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

A thesis submitted in partial fulfillment of the requirement for the Master's Degree in Applied Statistics

By

Mahmoud Rady Hamed Noah
Faculty of Commerce, Ain Shams University

Supervised by

Dr. Amr Ibrahim Abdelrahman Elatraby
Associate Professor – Department of Statistics, Mathematics and Insurance
Faculty of Commerce, Ain Shams University

Dr. Madiha Abdelghany Abo Elmagd
Assistant Professor – Department of Statistics, Mathematics and Insurance
Faculty of Commerce, Ain Shams University

2010

Approval Sheet

Title of Thesis : A Proposed Statistical Model for Estimating the Probability of Second Primary Cancer Occurrence with Application in Ain Shams University Hospitals.

Academic Degree : MBA in Applied Statistics.
Name of Student : Mahmoud Rady Hamed Noah.

This thesis submitted in partial fulfillment of the requirement for
The Master's Degree in Applied Statistics has been approved by:

Examination Committee

1- Prof. Ahmed Hassen Ahmed Youssef
Professor – Department of Applied Statistics and Econometrics
Institute of Statistical Studies and Research
Cairo University.
……..

2- Dr. Tolba Elsayed Zein Eldin
Associate Professor – Department of Statistics, Mathematics and Insurance
Faculty of Commerce
Ain Shams University.
……..

3- Dr. Amr Ibrahim Abdelrahman Elatraby
Associate Professor – Department of Statistics, Mathematics and Insurance
Faculty of Commerce
Ain Shams University.
……..

Date of Dissertation Defense / / 2010 Approval Date / / 2010

Acknowledgments

After the help of **God**, *many people have been a part of my gratitude; I feel deep gratitude to those people who have helped me: friends, professors, and colleagues.* **Dr. Amr Elatraby**, *first and foremost, has been the best advisor and professor I could have wished for, he was a great help for all his students. I am grateful to him for pushing me forward, Time after time. In the personal side, he did not hesitate to invite his students to become an extended part of his family. I appreciate this immensely. My gratitude also goes to* **Dr. Madiha Abdelghany** *for her great supervision, help and advice.*

Very grateful to **Prof. Ahmed Hassen** *for his valuable comments and accepting sharing in the examination committee. Many thanks to* **Dr. Tolba Zein Eldin** *for his valuable comments and accepting sharing in the examination committee, which was generous gesture from them.*

Special thanks are for **Prof. Saadia Montasser** *for answering my questions, without ever really having been bored.*

I was lucky enough to have the help of many good friends. Life would not be the same without my classmates Reham Sayed and Mariam Mahmoud.

I also appreciate the assistance I received from the officials of the radiotherapy department; Faculty of Medicine, Ain shams university hospitals for providing the data for this research, especially ***Prof. Atef Youssef****.*

Finally, I would like to thank those closest to me, whose presence helped make the completion of my graduation work possible. Most of all, I would like to thank my family, and especially my sister Afaf, for their absolute confidence in me. Knowing that they are always there helping me to overcome all difficulties.

Contents

Chapter Four: Linear Discriminant Analysis

Chapter Five: Logistic Regression and Discriminant Analysis.

Appendices

Arabic Summary

List of Tables

List of Figures

Summary

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

In this thesis, we use some classification methods to determine the social-demographic risk factors in addition to treatment by radiation which affected the second primary cancer occurrence and its probability for patients who were initially treated for first primary cancer stage I that have at least one year cancer-free after first primary cancer treatment. We consider about this cases because they have high risk to develop a second primary cancer. Treatment by radiation and social-demographic risk factors such that: age at first cancer, gender, living area nature, marital status, family history, smoking, education and obesity will be studied using the logistic regression model and the discriminant analysis model. We applied the logistic regression model (LR) to estimate the probability of having second primary cancer. The odds ratio analysis is used to compares whether the probability of having a second primary cancer is the same for the two groups for each factor. For testing the significance of the coefficients we used Wald test and likelihood ratio test. Hosmer and Lemeshow test and cross validation is considered to assess the fit of the model.

Linear discriminant analysis (LDA) is used as a comparative method with logistic regression model results.

Nature of the Problem:

Sometimes after the therapy; the patient is exposed to develop a second primary cancer that may make the patient feels depressed and hopeless at the therapy. We may assume that the age and smoking are factors causing the occurrence of second primary cancer but we need to test this kind of assumptions and to know how far these factors are responsible for causing the second primary cancer.

Objectives of the Study:

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; that may help doctors to focus on these affected factors and inform patients to avoid it, and give more care for second cancer's predicted patients. The usage of statistical methods to identify risk factors would help to identify the probability of second primary cancer occurrence.

This Study Proposes to:

a. Identifying 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; depending on the estimated probability and the used cut-point; that at 0.5 cut-point when the estimated probability exceeded 0.5 the patient will classify as an expected second cancer patient.

Logistic regression analysis method is used to estimate the optimal model which helps us to estimate the probability of second primary cancer. We used Wald test, likelihood ratio test, Hosmer-Lemeshow test, cross validation and ROC curve to verify the model.

Linear discriminant analysis (LDA) is used for compare with the logistic regression (LR) results. We show that in small samples size; the results of LDA and LR are close even if the normality assumptions did not exist and we also set some guidelines for recognizing these situations.

Source of Data and Variables:

From 1500 registered patients in Ain shams university hospitals, Cairo, Egypt, in different stages of cancer at 2006; 240 patients meet the study assumptions as follows:

a) Patients have a first primary cancer stage I.

b) Patients are at least one year cancer-free after first cancer treatment.

We studied data from 240 patients in the study.

Dependent variable: having a second primary cancer.

Independent variables:

a. Patient's age at first cancer,

b. Patient's gender,

c. Patient's marital status,

d. Living area,

e. The first cancer treatment by radiation,

f. Patient has cancer family history (first degree relatives),

g. Smoking history,

h. Patient was suffering from obesity before the first primary cancer, and

i. Patient's education.

Most of medical risk factors (i.e., radiation dose rate, chemotherapy dose rate, number of nodes of first cancer,

first cancer size) were not available at the hospitals records when the research was conducted.

Results

The logistic regression model and the Discriminant analysis show that the patients who smoke, with family history in cancer and married are more exposed to a second primary cancer occurrence.

The hit rate for the binary logistic regression full model classification and discriminant analysis full model classification is 80%, and 79.6 % for cross validation classification for the two models. This means that; the model has high classification accuracy and it is fit for predication; so we can depending on the logistic regression model for estimating the probability of the second primary cancer occurrence and the significant factors are important and play a role in determine this probability.

The hit rate for the forward Likelihood stepwise binary logistic regression method classification and Wilks' lambda stepwise discriminant analysis method classification is 80.4%, and 80 % for cross validation classification for the two models. This means that; the model has high classification accuracy more than the full model with only the significant factors and it

is fit for predication; so we recommended that depending on the stepwise logistic regression model for estimating the probability of the second primary cancer occurrence and the significant factors are important and play a role in determine this probability.

Chapter one is an overview on second primary cancer occurrences, causes, and types. Also this chapter views the importance and the objectives of the study.

In chapter two, we introduced an introduction to categorical data analysis, definition, types, properties, and methods to analyzing.

In chapter three, we present the logistic regression model; the Wald test, likelihood ratio test, Hosmer and Lemeshow test, cross validation methods and ROC curve.

In chapter four, we present an introduction to the discriminant analysis function; Walks' lambda statistic, The Eigenvalue, The Canonical Correlation and Press's Q statistic.

In chapter five, we apply the binary logistic regression analysis to estimate the probability of the occurrence of the second primary cancer; and discriminant analysis to estimate the probability of the occurrence of the second primary cancer,