A Proposed Statistical Model for Estimating the Probability of Second Primary Cancer Occurrence With Application in Ain Shams University Hospitals

Amr I. Abdelrahman
Professor – Dept. of Statistics, Mathematics and Insurance,
Faculty of Commerce, Ain Shams University

Mahmoud R. Noah
MBA Degree - Dept. of Statistics, Mathematics and Insurance,
Faculty of Commerce, Ain Shams University.

ABSTRACT

This study aimed to use some classification methods (i.e. logistic regression model and discriminant analysis) in determining social-demographic risk factors in addition to treatment by radiation which affected the second primary cancer occurrence and the probability of this occurrence for patients who were initially treated for first primary cancer stage I, at least 1 year cancer free after first primary cancer treatment. They have high risk to develop a second primary cancer. Treatment by radiation and social-demographic risk factors: age at first cancer, gender, area, marital status, family history, smoking, education and obesity were studied by using the logistic regression model and the discriminant analysis model. We applied the following methods; Logistic regression model to estimate the probability of having second primary cancer. The odds ratio analysis compare whether the probability of having a second primary cancer is the same for two groups for each factor. We used Wald test and likelihood ratio to test for the significance of the coefficients; in addition to Hosmer and Lemeshow test and cross validation to assess the fit of the model. Discriminant analysis used as a comparative method for logistic regression model results. Most of medical risk factors (i.e., radiation does rate, chemotherapy does rate, number of nodes of first cancer, first cancer size) were not available at the hospitals records when the research was conducted.

Keywords

Logistic regression model; Wald test; Odds Ratio; Cross-Validation; ROC (Receiver Operating Characteristic) curve; Discriminant analysis; Second primary cancer.

1. *Introduction*

Early detection and evaluation of the risk factors which might cause the occurrence of second primary cancer is very important. The prediction of risk factors is an important pivot of the war against cancer. The use of statistical methods to identify risk factors would help to identify the probability of second primary cancer occurrence.

We distinguish between two medical cases: a) Recurrence case: Cancer that has recurred (come back), usually after a period of time during which the cancer could not be detected. The cancer may come back to the same place as the original (primary) tumor or to another place in the body. Also called recurrent cancer, and

b) Second cancer: a new primary cancer in a person with a history of another cancer.

According to DeVita et al. (2008) Second cancers can reflect the late sequel of treatment, as well as the influence of lifestyle factors, environment exposures, host determinants and gene-gene interactions. The main life style factors are tobacco and alcohol; the environmental factors are: contaminants and viruses; and the host factors are gender, age, genetics, immune function and hormonal factors.

A statistical model is proposed to explain the association between the studied covariates and its effect on the probability of the second primary cancer occurrence. Data included 240 patients were have a first primary cancer stage I, and have at least one year free cancer after first cancer treatment. Covariates used in the analysis were Age at first Cancer, Gender, Marital status, Area the patient lives in, Treatment by Radiation, Family History, Smoking, Obesity, and Education status. This study proposes to:

a. Identify the independent variables that impact the second primary cancer occurrence group membership and propose a statistical model to explain the association between the studied covariates and second primary cancer occurrence.
b. Establishing a classification system using the logistic model to determine group membership

In Section 2, we present the logistic regression model to estimate the probability of occurrence of second primary cancer; the Wald test, likelihood ratio test, Hosmer-Lemeshow test, cross validation methods and ROC curve are also introduced in section 2, in addition to Discriminant analysis overview in Section 3. In Section 4, we apply the binary regression model to the data; SPSS 18.0 is used for the analysis and a comparison between the logistic model and Discriminant analysis results is introduced. Summary and conclusions are given in Section 5.

2. The Binary Logistic Regression Model

The logistic regression model has been used in many disciplines including medical studies. It has been used in the social research (Ingles et al., 2009; King and Zeng, 2002; Saijo et al., 2008; and Garcia-Ramirez et al., 2005), in market research (Neagu and Hoerl, 2005; Kleijnen et al., 2004; Barone et al., 2007; Sallis and Sharma, 2009; and Kirkos, 2009), also become an important tool at the commercial applications (Erhart et al., 2009; O'Leary , 2009; and Weber et al., 2008); and in medical studies(Sanchez et al., 2008; Kaufman et al., 2000; Rubino et al., 2003).

The dependent variable of the logistic model is classified into two basic types (Afifi et al., 2004);

a- Continuous Variable: can assume any value within a specified range.

b- Discrete Variable: can only assume certain values and there are usually "gaps" between values (categorical response has two main categories: success (occurrence) and fail (no occurrence)).

Everitt (1998) gave the following definition for logistic distribution: " the limiting probability distribution as n tends to infinity, of the average of the largest to smallest sample values, of random samples of size n from an exponential distribution".

The logistic distribution is given by

$$f(x)=\frac{\exp[(x-\alpha)/\beta]}{\beta\{1+\exp-[(x-\alpha)/\beta]\}^2} \qquad -\infty<x<\infty\ ,\ \beta>0$$

The location parameter α is the mean. The variance of the distribution is $\pi^2\beta^2/3$, its skewness is zero and its kurtosis is 4.2 The standard logistic distribution with $\alpha=0$, $\beta=1$, with cumulative probability function. F(x), and probability distribution, f(x), has the property

$$f(x) = F(x)\,[1 - F(x)]$$

See also, (Evans et al., 1993).

In order to simplify notation, we will use the quantity p(x)=E(Y/X=x) to represent the conditional mean of Y given x when the logistic distribution is used. The specific form of the logistic regression model we will use is as follows:

$$p(x)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}=\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

The logistic regression is a form of regression analysis used when the response variable is a binary variable (Altman, 1991 and Everitt, 1998). The method is based on the logistic transformation or logit proportion, namely;

$$Logit\,(p)=\ln\frac{p}{1-p}$$

Where ;

$$p = Pr\ (y = 1)$$
$$(1\text{-}p) = Pr\ (y=0)$$

As p tends to 0, Logit (p) tends to $-\infty$ and as p tends to 1, Logit (p) tends to ∞. The function Logit (p) is a sigmoid curve that is symmetric about p = 0.5

The logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the predictor and response variables is not a linear function in logistic regression.

The logistic regression function is the logit transformation of P, where;

$$Logit(P) = \ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots\ldots\ldots + \beta_i x_i + \ldots\ldots\ldots + \beta_n$$

Where p = pr (dependent variable = 1) and $x_1, x_2, \ldots, x_i, \ldots, x_n$ are the explanatory variables, β_0 = the constant of the equation and, β_i = the coefficient of the predictor variables i. Using the logistic transformation in this way overcomes problems that might arise if p was modeled directory as a linear function of the explanatory variables; in particular it avoids fitted probabilities outside the range (0, 1). The parameters in the model can be estimated by maximum likelihood estimation (Gujarati, 2003).

The slope coefficient β_i associated with an explanatory variable x_i represents the change in log odds for an increase of one unit in x_i.

To assess the significance of the logistic regression coefficients the Wald statistic and likelihood ratio test are used; Wald statistic puts the value of the slope in perspective to the estimated variability of the slope. According to Hosmer and Lemeshow (2003), the wald test is obtained by comparing the maximum likelihood estimate of the beta's, $\hat{\beta}_i$, to an estimate of its standard error. The resulting ratio under the hypothesis that $\beta_i = 0$, will follow a standard normal distribution. The statistic takes the form:

$$z = \left(\frac{\hat{\beta}_i}{s.e(\hat{\beta}_i)}\right) \quad \text{or} \quad z^2 = \left(\frac{\hat{\beta}_i}{s.e(\hat{\beta}_i)}\right)^2$$

Where $\hat{\beta}_i$ represents the estimated coefficient β_i and S.E ($\hat{\beta}_i$) is its standard error. According to Agresti (2007), this statistic has approximately a standard normal distribution. Equivalently, z^2 has approximately a chi-squared distribution with $df = 1$.

According to Afifi et al. (2004), if the estimated value of the slope is small and its estimated variability is large we do not have enough evidence to conclude that the slope is significantly different from zero and vice versa.

The likelihood ratio test for overall significance of the beta's coefficients for the independent variables in the model is used (Hosmer and lemeshow, 2000; Fienberg,1998). The test based on the statistic" G" under the null hypothesis that the beta's coefficients for the covariates in the model are equal to zero.

G statistic takes the form:

$$G = -2\ell n\left[\frac{Likelihood \cdot without\ the.\ variable}{Likelihood \cdot with \cdot the \cdot variable}\right]$$

The distribution of "G" is a chi-square with q degree-of-freedom, where q is the number of covariates in the logistic regression equation. Hauck and Donner (1977) and Jennings (1986) examined the performance of the Wald test and found that the test often failed to reject the null hypothesis when the coefficient was significant. They recommended that the likelihood ratio test to be used.

The likelihood statistic L is used to assess the fitness of the model. The sampling distribution of the – 2 log L has a chi-square distribution with q degrees of freedom under the null hypothesis that all regression coefficients of the model are zero (Fienberg, 1998). A significant p-value provides evidence that at least one of the regression coefficients for an explanatory variable is non zero.

Hosmer and Lemeshow (2000) developed a goodness-of-fit test for logistic regression models with binary responses. They proposed grouping based on the value of the estimated probabilities. This test is obtained by calculating the Pearson chi-square statistic from the 2×g table of observed and expected frequencies, where g is the number of groups. The statistic is written as;

$$x^2_{HW} = \sum_{i=1}^{g} \frac{(o_i - N_i\overline{\pi}_i^2)}{N_i\overline{\pi}_i(1-\overline{\pi}_i)}$$

Where;

N_i Is the number of observation in the i[th] group.

o_i Is the number of event outcomes in the i[th] group.

$\overline{\pi}_i$ Is the average estimated probability of an event outcome for the i[th] group.

The Hosmer and Lemeshow statistic is then compared to a chi-square distribution with (g – 2) degree of freedom. However, Christensen (1997) gave the following warnings about the Hosmer and Lemeshow goodness-of-fit test;

1. If too few groups are used to calculate the statistic (<5) it will always indicate that the model fits the data. That is why Hosmer and Lemeshow (2000) advocated that, before finally accepting that a

model fits; an analysis of the individual residuals and relevant diagnostic statistics be performed.

2. It is highly dependent on how the observations are grouped.
3. It is a conservative test.
4. It has low power to detect specific types of lack of fit (such as nonlinearity in an explanatory variable).

2.1 The odds ratio

The odds ratio is a measure of association for 2×2 contingency table (Agresti, 2007). In 2×2 tables the probability of "success" is π_1 in row 1 and π_2 in row 2. Within row 1, the odds of success are defined to be:

$$odds_1 = \frac{p_1}{1-p_1} \qquad and \qquad odds_2 = \frac{p_2}{1-p_2}$$

Evaritt (1998) and Agresti (2002) define the odds ratio in two groups of subjects as "the ratio of odds". Thus;

$$\theta = \frac{odds_1}{odds_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

For the binary regression model, the odd ratio is the exponent $\left(e^{\beta_i}\right)$ is the ratio of odds for a one-unit change in x_i (Hosmer and Lemeshow, 2000). The change in Log odds, and the corresponding change in the odds ratio, for a c units is estimated $\exp\left[c\cdot\hat{\beta}_i\right]$ (Fleiss, 1981). When the two groups of odds are identical then the odds ratio is equal to one.

The corresponding lower and upper confidence limits for odds ratio for a c units change are $\exp\left[c\,L_i\right]$ and $\exp\left[c\,U_i\right]$, respectively, for (c>0), or $\exp\left[c\,U_i\right]$ and $\exp\left[c\,L_i\right]$ respectively, for (c<0), where $\left(L_i, U_i\right)$; can be either the likelihood ratio-based confidence interval or the Wald confidence interval for β_i (Agresti, 2002 and The SAS system ,1995).

2.2 Cross Validation Techniques

Cross - validation is a general procedure used in statistical model building. It can be used to decide on the order of a statistical model including, time series models, regression models, mixture distribution models, and discrimination models (Chernick, 2008).

Cross validation is performed in different ways, some of them are:

1. Take two random subsets of the data. Models are fit or various statistical procedures are applied to the first subset and then are tested on the second subset.
2. Leave - one – out technique is performed by fitting to all but one observation and then testing on the remaining one and has also been

called "cross - validation by Efron" (1983), but it does not provide an adequate test.

3. Fit the model n times, each time leaving out a different observation and testing the model on estimating or predicting the observation left out each time. This provides a fair test by always testing an observations not used in the fit. It also is efficient in the use of the data for fitting the model since n − 1 observation are always used in the fit.

Hit ratio is the percentage of objects (individuals, respondents, firms, etc.) correctly classified by the logistic regression model. It is calculated as the number of objects in the diagonal of the classification matrix (H_O)divided by the total number of objects (N). Also known as the Percentage correctly classified (Hair et al. 2009).

This can be compared with the maximum chance and proportional chance criterion to determine the discriminating power of the function. Maximum chance criterion is the percentage of the total sample represented by the larger of the two groups (H_C). The proportional chance criterion is obtained from the actual occurrence of second primary cancer by the equation $p^2 + (1 - p)^2$, where p = proportion of individuals in group (having a second primary cancer) and 1-p = proportion of individuals in group (not having a second primary cancer).

According to Marcoulides (1997) the difference between H_O and H_C may be tested by the following statistic

$$z = \frac{H_O - H_C}{\sqrt{H_C(N - H_C)/N}}$$

Where the significance of z is found by comparison with a critical value from a standard normal distribution.

2.3 Classification Accuracy: The ROC curve

ROC (Receiver Operating Characteristic) analysis is being used as a method for evaluation and comparison of classifiers (Ferri et. Al. 2002). The ROC gives complete description of classification accuracy as given by the area under the ROC curve. The ROC curve originates from signal detection theory (Hosmer and Lemeshow, 2000); the curve shows how the receiver operates the existence of signal in the presence of noise.

The ROC curve plots the probability of detecting true signal (sensitivity) and false signal (1 – specificity) for an entire range of possible cut points.

The sensitivity and specificity of a classifier also depend on the definition of the cut-off point for the probability of predicted classes. In many situations, not all misclassifications have the same consequences, and misclassification costs have to be taken into account. A ROC curve demonstrates the trade-off between true positive rate and false positive rate in binary classification problems.
To draw a ROC curve, the true positive rate (TPR) and the false positive rate (FPR) are needed.

- TPR determines the performance of a classifier or a diagnostic test in classifying positive cases correctly among all positive samples available during the test.
- FPR, on the other hand, defines how many incorrect positive results, which are actually negative, there are among all negative samples available during the test.
- Because TPR is equivalent to sensitivity and FPR is equal to (1 – specificity), the ROC graph is sometimes called the sensitivity vs. (1 - specificity) plot.

The area under the ROC curve has become a particularly important measure for evaluating classifiers' performance because it is the average sensitivity over all possible specificities (Bradley 1997). The large area, the better classifier performs. If the area is 1.0, the classifier achieves both 100% sensitivity and 100% specificity. If the area is 0.5, then we have 50% sensitivity and 50% specificity, which is no better than flipping a coin. This single criterion can be compared for measuring the performance of different classifiers analyzing a dataset. (Hanley, 1982; Bamber, 1975)

After a classifier has been made, it is also useful to measure its calibration. Calibration evaluates the degree of correspondence between the estimated probabilities of a specific outcome resulting from a classifier and the outcomes predicted by domain experts. This can then be tested using goodness-of-fit statistics. This test examines the difference between the observed frequency and the expected frequency for groups of patients and can be used to determine if the classifier provides a good fit for the data. If the p-value is large, then the classifier is well calibrated and fits the data well. If the p-value is small, then the classifier is not well calibrated.

3. Discriminant analysis overview

Another method to classify categorical outcome is discriminant analysis types (Afifi et al., 2004). The main purpose of a discriminant function analysis is to predict group membership based on a linear

combination of the interval variables. The procedure begins with a set of observations where both group membership and the values of the interval variables are known. The end result of the procedure is a model that allows prediction of group membership when only the interval variables are known. A second purpose of discriminant function analysis is an understanding of the data set, as a careful examination of the prediction model that results from the procedure can give insight into the relationship between group membership and the variables used to predict group membership.
Discriminant analysis also has several assumptions (Press et al., 1978), such as

- Normally distributed: The predictor variable should be normally distributed.
- Homogeneity of variances: Variance with each group of independent variables should be equal.
- The relationship is linear in parameters.
- Absence of outliers.
- Independence: Each case should be independent of each other or not-collinear. Correlated data cannot be used in discriminant analysis.
- Adequate sample size: There must be at least two cases for each category of the dependent variable. However, it is recommended that there should be at least four or five times as many cases as independent variables.
- Interval data: In discriminant analysis, there should be an interval data for independent variable.
- Variance: No independents have a zero standard deviation in one or more of the groups formed by the dependent.
- Random error: Error terms are assumed to be randomly distributed.
- Absence of perfect multicollinearity: There should be no perfect multicollinearity between the independent variables. Assumes linearity: The discriminant functions should be linear and related to each other.

3.1 Estimating and Interpretation of the Discriminant Analysis Model.

Linear discriminant function: The Linear combination of the discriminating (independent) variable is called the discriminant function,
Hair et al. 2009 gave the following form for the linear function as:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \ldots\ldots\ldots + W_n X_{nk}$$

Where
Z_{jk} = discriminant Z score of a discriminant function j for object k.
a = intercept
W_i = discriminant coefficient for independent variable i.

X_{ik} = independent variable i for object k.

Unstandardized discriminant coefficients are simply like the regression beta, which is used to predict the discriminant score. Standardized discriminant coefficients are used to compare the relative importance of the independent variables.

For the two groups, there is one discriminant analysis function. For multivariate discriminant analysis there will be g-1 discriminant function.

In case an individual may belong to one of two populations. We considering how an individual can be classified into one of these populations on the basis of a measurement of one characteristic, say X. when we have a representative sample from each populations enabling us to estimate the distribution of X and their mean. Typically, these distribution be represented as in figure (1);Afifi et al. (2004)

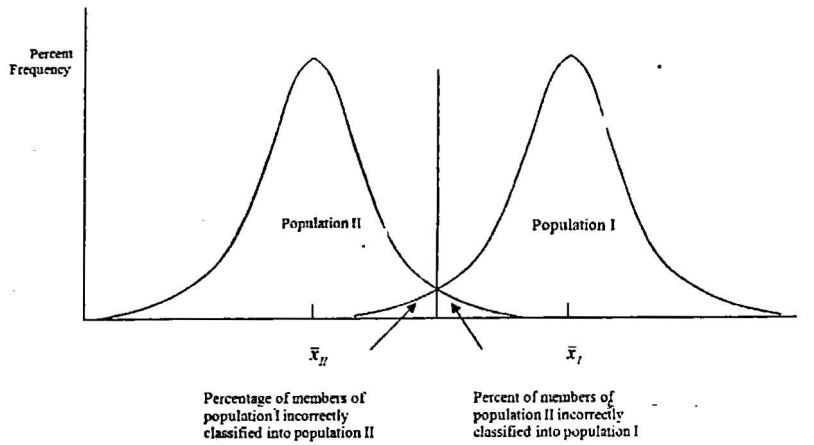

Figure (1): hypothetical frequency distributions of two populations showing percentage of cases incorrectly classified (Afifi et al. 2004; pp.247).

For all variables case and for each individual from each population, the value of Z is calculated when the frequency distributions of Z are plotted separately for each population, the results is illustrated in figure (2).

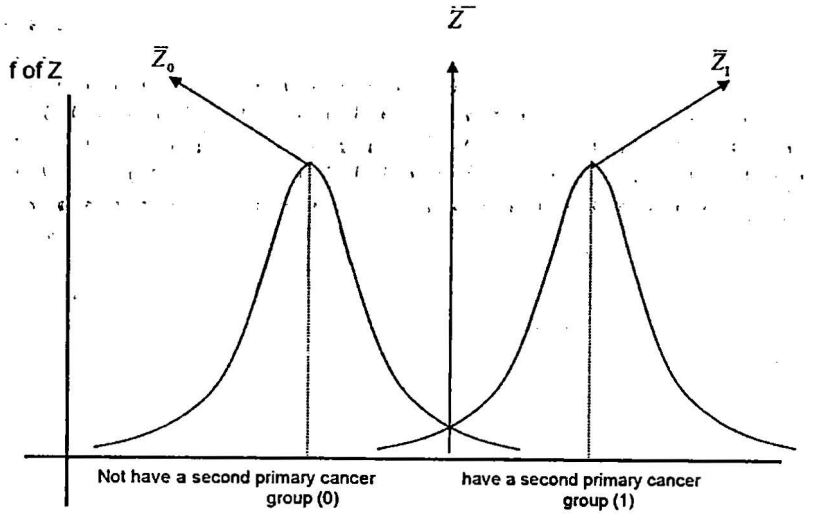

Figure (2): Discriminant Analysis with Two Groups, (Charles. 2008; pp.11).

Thus for the classification procedures assigned an individual to either group I or group II. Since there is always as probability of making the wrong classification we compute the probability that the individual has come from one group or the other. Such probability computes under the multivariate normal model. The formula is:

Probability of belonging to population I = $\frac{1}{1+e^{-Z}}$

And the probability of belonging to population II = $1-\frac{1}{1+e^{-Z}}$

In some programs these probabilities are called posterior probabilities (Afifi et al., 2004); since they express the probability of belong to a particular popu'ation posterior (i.e. after) performing the analysis.

The F test (Wilks' lambda) is used to test whether or not the discriminant model is significant as a whole. If the F test shows the overall significance of the model, then the individual variables are accessed to see which variable will move the significance from the group mean.

4. Statistical Analysis and Results

Data used for the analysis comprised of 1500 registered patients in Ain shams university hospitals, Cairo, Egypt, by different stages of cancer in 2006; 240 patients met the study assumptions were classified as:

1- **Patients have a first primary cancer stage I.**
2- **Patients are at least one year free cancer after first cancer treatment..**

The dependent variable (classification variable) used in the study was having a second primary cancer (0 for those not have a second primary cancer, 1 for those who have a second primary cancer), explanatory variables used in this study were: age at first cancer occurrence, gender(male-female), marital status(married –single), area (urban or rural), radiation treatment of first cancer(yes- no) ,family history of cancer (yes, no), smoking (yes-no), Obesity before first cancer (yes-no), and education (Yes-no)for patients above 18 or parents for patients less than 18 .

SPSS software package is used for the analysis. The maximum likelihood method is used to estimate the coefficient and its standard error in addition the Newton-Raphson method to solve the nonlinear equations for the logistic model maximum likelihood estimations, table 1 shows the SPSS output.

Table 1 : The estimated coefficient , its S.E and Wald test

Covariate	*Beta estimate*	*S.E*	*Wald*	*P-value*
Age (x_1)	*0.015*	*0.018*	*0.655*	*0.418*
Gender $(x_2$	*-0.497*	*0.582*	*0.730*	*0.393*
Marital Status (x_3)	*1.112*	*0.544*	*4.180*	*0.041*
Living Area (x_4)	*0.553*	*0.427*	*1.674*	*0.196*
Treatment.by.Radiation (x_5)	*-1.916*	*0.418*	*21.055*	*0.000*
Family.History (x_6)	*1:106*	*0.380*	*8.488*	*0.004*
Smoking (x_7)	*3.215*	*0.749*	*18.412*	*0.000*
Obesity (x_8)	*1.053*	*0.503*	*4.387*	*0.036*
Education (x_9)	*-1.753*	*0.394*	*19.760*	*0.000*
Constant	*-0.706*	*0.727*	*0.942*	*0.332*

At the .05 level of significant, Table 1 shows that "Education", "Smoking ", "family history and "marital status" were highly significant. The coefficients estimates are used to estimate the probability of the second primary cancer occurrence (Ashour and Abo Elfotouh 2005) as follows:

$$P(y=1 \mid x) = \frac{e^{=}}{1+e^{=}} \quad \text{or} \quad \frac{1}{1+\exp^{-=}}$$

Where ;

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_p x_p$$

Hence:

$$Z = -0.706 + 0.015\,x_1 - 0.497\,x_2 + 1.112\,x_3 + 0.553\,x_4 - 1.916\,x_5 + 1.106\,x_6 + 3.215\,x_7 + 1.053\,x_8 - 1.753\,x_9$$

The sign of the coefficients of the estimated logistic function in Table 1 above gives an explanation of the explanatory variables used, as given in Table 2.

Table 2: The sign analysis

Covariate	Codes	Sign	Explanation
Age at first Cancer	Quantitative	Positive	Older age increases the probability of second primary cancer
Gender	1 Male 0 Female	Negative	Male decreases the probability of second primary cancer
Marital status	1 Married 0 Single	Positive	Married increases the probability of second primary cancer
Area	1 Urban 0 Rural	Positive	Living in urban increases the probability of second primary cancer
Treatment by Radiation	1 Yes 0 No	Negative	Radiation decreases the probability of second primary cancer
Have a Family History	1 Yes 0 No	Positive	family history increases the probability of second primary cancer
Smoking	1 Yes 0 No	Positive	Smoking increases the probability of second primary cancer
Obesity	1 Yes 0 Not	Positive	obesity increases the probability of second primary cancer
Education	1 Educated 0 Illiterate	Negative	Education decreases the probability of second primary cancer

4.1 The odds Ratio Results

The following odds ratios were calculated using the formula;

$$\theta = \frac{odds_1}{odds_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

For every covariate used in the study, results are given in Table 3.

Table 3: Odds Ratios and 95% Confidence Intervals for Covariates

Variable	*Odds ratio*	*95% Confidence interval*
Gender	*0.608*	*{0.194 to 1.904}*
Marital Status	*3.041*	*{ 1.047 to 8.830}*
Area	*1.738*	*{0 .752 to 4.018}*
Radiation	*0.147*	*{0.065 to 0 .334}*
Family History	*3.022*	*{ 1.436 to 6.359}*
Smoking	*24.910*	*{ 5.735 to 108.193}*
Obesity	*2.856*	*{1.070 to 7.672}.*
Education	*0.173*	*{0.080 to 0 .375}*

From Table 3, it is evident that patients who smoke, patients with family history and married persons are highly susceptible for a second primary cancer occurrence.

Table 4 gives the classification table. Using the obtained Z function observations are classified as follows, using a prior probability of 0.50

Table 4: Classification Table

Observed		Predicted		
		Have.a.Second. Cancer		Percentage Correct
		Not have	Have	
Have.a.Second.Cancer	Not have	103	20	83.7
	Have	28	89	76.1
Overall Percentage				80.0

From Table 4, we conclude that;

- **a- 83.7% of all patients not have a second primary cancer are correctly classified, and 16.3% are incorrectly classified.**
- **b- 76.1% from all patients who have a second primary cancer are correctly classified, 23.9% are incorrectly classified.**
- **c- The overall correct percentage was 80% , which reflects the model's overall explanatory strength.**

4.2 Model Assessment fit of goodness for Logistic regression

The -2 log likelihood for the constant only model obtain by fitting the constant only model was 332.561;and the -2 log likelihood for the overall model was 209.266.
Thus the value of the likelihood ratio test is;

$$G = 332.561 - 209.266 = 123.195$$

And the p-value for the test is p [$\chi^2(9) > 123.195$] < 0.001which is highly significant at the α <0.001 level. The null hypothesis is rejected and we conclude that at least one and perhaps all betas' coefficient are different from zero.

The likelihood ratio tests for all covariates and for each covariate are given in Table 5.

Table 5: likelihood ratio test

Model	-2loglikelihood	G	P-value
Model with constant only	332.561		
Model with all covariates (full model)	209.266	123.295	<0.001
Model without family history	218.278	9.012	0.003
Model without smoking	230.896	21.63	<0.001
Model without education	214.002	4.736	0.030
Model without Age at first cancer	210.025	0.759	0.384
Model without Treatment by radiation	234.336	25.07	<0.001
Model without Gender	210.112	0.846	0.358
Model without Marital status	213.777	4.511	0.034
Model without area	211.079	1.813	0.178
Model without obesity	232.245	22.979	<0.001

From table 5 we note that the covariates (family history, smoking, education, treatment by radiation, marital status and obesity) are statistically significant; while the covariates (gender, age at first cancer and area and) are statistically non-significant.

The Wald test is obtained by comparing the maximum likelihood estimate of the beta's, $\hat{\beta}_i$, to its standard error. The resulting ratio, under the hypothesis that $\beta_i = 0$ are given in Table 5.

It is evident that the covariates (family history, smoking, education, treatment by radiation, marital status and obesity) are statistically significant; while the covariates (gender, age at first cancer and area) are

statistically not-significant.

Stepwise logistic regression analysis is used to reduce number of covariates. Results are summarized the results as in table 6.

Table 6: Stepwise Binary Logistic Regression Results

	B	S.E.	Wald	d.f	p-value	Exp(B) Odds ratio
Marital.status	1.516	0.430	12.432	1	0.000	4.556
Treatment.by.Radiation	-1.834	0.406	20.407	1	0.000	0.160
Have.a.Family.History	1.223	0.367	11.115	1	0.001	3.398
Smoking	2.829	0.523	29.250	1	0.000	16.931
Obesity	0.993	0.490	4.105	1	0.043	2.699
Education	-1.710	0.383	19.926	1	0.000	0.181
Constant	-0.259	0.495	0.273		0.601	0.772

And the logit is:

$$Z = -0.259 + 1.516\ (\text{Marital status}) - 1.834\ (\text{Radiation}) + 1.223\ (\text{Family History}) + 2.828\ (\text{Smoking}) + 0.993\ (\text{obesity}) - 1.710\ (\text{Education})$$

The Logit (Z) above indicates that: married patients are more susceptible to develop a second primary cancer ; treatment by radiation decreases the susceptibility; a patient with family history is more susceptible to develop second primary cancer; smokers are more susceptible than non-smokers, and educated patients are less susceptible to develop a second primary cancer.

The exponent (Exp (B)) in Table 6 is the odds ratio, thus:

1. The odds for married patients to single patients to develop second primary cancer are 4.556.
2. The odds for patients with family history to patients with no family history to develop second primary cancer is 3.398.
3. The odds for smokers to nonsmokers to develop second primary cancer is 16.931.

Table 7 gives the classification table. Using the obtained Z function observations are classified as follows, using a prior probability of 0.50.

Table 7 :.Classification Table

Observed		Predicted		
		Have.a.Second. Cancer		Percentage Correct
		Not have	Have	
Have.a.Second.Cancer	Not have	106	17	86.2
	Have	30	87	74.4
Overall Percentage				80.4

a- 86.2% of all patients not have a second primary cancer are correctly classified, and 13.8% are incorrectly classified.

b- 74.4% from all patients who have a second primary cancer are correctly classified, 25.6% are incorrectly classified.

c- The overall correct percentage was 80.4%, which reflects the model's overall explanatory strength.

Hence the hit ratio for the stepwise logistic model 80.4% is better than the full model hit ratio 80%; so we will depend on the stepwise model

The value of the Hosmer – Lemeshow goodness-of-fit statistic computed for the full model was C = 2.976 and the corresponding p-value computed from the chi-square distribution with 7 degree of freedom is 0.887 this indicates that the model seems to fit quite well.

4.3 Cross Validation Results

Using Efron (1983) leave-one-out Cross Validation goodness-of-fit statistic the results for the stepwise model was (using prior probability of .50) summarized in table 8.

Table 8: Cross Validation Results

Group	Actual no. of cases	Predicting group membership		%Correctly classified
		Having second primary cancer	Not having second primary cancer	
Having second primary cancer	117	88	29	80 %
Not having second primary cancer	123	19	104	

The classification matrix shows the accuracy of second primary cancer occurrence prediction in the cross validation leave-one-out sample as presented in table 8 above. In this sample of 240 patients, actual occurrence of second primary cancer was 117 patients; that 88 or 75.2% was correctly classified into group having a second primary cancer. Of the 123 patients that not having a second primary cancer, 104 or 84.6% was correctly classified into group not having a second primary cancer. The total correctly classified was 192 of 240 or 80%.

The maximum chance Criterion is 51.25% (123/240) and the proportional chance criterion is 50.031% $[(0.4875)^2 + (1-0.4875)^2]$ also. Because the percentage correctly classified was 79.6% (29.569% greater than proportional chance), Z test evident that difference are statistically significant

$$z = \frac{80 - 50.031}{\sqrt{50.031(240 - 50.031)/240}} = 4.762 \text{ (p-value <0.001).}$$

Using ROC curve for the classification accuracy, it is found that the area under the ROC curve, which ranges from zero to one, provides a measure of the model's ability to discriminant between those subjects who experience the response of interest versus those who do not.

Plotting sensitivity versus (1 – specificity) over all possible cut-points is shown in the Figure (3) below .The area under the ROC curve for the full model was is 0.870 this is considered excellent discrimination

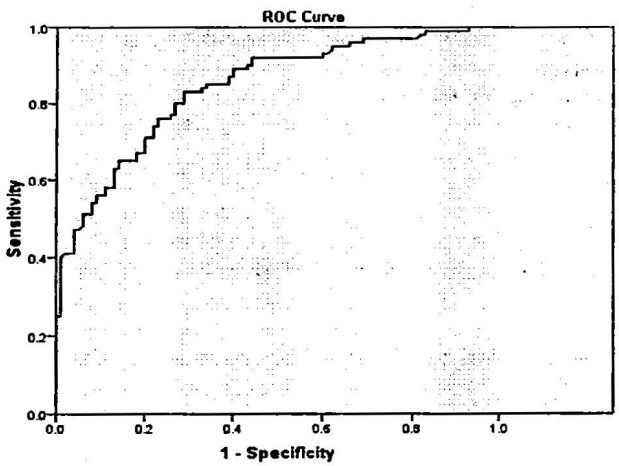

Figure (3): area under ROC curve

4.4 Comparison between binary logistic regression and the discriminant analysis results

	Binary logistic regression		Discriminant analysis	
	Full model	Stepwise model	Full model	stepwise model
Sensitivity	76.1%	74.4%	76.1%	75.2%
Specificity	83.74%	86.2%	83.74%	85.4%
Hit ratio	80%	80.4%	80%	80.4
Cross validation hit ratio	79.6%	80%	79.6%	80%
Area under ROC Curve	0.877	0.870	0.877	0.874

From the comparison above we conclude that:

1- The two methods results are two closed.
2- The stepwise models have highest specificity when the full models have the highest sensitivity.
3- The stepwise models have the highest hit ratio and highest cross-validation hit ratio
4- The full models have the biggest area under ROC Curve

We recommended that depending on the binary logistic regression stepwise model for

- The covariates haven't normal distribution which is a discriminant assumption.
- Have the highest hit ratio with a few covariates.

5. *Summary and Conclusions*

In this study, social-demographic risk factors of developing a second primary cancer using logistic regression model were studied. The social-demographic risk factors used are age at first cancer, gender, area the patient live in, marital status, family history, smoking, education and obesity in addition to treatment by radiation. The binary logistic regression model is used to estimate the probability of having second primary cancer. Significance testing for the logistic coefficients using Wald test and likelihood ratio show that smoking, family history, marital status, and education are the significant factors. The odds ratio for each covariate compare whether the probability of having a second primary cancer is the same for each covariate groups. The odds ratio for smokers to non-smokers ranges between 6 times to 47 times with confidence 95%. To assess the fitness of the model the maximum likelihood test and Hosmer and Lemeshow test are used. The logistic regression model proved to have a lower sensitivity level due to some other clinical risk factors not considered in this study.

The study concludes that: *married patients are more susceptible to develop a second primary cancer; treatment by radiation decreases the susceptibility; a patient with family history is more susceptible to develop second primary cancer; smokers are more susceptible than non-smokers, and educated patients are less susceptible to develop a second primary cancer, and patients with obesity before first cancer are more susceptible to develop a second primary cancer.*

- **The researcher recommends the following:**

1- **Establishing a National Cancer Association, this association i linked with a unit of cancer registry in all hospitals which treat cancer, registering of demographic and medical data of all cancer patients and update these statements as a regular, semi-annual or annual.**
2- **Replicating the same study with an increased sample size.**
3- **Replicating the same study to include repeated measures on the same patients, especially when some demographic factors change, and age develops.**
4- **Replicating the same study for each first primary cancer type in a separate study using the reached significant factors and add more medical risk factors (i.e., radiation does rate, chemotherapy does rate, number of nodes of first cancer, first cancer size) which were not available at the hospitals records when the research was conducted.**
5- **Applying Classification and R gression Tree (CART) and compare the results with the binary logistic regression model.**

References

1. Afifi, A., Clark, V. A., and May, S. (2004). Computer- Aided Multivariate Analysis. Fourth Edition, Champman and Hall/CRC.

2. Agresti, A. (2007). An Introduction to Categorical Data Analysis. Second Edition, Wiley, Inc., New York.

3. Agresti, A. (2002). Categorical Data Analysis. Second Edition, Wiley, Inc., New York.

4. Altman, D. G. (1991). Practical statistics for medical research. Champman and Hall, London.

5. Ashour, S., and Abo Elfotouh, S. (2005). Presentation and statistical analysis using SPSSWIN. Second Part, Advanced Applied Statistics, Institute of Statistical Studies and Research. Cairo University, Egypt (in Arabic).

6. Bamber, D. (1975) .The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of mathematical psychology, 12, pp. 387-415.

7. Barone S., Lombardo A. and Tarantino, P. (2007). A weighted logistic regression for conjoint analysis and Kansei engineering. Quality and Reliability Engineering International, Vol. 23 Issue 6, pp. 689 – 706, John Wiley & Sons, Ltd.

8. Bradley, A. P. (1997). Th use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition. Jul; 30(7): pp.1145-59.

9. Charles, M., Friel (2008). Notes on Discriminant Analysis. Criminal Justice Center, Sam Houston State University.

10. Chernick, M., R. (2008). Bootstrap Methods: A Guide for Practitioners and Researchers. Second Edition, Wiley, Inc., New York.

11. Christensen, R. (1997) Log-Linear models and logistic regression. Second Edition. Springer-verlag. New York.

12. DeVlta, Hellman, and Rosenberg's. (2008). Cancer Principles and practice of oncology. Eighth edition, vol. 6, wolters kluwer, lippincott Williams&wilkins.

13. Efron, B. (1983). Estimating the error rate of a prediction rule: improvements on cross validation. Journal of the American Statistical Association, 78, pp. 316-331.

14. Erhart, M., Hagquist C., Auquier P., Rajmil L., Power M., Ravens-Sieberer U. and the European KIDSCREEN Group. (2009). A comparison of Rasch item-fit and Cronbach's alpha item reduction analysis for the development of a Quality of Life scale for children and adolescents. Child: Care, Health and Development, Blackwell Publishing Ltd.

15. Evans, M. Hastings, N. and Peacock, B. (1993). Statistical Distributions. Second Edition, Wiley, New York.

16. Everitt, B. S. (1998). The Cambridge Dictionary of Statistics. Cambridge University Press.

17. Ferri, C., Flach P., Hernandez-Orallo J. (2002). Learning Decision Trees Using the Area under the ROC Curve. Nineteenth International Conference on Machine Learning (ICML 2002); Morgan Kaufmann; pp. 46-139.

18. Fienberg, S. E (1980). The analysis of cross-classified categorical data. Second Edition, The MIT Press, Cambridge, Massachusetts.

19. Fleiss, J., L. (1981). Statistical Methods For Rates And Proportions. Second Edition. John Wiley & Sons, Inc.

20. Gujarati, N.D. (2003). Basic Econometrics. Fourth Edition, Mcgraw -Hill.

21. Garcia-Ramirez M., Martinez, M., F. M., Balcazar F., E., Suarez-Balcazar Y., Albar M., Domínguez E. and Santolaya, F.,J. (2005). Psychosocial empowerment and social support factors associated with the employment status of immigrant welfare recipients. Journal of Community Psychology, Volume 33 Issue 6, Pages 673 – 690, Wiley Periodicals, Inc., A Wiley Company.

22. Hair, J. F., Anderson, R. E., Babin, B. J., and Black, W. C.(2009) Multivariate Data Analysis. Seventh Edition. Maxwell Macmillan International, New York.

23. Hanley, J.A. and McNeil, B., J. (1982). The meaning and the use of the Area under a receiver operating characteristic curve (Roc).' Radiology, 143, pp. 29-36.

24. Hauck, W.W., and Donner, A. (1977). Wlad's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association, 72, pp.851-853.

25. Hosmer, D. W., Lemeshow, S. (2000). Applied Logistic Regression, Second Edition, Wiley, Inc., New York.

26. Ingles, C., J.; Garcia-Fernandez, j., M.; Castejon,J., L. ; Valle Antonio, B., D., and Marzo, J., C. (2009). Reliability and validity evidence of scores on the Achievement Goal Tendencies Questionnaire in a sample of Spanish students of compulsory secondary education. Psychology in the Schools, Vol. 46 Issue 10, pp. 1048 – 1060, Wiley Periodicals, Inc., A Wiley Company

27. Jennings, D.E. (1986a). Judging inference adequacy in logistic regression. Journal of the American Statistical Association, 81, pp. 471-476.

28. Kaufman, E., L., Jacobson, J.; S. Hershman, D., L.; Desai M., and Neugut, A., I. (2008). Effect of Breast Cancer Radiotherapy and Cigarette Smoking on Risk of Second Primary Lung Cancer. Journal of clinical oncology, 26(3): pp. 392-398.

29. King G. and Zeng L. (2002). Estimating risk and rate levels, ratios and differences in case-control studies. Statistics in Medicine, Vol. 21 Issue 10, pp. 1409 – 1427, John Wiley & Sons, Ltd.

30. Kirkos, E., Spathis C., and Manolopoulos Y. (2009). Audit-firm group appointment: an artificial intelligence approach. Article on line in advance of print, Intelligent Systems in Accounting, Finance & Management, John Wiley & Sons, Ltd.

31. Kleijnen, M., De Ruyter K., Wetzels M. (2004). Consumer adoption of wireless services: Discovering the rules, while playing the game. Consumer adoption of wireless services: Discovering the rules, while playing the game. Journal of Interactive Marketing, Vol. 18 Issue 2, pp. 51 – 61, Wiley Periodicals, Inc., A Wiley Company, and Direct Marketing Educational Foundation, Inc.

32. Marcoulides, A. George., and Hershberger, L. Scott. (1997). Multivariate statistical methods: A first course. Lawrence Erlbaum Associates, Mahwah, New Jersey.

33. McLachlan, G. J. (2004). Discriminant analysis and statistical pattern recognition. New York: John Wiley & Sons.

34. Neagu R. and Hoerl R. (2005). A Six Sigma Approach to Predicting Corporate Defaults. Quality and Reliability Engineering International. Vol. 21 Issue 3, pp. 293-309, John Wiley & Sons, Ltd.

35. O'Leary, D., E. (2009). Downloads and citations in Intelligent Systems in Accounting: Finance and Management. Intelligent Systems in Accounting, Finance & Management, Vol. 16 Issue 1-2, pp. 21 – 31, John Wiley & Sons, Ltd.

36. Press, S. J. and S. Wilson (1978). Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association, Vol. 73: 699-705.

37. Rubino, C., De Vathaire, F., Shamsaldin, A., Labbe, M., and le M. G. (2003). Radiation dose, chemotherapy, hormonal treatment and risk of second cancer after breast cancer treatment. Britich Journal of Cancer; 89(5): pp. 840–846.

38. Saijo, Y., Ueno T., Hashimoto, Y. (2008). Twenty-four-hour shift work, depressive symptoms, and job dissatisfaction among Japanese firefighters. American Journal of Industrial Medicine, Vol. 51 Issue 5, pp. 380 – 391. Wiley-Liss, Inc., A Wiley Company.

39. Sallis,J., E. and Deo Sharma , D.(2009). Knowledge seeking in going abroad. Thunderbird International Business Review, Vol. 51 Issue 5, pp. 441 – 456, Wiley Periodicals, Inc., A Wiley Company.

40. Sanchez, L., A; Lana, A. B., Hidalgo, A. A.; Rodriguez, M., J. C.; Del Valle, M., D.; Cueto , A., b, Folgueras, M., C., Belyakova, E., C., Comendador , M. D., Lopez, M. L. (2008). Risk factors for second primary tumours in breast cancer survivors. European Journal of Cancer Prevention. 17(5): pp. 406-413.

41. The SAS System (1995) .Logistic regression examples using the SAS system. Version 6, First Edition. SAS institute Inc., Cary. NC, USA.

42. Weber, S., O.; and Michalik G., W., (2008). Incorporating sustainability criteria into credit risk management. Business Strategy and the Environment , Vol. 19 Issue 1, pp. 39 – 50, John Wiley & Sons, Ltd. and ERP Environment.

نموذج إحصائى مقترح لتقدير إحتمال الإصابه الأوليه الثانية بالسرطان بالتطبيق على مستشفيات جامعة عين شمس

د .عمرو إبراهيم عبد الرحمن الإتربى[1]
محمود راضى حامد[2]

ملخص البحث

تهدف هذه الدراسه إلى إستخدام بعض طرق التصنيف (مثل الإنحدار اللوجيستى و تحليل التمايز) لتحديد العوامل الديموجرافيه – الإجتماعيه و كذلك تأثير المعالجه بالإشعاع فى حدوث الإصابه الأوليه الثانية بالسرطان وإحتمال حدوث هذه الإصابه للمرضى المعالجون فى سرطان أولى أول من الدرجة الأولى وقضوا فترة لا تقل عن عام بدون سرطان بعد معالجة السرطان الأول . هؤلاء المرضى يكونوا ذات إحتمال عالى للإصابه بالسرطان الأولى الثانى . المعالجة بالإشعاع و العوامل الديموجرافيه – الإجتماعيه مثل : عمر المريض عند الإصابه بالسرطان الأول، النوع، الحالة الاجتماعية، التاريخ العائلى للإصابه بالسرطان، التدخين، التعليم والسمنة تم دراستهم بإستخدام نموذج الإنحدار اللوجيستى و نموذج تحليل التمايز، العديد من المتغيرات الطبية على سبيل المثال: حجم الورم الأول، عدد العقد فى الورم الأول، كمية الجرعة الأشعاعية، كمية الجرعة الكيميائية لم تضف إلى النموذج لعدم توافر بياناتها فى سجلات الحالات بمستشفيات جامعة عين شمس . تم تطبيق الأساليب الأتيه: نموذج الإنحدار اللوجيستى لتقدير إحتمال حدوث الإصابه الأوليه الثانية، مقياس الأفضلية النسبية لمقارنه ما إذا كانت إحتمال الإصابه الثانية متساوية لكل مجموعتين لكل عامل ثنائى، تم إستخدام إختبار wald و إختبار likelihood ratio لإختبار معنوية المعاملات المقدرة، بالإضافة إلى إختبار هوسمر – ليميشو وإسلوب cross validation لتقييم جودة تقديرات النموذج وكذلك نموذج تحليل التمايز كإسلوب مقارن لنتائج نموذج الإنحدار اللوجيستى .

النتائج : المرضى المدخنون يكونوا ٦ – ٤٧ مره اكثر احتمالية لحدوث إصابة ثانية كذلك المرضى المتزوجون يكونوا اكثر عرضة للإصابة الثانية، بينما المعالجون بالإشعاع يكونوا اقل خطورة لحدوث الإصابة الثانية .

---

١ أستاذ الإحصاء، قسم الإحصاء والرياضة والتأمين، كلية التجارة، جامعة عين شمس .
٢ ماجستير الإحصاء التطبيقى، قسم الاحصاء والرياضة والتأمين، كلية التجارة، جامعة عين شمس .