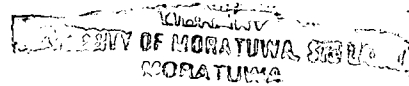


LB/DON/52/05
ma 08/01



Statistical Analysis and Modeling of Factors Influencing Lung Cancer



Final Report

University of Moratuwa, Sri Lanka
Electronic Theses & Dissertations

The Dissertation submitted for the

Degree of

MASTER OF SCIENCE

Department of Mathematics

Faculty of Engineering

University of Moratuwa

Sri Lanka

February, 2005

51° 05"
51043)



A. N. Ekanayake

Master of Science (Full Time)

02/8029

University of Moratuwa



83467

Thesis
83467

83467

DECLARATION

I certify that the dissertation entitled "**STATISTICAL ANALYSIS AND MODELING OF FACTORS INFLUENCING LUNG CANCERS IN SRI LANKA**" is entirely my own work. It has not been accepted for any degree and it is not being submitted for any other degree.

Candidate

A.N. Ekanayake

Signature : *A.N. Ekanayake*

Date : 01/03/05

Supervisor

Dr. M. Indralingam

Signature : *M. Indralingam*

Date : 1/3/05



University of Moratuwa, Sri Lanka
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

ABSTRACT

Statistics show that lung cancer occupies the third position among the incidence rates of cancers in Sri Lankan males and this rate is increasing yearly. This research is focused on two main areas. These are to find factors associated with lung cancers and study on time to death after detection of a lung cancer, known as the survival time.

Data collection was done at Cancer Institute, Maharagama (CIM) which is the largest hospital for treatment for the disease in Sri Lanka. Three sources of data have helped in this research study. First one was data in summary format at the CIM. Second was file belongs to each of the patients. Third was the patient's detail form, which is filled by a patient. All together two hundred and sixty two lung cancer patients have come to CIM, in the study period from 1st January to 31st December 2002.

Findings of this research are as follows. Smoking is the main risk factor for lung cancers. People who do occupations in areas uncovered for polluted air have high risk for lung cancer. There is a genetic effect for lung cancer. Consuming alcohol and chewing betel are also considerable factors for lung cancer. Having Tuberculosis is also risk factor for lung cancer. Among four types of lung cancer viz.; Adenocarcinoma, Squamous cell carcinoma, Small cell carcinoma and Large cell carcinoma, the most common types in Sri Lanka are Adenocarcinoma and Squamous cell carcinoma. Age, sex, religion and smoking habit of the patient have high relationship with those two types of lung cancer. A male person with age greater than 48 years having smoking habit is more susceptible to Squamous cell carcinoma than for Adenocarcinoma.

This research shows that the mean survival time of lung cancer patient is approximately 6 months. Treatment given at Cancer Institute, Stage of diagnosis and sex of the patient affect survival time. Treatment mixture reduces risk of death by half compared to single treatment. Our research shows that of a patient is diagnosed for a lung cancer in extended stage, he/she has eleven times more in risk of death than a patient with localized stage. Risk of death for males is three times more than females.

ACKNOWLEDGEMENT

It is with greatest respect and veneration that I express my sincere thanks to my main supervisor Dr. M Indralingam, the Coordinator of the postgraduate students of Department of Mathematics, University of Moratuwa. Also my sincere thanks to Proff. G.T.F. De Silva, Dr. R. Lokupitiya and Mr. T.M.J.A. Cooray, Senior Lectureres of Department of Mathematics, University of Moratuwa and Mrs. C.P.N. Attygalle, Institute of Technology, University of Moratuwa. I have to appreciate the assistance of Dr. Roshini Sooriarachchi, Senior Lecturer of Department of Statistics, University of Colombo. My special thanks for Dr. Murali Vallipurathan, Ministry of Health and Dr.R.V.Rabel, Medical Officer, University of Moratuwa who give me kind support in Medical field. Without the knowledge, advice and vast experience, which was imparted to me by my supervisors I may not have been able to complete the project successfully.

I would like to take this opportunity to thank specially for Proff (Mrs) N. Rathnayake, Director of Post Graduate unit of University of Moratuwa and ADB for granting me this scholarship to conduct this research and the head of the Department and all the staff members and non academic staff of the Department of Mathematics, University of Moratuwa. It is my obligation to thank all my friends for their support.

I would also like to thank Dr. Yasantha Ariyaratna, Director, Cancer Institute, Dr.(Mrs) Nirmala Gammanpila, and the staff of the record room and director's office of the Cancer Institute, Maharagama for providing me relevant information and data to carry out this research.

The assistance given by Dr. Thushara Fernando and the staff members of Planning division, Ministry of Health is gratefully acknowledged. Also a special thanks has to be given for the Medical Statistics Unit of Census and Statistics Department.

I gratefully acknowledge the support and encouragement given by my loving parents and my sister.

Finally special thank are due to my loving husband for his valuable advice, encouragement in all my endeavors in this research.

CONTENTS

	<i>Page no.</i>
DECLARATION	i
ABSTRACT	ii
ACKNOWLEDGEMENT	iii
CONTENTS	iv
LIST OF TABLES	vii
LIST OF FIGURE	ix
DEDICATION	xi
1. INTRODUCTION	
1.1 Background	1
1.2 What is a lung cancer?	1
1.2.1 Introduction	1
1.2.2 Understanding the cancer process	1
1.2.3 Risk factors of lung cancer	2
1.2.4 Types of lung cancer	4
1.2.5 Symptoms and signs of lung cancer	4
1.2.6 Diagnosing lung cancer	5
1.2.7 Stage explanation	5
1.2.8 Treatment for lung cancer	6
1.3 Objectives of the study	9
2. LITERATURE REVIEW	
2.1 Studies carried out in Sri Lanka	10
2.2 Studies carried out in other countries	14
3. METHODOLOGY AND TECHNIQUES USED	
3.1 Methodology	27
3.1.1 Literature survey	27
3.1.2 Study area and study population	28

3.1.3 Data collection	28
3.2 Techniques used	35
3.2.1 Descriptive methods	35
3.2.2 Univariate methods	36
3.2.3 Model fitting	37
3.2.4 Survival data analysis	44
3.2.5 Special software used in analyzing	45
4. DESCRIPTIVE STUDY ON LUNG CANCER PATIENTS	
4.1 Personal details of lung cancer patients in Sri Lanka	47
4.2 General information on Lung cancer patients	54
4.3 Habits of Lung Cancer Patients	58
4.4 Clinical Features of Lung Cancer Patients	60
4.5 Identify the common lung cancers in Sri Lanka according to cell types	62
4.6 Summary	67
5. LOGISTIC REGRESSION ANALYSIS TO LUNG CANCER PATIENTS	
5.1 Univariate analysis for determining variables affecting two common types of lung cancer in Sri Lanka	68
5.1.1 Testing the relationship between two types of lung cancer with personal information	69
5.1.2 Testing the relationship between two types of lung cancer with personal habits	72
5.2 Fitting logistic model for two types of lung cancer; Adenocarcinoma and Squamous cell carcinoma	73
5.2.1 Model fitting procedure	73
5.2.2 Parameter estimates of the fitted model	75
5.2.3 Goodness of fit of the model	76
5.3 Further areas of study	76
5.4 Summary	77
6. SURVIVAL DATA ANALYSIS FOR LUNG CANCER PATIENTS	
6.1 Some Definitions	78
6.2 Descriptive Methods of Comparing Survival Time for Different Groups of Individuals	81

6.3 Log-Rank test to compare survival times for different groups of individuals	100
6.4 Modeling survival data	102
6.4.1 The selection of the variables to be included in the model	103
6.4.2 Parameter estimates of the selected model	106
6.5 Model Diagnostics	107
6.5.1 Plots of the Cox-Snell residuals	107
6.5.2 Plots of Martingale residuals	108
6.5.3 Plot of Deviance residuals	110
6.6 Summery	110
7. CONCLUSIONS AND DISCUSSION	
7.1 Discussion	112
7.1.1 Descriptive Study on lung cancer	112
7.1.2 Association between common types of lung cancer viz. Adenocarcinoma and Squamous cell carcinoma and some variables	113
7.1.3 Statistical analysis of time to death after detecting lung cancer	114
7.2 Conclusions of the study	115
7.2.1 Conclusions of the general findings	115
7.2.2 Conclusions of analyzing time to death after detecting (survival time) a lung cancer	116
7.3 Recommendations	117
7.4 Problems of the study	117
7.5 Limitations of the study	118
7.6 Further studies	118
APPENDIX I	i
APPENDIX II	xxii
REFERENCES AND BIBLIOGRAPHY	xxvii

LIST OF TABLES

	<i>Page no.</i>
Table 3.1: Format of data summarizing at Cancer Institute	30
Table 3.2: Data collecting format	34
Table 3.3: Contingency Table of Factor A by Factor B	36
Table 4.1: Frequency distribution of patients district wise	49
Table 4.2: Frequency distribution of patients ethnic group wise	51
Table 4.3: Frequency distribution of religion	51
Table 4.4: Frequency distribution of no. of children to the lung cancer patient	52
Table 4.5: Frequency distribution of occupation of lung cancer patient	53
Table 4.6: Frequency distribution of monthly income of lung cancer patients	54
Table 4.7: Frequency distribution of getting lung cancer previously	55
Table 4.8: Frequency distribution of gene effect	56
Table 4.9: Frequency distribution of type of the hospital referring from	56
Table 4.10: Frequency distribution of having or had TB	57
Table 4.11: Frequency distribution of years with smoking	58
Table 4.12: Frequency distribution of smoking quantity per day	58
Table 4.13: Frequency distribution of consuming alcohol	59
Table 4.14: Frequency distribution of patients chewing betel	59
Table 4.15: Frequency distribution of laterality	60
Table 4.16: Frequency distribution of diagnostic status	60
Table 4.17: Frequency distribution of diagnostic evidence	61
Table 4.18: Frequency distribution of stage of diagnosis	61
Table 4.19: Frequency distribution of treatment given at cancer institute	62
Table 4.20: Frequency distribution according to cell type of lung cancer	62
Table 5.1: Associated demographic variables with two types of lung cancer	69
Table 5.2: Associated family background variables with two types of lung cancer	71
Table 5.3: Associated personal habits with two types of lung cancer	72
Table 5.4: Factors used for fitting logistic model	73
Table 5.5: Details of adding main effects to null model	74

Table 5.6: Details of adding main effects to model containing smoking status	75
Table 5.7: Parameter estimates of the fitted model	75
Table 6.1: Median survival times for type of the lung cancer	82
Table 6.2: Median survival times for stage of diagnosis	84
Table 6.3: Median survival times for treatment given at Cancer Institute	85
Table 6.4: Median survival times by sex of the patient	87
Table 6.5: Median survival times by age	88
Table 6.6: Median survival times for ethnicity	90
Table 6.7: Median survival times for religion	91
Table 6.8: Median survival times for smoking habit of the patient	93
Table 6.9: Median survival times for chewing betel habit of the patient	94
Table 6.10: Median survival times for consuming alcohol habit of the patient	96
Table 6.11: Median survival times for genetic effect of the patient	97
Table 6.12: Median survival times for presenting Tuberculosis of the patient	99
Table 6.13: Log-Rank test results for variables considering in survival analysis	100
Table 6.14: Resulting $-2\log L$ and degrees of freedom on fitting each variable separately	103
Table 6.15: Resulting $-2\log L$ and degrees of freedom on adding each term to the model adjusted for treatment given	104
Table 6.16: Resulting $-2\log L$ and degrees of freedom on adding each term to the model adjusted for treatment given and stage of diagnosis	104
Table 6.17: Resulting $-2\log L$ and degrees of freedom on adding each term to the model adjusted for treatment given, stage of diagnosis and sex of the patient	105
Table 6.18: Parameter estimates of the selected model	106

LIST OF FIGURES

	<i>Page no.</i>
Figure 4.1: Age distribution of lung cancer patients	47
Figure 4.2: Lung cancer patients by sex	48
Figure 4.3: Age distribution of lung cancer patients by sex	48
Figure 4.4: District distribution of lung cancer	50
Figure 4.5: Lung cancer patients by their marital state	52
Figure 4.6: Lung cancer patients by their living status	55
Figure 4.7: Lung cancer patients by transferred hospital	57
Figure 4.8: Lung cancer patients by smoking status	58
Figure 4.9: Lung cancer patients by cell type	62
Figure 4.10: Age distribution in two types; Adenocarcinoma and Squamous cell carcinoma	64
Figure 4.11: Age distribution of two types of lung cancers by sex	64
Figure 4.12: Common two types of lung cancer patients by sex	65
Figure 4.13: Two types of lung cancer patients by their smoking status	66
Figure 6.1: Kaplan-Meier Estimates of Survivor functions for type of the lung cancer	82
Figure 6.2: Log-Cumulative Hazard plot for type of lung cancer patients	83
Figure 6.3: Kaplan-Meier Estimates of Survivor functions for stage of diagnosis	84
Figure 6.4: Log-Cumulative Hazard plot for stage of diagnosis of lung cancer patients	85
Figure 6.5: Kaplan-Meier Estimates of Survivor functions for treatment given	86
Figure 6.6: Log-Cumulative Hazard plot for treatment given at Cancer Institute	86
Figure 6.7: Kaplan-Meier Estimates of Survivor functions by sex of the patient	87
Figure 6.8: Log-Cumulative Hazard plot by sex of the patient	88
Figure 6.9: Kaplan-Meier Estimates of Survivor functions by age of the patient	89
Figure 6.10: Log-Cumulative Hazard plot by age of the patient	89
Figure 6.11: Kaplan-Meier Estimates of Survivor functions for ethnicity	90
Figure 6.12: Log-Cumulative Hazard plot for ethnicity	91

Figure 6.13: Kaplan-Meier Estimates of Survivor functions for religion	92
Figure 6.14: Log-Cumulative Hazard plot for religion	92
Figure 6.15: Kaplan-Meier Estimates of Survivor functions for smoking status	93
Figure 6.16: Log-Cumulative Hazard plot for smoking status	94
Figure 6.17: Kaplan-Meier Estimates of Survivor functions for chewing betel	95
Figure 6.18: Log-Cumulative Hazard plot for chewing betel	95
Figure 6.19: Kaplan-Meier Estimates of Survivor functions for consuming alcohol	96
Figure 6.20: Log-Cumulative Hazard plot for consuming alcohol	97
Figure 6.21: Kaplan-Meier Estimates of Survivor functions for genetic effect	98
Figure 6.22: Log-Cumulative Hazard plot for genetic effect	98
Figure 6.23: Kaplan-Meier Estimates of Survivor functions for presenting TB	99
Figure 6.24: Log-Cumulative Hazard plot for presenting TB	100
Figure 6.25: Plot of Cox-Snell residuals	107
Figure 6.26: Log-Cumulative Hazard plot of the Cox-Snell residuals	108
Figure 6.27: Plot of Martingale residuals Vs Rank of survival time	108
Figure 6.29: Plot of Martingale residuals Vs treatment given	109
Figure 6.28: Plot of Martingale residuals Vs Stage of diagnosis	109
Figure 6.30: Plot of Martingale residuals Vs sex of the patient	109
Figure 6.31: Plot of Deviance residuals Vs Rank of survival time	110



DEDICATION

*To My Loving Parents, Sister
and
My Loving Husband*



University of Moratuwa, Sri Lanka
Electronic Theses & Dissertations
www.lib.mrt.ac.lk