African Journal of Basic & Applied Sciences 3 (3): 73-79, 2011
ISSN 2079-2034

Logistic Regression Approach to Lymphoma Cancer Data

[1]*Wan Muhamad Amir W Ahmad,* [1]*Nor Azlida Aleng,* [2]*Zalila Ali and* [1]*Siti Aisyah Abdullah*

[1]Jabatan Matematik, Fakulti Sains dan Teknologi, Malaysia,
Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu Malaysia
[2]Pusat Pengajian Sains Matematik, Universiti Sains Malaysia (USM), 11800 Minden Pulau Pinang, Malaysia

Abstract: Lymphoma is a cancer in the lymphatic cells of the immune system, called lymphocytes. Around 1,000 people worldwide are diagnosed with lymphoma every day. For those who are suffering with lymphocyte their lymph nodes become to swell with unusually swollen and possibly will be obvious at the surface of the body [6]. In this present paper, we study the properties of multiple logistic regressions using real lymphoma cancer datasets and present the numerical results regarding to the findings.

Key words: Multiple logistic regression · Categorical data variables and Receiver Operating Characteristic (ROC)

INTRODUCTION

Hodgkin's lymphoma also known as Hodgkin's disease. It is a type of lymphoma, which is a cancer originating from white blood cells called lymphocytes. It was named after Thomas Hodgkin (most prominent British physician in 1798 - 1866), who first described abnormalities in the lymph system in 1832 (Hellman). Hodgkin's disease is characterized by the orderly spread of disease from one lymph node group to another and by the development of systemic symptoms with advanced disease. When Hodgkin's cells are examined microscopically, it gives the sign of cancer. Hodgkin's lymphoma may be treated with andRadiation therapy, chemotherapy, andSurgery andand High-dose chemotherapy and radiation therapy with stem cell transplant [1].

Signs and Symptoms: Patients with Hodgkin's lymphoma usually have the following symptoms [2, 3]:

- Systemic symptoms such as low-grade fever; night sweats; unexplained weight loss of at least 10% of the patient's total body mass in six months or less, itchy skin due to increased levels of eosinophils in the bloodstream or fatigue.
- Lymph nodes: Swelling of the lymph nodes in the abdomen can cause bloating and discomfort. It also can block the sewer and cause constipation, nausea and vomiting.
- Splenomegaly: enlargement of the spleen occurs in about 30% of people with Hodgkin's lymphoma.
- Hepatomegaly: enlargement of the liver, due to liver involvement, is present in about 5% of cases.
- Pain following alcohol consumption: involved nodes are painful after alcohol consumption, though this phenomenon is very uncommon.
- Back pain: nonspecific back pain that cannot be localized or its cause determined by examination or scanning techniques which has been reported in some cases of Hodgkin's lymphoma.
- Red-coloured patches on the skin and easy bleeding due to low platelet count.

Cause: There are no guidelines for preventing Hodgkin's lymphoma because the cause is unknown or multifactorial [4]. Risk factors include:

- Ages: between 15 and 40 and over 55
- Sex: male
- Family history
- History of infectious mononucleosis or infection with Epstein-Barr virus, a causative agent of mononucleosis
- Weakened immune system, including infection with HIV or the presence of AIDS.
- Prolonged use of human growth hormone.
- Exotoxins, such as Agent Orange.

Corresponding Author: Wan Muhamad Amir W Ahmad, Jabatan Matematik, Fakulti Sains dan Teknologi, Malaysia, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu Malaysia.

NHL is a cancer of the fifth highest in the world today. In 2009, over 66,000 adults and children have been diagnosed with the NHL disease. 95 percent of cases involving adults aged 60 and above suffer from this disease. Men at higher risk of women suffering from this disease though the number of female patients who were in NHL were diagnosed increases. Among the 74,030 new cases of lymphoma in 2010, the diseases will affect 40,050 males and 33,980 females. Hodgkin lymphoma will account for 8,490 cases (4,670 males and 3,820 females) and 65,540 cases will be NHL (35,380 males and 30,160 females). Many NHL disease existed by white people against people of African and Asian Americans [1].

Based on research done by Gary, Martin, Mauro and Luigino [1, 5, 6] who are infected with HIV (PHIV) are at high risk for cancer, HL. In the expert study found that HIV-infected patients with standard incidence ratio (SIR) 5 of the 30 times higher than patients with AIDS Kaposi Sarcoma (KS). The study also showed that the risk of HL in PHIVs significantly higher than the general population. Most cases of HL that occur during active SHCS follow-related with the cell mixture (49%) or of lymphocytes (16%), which is associated with Epstein-Barr Virus (EBV).

Hahn, *et al.* [7] in his study of lymphoma cancer states that the diseases are characterized by presence of Reed-Sternberg cell, influenced by the history of other diseases and the HIV and HCV infection. Individuals with a history of other diseases or autoimmune diseases can affect the lymphoma cancer. Examples of diseases associated with the risk of lymphoma include: Diabetes type 1 and Rheumatoid Arthritis and the immunosuppressive therapy used to promote the acceptance of organ transplantation. In addition, the study also found that individuals who are infected with human immunodeficiency virus (HIV) and hepatitis C virus (HCV) at high risk for lymphoma [7].

Based on studies by Becker, Deeg and Nieters [8, 9] of factors for the disease of cancer has proved effective use of multiple logistic regression analysis in the medical field. Among the factors studied were gender, age, ethnicity, medical history and smoking. Age, gender, medical history and smoking are factors that are absolute and can be categorized with the value as 0 and 1. In general, the risk for lymphoma cancer face starting at the age of 35 years to 50 years and almost 50% of those who reach the age of 65 years of age. In these studies, he also found that more men at high risk for lymphoma cancer are more likely to develop heart disease. Simple logistic regression analysis was also used as a guide to continue the study with logistic regression analysis [10-12].

MATERIALS AND METHODS

Logistic regression is used to forecast a dichotomous variable from a set of predictor variables. With a dichotomous dependent variable, discriminant function analysis is usually employed if all of the predictors are continuous and nicely distributed; logit analysis is usually employed if all of the predictors are categorical; and logistic regression is frequently chosen if the predictor variables are a mix of continuous and categorical variables and/or if they are not nicely distributed [10]. Logistic regression has been especially popular with medical research in which the dependent variable is whether or not a patient has a disease [13]. In logistic regression, the dependent variable is binary or dichotomous; its only contains data coded as 1 or 0. The objective of logistic regression is to get the best fitting model to illustrate the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Let us consider that, there are p is independent variables which will be denoted as vector $x' = \{x_1, x_2, ..., x_p\}$. We assume that each of these variables is at least interval scaled. Let the conditional probability that the outcome is present be denoted by

$$P(Y = 1|x) = \pi(x) \quad (1)$$

Then the logit of the multiple regression models is given by the formula as follows:

$$\text{logit}(\pi) = b_0 + b_1x_1 + b_2b_2 + b_3x_3 + ... + b_px_p \quad (2)$$

The specific form of the logistics regression model is given by:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (3)$$

Where $\pi(x)$ is the probability of occurrence of the characteristic of interest. Since the model produce by logistic regression is nonlinear, the equations used to describe the outcomes are slightly more complex than those for multiple regression. This linear regression equation creates the logit transformation. This transformation is defined, in terms of $\pi(x)$, as:

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \sum_{j=1}^{p} b_j x_{ij} \quad (4)$$

The objective of this study is to discuss multiple logistic analyses, which can be used in the analysis of categorical or categorized data. Material of this study is a hypothetical sample which is composed of seven variables. Namely variables are as in Table 1. Multiple logistic regression technique was used in the analysis of relationship between variables. Data of 326 respondents (patients) were collected. The sample size calculations for the data are given by single proportion method.

Sample Size Required: The sample sizes required at analysis stage are as follows:

Anticipated population proportion (p) = 0.838
Level of significance = 5% (0.05)
Absolute precision (Δ) = ±5%
$= (1.96 / 0.05)^2(0.69)(1 - 0.69)$
= 326 respondents.

The sample of 200 respondents required at the analysis stage. For this analysis, we used 326 respondents.

RESULTS AND DISCUSSION

Perform Univariable Analysis: From Table 1(a) and Table 1(b) we found that *SEX, DLBCL* and *HPT* are important factors (p-value <0.05) at univariable analysis. Their crude (unadjusted) OR are 0.589, 5.136 and 2.318. At univariable analysis, we can interpret the output as

- The odds of developing Lymphoma cancer among female is 0.589 times compared to male.
- The odds of developing Lymphoma cancer among patients that having *DLBCL* is 5.136 times compared to non *DLBCL*.
- The odds of developing Lymphoma cancer among patients that having *HPT* is 2.318 times compared to non *HPT*.

Perform Variable Selection: We review all the results in univariable analysis and select the variables based on their p value (<0.05). In our cases, we select the three variables and they are *SEX, DLBCL* and *HPT*.

Multicollinearity Checking: In step 3, we check multicollinearity in order to assess which variables are correlated highly. There are two methods that can be used and they are as follows:

The first step is by referring to the Table 3, from that table, we found that correlation between all the variables is correlated lowly with one another's. The second step is by referring to the standard errors of the variables. Table 4 indicates that that the standard errors of the variables are small and this means that there are no multi-collinearity effects.

Checking for the Interaction: We have checked all possible two way interaction between independent variables in the equation and the results shows that the interaction part is not significant. So, we make decision not to include the interaction in the model.

Asses the Goodness-Of-Fit: In order to know the goodness-of-fit of the model, we tested the model with the three methods. They are

- The classification Table
- Hosmer-Lemeshow test
- Area under the ROC curve

The area under the ROC curve is 0.679, it is significantly different from 0.5(with p-value is less than 0.05). This information tells us that the model can discriminate 67.9% of the cases.

In order to assess the goodness-of-fit, we tested the hypothesis that data fit the model well versus data do not fit the model well. From the point of view of Hosmer and Lemeshow Test, we accept the hypothesis (with p = 0.660) that the model can predict well. P-value is greater than 0.05 indicated that the model is good enough to use.

The classification table in Table 7 shows that the overall percentage correct is quite good. 72.4% of can be predicted accurately by the model. Ideally, a valid test needs to be high sensitivity and high in specificity.

Establish Final Model:

$$\hat{g}(x) = -0.918 - 0.621\,SEX + 1.575\,DLBCL + 0.607\,HPT$$

From the Model We Found That:

- The odds of developing Lymphoma cancer among female is 1 times compared to male (Cl: 0.317 to 0.911 p-value < 0.05).
- The odds of developing Lymphoma cancer among patients that having *DLBCL* is 5 times compared to non *DLBCL*. (Cl: 2.631 to 8.867, p-value < 0.05).
- The odds of developing Lymphoma cancer among patients that having *HPT* is 2 times compared to non *HPT*. (Cl: 0.910 to 3.702, p-value < 0.05).

Table 1: Explanation of the Variables

Variables	Code	Explanation of the variables	Categorical
Lymphoma	Y	Type Of Hodgkin Lymphoma Cancer	0 = No
			1 = Yes
SEX	X1	Gender	0 = Male
			1 = Female
DLBCL	X2	Diffuse Large B-Cell Lymphoma	0 = No
			1 = Yes
HPT	X3	Hypertension status	0 = No
			1 = Yes
AGE	X4	Age Of Patients In Year	-
STATUS	X5	Patients Status	0 = alive
			1 = dead

Table 2(a): Variables *SEX* in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
SEX	-0.529	0.252	4.425	1	0.035	0.589
Constant	-0.527	0.146	13.081	1	0.000	0.591

Dependent variable: LYMPHOMA

Table 2(b): Variables *DLBCL* in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
DLBCL	1.636	0.301	29.459	1	0.000	5.136
Constant	-1.064	0.141	57.196	1	0.000	0.345

Dependent variable: LYMPHOMA

Table 2(c): Variables *HPT* in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
HPT	0.841	0.328	6.557	1	0.010	2.318
Constant	-0.841	0.130	41.953	1	0.000	0.431

Dependent variable: LYMPHOMA

Table 2(d): Variables *AGE* in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
AGE	-0.004	0.006	0.400	1	0.527	0.996
Constant	-0.556	0.278	3.986	1	0.046	0.574

Dependent variable: LYMPHOMA

Table 2(e): Variables *STATUS* in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
STATUS	0.011	0.339	0.001	1	0.973	1.011
Constant	-0.726	0.315	5.327	1	0.021	0.484

Dependent variable: LYMPHOMA

Table 3: Correlation Table (a)

Correlation between *SEX* and *HPT*		Correlation between *SEX* and *DLBCL*	
Pearson correlation	*r = 0.060*	Pearson correlation	*r = -0.003*
Sig. (2-tailed)	*0.277*	Sig. (2-tailed)	*0.953*
N	*326*	N	*326*

Table 3: Correlation Table (b)

Correlation between *SEX* and *STATUS*		Correlation between *SEX* and *AGE*	
Pearson correlation	*r =-0.064*	Pearson correlation	*r = -0.037*
Sig. (2-tailed)	*0.252*	Sig. (2-tailed)	*0.508*
N	*326*	N	*326*

Table 3: Correlation Table (c)

Correlation between *HPT* and *DLBCL*		Correlation between *HPT* and *STATUS*	
Pearson correlation	*r = 0.202*	Pearson correlation	*r = -0.072*
Sig. (2-tailed)	*0.000*	Sig. (2-tailed)	*0.195*
N	*326*	N	*326*

Table 3: Correlation Table (d)

Correlation between *HPT* and *AGE*		Correlation between *DLBCL* and *STATUS*	
Pearson correlation	*r = 0.250*	Pearson correlation	*r = -0.077*
Sig. (2-tailed)	*0.000*	Sig. (2-tailed)	*0.167*
N	326	N	*326*

Table 3: Correlation Table (e)

Correlation between *DLBCL* and *AGE*		Correlation between *STATUS* and *AGE*	
Pearson correlation	*r = 0.135*	Pearson correlation	*r = 0.014*
Sig. (2-tailed)	*0.014*	Sig. (2-tailed)	*0.807*
N	326	N	*326*

Table 4:Checking Interaction of the Variables in the Equation

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
HPT(1)	0.003	0.554	0.000	1	0.996	1.003
DLBCL(1)	1.418	0.412	11.844	1	0.001	4.130
SEX(1)	-0.704	0.326	4.675	1	0.031	0.495
DLBCL(1) by HPT(1)	1.050	0.844	1.548	1	0.213	2.857
HPT(1) by SEX(1)	0.671	0.768	0.763	1	0.382	1.956
DLBCL(1) by SEX(1)	-0.127	0.658	0.037	1	0.847	0.881
Constant	-0.853	0.178	22.988	1	0.000	0.426

Table 5: Summary under the Curve

Area		0.679
Std. Error[a]		0.032
Asymptotic Sig.[b]	Lower	0.000
Asymptotic 95% Confidence Interval	Bound	0.615
	Upper Bound	0.742

Table 6: Summary Hosmer and Lemeshow Test

	Chi-square	df	Sig.
Goodness-of-fit test	1.598	3	0.660

Table 7: Classification Table

		Predicted HL		
	Observed	Non-Hodgkin's Lymphoma	Hodgkin's Lymphoma	Percentage Correct
HL	Non-Hodgkin's Lymphoma	197	22	90.0
	Hodgkin's Lymphoma	68	39	36.4
Overall Percentage 72.4				

Table 8: Sensitivity and Specificity Table

Sensitivity 24/(60+24)% = 28.6%.	Percentage of occurrences correctly predicted
Specificity 126/ (126+10)% = 92.6%.	Percentage of nonoccurrence correctly predicted
False Positive Rate 10/34% = 29.4%.	Percentage of predicted occurrences which are incorrect
False Negative Rate 60/ 186% = 32.25%.	Percentage of predicted nonoccurrence which are incorrect

Table 9: Variables *Sex* and *Smoke* in the Equation

							95% C.I. for EXP(B)	
Variables	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
HPT(1)	0.607	0.358	2.877	1	0.090	1.835	0.910	3.702
DLBCL(1)	1.575	0.310	25.824	1	0.000	4.830	2.631	8.867
SEX(1)	-0.621	0.270	5.311	1	0.021	0.537	0.317	0.911
Constant	-0.918	0.169	29.518	1	0.000	0.400		

The estimate logit is given by the following expression:

Table 10: Logistic Regression Analysis of 326 Respondents

							95% C.I. for EXP(B)	
Variables	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
HPT(1)								
0 = No								
1 = Yes	0.607	0.358	2.877	1	0.090	1.835	0.910	3.702
DLBCL(1)								
0 = No								
1 = Yes	1.575	0.310	25.824	1	0.000	4.830	2.631	8.867
SEX(1)								
0 = Male								
1= Female	-0.621	0.270	5.311	1	0.021	0.537	0.317	0.911
Constant	-0.918	0.169	29.518	1	0.000	0.400		

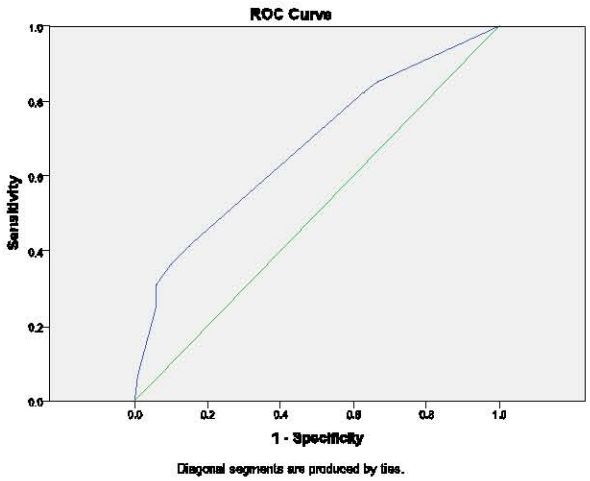

Fig. 1: Receiver Operating Characteristic (ROC) Curve.

RESULTS PRESENTATION DISCUSSION AND CONCLUSION

In this paper, we show that logistic regression is a very powerful analytical technique when the outcome variable is dichotomous. In order to know more about effectiveness of the logistic model we used the guideline as follow:

- Checking significance test for each predictor.
- Checking significance tests of the model against the null model.
- Descriptive and inferential goodness-of-fit indices.
- Predicted probabilities.

We hoped that this article give some guidelines and illustration of how logistic regression is applied to a dataset. As a research finding, we found three important things among patients and they are;

- The odds of developing Lymphoma cancer among female is 1 times compared to male (Cl: 0.317 to 0.911 p-value < 0.05).
- The odds of developing Lymphoma cancer among patients that having *DLBCL* is 5 times compared to non *DLBCL*. (Cl: 2.631 to 8.867, p-value < 0.05).
- The odds of developing Lymphoma cancer among patients that having *HPT* is 2 times compared to non *HPT*. (Cl: 0.910 to 3.702, p-value < 0.05).

REFERENCES

1. American Cancer Society. 2010. Cancer Facts and Figures 2010. *AmericanCancer Society*, 30 Ogos: 16.
2. Stein, RS. and D. Morgan, 2003. Hodgkin's Disease: Incidence of stages and results of therapy. Handbook of Cancer Chemotherapy, 6: 29-30.
3. The Lymphoma and Leukemia Society, 2010. Causes and risk factors of lymphoma. http://www.leukemia-lymphoma.org. [01 October 2010].

4. White, P. and A. Mamaroneck, 2005. Leukemia Facts and Statistics from Leukemia, Lymphoma, Myeloma, Facts 2010-2011. J. Leukemia and Lymphoma Society 13: 91-105.
5. Edelson, Ed. 2007. Doctors Report High Survival Rates for Hodgkin's Disease.MedicineNet.com. http://www.medicinenet.com/script/main/art. [5 Oktober 2010].
6. Clifford G.M., M. Rickenbach, M. Lise, L.D. Maso, M. Battegay, J. Bohlius, E.B.E. Amari, Karrer, G. Jundt, A. Bordoni and Ess S., Silvia, 2009. Hodgkin lymphoma in the Swiss HIV cohort study. Blood, 4 June 2009, 113(23): 5737-5742.
7. Hahn W.C. *et al.* 1999. Creation of Tumor Cells With Defined Genetic Elements.J. the National Cancer Institute, 400: 464-468.
8. National Comprehensive Cancer Network. 2009. Practice Guidelines in Oncology: Non-Hodgkin's Lymphoma. http://www.nccn.org/professionals/ physician_gls/ PDF/nhl.pdf. [28 September 2010].
9. Nikolaus Becker, Evelin Deeg and Alexandra Nieters, 2003. Population-based study oflymphoma in germany: rationale,study design and first results. Leukemia Res., 28: 713-724.
10. Long and Freese, 2006. Review of Regression Model for Categorical Dependent Variable using stat, second edition. The Stata J., 6: 273-278.
11. Lim, G.C.C., H. Yahaya and T.O. Lim, 2002. The first report of the national cancerregistry cancer incidence in Malaysia. National Cancer Registry. Kuala Lumpur. National Cancer Institute, 2010.
12. National Cancer Institute. Physician Data Query (PDQ). 2009. Adult Non-Hodgkin Lymphoma Treatment. http://www.cancer.gov/cancertopics/ pdq/treatment/adult-non-hodgkins/healthprofess ional. [5 September 2010].
13. David W. Hosmer and Stanley Lemeshow, 2000. Applied Logistic Regression, Second Edition John Wiley and Sons.