significant). Completeness was found to decline with time, as in most registries, which continue to identify late cases sometime after the initial diagnosis. Provided that there was no insistence on the very latest results, the results gave reassurance that published incidence rates were reasonably accurate.

Also, Curado et al [19] identified cancer registration data as a factor in the design and monitoring of cancer control activities and policies, and population based cancer registries. The aim was to demonstrate the value of cancer incidence data for low and middle income countries where health care facilities are limited or scarce, and cancer registration may be of low quality. In addition, they interpreted the messages that the quality indicators conveyed both for cancer registration and the health care system. The study concluded, low data quality signals, lack of collaboration among reporting source and the inability of the register to perform quality abstracting points to specific weaknesses of the cancer care system. Apparently, quality data is a determinant of cancer preventive measures and treatment in Africa and worldwide.

With a focus on the treatment of cancer, the question is, could there be factors to determine the survival or non-survival of cancer patients upon cancer detection? Does the age, sex, type of cancer and length of care and other factors have any statistically significant relationship with the survival of cancer patients admitted in a hospital or otherwise? The aim of this study is to build a predictive model for survivorship of cancer patients considering their age, sex, type of cancer and length of stay (in days) in the hospital. In addition, to examine which of the several types of cancer was considered more deadly.

Logistic regression is a statistical approach to prediction. The logistic regression analysis is used to model when the dependent variable is dichotomous (binary). It is used to describe data and to explain relationship between one dependent binary variable and one or more metric independent variables. It assumes that the dependent variable is a stochastic event. Logistic regression uses maximum likelihood estimation after transforming the predictor to a logit variable. By so doing, it estimates the odds of a certain event occurring [20 – 24].

2.0 Method

2.1 Data

Secondary data obtained from the medical records of 335 cancer patients who were attended to at Ladoke Akintola University of Technology Teaching Hospital, Osogbo, Osun State, Nigeria between 2004 and 2010 was used. In this study liver, lung and others (colon, colorectal, prostate, breast, skin) cancers were considered. The age, sex, types of cancer, length of stay (in days) in the hospital were the independent variables, while the survivorship (dead or alive) of the patients was the outcome of interest.

2.2 The Logistic Regression Model

Logistic regression was used to model the relationship between the outcome variable and the independent factors.
Suppose y represent a binary response variable from cancer patients with probabilities of two categories denoted by ρ and $1-\rho$ that can take the values 1 and 0. Let ρ be the probability (p) of dying from cancer and $1-\rho$ be the probability of not dying from cancer.
Thus,

$$p(y=1)=1-p(y=0)=\rho \tag{1}$$

$$In\left(\frac{\rho}{1-\rho}\right)=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n \tag{2}$$

The In symbol refers to a natural logarithm and $\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n$ is the familiar equation for the regression line. The ρ can be computed from the regression equation, knowing the regression equation, we could theoretically, calculate the expected probability that $Y=1$ for a given value of X .

Therefore, the probability of dying from cancer is given as

$$\rho=\frac{\exp^{(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n)}}{1+\exp^{(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n)}} \tag{3}$$

$$\rho=\frac{\exp^{\left(\beta_0+\sum_{i=1}^{n}\beta_ix_i\right)}}{1+\exp^{\left(\beta_0+\sum_{i=1}^{n}\beta_ix_i\right)}} \tag{4}$$

where β_0 serves as the bench mark for the equation and β_i is the coefficient of the exogenous variables x_i for $i=1,2,3,4$ with $x_1=Age(Years)$; $x_2=Sex$: Male(reference) and Female; $x_3=$ Length of care in hospital (Days) and $x_4=$ Type of cancer {Liver, Lung, Others(reference)}.

2.2.1 The Model Fit

The likelihood ratio test was used to fit the model with and without the predictors. To examine the overall fit of the model the Hosmer and Lemeshow (H-L) test was applied. The Maximum Likelihood (ML) is a way of finding the smallest possible deviance between the observed and predicted values using calculus. With the ML, the computer uses different iterations in which it tries different solutions until it gets the smallest possible deviance or best fit. Once it gets the best solution, it provides a final value for the deviance, the negative two log likelihood (-2LogLikelihood) which can be thought of as a Chi-Square value. If the H-L goodness of fit test statistic is greater than 0.05, we fail to reject the null hypothesis that there is no difference between the observed and the model-predicted values, implying that the model estimate fit the data at an acceptable level.

2.2.2 Test of significance for the model parameters

The likelihood ratio test is a Chi- Square difference test using the null or constant only model. Instead of using the deviance (-2LogLikelihood) in the H-L test to judge the overall fit of a model, however, another statistic is usually used that compares

the fit of the model with and without predictors. The likelihood ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_0). The likelihood – ratio test statistic equal

$$Z = \frac{\hat{\beta}}{Se_{\beta_i}} - 2\log\left(\frac{L_0}{L_1}\right) = -2\left[\log(L_0) - \log(L_1)\right] = -2(L_0 - L_1) \quad (5)$$

This log transformation of the likelihood function yields a Chi- Square statistic.

3.0 Result

Table 1 shows a summary statistics of secondary data used for this study.

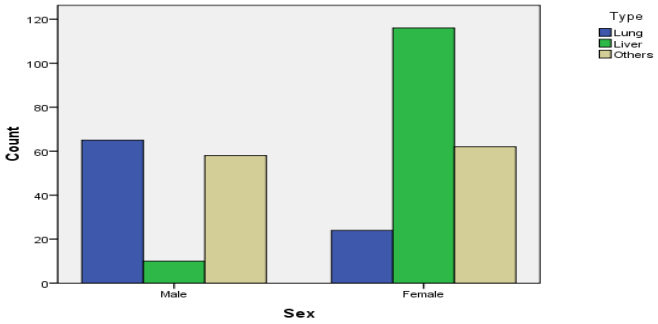
Table 1: Demographic profile of cancer patients in the studied sample

Characteristics	Categories	Frequency	Percentage in the population
Age	11-43	87	26.0
	44-53	77	23.0
	54-63	66	19.7
	64-90	105	31.3
Sex	Male	133	39.7
	Female	202	60.3
Type of cancer	Lung	89	26.6
	Liver	126	37.6
	Others*	120	35.5
Outcome	Survived	217	64.8
	Dead	118	35.2

*colon, colorectal, prostate, breast, skin etc

Source: *Compiled by the authors using data from LAUTECH teaching hospital, Oshogbo, Osun State, Nigeria*

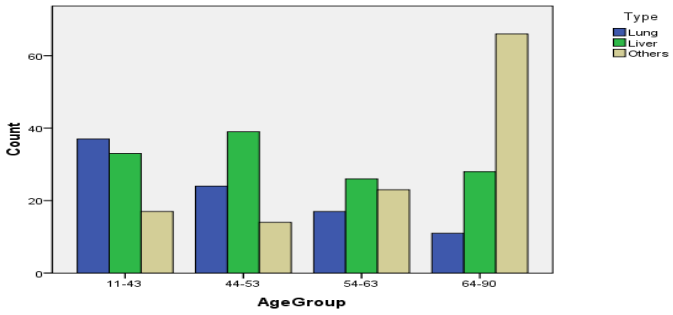
Figure 1 shows liver cancer was more prevalent with females (green), while, the incidence of lung cancer was more with the males (blue).



Source: *Compiled by the authors using data from LAUTECH teaching hospital, Oshogbo, Osun State, Nigeria*

Figure 1: Occurrence of cancer cases by type and sex

Figure 2 shows the frequency of cancer patients by type and age for the period under study. Most of the patients with colon, colorectal, prostate, breast, skin cancer were between ages 64 and 90 (brown) years. The incidence of lung and liver cancer was more prevalent amongst patients between 11 and 43 (blue) and 44 and 53 (green) years respectively.



Source: *Compiled by the authors using data from LAUTECH teaching hospital, Oshogbo, Osun State, Nigeria*

Figure 2: Occurrence of cancer cases by type and age of patients

A logistic regression model with outcome (dead, alive) as dependent variables and age, sex, type of cancer, and length of stay (in days) in the hospital as explanatory variables was fitted. These variables were then screened to obtain a parsimonious model. Table 2 shows the first step of logistic regression analysis for the cancer patients under study. The deviance $-2LL$ of the model was 408.931.

Table 2: The logistic regression model for cancer patients

Variables	B	S.E	Wald	Df	Sig.	Exp(B)
Age	-0.008	0.008	0.945	1	0.331	0.992
Sex	0.323	0.296	1.195	1	0.274	1.381
Length of Stay (Days)	0.004	0.004	0.828	1	0.363	1.004
Type of Cancer			14.938	2	0.001	
Lung	1.307	0.338	14.983	1	0.000	3.695
Liver	0.580	0.327	3.143	1	0.076	1.787

Key: B=Coefficients; S.E.= Standard Error; Df = Degrees of freedom;
Sig.=Significance; Exp(B)= Exponent(B)

The overall fit of the model was examined using the Hosmer and Lemeshow test, for the null hypothesis that there was no difference between the observed and the model-predicted values. The Chi-Square value was 6.739. The H-L statistic (0.565) was greater than 0.05, therefore, we did not reject the null hypothesis. This suggests that the model estimate fits the data at an acceptable 5% level of significance and can be used to predict the probability of dying by cancer patients. Table 3 gives a summary of the likelihood ratio test for the individual parameters. Therefore age, length of stay and sex were dropped from the model. Invariably only the type of cancer that a patient suffered from significantly contributes to the outcome of the prediction model.

Table 3: Summary for the likelihood ratio test for individual coefficients

Variables	-2LL	Df	Sig.	Change in -2LL
Age	410.342	1	0.185	1.411
Length of stay	409.654	1	0.116	0.732
Sex	408.931	1	0.836	2.380
Type of Cancer	423.207	3	0.006	14.276

Key: -2LL= -2LogLikelihood; Df = Degrees of freedom; Sig.=Significance

Table 4 shows the model with only type of cancer as the independent variable. The $-2LL$ for this model was 423.207. The predictive logistic regression model was $log\left(\frac{\rho}{1-\rho}\right) = -1.237 + 1.485x_1 + 0.508x_2$.

The odds of dying by lung cancer increased by 4.416 times that of other types of cancer considered, while, the odds of dying by liver cancer increased by 1.661 times that of other types of cancer considered.

Table 4: Parameter estimate for the model with only type of cancer as the independent variable

Variable	B	S.E	Wald	Df	Sig	Exp(B)
Type of cancer			24.677	2	0.006	
Lung	1.485	0.306	23.610	1	0.000	4.416
Liver	0.508	0.219	32.006	1	0.080	1.661

Key: B=Coefficients; S.E.= Standard Error; Df = Degrees of freedom; Sig.=Significance; Exp(B)= Exponent(B)

4.0 Discussion

The category of patients found with lung and liver cancer form an active part of any country's economy, such that, the gross domestic product (GDP), gross national product (GNP) and the per capita income, may be negatively affected. A war on cancer needs to be declared. There is an urgent need for a strong campaign against the causes of cancer to aid social and economic development especially in Nigeria and other developing countries.

Only the type of cancer suffered by patients contributed significantly to the prediction of the outcome of the model. This result supports the position of [15], [22] and [23] that the chance of survival depends on the type of cancer and extent of disease at the start of treatment. This study did not obtain information on the extent of disease at the start of treatment, future studies may consider other independent variables such as type of medical intervention (chemotherapy, surgery etc) and the quality of health services received by patients. All other exogenous factors namely, age, sex and length of stay in the hospital considered were insignificant. The odds of dying for patients with lung and liver cancer were 4.416 and 1.661 times that of patients with colon, colorectal, prostate, breast and skin cancer respectively.

The incidence of liver, lung, colon, colorectal, prostate, breast and skin cancer was prevalent among persons between ages 11 and 90 years, irrespective of sex. The patient's age, sex and length of stay in the hospital did not determine whether the patient will survive or not. Ultimately, lung cancer was found to be more deadly than all other types of cancer observed in the sample studied. This result is similar to the findings of [16] and [19]. The findings show an urgent need for aggressive measures against the causes of cancer specifically at grassroots.

As suggested by [10], a substantial proportion of the worldwide burden of cancer could be prevented through the application of existing cancer control knowledge and by implementing programs for tobacco control, vaccination (for liver and cervical cancers), and early detection and treatment, as well as public health campaigns promoting physical activity and a healthier dietary intake. Clinicians, public health professionals, and policy makers can play an active role in accelerating the application of such interventions globally. This study should enable the orientation of health policies targeted towards individuals, cancer patients and the health sector in the study area, Nigeria and other developing countries.

5.0 References

[1] GLOBOCAN. (2015, November 11). GLOBOCAN Estimated Incidence, Mortality and Prevalence Worldwide in 2012. http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx.

[2] Stewart, B. W. & Wild, C. P (2014) World Cancer Report. World Health Organization.

[3] Jemal A., Bray, F., Center, M.M., Ferlay, J., Ward, E. & Forman, D. (2011). Global cancer statistics. CA: A Cancer Journal for Clinicians, 61(2), 69–90.

[4] Kushi L.H., Doyle C., McCullough M., Rock C.L., Demark-Wahnefried W., Bandera E.V., Gapstur S., Patel A.V., Andrews K. & Gansler T. (2012). American cancer society guidelines on nutrition and physical activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity. CA: A Cancer Journal for Clinicians, 62 (1), 30–67.

[5] National Cancer Institute. (2015, July 4). Obesity and Cancer Risk. http://www.cancer.gov/about-cancer/causes-prevention/risk/obesity/obesity-fact-sheet#q3

[6] Parkin, D.M., Boyd, L. & Walker, L.C. (2012). The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. British Journal of Cancer, 105 (2), 77–81.

[7] World Cancer Research Fund International. (2015, November 11). Link between lifestyle and cancer risk. http://www.wcrf.org/int/link-between-lifestyle-cancer-risk.

[8] Sluis-Cremer G.K. & B.N. Bezuidenhout. (1989). Relation between asbestosis and bronchial cancer in Amphibole abestos miners. British Journal of Industrial Medicine, 46(8), 537-540.

[9] Olga Mzileni, Freddy Sitas, Krisela Steyn, Henri Carrara & Pieter Bekker. (1999). Lung cancer, tobacco and environmental factors in the African population of the Northern Province, South Africa. Tobacco Control, 8(4), 398-401.

[10] Alexis Elbaz, Brett J. Peterson , Ping Yang, Jay A. Van Gerpen, James H. Bower, Demetrius M. Maragonore, Shannon K. McDonnell, J. Eric Ahlskog & Walter A. Rocca. (2002). Nonfatal cancer preceding Parkinson's disease: a case-control study. Epidemiology, 13(2), 157-164.

[11] Aliza K. Fink & Timothy L. Lash. (2003). A null association between smoking during pregnancy and breast cancer using Massachusetts registry data (United States). Cancer Causes and Control, 14(5), 497-503.

[12] Xiao-Rong Wang, Ignatius T.S.Yu, Yuk Lan Chiu, Hong Qiu, Zhenming Fu, William Goggins, Joseph S.K. Au, Lap-Ah Tse & Tze-Wai Wong. (2009). Previous pulmonary disease and family cancer history increase the risk of lung cancer among Hong Kong women. Cancer Causes and Control, 20(5), 757-763.

[13] WHO (2015a). (2015, November 11). Cancer. http://www.who.int/mediacentre/factsheets/fs297/en/ .

[14] WHO (2015b). (2015, November 11). Global cancer rates could increase by 50% to 15 million by 2020. http://www.who.int/mediacentre/news/releases/2003/pr27/en/.

[15] Gijs Walraven. (2003). Prevention of cervical cancer in Africa: a daunting task?. African Journal of Reproductive Health, 7(2),7-12.

[16] Gotzsche P.C. & Jorgensen K.J. (2013). Screening for breast cancer with mammography. The Cochrane Database of Systematic Reviews, 6, CD001877.

[17] Robert S. Taylor. (2014). Religious conservatives and safe sex: reconciliation by non-public reason. American Political Thought, 3(2), 322-340.

[18] Maxwell D. Parkin, Henry Wabinga & Sarah Nambooze. (2001). Completeness in an African cancer registry. Cancer Causes and Control, 12(2), 147-152.

[19] Maria Paula Curado, Lydia Voti & Ana Maria Sortino-Rachou. (2009). Cancer registration data and quality indicators in low and middle income countries: their interpretation and potential use for the improvement of cancer care. Cancer cause and control, 20(5), 751-756.

[20] Agresti, A. (2002). Categorical Data Analysis. 2nd edition, New York, John Wiley.

[21] Field, A. (2009). Discovering Statistics Using SPSS. 3rd edition, London, Saga Publication.

[22] Gujarati, D. (2006). Essential of Econometrics. 3rd edition, New York, McGraw- Hill.

[23] Hosmer, D.W., Lemeshow, S.(2000). Applied Logistics Regression. 2nd edition, New York, Wiley.

[24] Krammer , J.S. (1991). The Logit Model for Economists. London: Edward Arnold Publishers