12-2010

Data mining and analysis of lung cancer data.

Guoxin Tang
University of Louisville

DATA MINING AND ANALYSIS OF LUNG CANCER DATA

By

Guoxin Tang
B.S., Tianjin University, China, 2001
M.S., Tianjin University, China, 2004
M.A., University of Louisville, 2007

A Dissertation
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

Department of Mathematics
University of Louisville
Louisville, Kentucky

December 2010

DATA MINING AND ANALYSIS OF LUNG CANCER DATA
By

Guoxin Tang
B.S., Tianjin University, China, 2001
M.S., Tianjin University, China, 2004
M.A., University of Louisville, 2007

A Dissertation approved on

September 24, 2010

by the following Dissertation Committee:

Dr. Patricia Cerrito (Dissertation Director)

Dr. Ryan Gill

Dr. Kiseop Lee

Dr. Ahmed Desoky

Dr. Mehmed M. Kantardzic

DEDICATION

This Dissertation is dedicated

to my parents Mr. Airen Tang and Mrs. Aiyu Wang

to my wife Yan Jin

and to my brother Lixin Tang

for their support, guidance and encouragements.

ACKNOWLEDGEMENT

I would like to thank my academic advisor Dr. Patricia Cerrito for her guidance and support throughout the research during the past four years. Her valuable comments on this project and opportunities of conferences are greatly appreciated.

I also thank all members of my dissertation committee, Dr. Ryan Gill, Dr. Kiseop Lee, Dr. Ahmed Desoky and Dr. Mehmed M. Kantardzic, for their feedback, their comments and their time for reviewing the work.

A special thanks to my brother, Lixin Tang, and my parents, Mrs. Aiyu Wand and Mr. Airen Tang. They have been a great source of motivation and support. Many thanks to my wife, Yan Jin, for her support and patience.

ABSTRACT

DATA MINING AND ANALYSIS OF LUNG CANCER DATA

Guoxin Tang

September 24, 2010

Lung cancer is the leading cause of cancer death in the United States and the world, with more than 1.3 million deaths worldwide per year. However, because of a lack of effective tools to diagnose Lung Cancer, more than half of all cases are diagnosed at an advanced stage, when surgical resection is unlikely to be feasible. The main purpose of this study is to examine the relationship between patient outcomes and conditions of the patients undergoing different treatments for lung cancer and to develop models to predict the mortality of lung cancer. This study will identify the demographic, finance, and clinical factors related to the diagnosis or mortality of Lung Cancer to help physicians and patients in their decision-making.

We combined Text Miner and Cluster analysis to identify the claim data for Lung Cancer and to determine the category of diagnosis, treatment procedures and medication treatments for those patients. Moreover, the claims data were used to define severity level and treatment categories. Compared with using diagnosis

codes directly, the combination of text mining and cluster analysis is more efficient and captures more useful information for further analysis. In order to analyze the mortality of Lung Cancer, we also found that survival analysis is appropriate to preprocess the data for the relationship between a predictor variable of interest and the time of an event. The proportional hazard model examined the effects of different treatment clusters using a hazard ratio and the proportional effect of a treatment cluster (treatment procedure or medication treatment) may vary with time. A decision tree was built to generate rules for identifying high risk lung cancer cases among the regular inpatient population.

Two primary data sets have been used in this study, the Nationwide Inpatient Sample (NIS) and the Thomson MedStat MarketScan data. Kernel density estimation was used for NIS to examine the relationship between Age, Length of stay, Diagnosis Categories, Total Cost and Lung Cancer by visualization. The Kaplan–Meier method and Cox proportional hazard model are used for the Medstat data to discover the relationship between the factors and the target variable for more detail. Time series and predictive modeling are used to predict the total cost for hospital decision making, the mortality of Lung cancer based on the historical data and to generate rules to identify the diagnosis of Lung cancer.

Older patients are more likely to have lung cancers that would lead to a higher probability of longer stay and higher costs for the treatment. Within 7 defined

clusters of diagnosis for Lung Cancer, the malignant neoplasm of lobe, bronchus or lung is under higher risk. Age, length of stay, admit type, clusters of diagnosis, and clusters of treatment procedures and Major Diagnostic Categories (MDC) were identified as significant factors for the mortality of lung cancer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I INTRODUCTION

Cancer researchers, clinicians and public notice are dedicated to develop or improve statistical models to predict the occurrence or the mortality of some cancers. Many risk prediction models have been developed for chronic disease since 1976[1]. People are more and more interested in the development of individual risk assessment methods of lung cancer, which also has been identified as an area of extraordinary opportunity by the National Cancer Institute. However, very few models have been developed to estimate the risk of lung cancer so far, in contrast to more prevalent modeling in the breast and certain other sites[2].

Cancer predictive models were applied to design, plan, and establish eligibility criteria for cancer intervention and screening trials that also have been used to identify individuals at high risk of cancer who may benefit from targeted screening or other interventions such as tamoxifen chemoprevention [3]. Cancer predictive models are also used to examine the population distribution, the cost and the impact of interventions. They are used in clinical decision making to help physicians and patients determine the current stage of the cancer and appropriate screening regimens, medication, and/or interventions [3].

In addition, personalized medicine will play an important role in the treatment of Lung cancer in the next few years. Recent advances in drug development, pharmacogenomics, and the molecular characterization of tumors have brought the opportunity for individualized selection of treatment based on the characteristics of the patient and the tumor. As a genetic disease, it is useful to know how many genes are altered in solid tumors associated with Lung Cancer. With the advances in massively parallel sequencing technologies, new cancer genomes are being sequenced at an astonishing rate. As more cancer genomes are sequenced, it will be possible to identify new genes and key mutations [4] [5]. However, as Anirban Mahapatra mentioned, molecular characterization of lung cancer tumors will continue to be difficult because of the heterogeneity of the constituent cells [5].

The main purpose of this study is to investigate the relationship between the demographic factors, finance, clinical conditions and the diagnosis of Lung Cancer. Specific lung cancer risk prediction models based on available variables were developed, evaluated and used to predict the cost and mortality for the patients and to build rules for identifying diagnoses of Cancer, which will identify high-risk individuals and help the physician's clinical decision making.

This research focuses on the Inpatient Sample (NIS) (inpatient claims only)[6] and Medstat MarketScan Databases (Inpatient, outpatient, and pharmacy claims) [7]. The NIS database includes five years of data, 2000- 2004, and about 8 million records for each year. The Thomson MedStat MarketScan data containing all patient claims for 40 million people followed for the years 2000-2001. We use publicly available data throughout this study.

Each observation includes fifteen columns of Diagnosis codes and fifteen columns of Procedure codes in these data sets. One inpatient admission consists of one or more observations. It is possible that there are multiple admissions for one patient. Here, any diagnosis or procedure code can appear in any one of the fifteen columns. In order to identify the claims for Lung cancer, we work with all fifteen columns to bring all of them into one column as a string of codes, using the CATX function, which concatenates character strings, removes leading and trailing blanks, and inserts separators.

This dissertation is divided into nine chapters. In chapter 2, the background and current status of lung cancer in the world and USA will be explored as well as how Lung Cancer has been staged, which shows the severity level of the cancer and the survival rate for each stage. In chapters 3, 4, 5 and 6, we will give some background, concepts and theories for data mining, cluster analysis, survival analysis, time series and predictive modeling, respectively. In chapter 7, some methods mentioned above will be used to analyze the Nationwide Inpatient Sample (NIS) database. Lung Cancer claims data will be identified by using the Text Mining methodology and Kernel Density Estimation (KDE procedure), which was used to examine the relation of lung disease by Age, Length of Stay and

Total Charges. Then, cluster analysis will be used to define diagnosis categories. Furthermore, an ARIMA model with inflation rate as the dynamic regressor and a logistic regression model will be used to predict the cost and mortality of lung cancer, respectively. In chapter 8, survival analysis and predictive modeling will be applied to the Medstat MarketScan Databases and the four main treatment procedures will be defined. Also, a Cox Proportion Hazard Model gives us the hazard ratio of each variable to explain how it relates to the mortality of Lung Cancer. Then, Decision trees give us the rules for identifying Lung Cancer for reference and the Neural Network model is optimal for the prediction of the mortality of lung cancer. Chapter 9 concludes and summarizes all the findings.

CHAPTER II BACKGROUND

2.1. Lung Cancer

At the end of the 20th century, lung cancer had become the first worldwide cancer in deaths and the number is rising every year. Lung cancer was a rare disease at the beginning of that century. Table 2.1 shows the estimated number of cases and deaths for 26 different types of cancer in men and women together with the age, standardized incidence and mortality rates and the cumulative risk (%) between ages 0 and 64 during 2002. Lung cancer is the main cancer with 1.5 million cases and 1.18 million deaths in the world yearly [8].

There were 1.35 million new cases of lung cancer, which was about 12.4% of all new cancers. Furthermore, at the beginning, more than half of the cases occurred in the developing countries of the world; now it is estimated that 69% were in developed countries since 1980. It is the most common cancer for men, with the highest rates observed in North America and Europe. For women, the incidence rates are lower compared to men (globally, the rate is 12.1 per 100,000 women compared with 35.5 per 100,000 in men). The highest rates for women are in North America and Northern Europe. (D. Max Parkin, Freddie Bray, Global Cancer Statistics, 81) [8].

Table 2.1 Incidence and Mortality by Sex and Cancer site worldwide, 2002

	Incidence				Mortality			
	Males		Females		Males		Females	
	Cases		Cases	Cumulative risk(age 0-64)	Deaths	Cumulative risk(age 0-64)	Deaths	Cumulative risk(age 0-64)
Oral Cavity	175,916	0.4	98,373	0.2	80,736	0.2	46,723	0.1
Nasopharynx	55,796	0.1	24,247	0.1	34,913	0.1	15,419	0.0
Other pharynx	106,219	0.3	24,077	0.1	67,964	0.2	16,029	0.0
Esophagus	315,394	0.6	146,723	0.3	261,162	0.5	124,730	0.2
Stomach	603,419	1.2	330,518	0.5	446,052	0.8	254,297	0.4
Colon/rectum	550,465	0.9	472,687	0.7	278,446	0.4	250,532	0.3
Liver	442,119	1.0	184,043	0.3	416,882	0.9	181,439	0.3
Pancreas	124,841	0.2	107,465	0.1	119,544	0.2	107,479	0.1
Larynx	139,230	0.3	20,011	0	78,629	0.2	11,327	0
Lung	965,241	1.7	386,891	0.6	848,132	1.4	330,786	0.5
Melanoma of Skin	79,043	0.2	81,134	0.2	21,952	0	18,829	0
Breast			1,151,298	2.6			410,712	0.9
Cervix uteri			493,243	1.3			273,505	0.7
Corpus uteri			198,783	0.4			50,327	0.1
Ovary			204,499	0.5			124,860	0.2
Prostate	679,023	0.8			221,002	0.1		
Testis	48,613	0.1			8,878	0		
Kidney	129,223	0.3	79,257	0.1	62,696	0.1	39,199	0.1
Bladder	273,858	0.4	82,699	0.1	108,310	0.1	36,699	0
Brain,nervous system	108,221	0.2	81,264	0.2	80,034	0.2	61,616	0.1
Thyroid	37,424	0.1	103,589	0.2	11,297	0	24,078	0
Non-Hodgkin lymphoma	175,123	0.3	125,448	0.2	98,865	0.2	72,955	0.1
Hodgkin Disease	38,218	0.1	24,111	0.1	14,460	0	8,352	0
Multiple	46,512	0.1	39,192	0.1	32,696	0.1	29,839	0

	Incidence				Mortality			
	Males		Females		Males		Females	
	Cases		Cases	Cumulative risk(age 0-64)	Deaths	Cumulative risk(age 0-64)	Deaths	Cumulative risk(age 0-64)
myeloma								
Leukemia	171,037	0.3	129,485	0.2	125,142	0.2	97,364	0.2

2.2 The Causes of Lung Cancer

The etiology of Lung cancer is still not completely clear, and a large amount of data show that lung cancer risk factors include smoking, asbestos, radon, arsenic, ionizing radiation, halogen alkene and polycyclic aromatic compounds, nickel, etc.

Cigarette smoking is a well-established cause of lung cancer. Carcinogens are released when burning cigarettes. Long-term smoking can cause a phosphorus shape of bronchial epithelium cells, which leads to squamous cell carcinoma or undifferentiated small cell carcinoma. Among those patients who did not smoke, the adenocarcinoma are more common than lung cancer. In fact, an estimated 87% of all lung cancers can be attributed to cigarette smoking alone [9]. Although smoking is by far the leading cause of lung cancer, the disease has several other causes. Lung cancers are still directly or indirectly related to tobacco use from cigars, pipes, and secondhand cigarette smoke, but several other risk factors act independently or synergistically with tobacco to cause lung cancer. Occupational and environmental exposures, such as asbestos, arsenic, secondhand smoke, and radon also increase the risk of lung cancer.

2.3. Diagnosis and Staging of Lung Cancer [10]

There are no specific symptoms in the early stage of lung cancer, only for general respiratory diseases with symptoms such as cough, phlegm with blood, low fever, chest pain, etc, which are often ignored. Here, we list some symptoms that occur during advanced stages of lung cancer.

a) The edema of face or neck.

b) Hoarse voice, a common symptom.

c) Difficulty in breathing.

It easily causes metastases for Lung cancer. If it metastasizes to the brain, it would cause a persistent headache, which is not obviously different from ordinary tension headaches. If the cancer metastasizes to the bone, it will cause damage to the bone. The most difficult cancer metastasis is to the spine, which follows with back pain.

a) Methods of Diagnosis

What method of diagnosis of suspected lung cancer should be applied depends on the type of lung cancer, the size and location of the primary tumor, the presence of metastasis, and the overall clinical status of the patient.

In order to detect and diagnose lung cancer, the routine disease history and physical examination is the important first step, which might provide some signs of the cancer. The patient's symptoms that include observing possible indicators as discussed above could also carry information about the possibility of lung disease[9].

Here, we list the several available methods applied to detect lung disease (Information is from Thompson Cancer Survival Center) [9][10].

- Diagnostic Imaging

  The first step in determining whether a mass is cancer or benign is a diagnostic image.

  - *X-Ray*

    X-ray examination is the most common cancer diagnosis method. The X-ray examination can detect the position and size of the potential lung masses, but cannot determine if they are cancerous or benign. For early lung cancer cases, x-rays cannot detect small, potential tumors, but can show a partial obstruction due to the bronchial emphysema, atelectasis or adjacent parts of the lesion.

    - *CT Scan*

      CT stands for computerized tomography scan, which is a series of X-rays combined by a computer in a cross-sectional view and can provide a more-detailed image of the lung compared with X-rays. It is performed with injected contrast material to highlight lung tissue and suspicious masses.

    - *MRI*

      Magnetic resonance imaging (or MRI) scans use magnetism, radio waves and computer image manipulation to produce an extremely detailed image without radiation. Compared to the CAT scan, there is no damage to the human body since an MRI does not use X-rays. The MRI can provide imaging for soft tissue based on multiple planes, which the CAT scan cannot match. There are many imaging methods and parameter selections for MRI that make more applications for the MRI. Changing the radio frequency of pulse, repetition of program and echo time will change the imaging.

- Biopsies

  Biopsies are procedures where a small amount of a suspicious mass is removed for examination. There are three main types of biopsy for suspected lung masses(Thompson Cancer Survival Center) [9]:

  - *Surgical*

    A surgical biopsy is a procedure of gaining a small sample of a suspected mass. For such purposes, the patient's chest needs to be opened, and then the sample is analyzed by a pathologist.

  - *Bronchoscopy*

    From the patient's mouth or nose, a bronchoscope is inserted to the suspected area, passed through the trachea and bronchial tubes. Through this method, the physician could examine the lung mass to determine the stage of lung cancer. Usually, there is a sampling device to gain a small sample of the suspected mass for analysis.

  - *Needle Aspiration or Core Biopsy*

    In needle aspiration, a thin needle is inserted into the suspected mass and a small sample was gained for analysis.

- Sputum cytology

  In sputum cytology, the patient's sputum was examined by a pathologist under a microscope. The cancer cells can be found in most primary lung cancer patients' sputum. Therefore, the phlegm cytology is a simple and effective method for lung cancer diagnosis.

b) Lung Cancer Staging

Lung Cancer Staging is the evaluation or measurement of the size and spread of a lung cancer. Different lung cancer treatments are used at various stages of the cancer. For example, the early stage of cancer can be treated surgically while higher stages of cancer use chemotherapy and radiation in a combined therapy. The treatment and prognosis of the patients with lung cancer in general may depend largely on the stage and cancer cell types. There are two main types of lung cancer cell by the size and appearance of the malignant cells. The stage systems for non-small cell lung carcinoma (NSCLC) and small-cell lung carcinoma (SCLC) are different.

Lung cancer can be divided into four stages for NSCLC [9][11].(Thompson Cancer Survival Center)

1) *NSCLC stage I*

   The cancer is small and it has not spread into any lymph nodes or any other part of the body.

2) *NSCLC stage II*

   The cancer is in the chest area. However, it has spread into some other areas such as the chest wall, the muscle under the lung, or the phrenic, or the layers that cover the heart.

3) *NSCLC stage III*

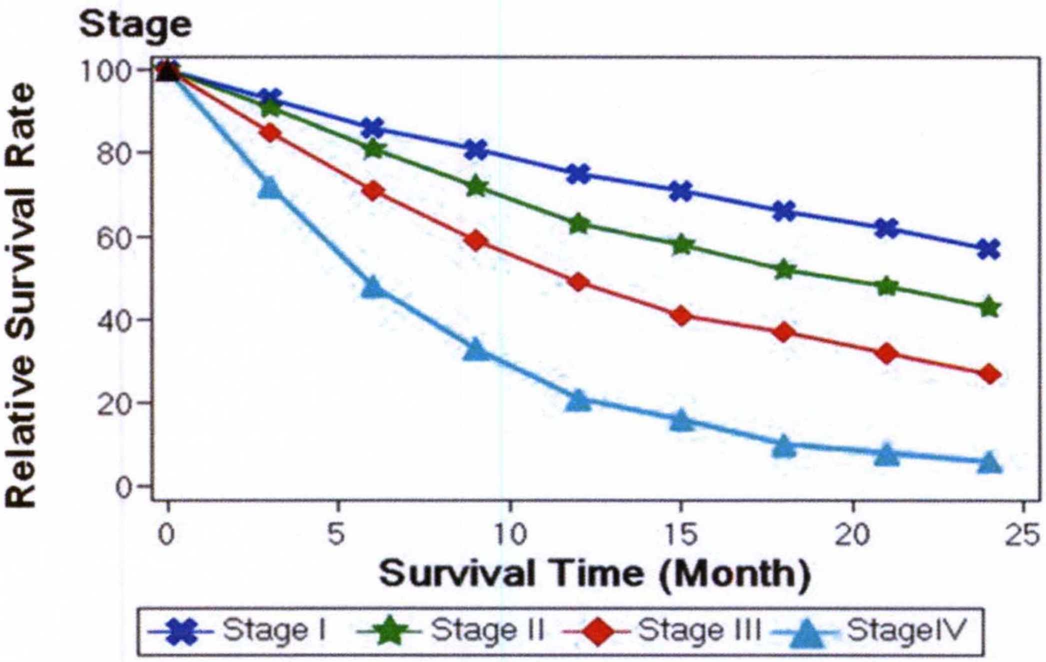
The cancer is still in the chest area, and the tumors are larger and more invasive. Furthermore, it has spread into lymph nodes on the opposite side of the chest.

4) *NSCLC stage IV*

The cancer has spread to other parts of the body, such as the liver, bones or has caused a fluid collection around the lung or heart that contains cancer cells.

The following figure shows the two year relative survival rate for each stage we discussed above.

Figure 2.1. Lung Cancer Survival by Stage [12]



2.4. Treatment of Lung Cancer

Nowadays, there are several treatments for Lung Cancer as follows.

- Surgery

  The most common treatment is Surgery. It could be used to remove those tumors for the patient in stages I and II. It also could be used for removal of a lung lobe or the entire lung. However, it is seldom used on small cell lung cancer because the disease has usually spread beyond the lung by the time it is detected and diagnosed.

- Chemotherapy

  Chemotherapy is a chemical treatment which could be used for both the non-small cell and small cell lung cancers.

- External Radiation Therapy

  CT is used to identify the location of the tumor, which could be applied before the treatment is scheduled or at the time of treatment. On subsequent visits, the tumor is radiated from different angles to maximize the dose delivered to the tumor with minimum impact on surrounding healthy tissue.

- Photodynamic therapy

  Photodynamic therapy is used for patients to limit side effects and to keep healthy lung tissue in the outpatient. However, it has not been used widely because it cannot penetrate deeply into the lung tissue.

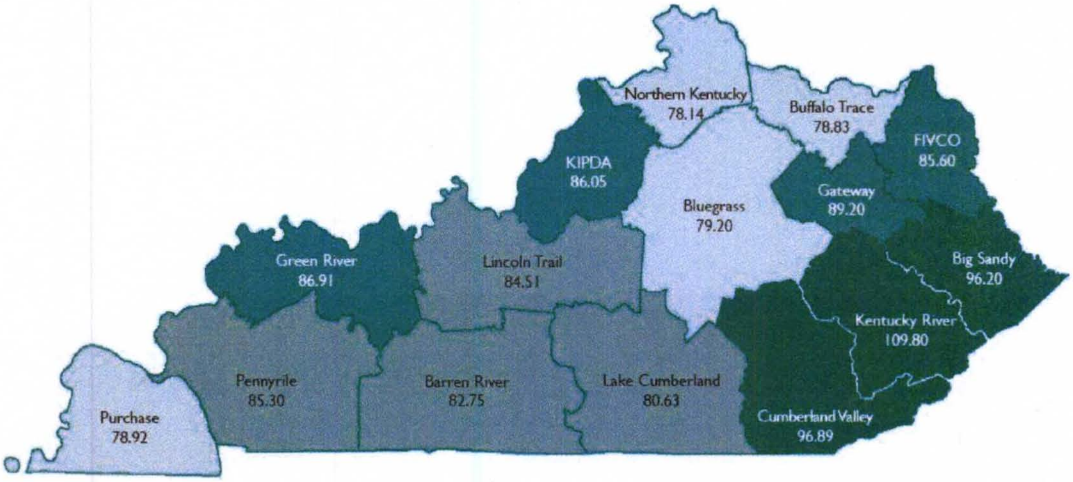
2.5. The distribution of Lung Cancer in US

Lung cancer is a devastating disease. Not only is it the most common cancer in the United States and in Kentucky, it also claims more lives than any other cancer.

Kentucky's lung cancer mortality rate is the highest of all states in the nation. The American Cancer Society estimated that in the year 2000, lung cancer accounted for 14% of all newly diagnosed cancers and 28% of all cancer deaths, killing more than 150,000 people. In Kentucky alone, more than 3,000 people die from lung cancer every year.

Kentucky has distinct geographic and lifestyle regions (e.g., western Kentucky differs substantially from eastern Kentucky, just as the state's rural areas tend to differ from its urban ones). Each region has different lung cancer rates. Figure 2.3 shows lung cancer incidence rates across Kentucky's 15 Area Development Districts for the period 1994–98. The highest rates are clustered in the Big Sandy, Kentucky River and Cumberland Valley Districts[13].

Figure 2.2 Age-adjusted lung cancer incidence rates (cases per 100,000 people) by Area Development District in Kentucky, 1994 –98 (Source: Kentucky Cancer Registry). Here, the color of light gray represents rates less than 80, the color of gray represents rates between 80 and 85.5, the color of light black represents rates between 85.5 and 90 and the black is the rate greater than 90 [13].



2.6. ICD9 Codes

The *International Statistical Classification of Diseases and Related Health Problems* (most commonly known by the abbreviation ICD) introduced the ICD9 codes to classify diseases, also including a variety of signs, symptoms, abnormal findings, complaints, and so on. Every health condition can be assigned to a unique category with up to six characters. Such categories can include a set of similar diseases that can be converted to a diagnosis.

The ICD 9 code of 162 means malignant neoplasm of trachea, bronchus, and lung. Below are all of the 4-digit codes associated with lung diseases [14].:

162.0 Trachea (Cartilage of trachea, Mucosa of trachea).

162.2 Main bronchus (Carina, Hilus of lung).

162.3 Upper lobe, bronchus or lung.

162.4 Middle lobe, bronchus or lung.

162.5 Lower lobe, bronchus or lung.

162.8 Other parts of bronchus or lung (Malignant neoplasm of contiguous or overlapping sites of bronchus or lung whose point of origin cannot be determined).

162.9 Bronchus and lung, unspecified.

The data sets we used to analyze in this paper all include ICD9 codes and DRG codes (Diagnosis-Related Group); however, DRG codes are based on ICD9 diagnosis codes, procedure codes, some patient demographic factors and the presence of complications or co-morbidities. Hence, we just use the ICD9 code to identify the lung cancer population.

2.7. Summary

The most important reason that Lung cancer is the leading cause of cancer deaths is that the symptoms of lung cancer are very often lacking or occur only late in the course of the disease. The prognosis of Lung cancer patients is very dependent on how advanced their disease is and what type it is. Developing better methodologies for distinguishing between lung cancer and other lung diseases or diagnosing early in the course of the disease will help us offer greater hope for patients.

In this dissertation, we will examine the relationship between all the factors for the lung cancer population. Several clusters for diagnosis, treatment procedures and medications will be defined based on different severity levels.

CHAPTER III DATA MINING AND TEXT MINING

3.1. Data Mining

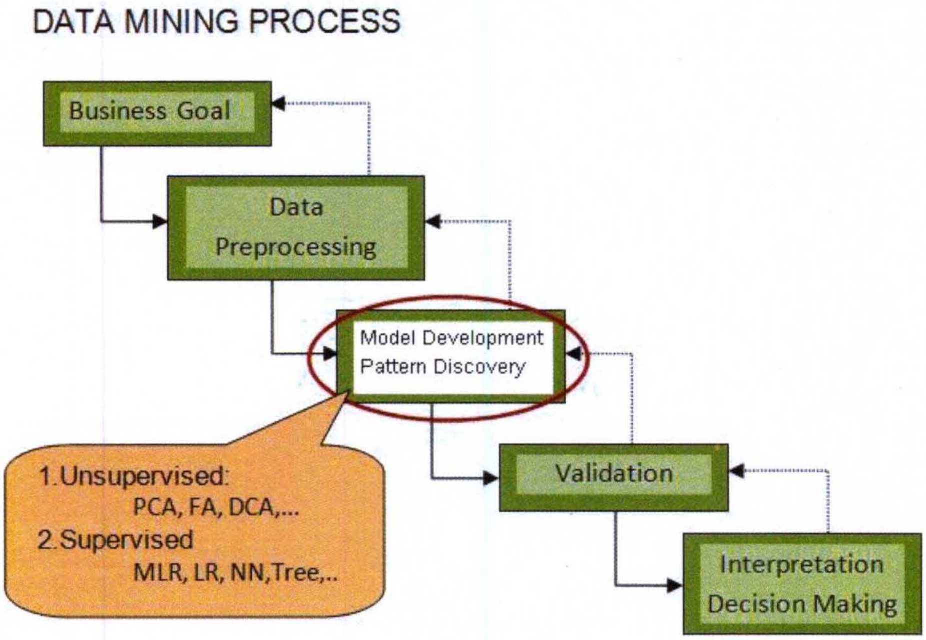
The purpose of this chapter is to introduce data mining and text mining concepts, the general data mining process, and the most commonly used data mining techniques.

Data mining is one of the fastest growing fields developed with artificial intelligence and database technology in recent years. Its core function is to extract useful information from huge data sets or a data warehouse and to deal with various complex data. Data mining is the process of selecting data, exploring data and building models using vast data stores to uncover previously unknown patterns [15]. As advanced information processing technology, the distinction between data mining and traditional data analysis is that it is a data discovery process, and in most cases, there is no hypothesis or a premise condition. Data mining can be used for market analysis, risk analysis, defect analysis and management, and so on.

3.2 Data Mining Process

Data mining is an iterative process that typically involves the following steps displayed in Figure 3.1, which also shows the phases of the Data Mining standard process[16].

Figure 3.1. General data mining process



Step Identification of Problem and Defining the Business Goal

Identification is the initial step in the modeling process. This step determines what the goal of the model will be. Usually, this step begins with an understanding of the nature of the problem and how a predictive model may aid in solving the problem. The business goal is the main target of data mining, and determines the main direction and the process of data mining. With well-identified problems and a clear definition of the business goal, data mining will lead to measurable outcomes.

Step Data Preprocessing

Data processing is generally considered the most important and time intensive step in the modeling process. The variable data preprocessing and variable selection step involves selecting the appropriate variables to include in the model. That is, the key to successful data mining is to use the appropriate data where some methods are used to distinguish the valuable information to collect, eliminate redundant data, and reduce some of the noise or random variation in the data. Often, this step involves substantial data transformation in order to set up the data so that it is used in the model in the most effective manner. Generally, data processing should also include data cleaning and data preprocessing. This cleaning and preprocessing is defined as modeling data preparation, including data sampling and data transfer, and dealing with missing data. Sampling is applied in extremely large databases because it significantly reduces the model training time. Additional data are generally separated into several parts; some of the data are used to train the model, while others are used to test and to validate the model. Data conversion is used to guarantee the quality and availability of data. For example, for some predictive models, the data need to be discrete and normalized from continuous data.

Step Model Development/ Pattern Discovery

Many different types of data mining technology and modeling techniques, such as classification, clustering, regression, neural networks, and decision trees are employed to highlight previously hidden relationships amongst the data.

a) Classification analysis: discovery of a predictive learning process that classifies important and relevant information into one of several predefined classes.
b) Clustering analysis: Cluster analysis is a set of methodologies for automatic classification of samples into a number of groups using a measure of association, so that the samples in one group are similar and samples belonging to different groups are not similar [17].. In the process, clustering analysis is to identify a finite group of clusters to describe the entire data.
c) Time series analysis and forecasting: Time series analysis, similar to correlation analysis, whose purpose is to find the inner relation between data values, comprises methods that attempt to understand the underlying context of the data, measured typically at successive times, or to make predictions.

Step Model Validation

Once the model is built using the training data sample, the model is subjected to validation by independent datasets, which is an important requirement in data mining. This step involves evaluating the model and deciding if the model was successful both in determining which variables are the predictors, and the degree of confidence in selecting the variable the model demonstrates. If multiple models were developed, the models are evaluated together and the best model is typically chosen.

Step Interpretation and Decision Making

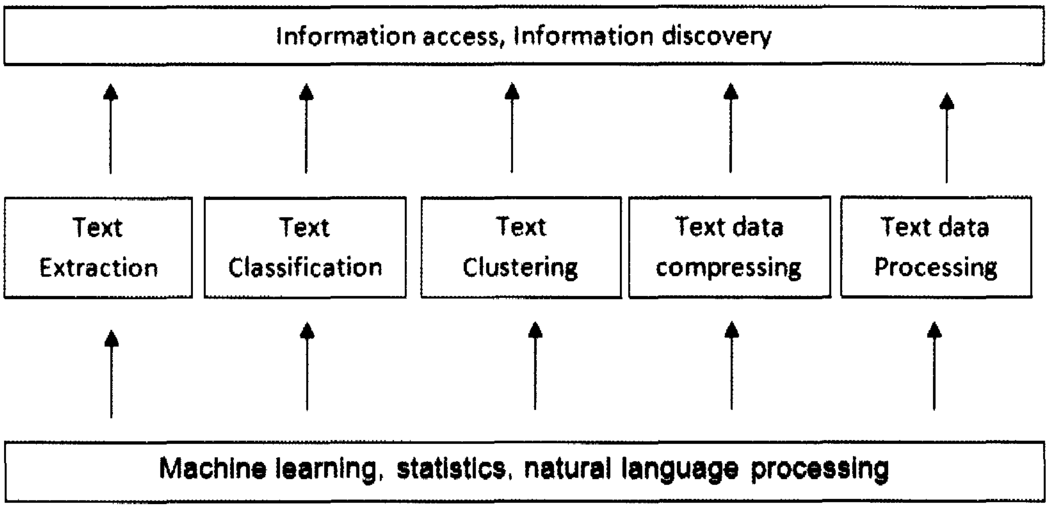
Data mining models are used to help in decision making. Usually, the simple models are more interpretable; the results of the model are used to develop a series of "rules" used for decision-makingand specific techniques are needed to validate the results to interpret the high-dimensional models.

3.3. Text Mining

Text Mining is a technology for discovering and extracting information from a wide variety of unstructured text documents in a collection by uncovering any themes and concepts that are contained in the collection [18]. Text Mining consists of many technologies such as information retrieval and information extraction, natural language processing and data mining technology. It is a multidisciplinary mixed area. It cooperates with information technology, text analysis, statistics, pattern recognition, data visualization, database technology and machine learning.

Text mining can give a general understanding of the documents as shown in Figure 3.2. It has the following components: the bottom part represents the base areas of Text mining, including machine learning, mathematical statistics, and natural language processing. The middle part introduces the basic text mining technologies of text extraction, classification, clustering, text data compressing and processing. Based on those two parts, the top one is the application for text mining, including information access and discovery. Information access includes information retrieval, filtering, and reporting. Information discovery includes data analysis and data prediction.

Figure 3.2 The concept and technologies of Text Mining [18].


Information access, Information discovery
Text Extraction
Text Classification
Text Clustering
Text data compressing
Text data Processing
Machine learning, statistics, natural language processing

In order to apply text mining methods, text needs to be broken into components such as words, phrases, multiword terms, entities, punctuation marks, and terms in foreign languages. Here, a multiword term is defined as a group of words to be processed as a single term and entities include items such as names, addresses, companies, and measurements [19].

3.4 Dimension Reduction

Because text mining is always applied to extract useful information from unstructured text documents, which are multivariate, it is difficult and time-consuming to use the entire data (including all the terms). A term-document frequency matrix is generated by paring the document collection. The rows and columns represent the terms and documents, respectively. Each entry of the matrix is the number of times that a term appears in a document. Dimension reduction will be necessary and applied to improve the model performance and efficiency with high dimensional data. Here, we introduce singular value decomposition, which will be used in this paper.

Singular Value Decomposition (SVD) is a proven mathematical method that defines concepts by projecting each document into a reduced dimensional space. The more similar documents will be closer in the reduced space; on the other hand, the more dissimilar documents will be farther in the reduced space. This is the main idea that SVD explains. The following is the mathematical definition for SVD.

The SVD of matrix G, the sparse term-by-document frequency matrix having n documents and p terms, is defined by the equation [20].

$$G = U_{n \times n} \Sigma_{n \times p} V^{T}_{p \times p},$$

where the columns of U are orthogonal eigenvectors of GG^{T}, the columns of V are orthogonal eigenvectors of $G^{T}G$, and Σ is a diagonal matrix of singular values, which are the square roots of the eigenvalues of GG^{T}. For SVD reduction, the matrix G consists of columns of V that correspond to the largest singular values. Rows of U are the number of terms and rows of V are the number of documents [19].

To get the SVD values, we need to calculate V^{T} and Σ by using $G^{T}G$.

$$G^{T}G = V\Sigma^{2}V^{T}$$

and

$$U = GV\Sigma^{-1}$$

3.5 Clustering

In order to use the clustering analysis, the documents need to be represented and reduced as concepts in multidimensional space. In classification, the category structure is known, and in cluster analysis, the category structure is unknown. The objective in this case is to discover a category structure based on the data (observations).

There are two clustering techniques in Text Mining. One method uses a hierarchical clustering algorithm where each document is placed in a specific sub-tree. The other method uses the expectation-maximization (EM) algorithm, which approximates the observed distributions of values based on the combination of different distributions of different clusters.

- The expectation-maximization (EM) algorithm for clustering [19]

Assume there are *k* clusters with density functions f_i, $i = 1,2,....k$, and there are n variables in the data set. Then the mixture model probability density function at point x is:

$$P(x) = \sum_{i=1}^{k} w_i f_i(x \mid u_i, \Sigma_i)$$

where w_i is the proportion of data that belong to cluster i, and u_i and Σ_i are the mean vector and covariance matrix for cluster i.

For each cluster, the n dimensional Gaussian distribution is

$$f_i(x \mid u_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^d \mid \Sigma_i \mid}} \exp(-\frac{1}{2}(x - u_i)^T (\Sigma_i)^{-1}(x - u_i))$$

The expectation-maximization clustering is a process which has the following basic steps:

1: Obtain initial parameter estimates.

2: Apply the standard EM algorithm to find new clusters.

3: Update parameter estimates.

4. Repeat step 2 and 3 until the cluster membership stabilizes.

For each observation x in the data set at iteration j, the parameter estimates of the standard EM algorithm are computed as follows [19].

1. Compute the membership probability of x in each cluster $h = 1,...,k$.

$$w_h^j(x) = \frac{w_h^j f_h(x \mid u_h^j, \sum_h^j)}{\sum w_i^j f_i(x \mid u_i^j, \sum_i^j)}$$

2. Update mixture model parameters for each cluster $h = 1,...,k$.

$$w_h^{j+1} = \sum w_h^j(x)$$

$$u_h^{j+1} = \frac{\sum_x w_h^j(x)x}{\sum_x w_h^j(x)} \sum\nolimits_h^{j+1} = \frac{\sum_x w_h^j(x)(x - u_h^{j+1})(x - u_h^{j+1})^T}{\sum_x w_h^j(x)}$$

The iterative process stops if

$$\left|L(\varphi^{j}) - L(\varphi^{j+1})\right| \leq \varepsilon$$

where $\varepsilon > 0$, and

$$L(\varphi) = \sum_{x} \log\left[\sum_{h=1}^{k} w_h f_h(x \mid u_h, \textstyle\sum_h \;)\right].$$

The EM clustering algorithm computes the probabilities of clusters according to the probability distributions, instead of maximizing the difference in mean for the variables by assigning observations to clusters. And the EM algorithm could be used for categorical or continuous variables.

3.6 Summary

Data mining is a powerful technology to extract information from the huge potential of unknown and potentially useful information from large databases. With this strong analysis technique, the original raw data are changed into valuable information, which may provide a competitive advantage for the decision-makers.

Text Mining uses unstructured textual information and examines it in an attempt to discover structure and implicit meanings hidden within the text. SAS Text Miner provides a rich suite of tools for discovering and extracting knowledge from text documents. It transforms textual data into a usable, intelligible format that facilitates classifying documents, finding explicit relationships or associations between documents, and clustering documents into categories.

CHAPTER IV SURVIVAL ANALYSIS

4.1. The History of Survival Analysis

The origin of survival analysis can be traced back to several centuries ago. Based on the interest in reliability (or failure time) of military equipment, survival analysis developed quickly. At the end of World War II, these newly developed statistical methods quickly spread to industry from strict mortality data research to failure time research. With the development of survival analysis, nonparametric or semi-parametric methods took the place of parametric models in dealing with the clinical trials in medical research [21].

Next, we introduce several distribution functions in the following sections, which are the fundamental tools of survival analysis [22].

4.2. The Cumulative Distribution Function [23]

Suppose a non-negative random variable T denotes the lifetimes of individuals in some population. The *c.d.f* of T, $F_T(t)$, is defined by

$$F_T(t) = P_T(T \leq t).$$

For example, if T represents the age of first occurrence of a chronic disease, then F_T *(t) is* the distribution of age for the disease. In survival analysis, the complementary function is more commonly used, which is denoted as the survival distribution function discussed in the following sections.

4.3. The Probability Density Function

The *p.d.f* of T, $f_T(t)$, is defined by

$$f_T(t) = \frac{d(F_T(t))}{dt}.$$

Here, $f_T(t)$ is the absolute instantaneous rate of death (disease).

4.4. The Survival Function

The Survival function is the complementary function of the cumulative distribution function. Let T ≥ 0 have a *p.d.f* $f_T(t)$ and *c.d.f* $F_T(t)$. Then, the survival function is denoted by the following form:

$$S_T(t) = P\{T > t\} = 1 - F_T(t) = \int^{\infty} f_T(x)dx.$$

Conversely, the *pdf* can be expressed as

$$f_T(t) = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t)}{\Delta t} = \frac{dF_T(t)}{dt} = -\frac{dS_T(t)}{dt}.$$

4.5. The Hazard Function

The hazard function $h_T(t)$, which is the relative failure (death) rate at time t is given by the following [24]:

$$h_T(t) = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

$$= \frac{f_T(t)}{1 - F_T(t)}$$

$$= \frac{f_T(t)}{S_T(t)}$$

It is easily verified that $h_T(t)$ specifies the distribution of T, since

$$h_T(t) = -\frac{dS_T(t)/dt}{S_T(t)} = -\frac{d\log(S_T(t))}{dt}.$$

Integrating $h_T(u)$ over (0,t) gives the cumulative hazard function $H_T(t)$:

$$H_T(t) = \int_0^t h_T(u)du = -\log(S_T(t)),$$

which is the inverse function of the exponential function. Thus,

$$S_T(t) = \exp(-H_T(t)) = \exp(\ -\int_0^t h_T(u)du\ ).$$

Hence, the *p.d.f* of T can be expressed as

$$f_T(t) = h_T(t)\exp(\ -\int_0^t h_T(u)du\ ).$$

Note that $H\ (\infty\ ) = \int_0^\infty h_T(t)dt = \infty$.

The relationship is clear between these functions. The *p.d.f* is the derivative of the *c.d.f*, and the *c.d.f* is the integral of the *p.d.f.* The survival function is simply 1 minus the *c.d.f,* and the hazard function is simply the *p.d.f* divided by the survival function [25].

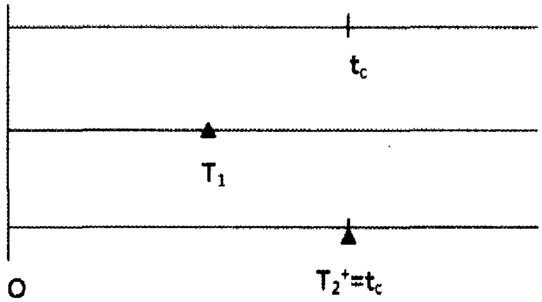
4.6. Censoring

In order to apply survival analysis, the beginning and the end of the study need to be carefully defined. For example, for a complete observation in our lung cancer study, the survival time may begin on the day a patient is diagnosed with lung cancer and end when that patient dies because of the cancer. Here, this patient is an uncensored subject, where the event occurs during the time period of observation. In a data set, some observations end because of the occurrence of the event, and others have no event by the end of the observation period, which would be called censored observations. For example, the patient might be still alive after five years from lung cancer and do not know when he died or died due to an unrelated cause. Therefore, we need some techniques to deal with these observations due to different reasons.

Right censoring is the most common method for dealing with incomplete data. A right censored subject is one that is no longer observed before the event occurrence. It allows subjects to contribute to the model until they are no longer able to contribute (end of the study, or withdrawal), or they have an event. Figure 4.1 gives us the basic information about the right censoring method.

Let $T_1, T_2, ..., T_n$ be independent and identically distributed n subjects with distribution function $F_T(t)$ in a study. Notice that some subjects have events early during the study period, some have events at the end of the study period and others have no event during the entire period. For example, for the patients with lung cancer, some died during the observation period and most of them are simply right censored at the end. Therefore, we terminate the study at a pre-specified time t_c (for our case, the end of two years).

Figure 4.1 Right censoring [24].



Let t_c be denoted as the fixed censoring time [24]. Instead of observing the T_i, we observe $Y_1, Y_2, ..., Y_n$ where

$$Y_i = \min(T_i, t_c) = \begin{cases} T_i & if\ T_i \le t_c \\ t_c & if\ t_c < T_i \end{cases}.$$

A binary random variable *δ* is used to denote if an event time is observed or censored,

$$\delta = \begin{cases} 1 & if\, T \le t_c \\ 0 & if\, t_c < T \end{cases}.$$

Note that (*δ* =0 and $T \le t_c$) implies that the event time was precisely $T = t_c$, which occurs with zero probability if *T* is a continuous variable.

With maximum likelihood estimation, the joint likelihood of the pair (Y ,δ) could be calculated as follows. For $y < t_c$,

$$P(Y \le y) = P(T \le y) = F(y),$$

and

$$P(\delta = 1 \mid Y \le y) = 1.$$

Therefore, the likelihood for Y=$y < t_c$ and δ =1 is the density $f(y)$. For $y = t_c$ and δ =0, the likelihood for this event is the probability

$$P(\delta = 0,\ y = t_c) = P(T > t_c) = S(t_c).$$

Hence, the likelihood function for the n independent and identically distributed random pairs (Y_i,δ_i) is given by [24],

$$L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(t_c)^{1-\delta_i} .$$

4.7. The Kaplan–Meier method

The Survival curve is the graph of the survival function *S(t)* based on time *t.* The Kaplan–Meier method can be used to estimate this curve from the observed survival times without the assumption of an underlying probability distribution [25].

Let t_i denote an ordered observed value. The empirical survivor function (*esf*), denoted by $S_n(t)$, is defined to be

$$S_n(t) = \frac{\#of \quad observation > t}{n} = \frac{\#\{t_i > t\}}{n}$$

For each of *n* individuals, the pair (Y_i, δ_i) is denoted as follows

$$Y_i = \min(T_i, C_i)$$

and

$$\delta_i = \begin{cases} 1 & if \ T_i \leq C_i \\ 0 & if \ C_i < T_i \end{cases} .$$

The people at risk of an event at the beginning of the interval y_i are those people who survive (no event occurred) the previous interval y_{i-1} .

Let $R(t)$ denote the risk set just before time t and let

n_i = # in $R(y_{(i)})$

= *# no event (and not censored) just before* $y_{(i)}$

d_i =# *of event occurred at time* $y_{(i)}$

p_i = *P*(no event through y_i | no event at beginning y_i)

= $P(T > y_{(i)} \mid T > y_{(i-1)})$

$q_i = 1 - p_i$

=*P(event occurred in* y_i | alive at the beginning y_i *).*

Recall the general multiplication rule for joint events A_1 and A_2:

$$P(A_1 \cap A_2) = P(A_2 \mid A_1)P(A_1).$$

The survival function can be expressed as

$$S(t) = P(T > t)$$

$$= \prod_{y_{(i)}} p_i \quad .$$

Here, p_1 is the proportion surviving the first period, p_2 is the proportion surviving beyond the second period conditional on having survived up to the second period, and so on.

The estimates of p_i and q_i are

$$\hat{q}_i = \frac{d_i}{n_i}$$

and

$$\begin{aligned}\hat{p}_i &= 1-\hat{q}_i \\ &= 1-\frac{d_i}{n_i} \\ &= \frac{n_i-d_i}{n_i}\end{aligned}$$

The K-M estimator of the survivor function is

$$\begin{aligned}\hat{S}(t) &= \prod_{y_{(i)}} \hat{p}_i \\ &= \prod_{y_{(i)}} (\frac{n_i-d_i}{n_i}) \\ &= \prod_{i=1}^{k} (\frac{n_i-d_i}{n_i})\end{aligned}$$

where $y_{(k)} \leq t < y_{(k+1)}$.

4.8. Cox's proportional hazards model (Cox regression) [24][26]

The Cox proportional hazard model is the most widely used method in survival analysis, especially in medical or clinical studies. It is an extension of the logistic regression model with conditions. For example, the baseline hazard function is equivalent to the intercept in the logistic regression.

For the Cox proportional hazards model, the hazard function is

$$h(t \mid X) = h_0(t) \cdot e^{\beta X},$$

where $h_0(t)$ is an unspecified baseline hazard function, which is estimated at the mean values of the variables. Here

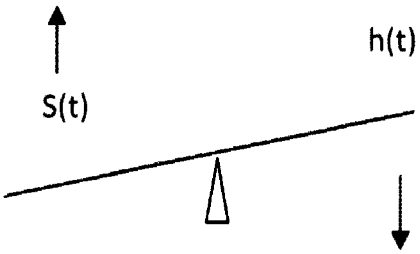
$$\beta X = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p .$$

Then we have

$$\log(h(t \mid X)) = \log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

One of the advantages for the Cox hazard model is the non-parametric baseline function $h_0(t)$. Here, e^{β_i} is the hazards ratio, which is assumed constant with respect to time *t,* which is similar to the odds ratio in logistic regression.

Figure 4.2. The relationship between hazard and survival [24]



If the hazard ratio is less than one, then the ratio of corresponding survival probabilities is larger than one. Hence, the treatment group has a larger probability of survival at any given time *t*, after adjusting for the other covariates.

For any (Proportion Hazard) PH model, the survival function at time t, given a set of predictors X is

$$
\begin{aligned}
S(t \mid X) &= \exp(-\int_0^t h(u \mid X)\,du) \\
&= \exp(-\exp(\beta X)\int_0^t h_0(u)\,du) \\
&= (\exp(-\int_0^t h_0(u)\,du))^{\exp(\beta X)} \\
&=(S_0(t))^{\exp(\beta X)}
\end{aligned}
$$

where $S_0(t)$ denotes the baseline survivor function. The p.d.f. at time t, given a set of predictors X is

$$f(t \mid X) = h_0(t)\exp(\beta X)\,(S_0(t))^{\exp(\beta X)}.$$

In Cox's model, there is no assumption about the distribution of the hazard except that it is assumed that the hazard ratio does not change over time.

4.9. Summary

The survival curve describes the relationship between the probability of survival and time. The Kaplan- Meier method is used to estimate the survival curve from all of the observations available. Cox's proportional hazards modeling gives us the hazard ratio, which explains the risk of the event for a certain variable. No assumptions of parametric distribution for the survival time make Cox regression more attractive. Using the partial likelihood function makes the Cox regression model more flexible to examine the dependent variables.

Survival analysis is appropriate for outcomes that occur during follow-up of patients. The outcome may be death or another event, such as the recurrence of disease in cancer, or a complication after implantation of a heart valve.

We will be applying Kaplan–Meier method and Cox regression model to examine the relationship between the patient demographic factors, the diagnosis, treatment procedures and mortality of lung cancer.

CHAPTER V TIME SERIES MODELS AND PREDICTIVE MODELS

5.1. Time Series Models

A Time series is a set of measurements for a sequence of random events according to the time. It is a prediction method using historical time series to forecast the future value by statistical analysis and a mathematical model. The ordinary regression model is the foundation of time series analysis. The independent variables include the lag value from itself and other dependent variables, including their lag values.

There are three basic models for Time Series Models: the autoregressive model (AR), moving average model (MA) and autoregressive moving average model (ARMA). When regular differencing is applied, The Autoregressive Integrated Moving Average Model (ARIMA) is a combined model together with AR and MA [27].

Let $\{Y_t, t = 0, \pm 1, \pm 2, ...\}$ be a time series and $\{\varepsilon_t, t = 0, \pm 1, \pm 2, ...\}$ be the error (white noise) with mean zero and variance σ^2,

$$E(\varepsilon_t) = 0 \qquad (5.1)$$

$$E(\varepsilon_t^2) = \sigma^2 \qquad (5.2)$$

with the ε's uncorrelated with time,

$$E(\varepsilon_i, \varepsilon_j) = 0 \text{ if } i \neq j \quad (5.3)$$

The jth autocovariance of Y_t is

$$\gamma_{jt} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(y_t - u_t)(y_{t-j} - u_{t-j})f_{Y_t,Y_{t-1},\ldots,Y_{t-j}}(y_t, y_{t-1},\ldots y_{t-j})dy_t dy_{t-1}\ldots dy_{t-j}$$

$$= E(Y_t - u_t)(Y_{t-j} - u_{t-j}) \qquad (5.4)$$

where u_t is the mean of Y_t and $f_{Y_t,Y_{t-1},\ldots Y_{t-j}}$ is the joint density function of Y_t, $Y_{t-1},\ldots Y_{t-j}$.

We also define the backward shift operator B here by $BY_t = Y_{t-1}$. We also have

$$B^k Y_t = Y_{t-k}. \qquad (5.5)$$

The backward difference operator ∇, is defined by

$$\nabla Y_t = Y_t - Y_{t-1} = (1-B)Y_t. \qquad (5.6)$$

5.1.1 Autoregressive Model

The basic assumption for time series models is that the data have an internal structure, such as autocorrelation, trend or seasonal variation, which is the forecasting methods' purpose. The *AutoRegressive* (AR) [28] model can be used to forecast future values. A *p*th-order autoregressive, AR(*p*), model is defined as

$$Y_t = a_t + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t \qquad (5.7)$$

where ε_t is the prediction error, and $\phi_1,\ldots,\phi_p$ are the unknown autoregressive coefficients. The order of the model tells how many lagged past values are included. The simplest AR model is the first-order *AR(1)*.

Here, *AR(p)* could be written as the following:

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) Y_t = \varepsilon_t$$

$$\phi(B) Y_t = \varepsilon_t \qquad (5.8)$$

An *AR(p)* process is covariance stationary if and only if all roots of its *characteristic polynomial*:

$$1 - \phi_1 y - \phi_2 y^2 - \phi_3 y^3 - \ldots - \phi_p y^p = 0 \qquad (5.9)$$

lie outside the unit circle. If *y = 1* is a solution of the characteristic polynomial (5.6), then the process has a unit root. The presence of a unit root causes the autocovariances to vary over time. Therefore, for any models with autoregression or other stationary methods, the data need to be differenced at the first lag. The autoregression model can also be viewed as a special case of the multiple regression models, where the independent or predictor variables are the past values of the process itself.

5.1.2. Autoregressive Moving Average Model

The *moving average* (MA) is a model in which the time series is regarded as a moving average (unevenly weighted) of a random shock series ε_t. The moving average model of order *q,* or *MA(q),* is given by

$$Y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \dots - \theta_q\varepsilon_{t-q}$$

$$= (1 - \theta_1 B - \dots - \theta_q B^q)\varepsilon_t$$

$$= \theta(B)\varepsilon_t. \quad (5.10)$$

where $\{\varepsilon_t\}$ satisfies (5.1), (5.2) and (5.3) and θ_1, θ_2,..., θ_q are the unknown coefficients.

If the parameters of moving average satisfy some certain conditions such that the model is invertible, there exists a duality between the moving average process and the autoregressive process, which is that the moving average model can be rewritten into an autoregressive form.

Moreover, if combined with AR models, the moving average model and autoregressive model form a very powerful tool, the autoregressive moving average (ARMA) model. In the ARMA model, the current value of the time series Y_t is expressed linearly in terms of its past values and previous values of the white noise (error).

The *autoregressive moving average* model of order (p,q), or ARMA (p,q), is written as

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} \quad (5.11)$$

where the coefficients are the same as in (5.7) and (5.10).

or, ARMA (p,q) could be written as in lag operator form:

$$(1-\phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)Y_t = (1-\theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)\varepsilon_t;$$

that is,

$$\phi(B)Y_t = \theta(B)\varepsilon_t.$$

5.1.3. Autoregressive Integrated Moving Average Model

In the 1970s, Box and Jenkins introduced a general model, *autoregressive integrated moving average* (ARIMA), which contains three parts, autoregressive, moving average, and differencing. Specifically, it has three types of parameters: the autoregressive parameters ($\phi_1,\ldots,\phi_p$), the number of differencing passes at lag 1 (*d*), and the moving average parameters $(\theta_1,\ldots,\theta_q)$. In the notation introduced by Box and Jenkins, a series that needs to be differenced *d* times at lag 1 and afterwards has orders *p* and *q* of the AR and MA components, respectively, is denoted by ARIMA(p,d,q) and can be conveniently written as

$$\phi(B)\nabla^d Y_t = \theta(B)\varepsilon_t \qquad (5.12)$$

where the ∇ is backward difference operator, and B is the backward shift operator; that is, $B^h x_t \equiv x_{t-h}$. Note that ARIMA(*p*, 0, *q*) is simply an ARMA(*p*, *q*) process.

5.1.4. Model Identification

Time series modeling of ARMA and related models proceed by a series of well-defined steps. The first step of this process is model identification to specify the appropriate structure (AR, MA, ARMA or ARIMA) and the order of the model.

The *sample autocorrelation function* (ACF) and the *partial autocorrelation function* (PACF) plots or a goodness-of-fit statistic or information criterion are used to identify or select the best model for prediction. In general, increasing the complexity of the model structure, for example, increasing the number of variables, just provides an artificial improvement of fitness.

5.2 Predictive Models

Multiple linear regression modeling and Logistic regression modeling are all regression methods. The difference between them is that the former is designed to predict an interval-valued target variable and the latter is designed to predict a categorically-valued target variable. Regression modeling is used to examine the relationships between the input variables (independent variables) and the target variable (dependent variable) and to determine the best predictors for the target. In other words, the objective of regression analysis is to determine the optimal balance of choosing the model with the smallest error and the fewest number of parameters [29].

5.2.1. Logistic Regression Modeling

Logistic regression models are used to predict the response variable by redefining the variable through the use of an indicator variable. Since the predicted value is used to estimate the probability of the target event and will not result in the desired interval in predicting the probabilities that must be between the interval of zero and one, the linear regression should not be applied to the categorical target variable[30][31].

The logistic regression model is a widely used statistical technique for binary medical outcomes. The model is flexible in that it can incorporate categorical and continuous predictors, non-linear transformations, and interaction terms. Many of the principles of traditional regression also apply for logistic regression.

In order to obtain the best predictive performance, it is important that the model be correctly specified. Correct specification of the model involves selecting the correct set of input variables in the model and the appropriate error distribution that matches the range of values of the target variable. However, in logistic regression modeling, it is also extremely important to select the correct link function that conforms to the range of values of the target variable.

For k explanatory variables and $i = 1,2,...,n$ observations, the model is

$$\text{Log}[\frac{p_i}{1-p_i}] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \qquad (5.15)$$

where p_i is the probability for $y_i = 1$. Here, *log* is the natural logorithm, α is the intercept, and β_i are the estimated regression coefficients.

Unlike the usual linear regression model, there is no random white noise or error in the equation for the Logistic model, which does not mean that the model is deterministic because of the probabilistic relationship between p_i and y_i.

We can solve the *logit* equation for p_i,

$$p_i = \frac{\exp(\alpha + \beta_1 x_{i1} + ... + \beta_k x_{ik})}{1 + \exp(\alpha + \beta_1 x_{i1} + ... + \beta_k x_{ik})}$$

$$= \frac{1}{1 + \exp(-\alpha - \beta_1 x_{i1} - ... - \beta_k x_{ik})}. \qquad (5.16)$$

5.2.2. Decision Tree

The goal of decision tree modeling is to identify various target groups based on the values from a set of input variables to provide insight into the decision process. Each split is performed from the values of one of the input variables that best partitions the target values. For categorical target variables, the model is called a classification tree, in which the leaves that contain the target proportions can be interpreted as predicted probabilities or predicted proportions of the categorically-valued target variable. For interval-valued target variables, the model is called a regression tree, where the leaves that contain the target means can be interpreted as predicted values of the target variable[32].

The Standard Decision Tree Algorithms

The AID, CHAID and CART methods are the most widely used algorithms for decision tree modeling [33].

Automatic Interaction Detection (AID) was used to fit trees to predict a quantitative variable. Stepwise splitting is the foundation of the algorithm. It begins with a single cluster and splits this cluster into two or more clusters based on some defined rules. Here, each variable is examined for splitting as follows [34] :

a) Sort all the *n* observations on the variable and examine all *n*-1 ways to split the cluster into two or more groups.

b) Calculate the sum of squares about the mean of the cluster on the dependent variable for each possible split.

c) Choose the best of the *n*-1 splits to represent the variable's contribution and repeat for all the variables.

The CHi-squared Automatic Interaction Detector (CHAID) algorithm is only for categorically-valued input variables, and it cannot be used for ordinal-valued input variables.

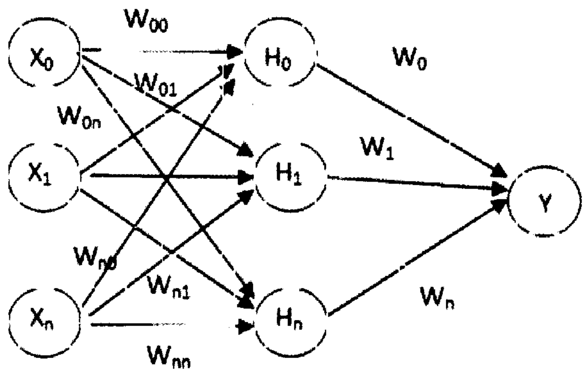
The Classification and Regression Tree (CART) method is based on statistically optimal splitting of the observations into pairs of smaller subgroups. The Gini index is used to measure the probability of a randomly selected element would be incorrectly labeled if it was randomly labeled based on the distribution of labels. Here CART uses the Gini index to measure the impurity at a node, and then choose the split to maximize the reduction in impurity [32].

5.2.3. Neural Network

There are many applications of neural networks to industrial applications, including predictions of occurrence or cost and target classification. In general, neural networks have several layers with interconnected nodes where each node is determined by a non-linear function of the inputs.

There are three types of nodes. The following Figure 5.1 shows the basic information of the *feedforward* neural network. The input unit represents the input variables, where each input variable has its own weights in the input layer. The hidden units perform an internal, nonlinear transformation and the output units generate the predicted values, and then compute the error that is the difference between the predicted values and the values of the output units in the output layer.

Figure 5.1 The diagram for multilayer feed-forward network.



The general Neural Network model is as following:

$$Y = f(H_0W_0 + H_1W_1 + ... + H_nW_n + \varepsilon) \qquad (5.17)$$

where

$$H_0 = f_0(w_{00}X_0 + w_{10}X_1 + ... + w_{n0}X_n),$$

$$...,$$

$$H_n = f_n(w_{0n}X_0 + w_{1n}X_1 + ... + w_{nn}X_n)$$

There are several general kinds of transfer functions commonly used:

- Identity Function does not change the value of the inputs, which has the same range as the inputs.
- Sigmoid Functions are S-shaped functions such as the logistic and hyperbolic tangent functions.
- Softmax Function is a multiple logistic function.
- Value Functions are bounded bell-shaped functions such as the Gaussian function.
- Exponential Functions.

In this paper, the transfer function is given by

$$f(x) = \frac{1}{1+e^{-x}}.$$

A big drawback to neural network modeling is that it is very difficult to determine the contribution of the input variables to the target variable in the model. Unlike traditional regression parameter estimates, the weight estimates do not indicate the effect, magnitude or the rate of change in the relationship between the target variable and the input variables.

5.3 Summary

Time series analysis is a prediction method using historical time series to forecast the future value by statistical analysis and mathematical models, which determines if the data taken over time has an internal structure (such as autocorrelation, trend or seasonal variation). A predictive model is made up of a number of predictors, which are variable factors that are likely to influence future behavior or results. These models can be used to help physicians and health policy makers in their decision-making on the screening and treatment of disease in high-risk patient groups.

CHAPTER VI DATA ANALYSIS FOR NIS DATABASE

6.1 Overview of NIS Data

The Nationwide Inpatient Sample (NIS) is a powerful database of hospital inpatient stays, which includes healthcare cost and utilization. It could be used to identify or analyze national trends in health care cost, utilization, quality, and outcomes by researchers or decision-makers.

The NIS contains clinical and resource use information included in a typical discharge abstract, with safeguards to protect the privacy of individual patients, physicians, and hospitals (as required by data sources). The NIS can be weighted to produce national estimates. The new version of NIS contains severity adjustment data elements, such as APR-DRGs, APS-DRGs, Disease Staging, and AHRQ Co-morbidity Indicators. The Diagnosis and Procedure Groups Files are also added to the 2005 version of NIS. Access to the NIS is open to users who sign data use agreements. Uses are limited to research and aggregate statistical reporting.

The NIS is a uniform, multi-state database that promotes comparative studies of health care services and will support health care policy research on a variety of topics including:

- Use and cost of hospital services
- Medical practice variation
- Health care cost inflation
- Hospital financial distress
- Analyses of States and communities
- Medical treatment effectiveness
- Quality of care
- Impact of health policy changes
- Access to care
- Diffusion of medical technology
- Utilization of health services by special populations.

There is much clinical and nonclinical information for each hospital stay in NIS, such as:

- Primary and secondary diagnoses
- Primary and secondary procedures
- Admission and discharge status

- Patient demographics (e.g., gender, age, race, median income for ZIP Code)
- Expected payment source
- Total charges
- Length of stay
- Hospital characteristics (e.g., ownership, size, teaching status).

Some of the variables we will work with in the NIS data include patient demographics:

- Age (in years)
- Female (0=male, 1=female)
- Race(1=White, 2=Black, 3=Hispanic, 4=Asian/Pacific Islander, 5=Native American, 6=Other)
- DRG
- Patient diagnoses in ICD9 codes (DX1-DX15, fifteen columns)
- Patient procedures in ICD9 codes (PR1-PR15, fifteen columns)
- TOTCHG (Total Charges)
- LOS (Length of Stay)

In order to work with these variables, there are some preprocessing issues, especially to work with 15 columns of diagnosis and procedure codes.

6.2 Data preprocessing

The lung cancer data are from the NIS, and we had five years of data, 2000 to 2004. Here, we first put all the five years of data into one data set and create a binary variable to label lung cancer according to diagnosis codes. In order to simplify the process of discovery, we first concatenate all 15 columns of variables into one text string using the CATX statement. This code put all possible diagnosis codes into one text string, and defined a second string containing all possible procedure codes using the CATX statement.

To find those patients with Lung Cancer, the RXMATCH function was used. The RXMATCH looked for the initial code of ‘162’ that found all patients with a diagnosis code related to lung disease. Because ‘162’ can occur in other codes that are not related to lung cancer, such as ‘216.2’, we use four digits of code rather than three to avoid catching ‘216.2’. The code used was the following:

```
data nis.lungcancer_nis_00to04;
set nis. nis_00to04;
lungcancer=0;
diagnoses=catx(' ',dx1, dx2, dx3, dx4, dx5, dx6, dx7, dx8, dx9,
dx10, dx11, dx12, dx13, dx14, dx15);
procedures=catx(' ',pr1, pr2, pr3, pr4, pr5, pr6, pr7, pr8, pr9,
pr10, pr11, pr12, pr13, pr14, pr15) ;
if (rxmatch('1620',diagnoses)>0) then lungcancer=1;
if (rxmatch('1621',diagnoses)>0) then lungcancer=1;
if (rxmatch('1622',diagnoses)>0) then lungcancer=1;
if (rxmatch('1622',diagnoses)>0) then lungcancer=1;
if (rxmatch('1623',diagnoses)>0) then lungcancer=1;
if (rxmatch('1624',diagnoses)>0) then lungcancer=1;
if (rxmatch('1625',diagnoses)>0) then lungcancer=1;
if (rxmatch('1626',diagnoses)>0) then lungcancer=1;
if (rxmatch('1627',diagnoses)>0) then lungcancer=1;
if (rxmatch('1628',diagnoses)>0) then lungcancer=1;
if (rxmatch('1629',diagnoses)>0) then lungcancer=1;
run;
```

Recall that the ICD 9 code of 162 means malignant neoplasm of trachea, bronchus, and lung. Below are all of the 4-digit codes associated with lung diseases:

162.0 Trachea (Cartilage of trachea, Mucosa of trachea).

162.2 Main bronchus (Carina, Hilus of lung).

162.3 Upper lobe, bronchus or lung.

162.4 Middle lobe, bronchus or lung.

162.5 Lower lobe, bronchus or lung.

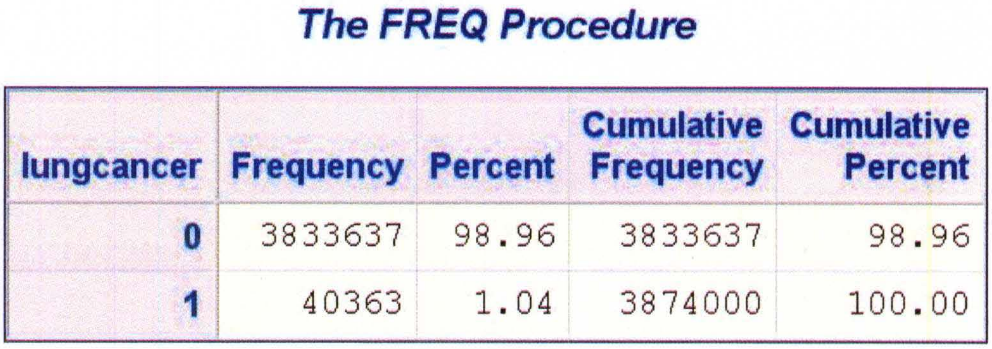
162.8 Other parts of bronchus or lung (Malignant neoplasm of contiguous or overlapping sites of bronchus or lung whose point of origin cannot be determined).

162.9 Bronchus and lung, unspecified.

6.3. Data Visualization

There are a total of 40,363 observations for the five years of data related to lung disease out of 3,833,637 records. Note that approximately 1.04% of the inpatient population has a diagnosis of lung disease (shown in Table 6.1 and Figure 6.1). It is clear that lung cancer is a small sample of patients compared with the total size of the data set.

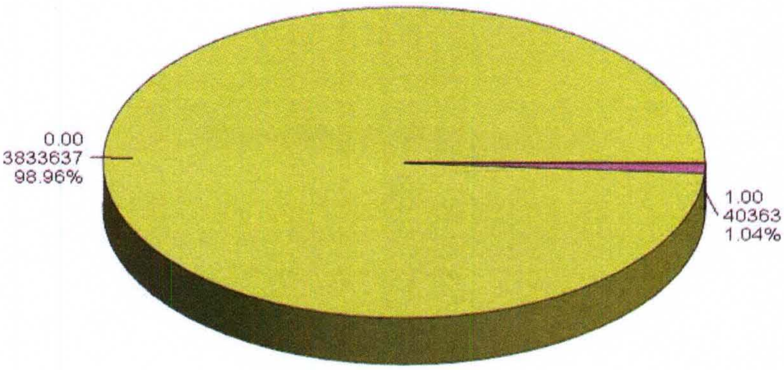
Table 6.1 The frequency of lung cancer.



The FREQ Procedure

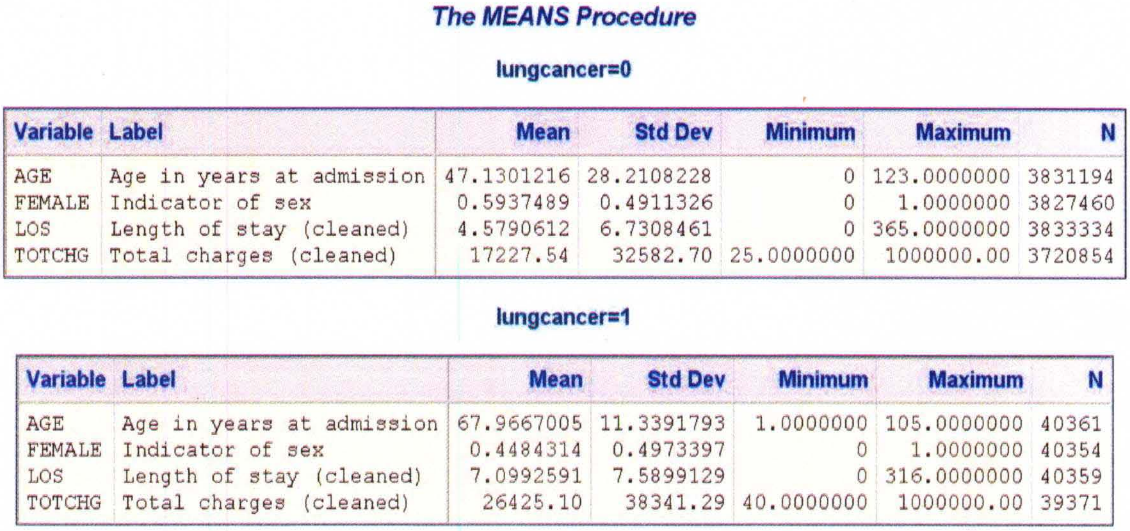
lungcancer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	3833637	98.96	3833637	98.96
1	40363	1.04	3874000	100.00

Figure 6.1 The pie chart for the proportion of lung cancer



The data summary of some of the variables is given in Table 6.2. Note that the average age for a patient with lung cancer is about 68, about 21 years more than the average age for those without lung cancer. Males have a higher probability of having lung cancer compared with females. The patients with lung cancer have a higher probability of staying about 7 days in the hospital compared to those without lung cancer. Obviously, they also have higher costs.

Table 6.2 Summary of Age, Gender, Length of Stay and Total Charge



The MEANS Procedure

lungcancer=0

Variable	Label	Mean	Std Dev	Minimum	Maximum	N
AGE	Age in years at admission	47.1301216	28.2108228	0	123.0000000	3831194
FEMALE	Indicator of sex	0.5937489	0.4911326	0	1.0000000	3827460
LOS	Length of stay (cleaned)	4.5790612	6.7308461	0	365.0000000	3833334
TOTCHG	Total charges (cleaned)	17227.54	32582.70	25.0000000	1000000.00	3720854

lungcancer=1

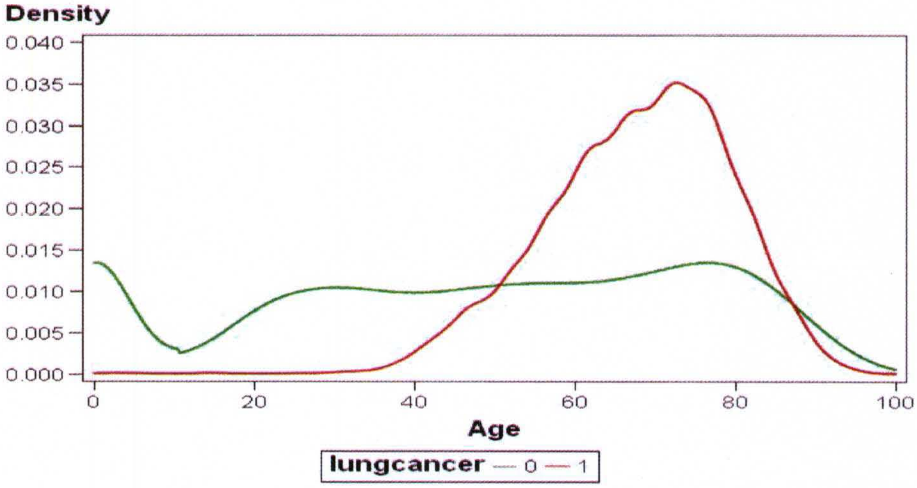
Variable	Label	Mean	Std Dev	Minimum	Maximum	N
AGE	Age in years at admission	67.9667005	11.3391793	1.0000000	105.0000000	40361
FEMALE	Indicator of sex	0.4484314	0.4973397	0	1.0000000	40354
LOS	Length of stay (cleaned)	7.0992591	7.5899129	0	316.0000000	40359
TOTCHG	Total charges (cleaned)	26425.10	38341.29	40.0000000	1000000.00	39371

Data visualization can be used to extract useful knowledge from large and complex datasets. The visualization can be used to build a narrative concerning the data. Kernel density estimation provides information about the entire population distribution rather than to rely on means and variances. Then, the Kernel Density Estimation (KDE procedure) was used to examine the lung disease by Age, Length of Stay and Total Charges, which showed the relationships among these outcomes by using data visualization.

First, we use PROC KDE to examine the variables in relationship to the data with kernel density. The main advantage of using kernel density estimation is that the graphs can be overlaid for more direct comparisons. For example, we consider the relationship of lung cancer to Age, Length of Stay and Costs.

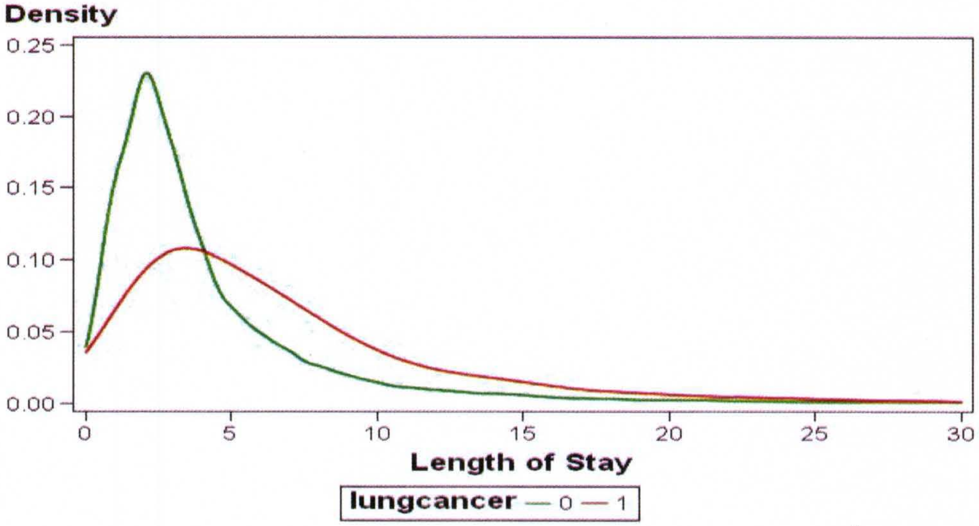
```
proc sort data=medstat.inpatientadm
out=work.sortedinpatientadm;
by lungcancer;
proc kde data=work.sortedinpatientadm;
univar age/gridl=0 gridu=100
out=medstat.kdeinpatientadmage;
by lungcancer;
run;
proc kde data=work.sortedinpatientadm;
univar days/gridl=0 gridu=500
out=medstat.kdeinpatientadmdays;
by lungcancer;
run;
proc kde data=work.sortedinpatientadm;
univar totpay/gridl=-100000 gridu=300000
out=medstat.kdeinpatientadmtotpay;
by lungcancer;
run;
```

Figure 6.2 The Kernel Density of Lung cancers by Age using Kernel Density Estimation.



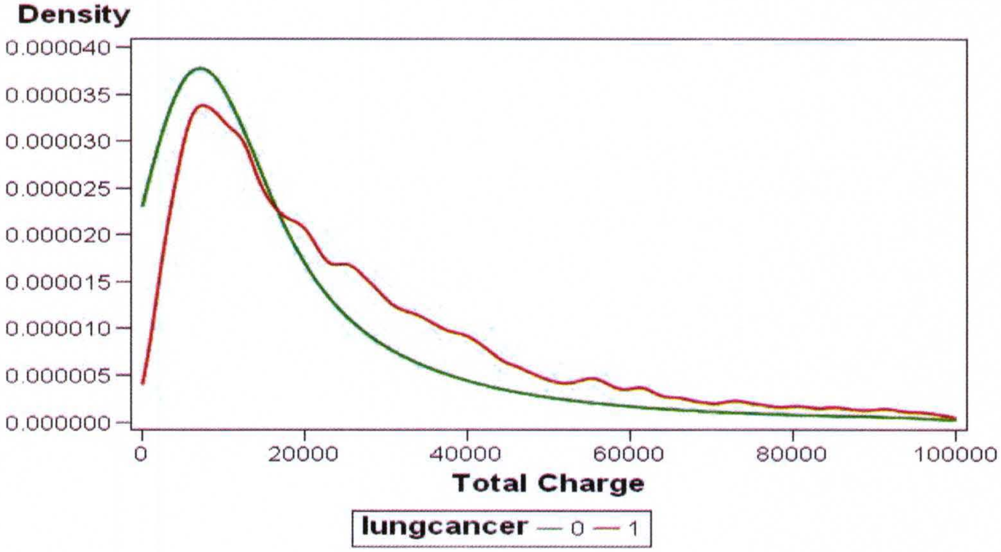
Note that the patients without lung cancer have a relatively constant likelihood of an inpatient event regardless of age (except for the interval of 0 to 20 and 60 to 80, where there is a slight change. However, patients with lung cancer increase inpatient events starting at age 38, accelerating at age 45, and decreasing at age 78.

Figure 6.3 The Density of Lung Cancer by LOS (Length of Stay) using Kernel Density Estimation.



Those with lung cancer have a higher probability of a stay of 6 or more days, and a lower probability of staying 5 or fewer days compared to patients without lung cancer.

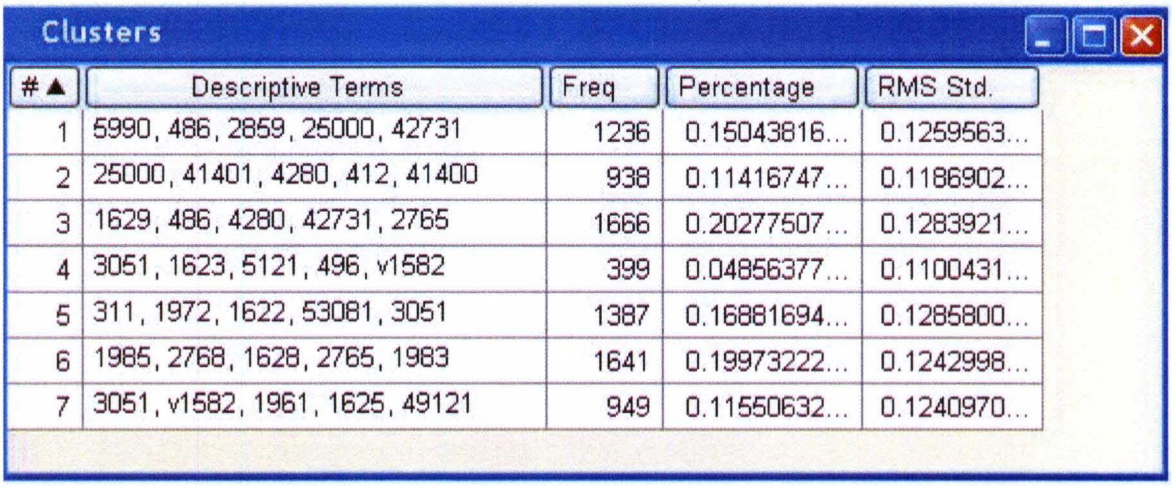
Figure 6.4 The density of Lung Cancer by Total Charge using Kernel Density Estimation.



Note that there is an intersection point of costs for patients at around 19,000, indicating that there is a higher probability of higher cost if the patient has lung cancer. From the summary table, we know that the average cost for a patient with lung cancer is around 26,425.

Next, we considered the diagnosis codes and examined more in-depth the types of complications that patients have in relation to lung cancer. Recall that there are 8,216 patients with a diagnosis of lung cancer for the year 2004. Patients can be represented in multiple categories. Since there is no ranking between patients based on claims data, text mining can be used to determine clusters and classify the patients in those clusters based on the conditions from the claims. Here, Text Miner in Enterprise Miner was used to examine the data according to text strings of patient conditions. In order to perform text analysis on the lung cancer data, the Text Miner node in Enterprise Miner was used to examine the data according to text strings of patient conditions defined above. Cluster analysis was used to find the categories of documents. To define text clusters, we limit the number of terms to ten to describe clusters. We use the standard defaults of Expectation Maximization and Singular Value Decomposition. For example, the text analysis defined seven different clusters in the data that were given in Table 6.3. In order to compare outcomes by text clusters, we merge the cluster descriptions and the cluster numbers into the original dataset. We use kernel density estimation to make a comparison of age, length of stay and cost by clusters.

Table 6.3 Cluster table for diagnosis strings.



# ▲	Descriptive Terms	Freq	Percentage	RMS Std.
1	5990, 486, 2859, 25000, 42731	1236	0.15043816...	0.1259563...
2	25000, 41401, 4280, 412, 41400	938	0.11416747...	0.1186902...
3	1629, 486, 4280, 42731, 2765	1666	0.20277507...	0.1283921...
4	3051, 1623, 5121, 496, v1582	399	0.04856377...	0.1100431...
5	311, 1972, 1622, 53081, 3051	1387	0.16881694...	0.1285800...
6	1985, 2768, 1628, 2765, 1983	1641	0.19973222...	0.1242998...
7	3051, v1582, 1961, 1625, 49121	949	0.11550632...	0.1240970...

Table 6.4 shows the translations of these clusters. These code translations are provided at http://icd9cm.chrisendres.com/.

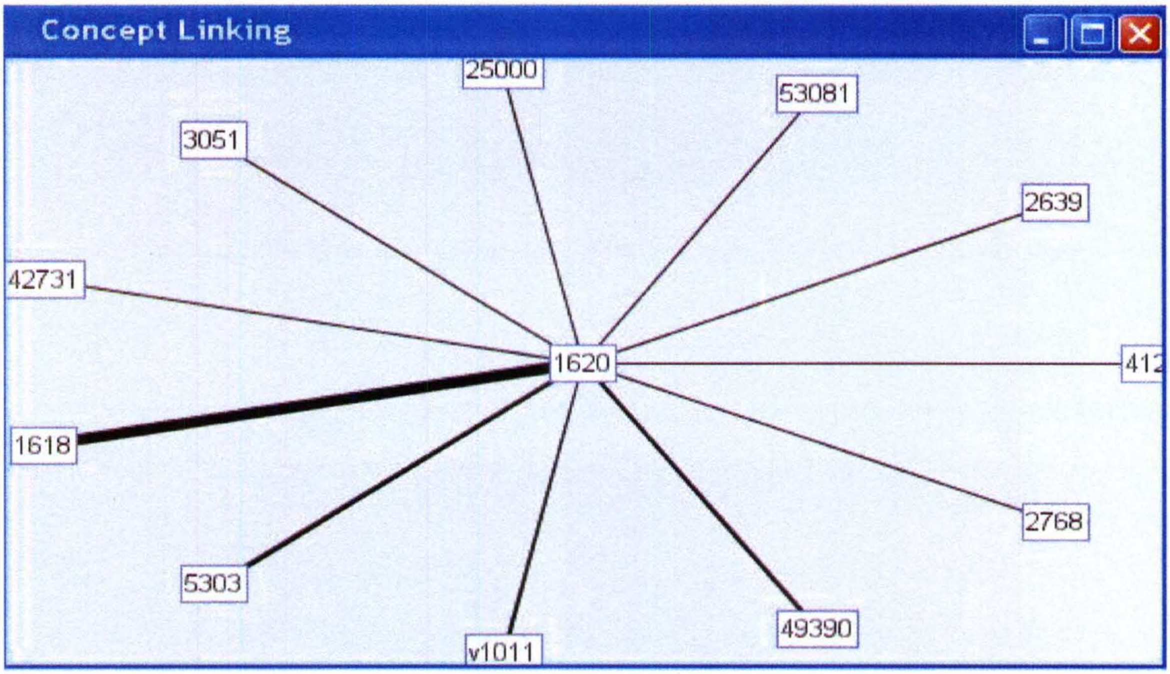
Table 6.4 Translation for the clusters

Cluster #	Description	Label
1	Unspecified Urinary tract infection, Pneumonia , Unspecified Anemia, Diabetes mellitus without mention of complication, Atrial fibrillation	Diabetes and Heart Problems
2	Diabetes mellitus without mention of complication, Coronary atherosclerosis, Unspecified Congestive heart failure, Old myocardial infarction	Diabetes and Heart Problems (CHF)
3	Unspecified Bronchus and lung, Pneumonia , Unspecified Congestive heart failure, Atrial fibrillation, Volume depletion	COPD and Heart problems
4	Tobacco use disorder, Upper lobe, bronchus or lung, Iatrogenic pneumothorax, Chronic airway obstruction, History of tobacco use	COPD and smoking
5	Depressive disorder, Pleura, Main bronchus, Esophageal reflux, Tobacco use disorder	Depression
6	Secondary malignant neoplasm of Bone, bone marrow, Brain and spinal cord , Hypopotassemia,	Metastasizing Cancer

Cluster #	Description	Label
	Malignant neoplasm of Other parts of bronchus or lung,	
7	Tobacco use disorder, History of tobacco use, Secondary and unspecified malignant neoplasm of Intrathoracic lymph nodes, Malignant neoplasm of Lower lobe, bronchus or lung, Chronic bronchitis With (acute) exacerbation	COPD and cancer in the lymph nodes

We want to examine the relationship between lung cancer and other diseases. The concept links could show how different terms are related in the documents. Hence, we use concept links of 1620; the links for 1622, 1623,1624,1625,1628 and 1629 are similar.

Figure 6.5 Concept links for 1620, Malignant Neoplasm of Trachea



Note that most of the links are to code 1618 (shown with the widest line), malignant neoplasm of other specified sites of larynx. The other large links are to 5303 (Stricture and stenosis of esophagus), v1011 (Personal history of malignant neoplasm of Bronchus and lung), and 49390 (Asthma).

Again, kernel density estimation was use to make a comparison of age, length of stay and cost by clusters. The code was the following:

```
data emws1.clusternis (keep=_cluster_ _freq_ _rmsstd_
clus_desc);
set emws1.text_cluster;
run;
data emws1.desccopynis (drop=_svd_1-_svd_500
_roll_1-_roll_1000 prob1-prob500);

set emws1.text_documents;

run;

proc sort data=emws1.clusternis;

by _cluster_;

proc sort data=emws1.desccopynis;

by _cluster_;

data emws1.nistextranks;

merge emws1.clusternis emws1.desccopynis;

by _CLUSTER_;

run;

proc kde data=emws1.nistextranks;

univar totchg/gridl=0 gridu=100000

out=emws1.kdecostbycluster;

by _cluster_;

run;

proc kde data=emws1.nistextranks;

univar age/gridl=0 gridu=100

out=emws1.kdeagebycluster;

by _cluster_;

run;

proc kde data=emws1.nistextranks;

univar los/gridl=0 gridu=35

out=emws1.kdelosbycluster;

by _cluster_;

run;
```

The average cost for cluster 6 is large compared to other clusters. There is no big difference between clusters 1, 2, 3 and 7, which mean that they have similar severity conditions. Cluster 5 has a slightly higher probability of a higher cost than cluster 4 (Figure 6.6).

Figure 6.6 Kernel Density Estimate for Total Charges by Clusters.

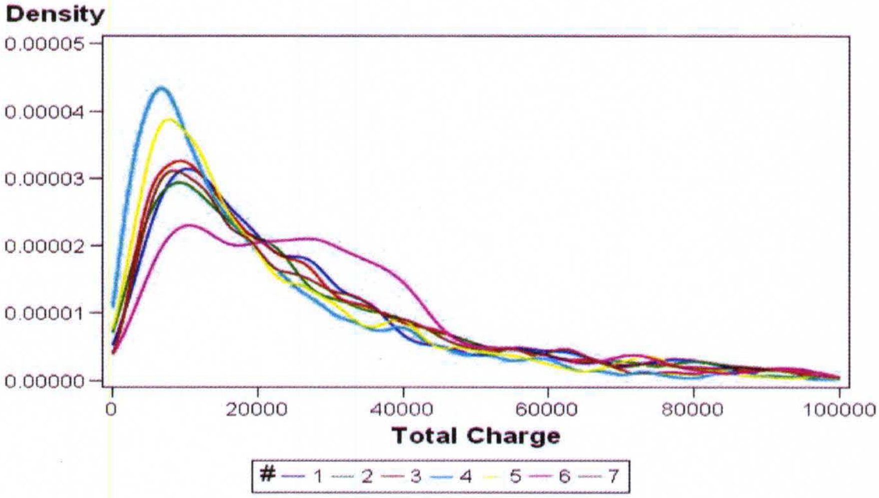
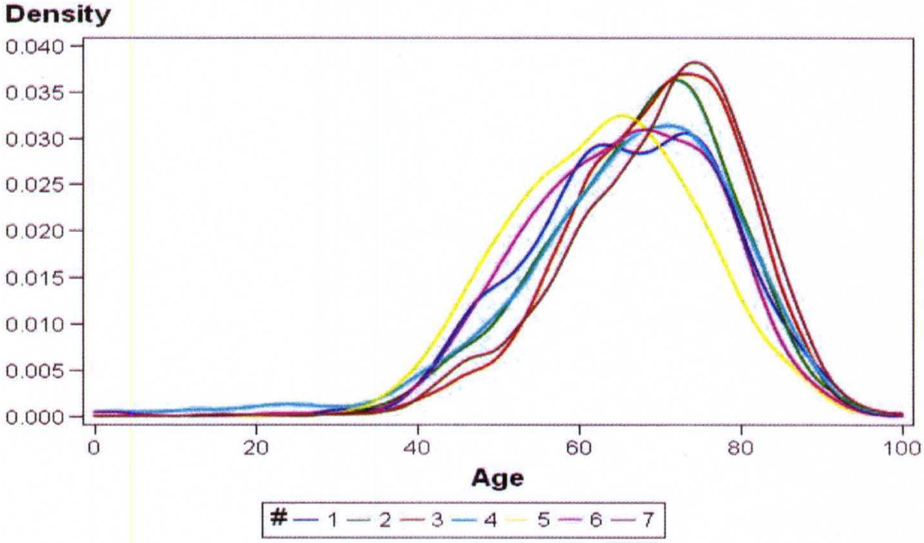
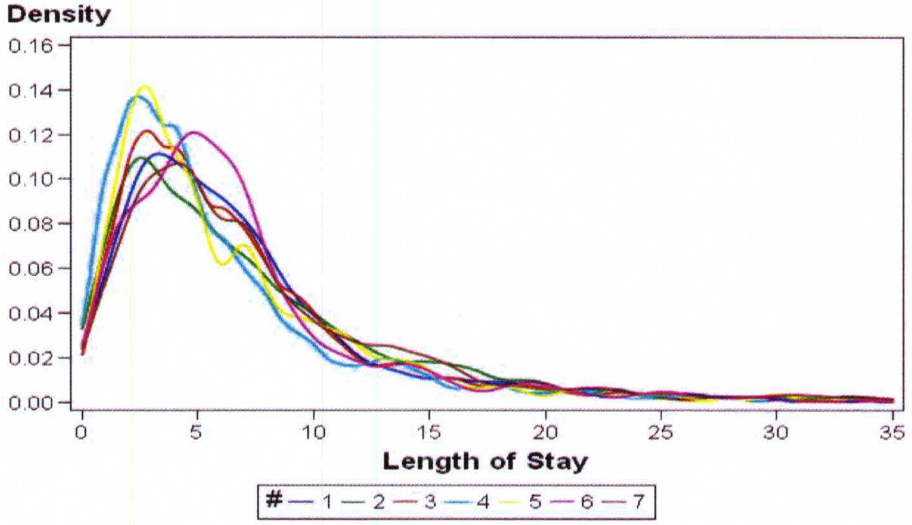


Figure 6.7 Kernel Density Estimate for Age by Clusters.



For the average age of each cluster, note that cluster 5 has the smallest average age, around 61, compared to other clusters. Clusters 1, 4 and 6 have a similar average age of 70. Similarly, clusters 2, 3 and 7 have an average age of 75.

Figure 6.8 Kernel Density Estimate for Length of Stay by Clusters.



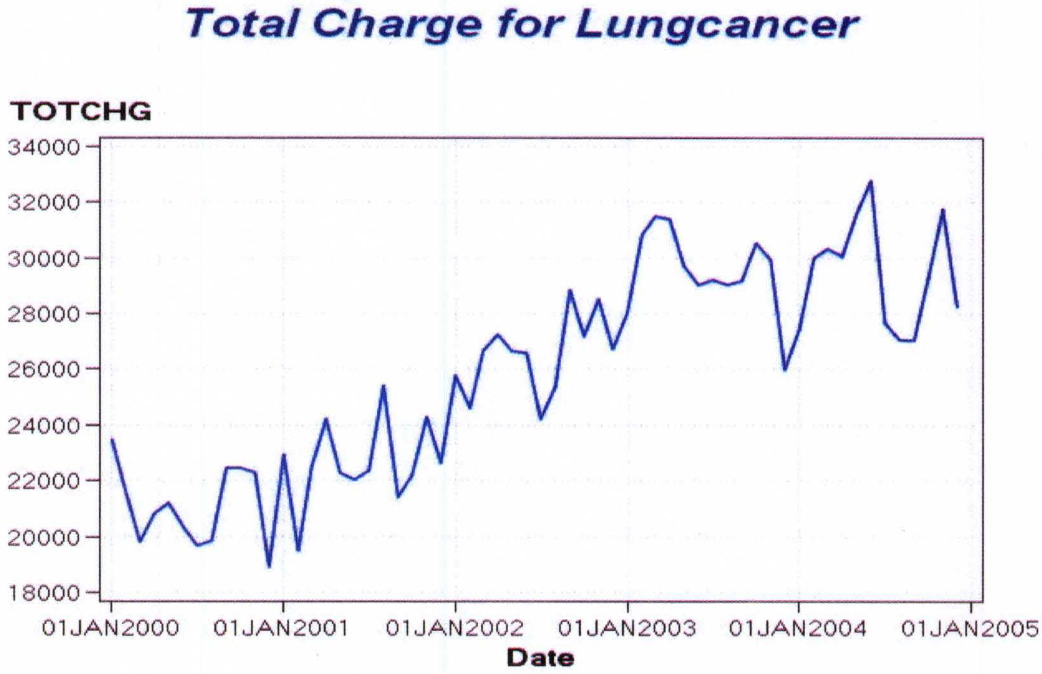
Note that cluster 6 has a higher probability of a longer stay compared to the others. It would seem reasonable that patients at higher risk will stay longer and have higher cost.

6.4. Time Series and Forecasting

Next, we want to investigate and forecast the total costs of treatment on lung cancer to determine the future costs based on the inflation rate [35] with consideration of the patient outcomes and conditions of the patients undergoing different treatments to help the hospital build a good and effective financial system.

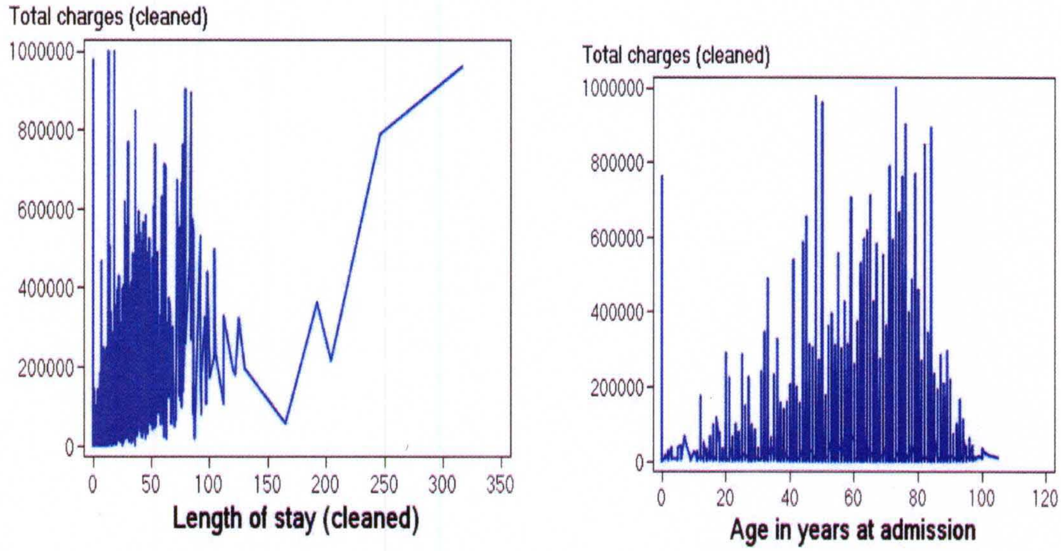
Consider Figure 6.9, for example, which shows the trend of the total cost for lung cancer over the period from January, 2000 to December, 2004 with 60 monthly average charges. For the 4-year period, the price increases from $20,000 in 2000 to $32,000 in 2004. Time series analysis is an appropriate method for this purpose.

Figure 6.9 The trend of Total Charge from Jan 2000 to Dec 2004.



Time series models were also used to analyze the Total Cost, Age, Length of Stay and Inflation Rate. We used some time series features in Enterprise Guide to create time series data. Here, we accumulated using the average. After accumulating, the number of records decreased to 60 with the time interval of month. The inflation rate data over the same period were collected from the website, inflationrate.com, and added to those monthly average data. Different models were considered with or without the inflation rate. Enterprise Guide was used to create a SAS dataset. Then, Time Series models were used to examine the data. It made the information of price more visible with respect to date.

Figure 6.10 The plots of Total Charges by Los and Age.

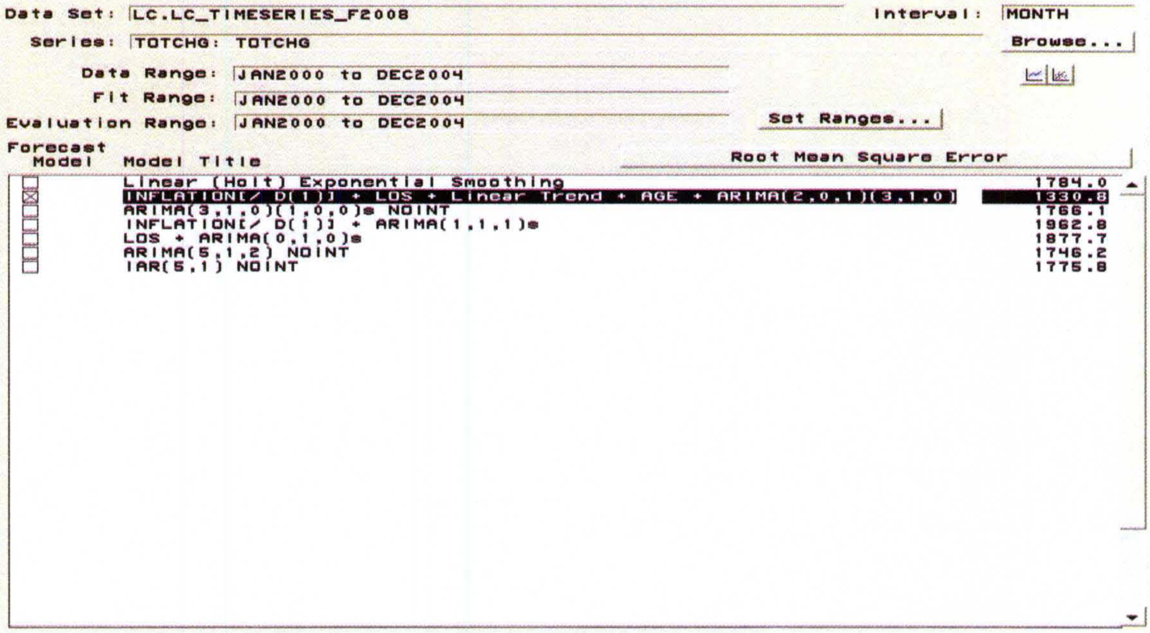


Note that there is an increasing trend for the mean of Total charges while Length of stay is approximately increasing. However, this trend is not very clear when the number of days is between 100 and 200 because of the insufficient data in this study. The plot of total charge and age shows that the amount of charges of patients with age from 50 to 80 is much higher than those of patients with age less than 45. There exists some relationship between them, which is not clear just by the information from these graphs.

The inflation rate was selected as a dynamic regressor, specifying a denominator factor with a simple order of 1, which represents a shifting of the inflation rate by one time unit, implying that the inflation rate leads the total cost by one month. Then, Age and Length of Stay were selected as regressors to predict the total cost. Since the actual data are increasing as shown in Figure 6.9, we added the linear trend in the model. ARIMA (3, 1, 0) and ARMA (2, 1) were selected as the seasonal model and error model, respectively.

We applied all possible models on the data by switching values for p and d. Also, other models were applied in order to choose the best fit model for the data. We compared all models by looking at the Root Mean Square Error and R Square. All these diagnostic measures show that INFLATION[/D(1)]+LOS+LINEAR TREAD +AGE+ARIMA(2,0,1)(3,1,0)s is the best model for our data. Figure 6.11 shows the list of models used. The smallest Root Mean Square Error is 1330.8.

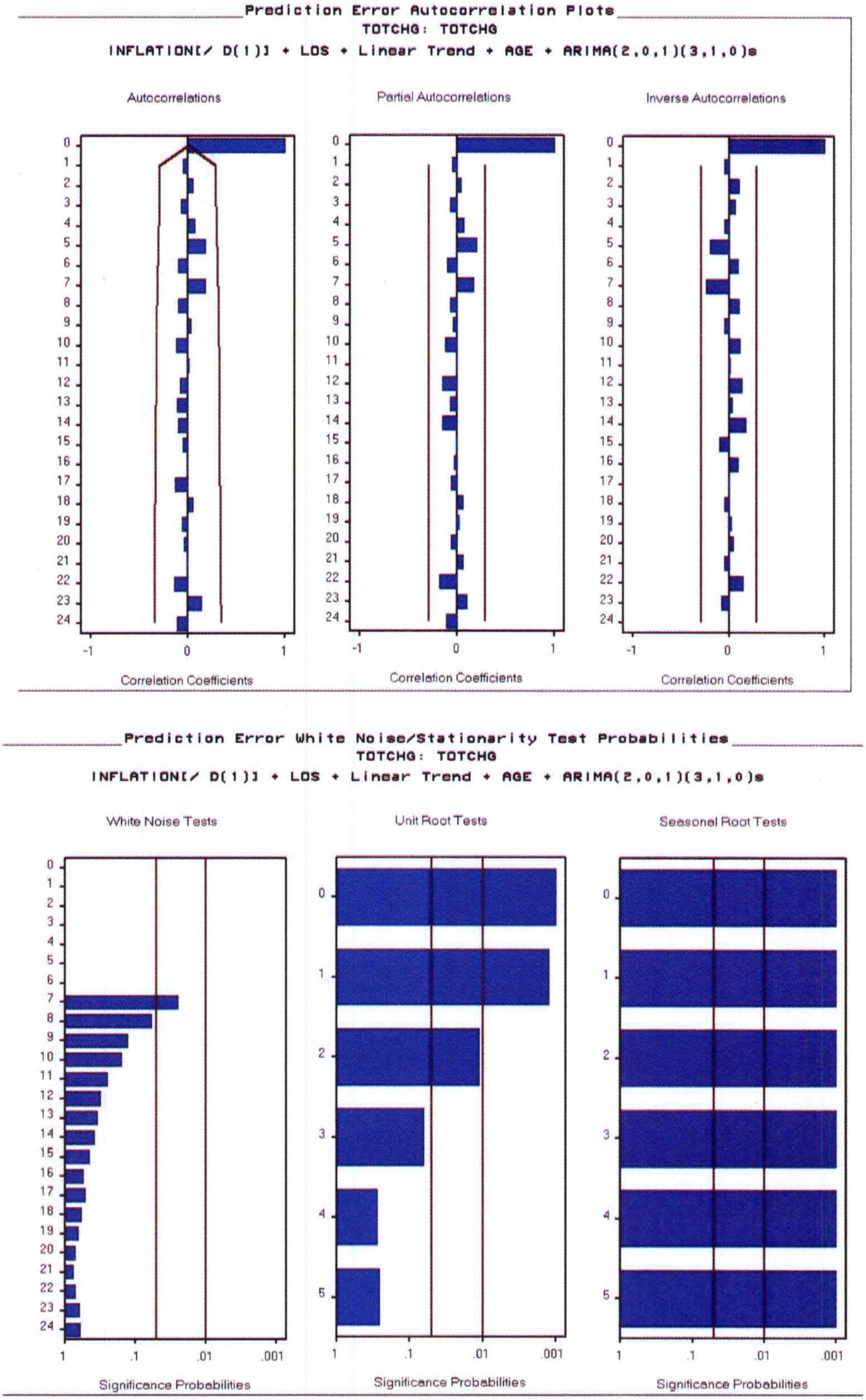
Figure 6.11 Root Mean Square Error for Different Models Used



Data Set: LC.LC_TIMESERIES_F2008
Interval: MONTH
Series: TOTCHG: TOTCHG
Data Range: JAN2000 to DEC2004
Fit Range: JAN2000 to DEC2004
Evaluation Range: JAN2000 to DEC2004

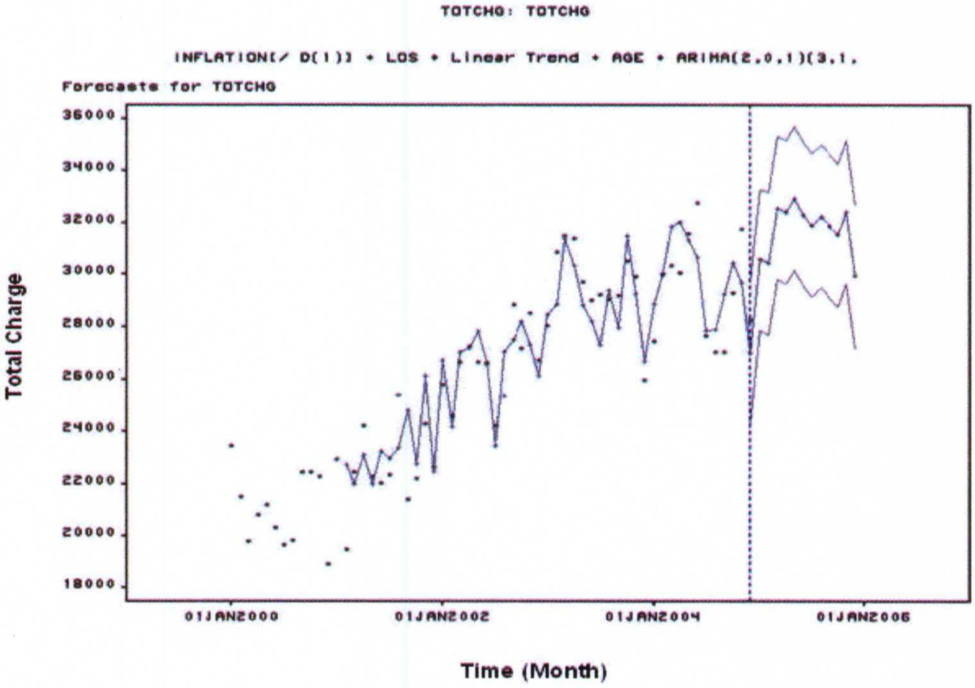
Forecast Model	Model Title	Root Mean Square Error
□	Linear (Holt) Exponential Smoothing	1784.0
☒	INFLATION[/ D(1)] + LOS + Linear Trend + AGE + ARIMA(2,0,1)(3,1,0)	1330.8
□	ARIMA(3,1,0)(1,0,0)s NOINT	1766.1
□	INFLATION[/ D(1)] + ARIMA(1,1,1)s	1962.8
□	LOS + ARIMA(0,1,0)s	1877.7
□	ARIMA(5,1,2) NOINT	1746.2
□	IAR(5,1) NOINT	1775.8

Figure 6.12 Prediction Error Autocorrelation Plots and White Noise



The autocorrelation function (ACF) was used to examine the seasonality by calculating and plotting the residuals, which are the difference from each data point to the mean. The null hypothesis for the ACF is that each time series observation is not correlated to others. A criterion for ACF to test the autocorrelation is whether there are residuals that are greater than two standard deviations away from the mean, then it indicates the statistically significant autocorrelation [36]. The partial autocorrelation function (PACF) is also used to detect trends and seasonality. From Figure 6.12, all ACF and PACF lags fall below significant levels, which means that the autocorrelation has been eliminated. Thus, we conclude that this model is an adequate model based on the white noise check, autocorrelation plot and the smallest Root Mean Square Error.

Figure 6.13 The forecast of Total Cost based on ARIMA model with Regressors.



Note that this model fits the data well. The predicted total charges for the next 12 months will still keep increasing, averaged at $31,500. There was a large drop at the end of 2004, and the predicted charge after 2004 would increase to the highest level, and then decrease a little.

6.5. Logistic Regression Model for Mortality

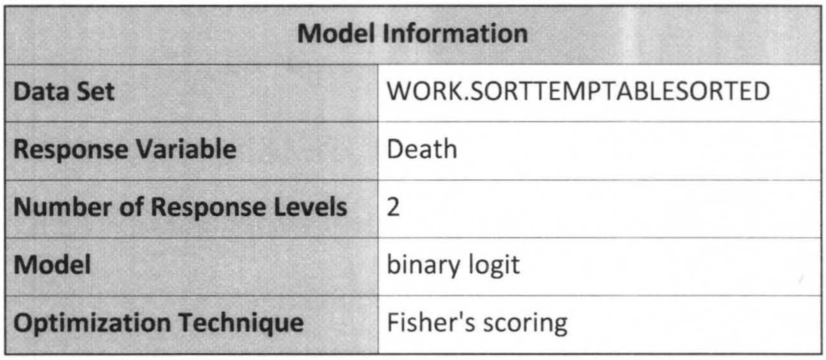
Lung Cancer is the leading cause of cancer deaths in the world. Therefore, it will be helpful for decision-making to examine the relationship between the death and conditions of the patients undergoing different treatments. Results show that the mortality of lung cancer was highly related to Age, Length of stay, 7 clusters of diagnosis and 9 clusters of procedures, and we also found the Logistic model for the mortality caused by lung cancer.

We filtered the patients who have lung diseases using the ICD9 diagnosis codes. We used SAS Enterprise Guide and the CATX and RXMATCH statements along with other functions in several lines of code to get a summary of the codes defining Lung cancer. After preprocessing the data, we had 8216 patient records involving Lung cancer for 2004.

Regression analysis can characterize the relationship between a response variable and one or more predictor variables. In linear regression, the response variable is continuous. In logistic regression, the response variable is categorical. The Logistic regression model uses the predictor variables, which can be categorical or continuous, to predict the probability of specific outcomes.

Here, we focus on the relationship between Death and demographic factors such as Age in years at admission, financial factors such as Total charges, Median household income for each patient, and other clinical conditions of the patients, Admission type, Elective versus non-elective admission, Length of stay, MDC (define what MDC represents), diagnosis and treatment procedures. Here, the variable, Death, was selected as the Dependent variable, and Age, Los, Totchg and Zipinc_qrtl were continuous variables. The Atype, Elective, MDC, Cluster of diagnoses and Cluster of Procedures were chosen as classification variables. The following table 6.5 gave us some basic information about the data for this study. A Logistic regression model with Fisher's scoring optimization technique was used to determine the relationship between those above variables.

Table 6.5 Basic information about the data set.



Model Information	
Data Set	WORK.SORTTEMPTABLESORTED
Response Variable	Death
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	8216
Number of Observations Used	5382

Response Profile		
Ordered Value	**Death**	**Total Frequency**
1	0	4677
2	1	705

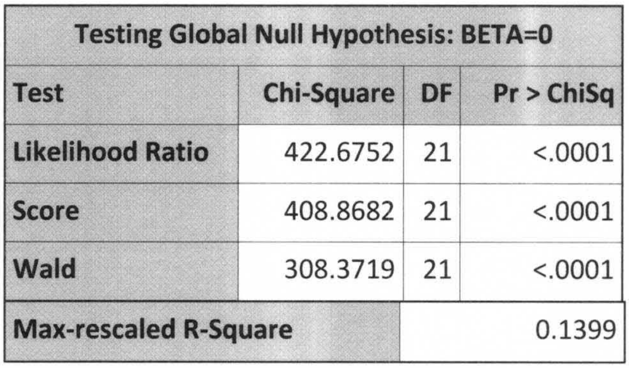
Probability modeled is Death=0.

Note that there were in total 8216 patient records, 705 patients who died and 2834 records deleted due to missing values for the response or explanatory variables.

We evaluated the significance of all variables in the model by using the full model fitted method based on the variables we discussed above.

Table 6.6 Model fit statistics.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	4181.322	3800.647
SC	4187.913	3945.645
-2 Log L	4179.322	3756.647



Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	422.6752	21	<.0001
Score	408.8682	21	<.0001
Wald	308.3719	21	<.0001
Max-rescaled R-Square			0.1399

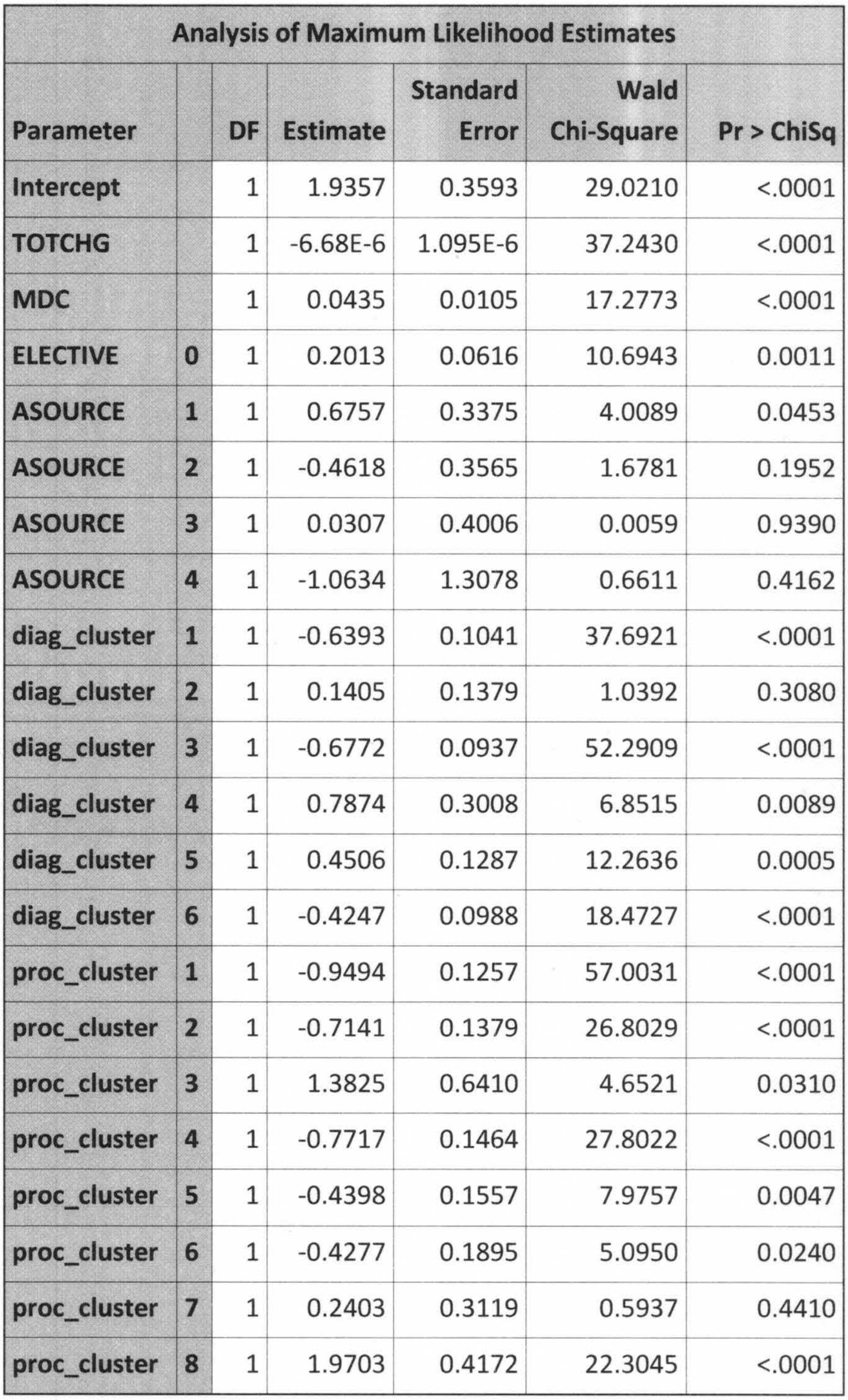
In the model fit statistics table, the various criteria (-2 Log L, AIC, SC) were calculated based on the likelihood for fitting a model with intercepts only and for fitting a model with intercepts and explanatory variables. In the Testing Global Null Hypothesis table to test the null hypothesis that all regression coefficients are zero, all three tests rejected the null hypothesis. That is, the model is not a constant. Here, the Max_rescaled R-square is 0.1399.

Table 6.7 Summary of Stepwise Selection

Summary of Stepwise Selection						
	Effect	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
Step	Entered					
1	diag_cluster	6	1	159.4608	<.0001	Cluster ID of diagnosis code
2	ASOURCE	4	2	83.2749	<.0001	Admission source (uniform)
3	proc_cluster	8	3	89.1558	<.0001	Cluster ID of treatment procedures
4	TOTCHG	1	4	53.8284	<.0001	Total charges (cleaned)
5	MDC	1	5	16.7903	<.0001	MDC in effect on discharge date
6	ELECTIVE	1	6	10.7637	0.0010	Elective versus non-elective admission

Table 6.7 listed all the variables selected by the model with the stepwise selection method. The diag_cluster, Asource, proc_cluster, MDC, Totchg and Elective are significant at the 0.05 level in the model.

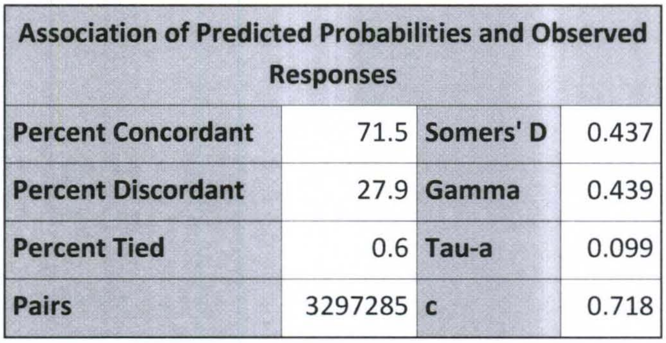
Table 6.8 The analysis of MLEs.



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.9357	0.3593	29.0210	<.0001
TOTCHG		1	-6.68E-6	1.095E-6	37.2430	<.0001
MDC		1	0.0435	0.0105	17.2773	<.0001
ELECTIVE	0	1	0.2013	0.0616	10.6943	0.0011
ASOURCE	1	1	0.6757	0.3375	4.0089	0.0453
ASOURCE	2	1	-0.4618	0.3565	1.6781	0.1952
ASOURCE	3	1	0.0307	0.4006	0.0059	0.9390
ASOURCE	4	1	-1.0634	1.3078	0.6611	0.4162
diag_cluster	1	1	-0.6393	0.1041	37.6921	<.0001
diag_cluster	2	1	0.1405	0.1379	1.0392	0.3080
diag_cluster	3	1	-0.6772	0.0937	52.2909	<.0001
diag_cluster	4	1	0.7874	0.3008	6.8515	0.0089
diag_cluster	5	1	0.4506	0.1287	12.2636	0.0005
diag_cluster	6	1	-0.4247	0.0988	18.4727	<.0001
proc_cluster	1	1	-0.9494	0.1257	57.0031	<.0001
proc_cluster	2	1	-0.7141	0.1379	26.8029	<.0001
proc_cluster	3	1	1.3825	0.6410	4.6521	0.0310
proc_cluster	4	1	-0.7717	0.1464	27.8022	<.0001
proc_cluster	5	1	-0.4398	0.1557	7.9757	0.0047
proc_cluster	6	1	-0.4277	0.1895	5.0950	0.0240
proc_cluster	7	1	0.2403	0.3119	0.5937	0.4410
proc_cluster	8	1	1.9703	0.4172	22.3045	<.0001

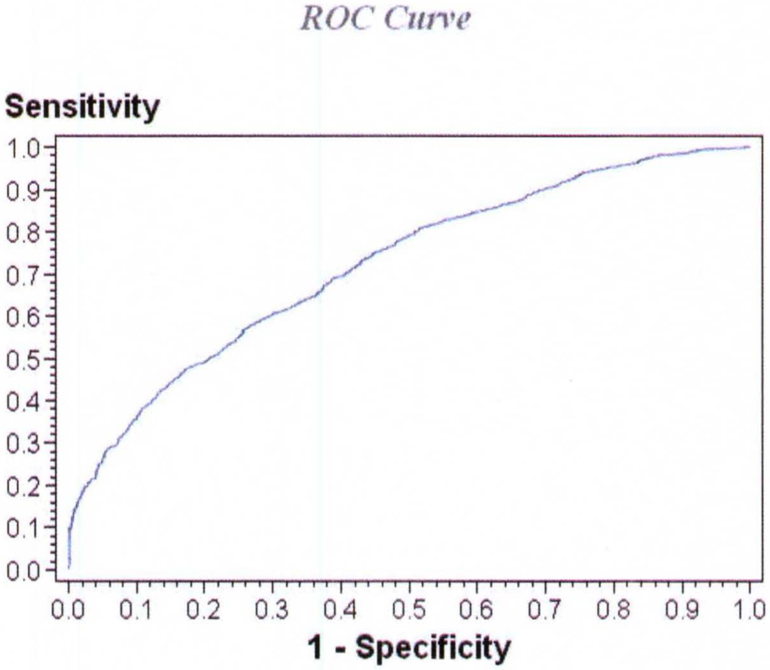
Since these variables are categorical, in Table 6.8, we have the regression coefficients for all the variables to predict the dependent variable by the maximum likelihood method.

Table 6.9 The Association of Predicted Probabilities and Observed Responses



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	71.5	Somers' D	0.437
Percent Discordant	27.9	Gamma	0.439
Percent Tied	0.6	Tau-a	0.099
Pairs	3297285	c	0.718

Figure 6.14 ROC Curve



Note that the area under the ROC curve is also given by the statistic c in the Association of Predicted Probabilities and Observed Responses table. In this study, the area under the ROC curve is 0.718. For model fitness, the closer the ROC plot is to the upper left corner (i.e. the area under the curve is close to 1), the more accurate of the model. However, since there are too many factors related to the patient's death for our study and lack of patient information, the logistic regression model is appropriate to predict the mortality of the lung cancer for this NIS database.

6.6 Summary

Kernel Density Estimation was used to compare graphs that can be overlaid to give us more information. Here, we might conclude that older patients are more likely to have lung cancers that would lead to a higher probability of longer stay and higher costs for the treatment procedure. With text analysis on the diagnosis codes and KDE, it shows that malignant neoplasm of lobe, bronchus or lung is of higher risk and has a higher cost compared to other lung cancers.

The ARIMA model with ordinary and dynamic regressors was used to analyze the hospital's financial data. It provides the hospital with the ability to predict total charges of lung cancer based on previous costs. The ordinary and dynamic regressors model showed the effect of the length of stay and age on the predicted values of total charges.

Then, the Logistic model was used to examine the relationship between death and conditions of patients with lung diseases. In this paper, we just focused on Age in years at admission, Admission type, Admission source, Elective versus non-elective admission, Length of stay, Total charges, Median household income for patient MDC, and diagnosis and treatment procedures. By using the stepwise selection method with level 0.05, we removed Admission type, Age, LOS and Median household income for patient, which are not statistically significant. With this model, we could calculate the probability of death based on the given patient condition.

CHAPTER VII DATA ANALYSIS AND MODELING FOR MEDSTAT DATA

7.1. MarketScan Database Overview

The MarketScan Databases include the information about clinical utilization, costs, plans and membership across inpatient, outpatient, prescription drugs at member level and service level, which makes it possible for data analysis concerning outcomes. There are links between those tables such that the claims, drugs and patient information are on the same level. Historically, there are more than 500 million claims records available in the MarketScan Databases.

The MarketScan Databases have the following tables:

- Medical/surgical claims consisting of three tables:
  1. Inpatient Admissions Table (I)
  2. Inpatient Services Table (S)
  3. Outpatient Services Table (O)
- Aggregated Populations Table (P)
- Outpatient Pharmaceutical Claims Table(D)
- Enrollment Summary Table (E) and Enrollment Detail Table (T)

Here we just introduce the inpatient table used in this study.

Inpatient Admissions Table (I)

The Inpatient Admissions Table contains the records with summarized information about each hospital admission case, which is combined by all of the associated service records. Contained in the cost information is the sum of all the services associated with the same admission case. The inpatient table also includes the principal procedure, principal diagnosis, Major Diagnostic Category (MDC), and Diagnosis Related Group (DRG), etc. The following table gives all the variables available in the admission table. We will specify those variables as the inputs for our analysis in later sections.

Table 7.1. List of variables in Inpatient admission table.

Name	Long Name	Name	Long Name	Name	Long Name
ADMDATE	Date of Admission	EESTATU	Employment Status	PROC2	Procedure 2
ADMTYP	Admission Type	EGEOLOC	Geographic Location Employee	PROC3	Procedure 3
AGE	Age of Patient	EIDGLAG	Enrollee ID Derivation Flag	PROC4	Procedure 4
AGEGRP	Age Group	EMPCTY	County Employee	PROC5	Procedure 5
CASEID	Case and Services Link	EMPREL	Relation to Employee	PROC6	Procedure 6
DATATYP	Data Type	EMPZIP	Zipcode Employee 3 Digit	PROC7	Procedure 7
DAYS	Length of Stay	ENRFLAG	Enrollment Flag	PROC8	Procedure 8
DOBYR	Patient Birth Year	ENROLID	Enrollee ID	PROC9	Procedure 9
DRG	Diagnosis Related Group	HOSPCTY	County Hospital	PROC10	Procedure 10
DSTATUS	Discharge Status	HOSPPAY	Payments Hospital	PROC11	Procedure 11
DX1	Diagnosis 1	HOSPZIP	Zipcode Hospital 3 Digit	PROC12	Procedure 12
DX2	Diagnosis 2	INDSTRY	Industry	PROC13	Procedure 13

DX3	Diagnosis 3	LASTADM	Days from Prior Discharge	PROC14	Procedure 14
DX4	Diagnosis 4	MDC	Major Diagnostic Category	PROC15	Procedure 15
DX5	Diagnosis 5	MSA	Metropolitan Statistical Area	REGION	Region
DX6	Diagnosis 6	NEXTADM	Days to Next Admission	RX	Cohort Drug Indicator
DX7	Diagnosis 7	PATFLAG	Patient Indistinct Flag	SEQNUM	Sequence Number
DX8	Diagnosis 8	PATID	Patient ID	SEX	Gender of Patient
DX9	Diagnosis 9	PDX	Diagnosis Principal	STATE	State Hospital
DX10	Diagnosis 10	PHYFLAG	Physician Specialty Coding Flag	TOTNET	Payments Net Case
DX11	Diagnosis 11	PHYSID	Physician ID	TOTPAY	Payments Total Case
DX12	Diagnosis 12	PHYSPAY	Payments Physician	TRIMLOS	Trim Flag Length of Stay
DX13	Diagnosis 13	PLANKEY	Benefit Plan Link	TRIMPDM	Trim Flag Per Diem
DX14	Diagnosis 14	PLANTYP	Plan Indicator	UNIHOSP	Hospital ID MDST
DX15	Diagnosis 15	PPROC	Procedure Principal	VERSION	Version
EECLASS	Employee Classification	PROC1	Procedure 1	WGTKEY	MarketScan National Weight Link
				YEAR	Date Year Incurred

7.2. Variables and their values

The variable, DSTATUS, means Discharge Status, which is the status of the patient upon discharge from the hospital. The possible values are as follows:

01: Discharged to home self-care

02: Transfer to short-term hospital

03: Transfer to SNF

04: Transfer to ICF

05: Transfer to other facility

06: Discharged home under care

07: Left against medical advice

08-19: Other alive status

20-29: Died

30-39: Not Yet discharged/Transferred

40-42: Other died status

50: Discharged to home (from Hospice)

51: Transfer to medical facility (from Hospice)

61: Transferred to Medicare approved swing-bed

71: Transfer/referred to other facility for outpatient services

72: Transfer/referred to this facility for outpatient services

99: Transfer, identified through Hospital ID MDST change

Missing: Invalid

The variable DAYS means Length of Stay, which is the number of overnight stays for a hospital admission.

The variable, AGEGRP, means Age Group, which is a value identifying the patient or members age group. The following are the possible values:

1: 0-17

2: 18-34

3: 35-44

4: 45-54

5: 55-64

6: 65 and older

The variable ADMTYP is Admission Type, which means the type of hospital admission with following values:

1: Surgical

2: Medical

3: Maternity & Newborn

4: Psych & Substance Abuse

5: Unknown

The variable SEX is the gender of patient with values:

1: Male

2: Female

The variable, LASTADM, is the number of days between a patient's previous discharge and their current admission date.

The variable, NEXTADM, is the number of days between a patient's current discharge and their next admission date.

The variable, MDC, represents the Major Diagnostic category, which is a set of Body-system or disease related groupings of clinical conditions, based on diagnosis codes. See Appendix A for the MDC values.

The variable, REGION, is the Geographic Region of an employee's residence.

1: Northeast

2: North Central

3: South

4: West

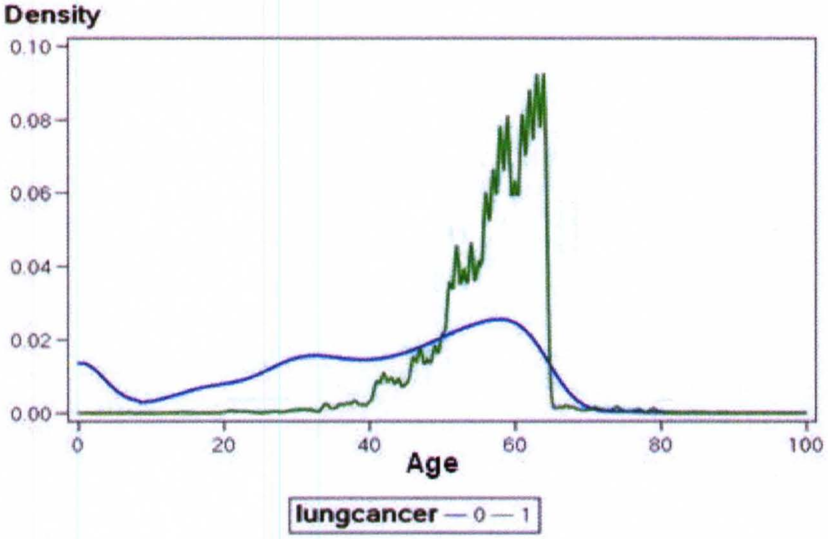
5: Unknown

7.3. Analysis of MarketScan Data

We next examine the lung cancer data from the Medstat MarketScan database, and we used two years of data, 2000 and 2001. Each year of data included medical and surgical claims, aggregated populations and enrollment information. Here, we consider inpatient cases first. There are a total of 4,718 observations for two years related to lung disease out of 800,000 records. Note that approximately 1.05% of the inpatient population has a diagnosis of lung disease.

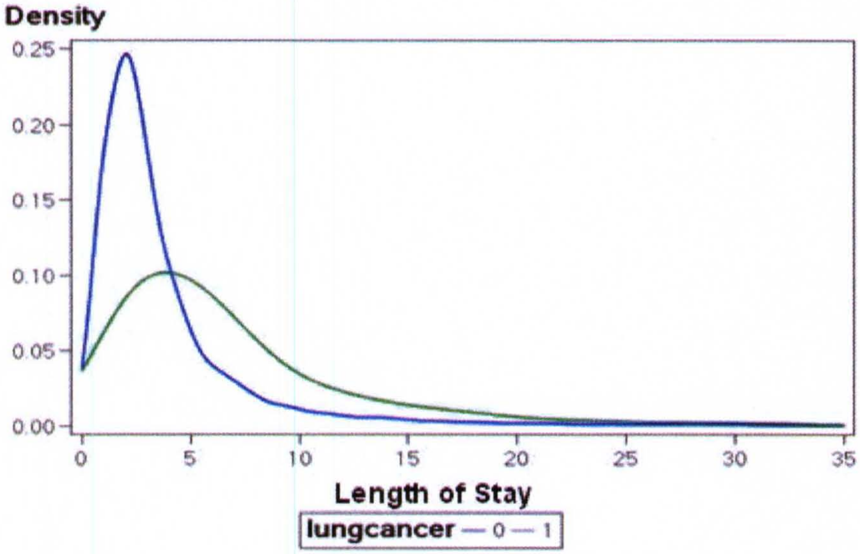
First, we use PROC KDE to examine the variables in relationship to the data using kernel density estimation. The main advantage of using kernel density estimation is that the graphs can be overlayed for more direct comparisons. For example, we consider the relationship of lung cancer to Age, Length of Stay and Costs.

Figure 7.1 The Kernel Density of Lung Cancer by Age using KDE.



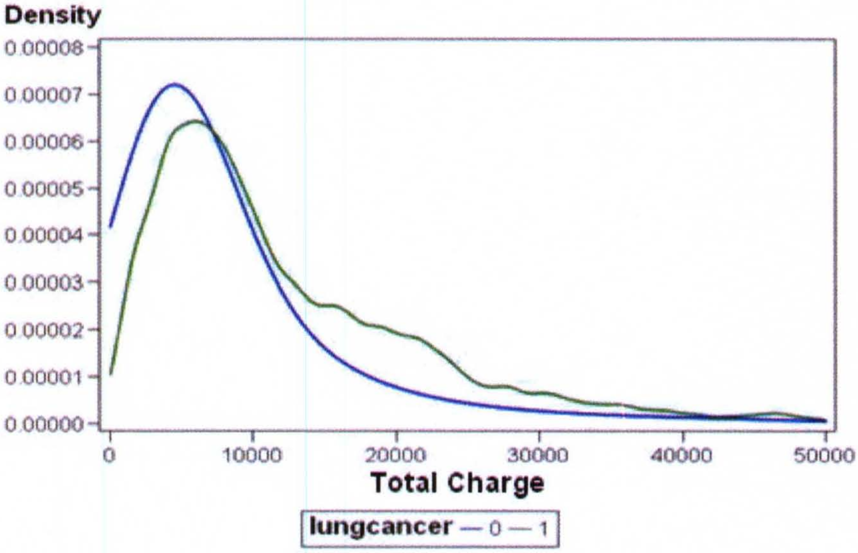
The study was limited to The MarketScan Commercial population (who are covered by employer-sponsored private health insurance). Most of the members in this database are 65 or younger [37]. Note that the patients without lung diseases have a relatively constant likelihood of an inpatient event regardless of age (except for the interval of 0 to 20 and 60 to 80, where there is a slight change because of the population limitation). However, patients with lung diseases increase inpatient events starting at age 38, accelerating at age 45. We could not make any conclusions for those 65 or older because the available data contained only privately insured individuals instead of Medicare.

Figure 7.2. The Density of Lung cancer by Days (Length of Stay) using KDE



Those with lung diseases have a higher probability of a stay of 4 or more days, and a lower probability of staying 3 or fewer days compared to patients without lung diseases.

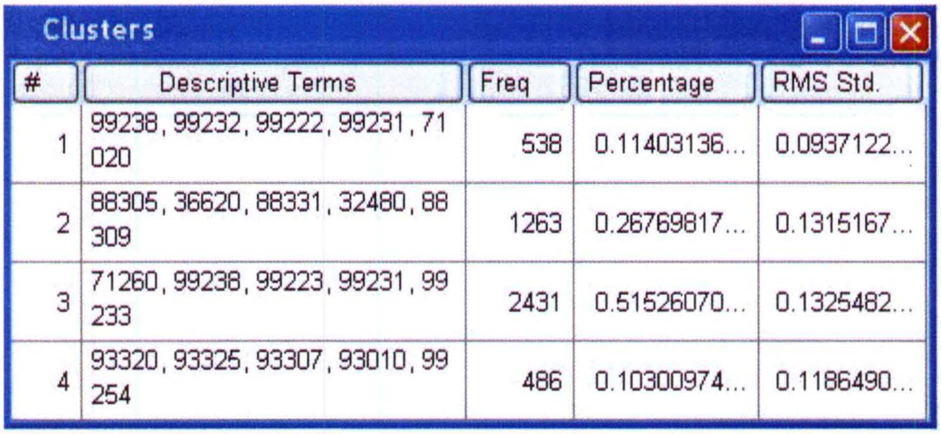
Figure 7.3. The density of Lung cancer by Total Charge using KDE



Note that there is an intersection point of costs for patients at around 7,000, indicating that there is a higher probability of higher cost if the patient has lung disease.

Next, we considered the procedure codes and examined more in-depth the types of treatments that patients have in relation to lung cancer. Recall that there are 4,718 patients with a diagnosis of lung cancer. The procedures of treatment can be represented in multiple categories. Here, Text Miner in Enterprise Miner was used to examine the data according to text strings of treatment procedures. Cluster analysis was used to find the categories of documents. For example, the text analysis defined four different clusters in the data that are given in Table 7.2.

Table 7.2. Cluster table for procedure strings.



#	Descriptive Terms	Freq	Percentage	RMS Std.
1	99238, 99232, 99222, 99231, 71020	538	0.11403136...	0.0937122...
2	88305, 36620, 88331, 32480, 88309	1263	0.26769817...	0.1315167...
3	71260, 99238, 99223, 99231, 99233	2431	0.51526070...	0.1325482...
4	93320, 93325, 93307, 93010, 99254	486	0.10300974...	0.1186490...

The following Table 7.3 shows the translations of these clusters by using the CPT codes.

Table 7.3 Translation for the clusters

Cluster #	Description	Label
1	Initial and subsequent hospital care, Radiologic examination, chest, two views, frontal and lateral	Screening by Scan (X-Ray or Other)
2	Level IV and VI- Surgical pathology, gross and microscopic examination, Pathology consultation during surgery; first tissue block, with frozen section(s), single specimen	Biopsy Examination
3	Initial and subsequent hospital care, Computed tomography, thorax; with contrast material(s)	MRI
4	Doppler echocardiography, Inpatient consultation for a new or established patient	Doppler

According to the results of the NIS data, cluster 6, Secondary malignant neoplasm of Bone, bone marrow, Brain, is related to Biopsy Examination, with a higher average cost. Cluster 4, COPD and smoking, are related to Screening by Scan (X-Ray or Other).

Again, kernel density estimation was used to make a comparison of age, length of stay and cost by clusters. The average cost for cluster 2 is greater compared to other clusters. There is no big difference between clusters 3 and 4, which means that they have similar severity conditions. Cluster 1 has a slightly higher probability of a lower cost than other clusters (Figure 7.4).

Figure 7.4 Kernel Density Estimate for Total Charges by Clusters.

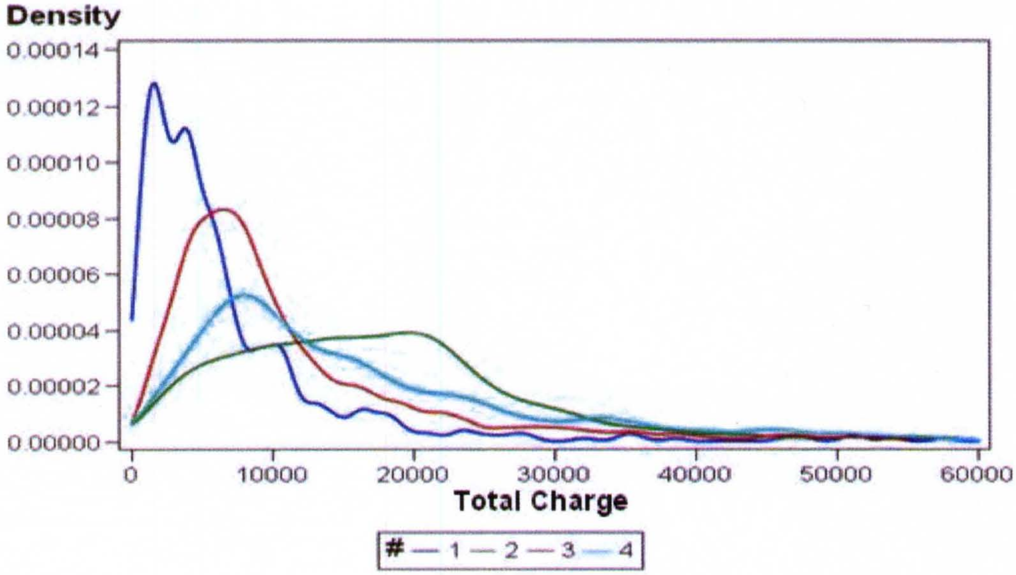
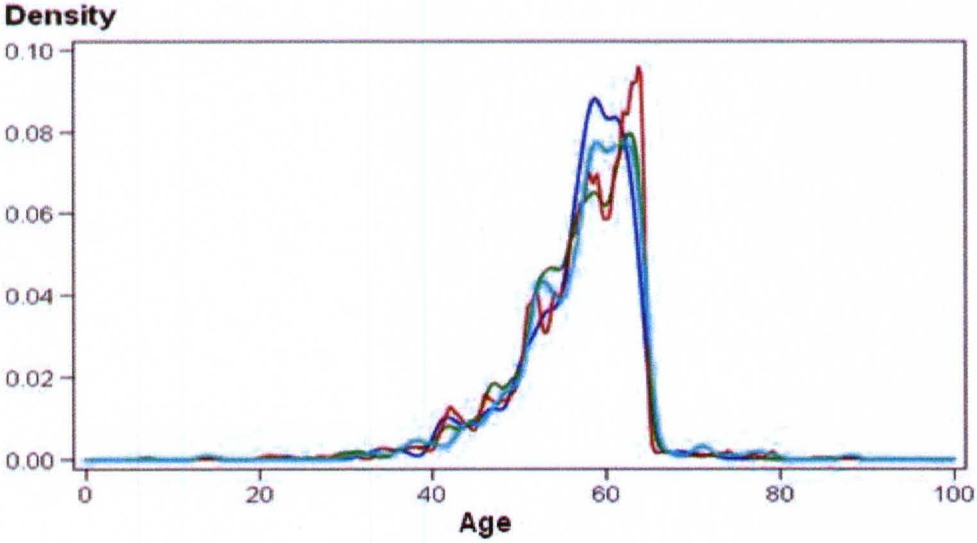
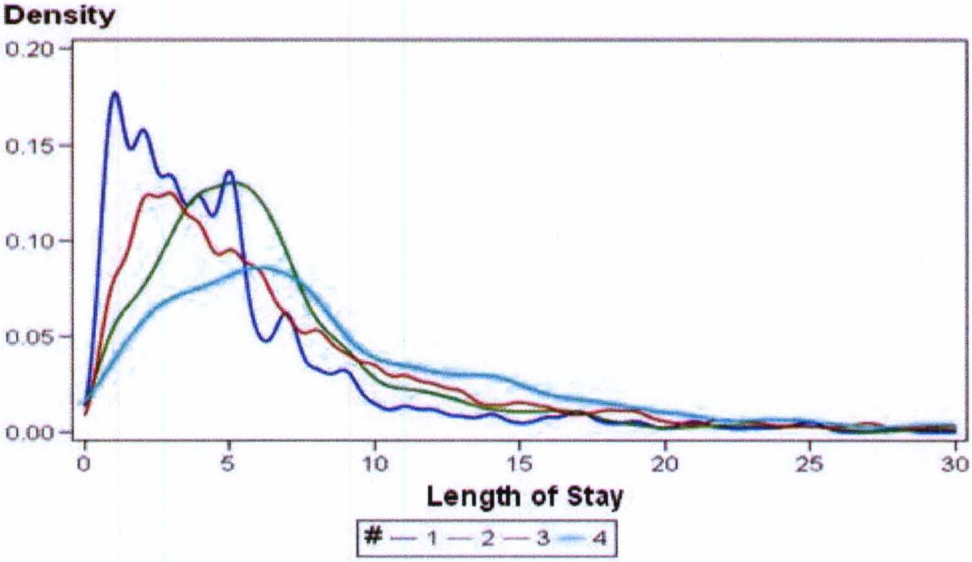


Figure 7.5 Kernel Density Estimate for Age by Clusters.



For the average age of each cluster, note that all four clusters have similar shapes, which indicates that the average age for each cluster is 60.

Figure 7.6 Kernel Density Estimate for Length of Stay by Clusters.



Note that clusters 2 and 4 have a higher probability of a longer stay compared to the others. It would seem reasonable that patients at higher risk will stay longer and have higher cost.

7.4 Medication

The purpose of this section is to investigate top medications for lung cancer based on the Medstat lung cancer population. Based on the medication information, we examined the relationship between the medication and the mortality. In the Medstat database we have, the prescription data are available in ten different data files. We need to filter the data for the medications to study the treatment of lung cancer.

In the available data set, there are two variables specifying the drug used in the treatment, the NDC number (NDCNUM) and the generic product ID (GENERID). The converter for NDC numbers is available at the FDA website [38]. For each drug, there are multiple NDC numbers due to their dosage and type, and a few Generic product ID values. Therefore, it is faster and more general to use generic IDs. We found the Generic IDs (names) of each drug using NDC numbers as well as the drug class defined. Those ten data sets were filtered by the lung cancer population at the beginning to reduce the process time and then were appended as one table at the end.

We combined all datasets resulting from the above procedure to have one file containing all RX data for our study. We use programming similar to what we used for procedures, but this time for antibiotic so as to follow the antibiotics prescribed for each patient with the date. In other words, we have a one-to-many relationship between patients and their prescriptions.

As a result, we have information about patients with lung cancer, including their procedures and prescriptions in the same dataset. In this section, we will focus on medication, so we only keep the drug column of the joined table with patient IDs.

Table 7.4 The top Medicines sorted by prescriptions.

Generic Name	Drug Class	Prescriptions	Percent	Cumulative Percent
Dexamethasone	STEROIDS	2428	2.19	2.19
Hydrocodone/Acetaminophen	PAIN MGMT - NARCOTIC	1806	1.63	3.82
Levofloxacin	ANTIBIOTICS	1800	1.62	5.45
Oxycodone w/ Acetaminophen	PAIN MGMT - NARCOTIC	1679	1.52	6.96
Prochlorperazine Maleate	MENTAL HEALTH - MISC	1631	1.47	8.44
Propoxyphene-N w/ APAP	PAIN MGMT - NARCOTIC	1397	1.26	9.7
Lorazepam	MENTAL HLTH - ANXTY	1279	1.15	10.85
Azithromycin	ANTIBIOTICS	1200	1.08	11.93
Albuterol Inhal Aerosol	ASTHMA	1173	1.06	12.99
Prednisone	STEROIDS	1076	0.97	13.96
Megestrol Acetate Susp	CHEMOTHERAPY	1047	0.95	14.91
Omeprazole Cap Delayed Release	STOMACH - PPI	1036	0.94	15.84
Famotidine	STOMACH - H2	911	0.82	16.67
Ipratropium-Albuterol Aerosol	ASTHMA	862	0.78	17.45
Cephalexin Cap	ANTIBIOTICS	860	0.78	18.22
Ciprofloxacin HCl	ANTIBIOTICS	834	0.75	18.97
Potassium Chloride Microencapsulated Crys CR	NUTRITIONAL/ VITAMIN	828	0.75	19.72

The above top 17 medications already account for 19.72% used out of more than 2000 medications. Below, we list some descriptions for some medications from the above table.

1. Dexamethasone/ Prednisone tablets are used for the treatment of many different conditions, such as allergic disorders, or breathing disorders, which also prevent nausea and vomiting caused by the chemotherapy of cancer.

2. Hydrocodone is a narcotic pain-reliever and a cough suppressant and Acetaminophen is a non-narcotic analgesic (pain reliever) and antipyretic (fever reducer).

3. Levofloxacin is a quinolone antibiotic used in lung or other area where infections are caused by certain bacteria.

4. Oxycodone is used to treat moderate to severe pain. It is stronger than Hydrocodone.

5. Prochlorperazine Maleate is used to treat severe nausea and vomiting from various causes such as anti-cancer treatment.

6. Propoxyphene is a pain reliever.

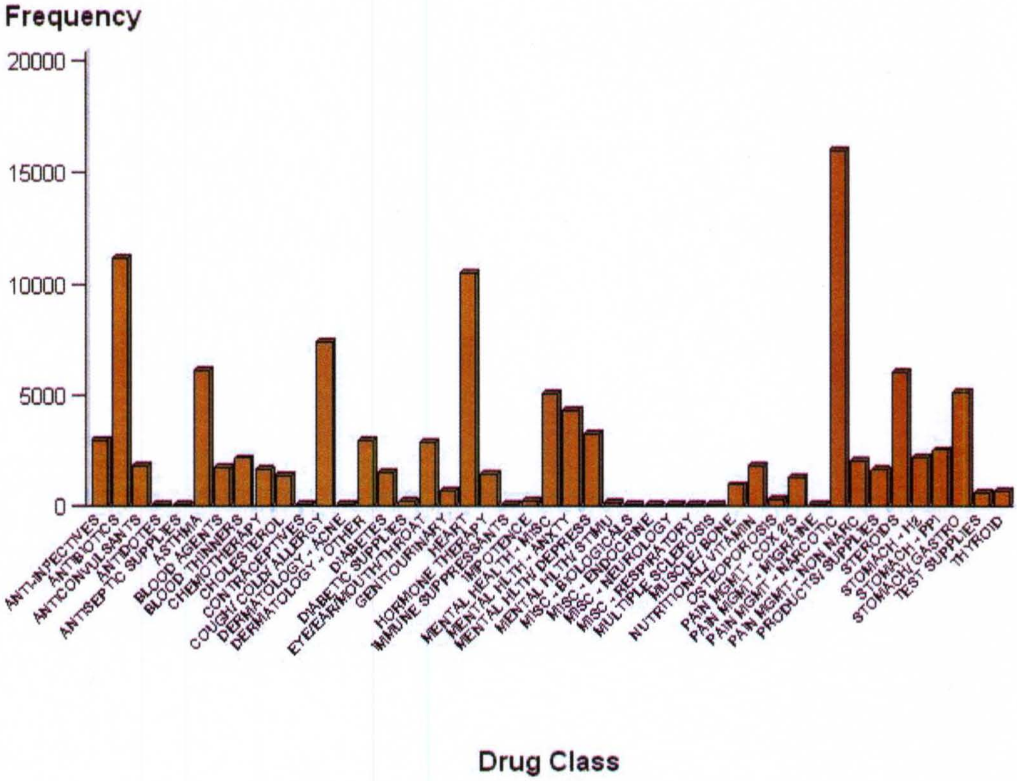
7. Lorazepam is used to reduce nervous tension.

8. Azithromycin/ Zithromax , as Levofloxacin , is used for the treatment of mild to moderate infections caused by certain bacteria.

9. Inhalation Aerosol is used in chronic obstructive pulmonary disease (COPD), a type of lung disease leading to breathing difficulties.

10. Megestrol Oral Suspension is used for weight loss caused by malignancies, systemic infections, and so on.

Figure 7.7 The distribution of Drug class



Note that the most commonly used drug classes of treatment for lung cancer are pain management-Narcotic, Antibiotics, Heart, Cough/Cold/Allergy, and so on. More details about the frequencies for each of the drug classes are listed in the following table.

Table 7.5 The frequency for each Drug class

Drug Class	Frequency
PAIN MGMT - NARCOTIC	16047
ANTIBIOTICS	11186
HEART	10485
COUGH/ COLD/ ALLERGY	7405
ASTHMA	6140
STEROIDS	6049
STOMACH/ GASTRO	5180
MENTAL HEALTH - MISC	5073
MENTAL HLTH - ANXTY	4349
MENTAL HLTH/ DEPRESS	3275
ANTI-INFECTIVES	2993
DERMATOLOGY - OTHER	2966
EYE/EAR/MOUTH/THROAT	2874
STOMACH - PPI	2542
STOMACH - H2	2211
BLOOD THINNERS	2169
PAIN MGMT - NON NARC	2029
NUTRITIONAL/ VITAMIN	1846
ANTICONVULSANTS	1825
BLOOD AGENTS	1723
PRODUCTS/ SUPPLIES	1660
CHEMOTHERAPY	1644
DIABETES	1532
HORMONE THERAPY	1453
CHOLESTEROL	1399
PAIN MGMT - COX 2'S	1303
MUSCLE/ BONE	961
THYROID	721
GENITOURINARY	696
TEST SUPPLIES	625
OSTEOPOROSIS	320
DIABETIC SUPPLIES	222
IMPOTENCE	211
MENTAL HLTH/ STIMU	134
PAIN MGMT - MIGRAINE	102
DERMATOLOGY - ACNE	98
MISC - NEUROLOGY	82
IMMUNE SUPPRESSANTS	78
CONTRACEPTIVES	50
MISC - ENDOCRINE	19
ANTISEPTIC SUPPLIES	13
MISC - RESPIRATORY	12
MISC - BIOLOGICALS	5
ANTIDOTES	1
MULTIPLE SCLEROSIS	1

As we see, there are many patients treated with pain management and antibiotics, in addition to what was recommended in the literature for Chemotherapy.

Next, we will examine the medication treatment for lung cancer for each record. However, there are multiple medication treatments for each case. Some occur during the inpatient period and others occur during the follow up period. Hence, we need to combine all the medications together that are related to the same case. That is, we will have one column with all the medication procedures for certain cases. The following codes will perform this purpose.

```
proc sql;
create table inp_sur_model1 as
select distinct a.*,b.ndcnum, b.generid, b.svcdate
from medstat.rs2 as a
left join inp_sur_model as b
on (a.caseid=b.caseid and a.admdate=b.admdate);
quit;

proc sql;
create table ndccode as
select distinct ndcnum
from medstat.inp_sur_model1;
quit;

proc sql;
create table gpi as
select distinct a.*, b.gennam
from ndccode as a
left join medstat.inp_sur_drug_class as b
```

```
on (a.ndcnum=b.ndcnum);
quit;

proc sql;
create table inp_sur_model2 as
select distinct a.caseid,a.admdate, b.gennam
from medstat.inp_sur_model1 as a
left join gpi as b
on (a.ndcnum=b.ndcnum);
quit;

proc sort data=inp_sur_model2;
by caseid admdate;
run;

proc Transpose
data=inp_sur_model2
out=tran_p (drop=_name_ _label_)
prefix=proc_;
var gennam ;
by caseid admdate;
run;
data concat_p ( keep= caseid admdate gennam );
length gennam $32000 ;
set tran_p ;
array rxconcat_p {*} proc_: ;
gennam = left( trim( proc_1 )) ;
do i = 2 to dim( rxconcat_p ) ;
gennam = left( trim( gennam)) || ' ' || left(trim( rxconcat_p[i] )) ;
end ;
run ;

proc sql ;
select max( length( gennam )) into :gennam_LEN from
concat_p ;
quit ;

data concat_p1 ;
```

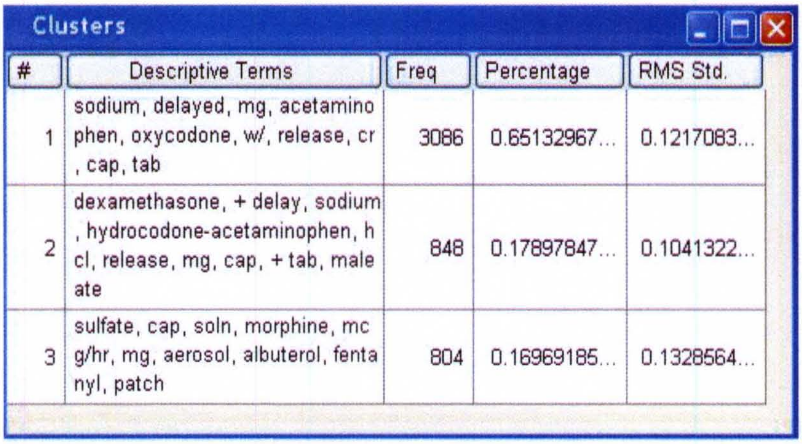
```
length gennam $ &gennam_LEN ;
set concat_p ;
run ;

proc sql;
create table inp_sur_model3 as
select distinct *
from medstat.inp_sur_model1 (drop=ndcnum generid svcdate);
quit;

proc sql;
create table inp_sur_model4 as
select distinct a.*, b.gennam
from inp_sur_model3 as a
left join concat_p1 as b
on (a.caseid=b.caseid and a.admdate=b.admdate);
quit;
```

Here, three medication clusters were obtained based on the medication procedures by using text mining techniques. The clusters are shown in Table 7.6. Note that the first cluster is related to Acetaminophen and Oxycodone used for pain reliever category. The second cluster is about the medication of Dexamethasone, Prochlorperazine Maleate, which is used for preventing nausea and vomiting caused by cancer chemotherapy. The last cluster is Morphine, Albuterol and Fentanyl Patch for pain relief and breathing problems.

Table 7.6 Clusters for medication procedures.



#	Descriptive Terms	Freq	Percentage	RMS Std.
1	sodium, delayed, mg, acetaminophen, oxycodone, w/, release, cr, cap, tab	3086	0.65132967...	0.1217083...
2	dexamethasone, + delay, sodium, hydrocodone-acetaminophen, hcl, release, mg, cap, + tab, maleate	848	0.17897847...	0.1041322...
3	sulfate, cap, soln, morphine, mcg/hr, mg, aerosol, albuterol, fentanyl, patch	804	0.16969185...	0.1328564...

7.5. Survival Analysis

The typical goal in survival analysis is to characterize the distribution of survival time for a given population, to compare this survival time among different groups, or to study the relationship between the survival time and some concomitant variables [39].

The data set analyzed in this study contains the survival times of patients with lung cancer diagnosed. Here, the survival time is defined as the time from the admission date to death. The event of interest is mortality by lung cancer, and interest lies in whether the survival distributions differ between the different factors.

Here, we define the Discharge status by using the following code.

```
data MEDSTAT. inp_sur_analysis (compress=yes);
set MEDSTAT.INP_ADM_TIME;
Status=0;
if dstatus=20 or dstatus=42 then Status=1;
  else if dstatus='' then status= 2;
else status=0;
run;
```

Note that the patient dies if the status variable is 1. Values 0 and 2 indicate that the patient does not die or has a missing value, respectively. Then, we use the following code to generate the survival time, so that we can use survival analysis to examine the data.

```
proc sql;
create table inp_sur_analysis1(compress=yes)
as select distinct a.*, b.death_date
from inp_sur_analysis as a
left join (select distinct caseid, admdate, max(dischgdate) as
death_date format=mmddyy10.0
from inp_sur_analysis
where status=1
group by caseid, admdate) as b
on (a.caseid=b.caseid and a.admdate=b.admdate);
quit;
```

```
PROC SQL;
 CREATE TABLE MEDSTAT.inp_sur_analysis_cluster AS SELECT distinct
B.*,A._cluster_
 FROM inp_sur_analysis1 AS B
        INNER JOIN MEDSTAT.medtextranks AS A  ON (A.caseid = B.caseid);
QUIT;
```

```
data inp_sur1;
set MEDSTAT.inp_sur_analysis_cluster;
if status=1 then time=death_date-admdate;
if status ne 1 then time='31dec2002'd-admdate;
run;
```

The time variable is defined as the days during which the patient was under observation until the event or the end of the study period, which may be either a death or a censoring.

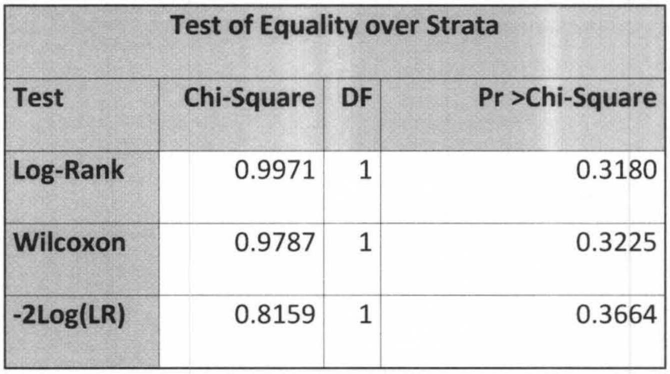
7.5.1 Kaplan-Meier method

There are two methods to produce estimates of survival functions, the Kaplan-Meier method and the life-table. The former is based on actual survival time, which is more suitable for smaller data sets with precisely measured event times and the latter is a time interval grouped by survival time, which is better for large data sets [27]. Another important advantage of the Kaplan-Meier method is that it considers the censored data from the sample before the event occurred.

In our research, there are 4,718 patients with a diagnosis of lung cancer and the survival time is defined as the time from the admission date to death. The data set contains the following variables: Time, Status, Sex, Admtyp, Agegrp, Age, Days, Previous admission within 6 months, 4 clusters based on procedures codes and so on. The Time is used as survival time in days, and the Status variable has the value 1 for uncensored observations and 0, 2 for censored observations. The other variables have the values listed above. Therefore, the

Kaplan-Meier method and right censored method are used to examine the distribution of the survival time with different strata variables.

Table 7.7 Test of Equality over Strata of Sex



Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr >Chi-Square**
Log-Rank	0.9971	1	0.3180
Wilcoxon	0.9787	1	0.3225
-2Log(LR)	0.8159	1	0.3664

This table contains rank and likelihood-based statistics for testing the homogeneity of survivor functions across strata. The rank tests for homogeneity indicate that there is no significant difference between male and female (P values are more than 0.3).

Figure 7.8. Life Tables: Survivor Distribution Plot for Sex

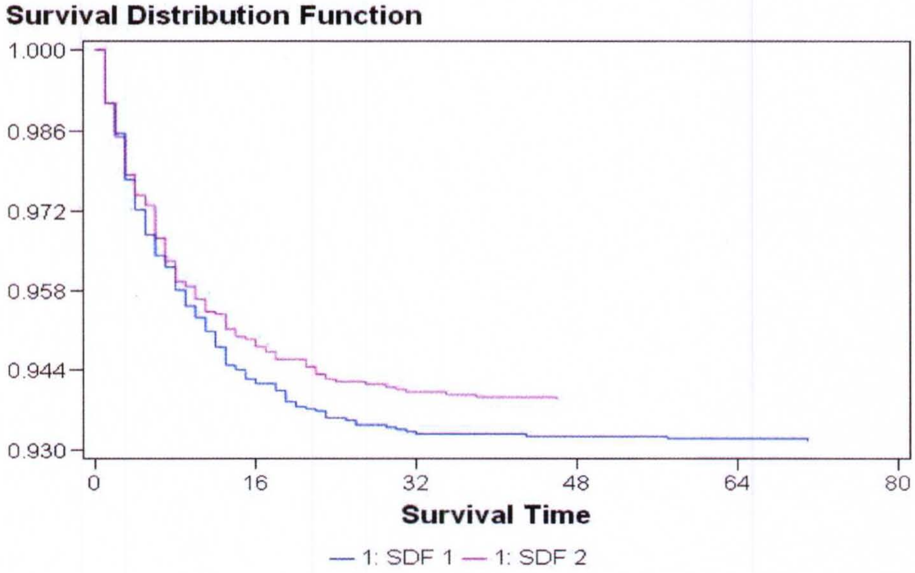


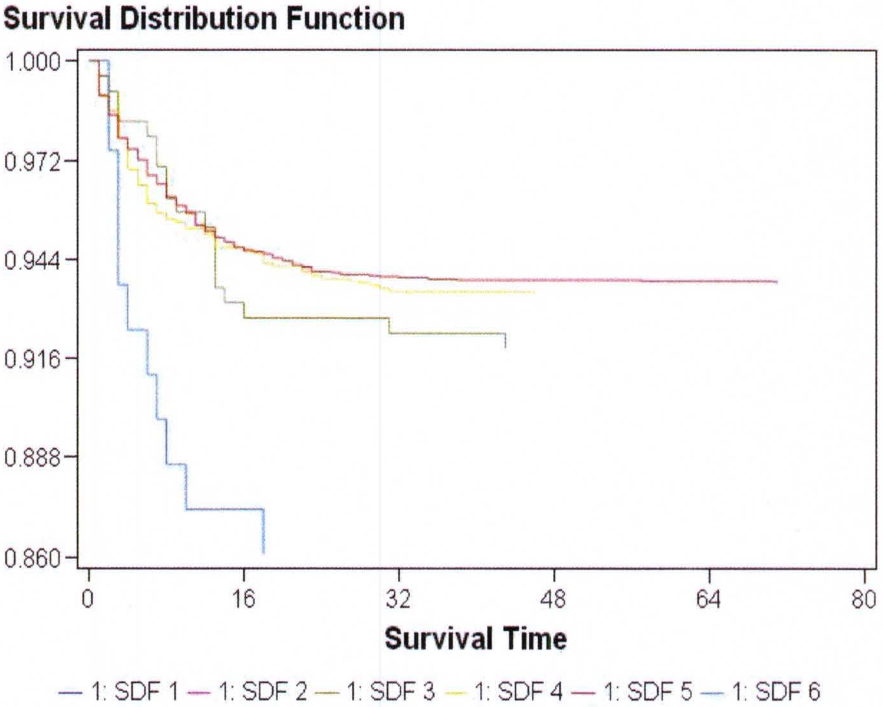
Figure 7.7 displays the survivor function against time for each gender, where SDF 1 was the survivor distribution function for male and SDF 2 was the function for female. There is no significant difference between them at the beginning, which is in accord with a Test of Equality. The gap between the two curves distinguishes between the survivals distributions after survival time of 13, where the curve for Male decreases after the curve for Female. The difference in displayed survival curves reinforces the conclusions that female patients live longer than males.

Table 7.8. Test of Equality over Strata of Age group

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr >Chi-Square**
Log-Rank	11.9045	5	0.0361
Wilcoxon	11.9283	5	0.0358
-2Log(LR)	12.3412	5	0.0304

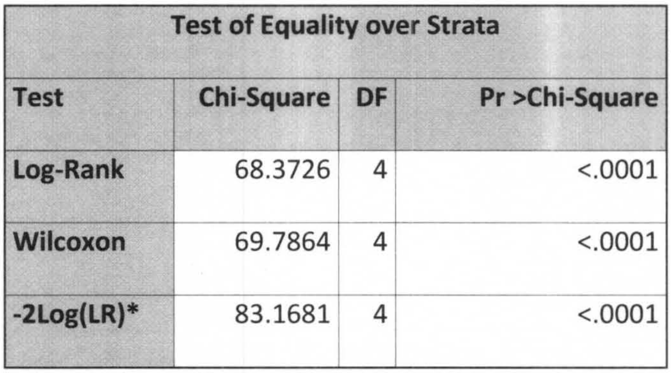
Note that the rank tests for homogeneity indicate that the Age group is significant of mortality by lung cancer.

Figure 7.9. Life Tables: Survivor Distribution Plot for Age Group



Note that it is clear that the six age groups were divided into two categories. Patients with age from 1 to 34 belong to the first category with very low mortality, and those with age 35 or more belong to the second category, which has a higher probability of death. Furthermore, the probability of survival for patients at age 65 or older decreases quickly from 0.888 after 10 days.

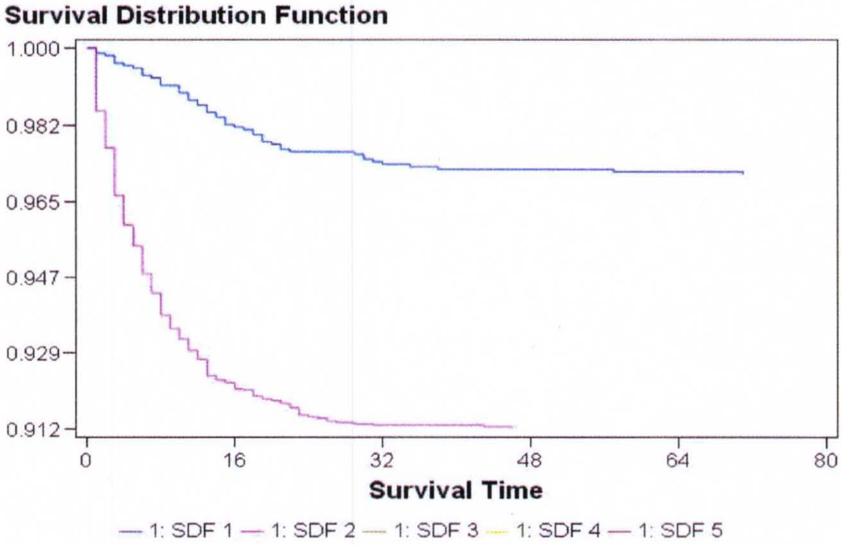
Table 7.9. Test of Equality over Strata of Admission Type



Test of Equality over Strata			
Test	Chi-Square	DF	Pr >Chi-Square
Log-Rank	68.3726	4	<.0001
Wilcoxon	69.7864	4	<.0001
-2Log(LR)*	83.1681	4	<.0001

All three tests show a significant difference between the admission types (Type of hospital admission: Surgical, Medical, Maternity & Newborn, Psych & Substance Abuse, or Unknown) (p- values are less than 0.001). It shows that the admission type is an indicator for the mortality of lung cancer. From the following figure, we will see which admission is more closely associated with mortality.

Figure 7.10. Life Tables: Survivor Distribution Plot for Admission Type

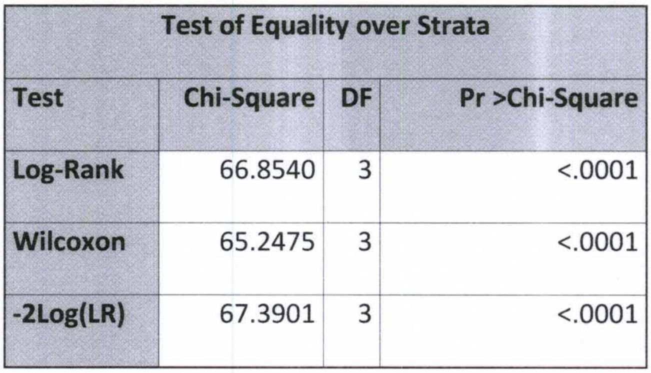


Note that there is no death among patients registered as Maternity & Newborn and Psych & Substance Abuse. During the first 16 days, the probabilities of survival for patients registered as Surgical type and Medical type decrease to 0.98 and 0.92, respectively.

From what we discussed above, we defined four clusters according to the treatment procedure for lung cancer. They were: Cluster 1: Screening by Scan (X-Ray or Other), Cluster 2: Biopsy Examination, Cluster 3: MRI and Cluster 4: Doppler. We examined the relationship between the treatment categories and mortality. First, we extracted the cluster information from the data set we used to define the clusters and merged it with the adjusted admission table using the following code.

```
PROC SQL;
 CREATE TABLE MEDSTAT.inp_sur_analysis_cluster AS SELECT distinct
B.DOBYR, B.ADMDATE FORMAT=MMDDYY10., B.AGE, B.CASEID, B.DAYS,
B.DRG, B.TOTNET, B.TOTPAY, B.ADMTYP, B.DSTATUS, B.PATID, B.SEX,
B.lungcancer, B.diagnoses, B.procedures,B.AGEGRP,A._cluster_,
B.time, B. status
 FROM medstat.inp_sur_analysis AS B
        INNER JOIN MEDSTAT.medtextranks AS A  ON (A.caseid =
B.caseid);
QUIT;
```

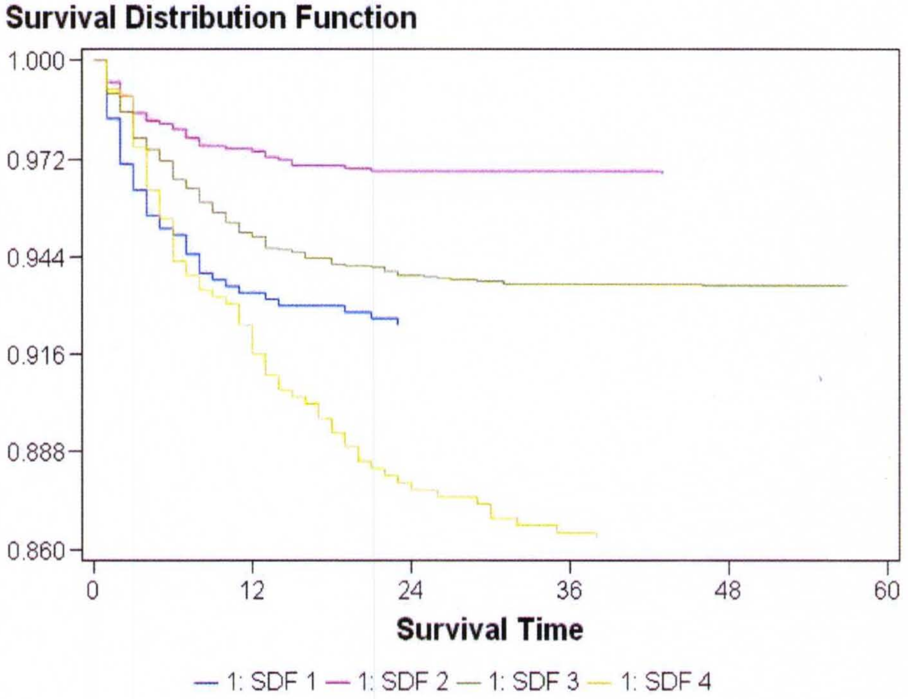
Table 7.10. Test of Equality over Strata of Treatment Clusters



Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr >Chi-Square**
Log-Rank	66.8540	3	<.0001
Wilcoxon	65.2475	3	<.0001
-2Log(LR)	67.3901	3	<.0001

Again, all three tests show a significant difference between the clusters (p-values are less than 0.001). Different clusters are related to different mortality rates of lung cancer.

Figure 7.11. Life Tables: Survivor Distribution Plot for Treatment clusters



Because these clusters were based on the treatment procedures, they indicated the severity level of the patient condition. Note that cluster 4 has a higher probability of mortality than the other three clusters, with a survival rate of 0.86 at the first 60 days. The survival rates for clusters 1, 2 and 3 are 0.92, 0.97 and 0.93, respectively. Furthermore, these four clusters might be divided into three categories. Clusters 1 and 3 have a similar survivor distribution according to survival time. The gaps between the categories become much clearer after 25 days.

7.5.2 Cox Proportional Hazard Model [34]

Survival analysis is appropriate for outcomes that occur during follow-up of patients, such as death because of cancer or recurrence of disease in cancer. In medical studies, the Cox proportional hazard model is the most often used method.

We used the following code to create some binary variables, including Admit type, Region, Discharge status, and clusters of treatment procedures.

```
data medstat.inp_sur;
set inp_surl;typ_surg=0; typ_med=0; typ_oth=0; prev_adm_180=0;
readm_90=0; Region_NE=0; Region_NC=0; Region_S=0;
egion_W=0;Region_un=0;
Dischg_home=0; dischg_transfer=0;
dischg_died=0;dischg_oth=0;age_grp1=0; age_grp2=0;
age_grp3=0;age_grp4=0;age_grp5=0;age_grp6=0;
cluster_Screening=0; cluster_Biopsy=0; luster_MRI=0;
cluster_Doppler=0;
cluster_oth=0;
if admtyp='1' then typ_surg=1;
else if admtyp='2' then typ_med=1;
else typ_oth=1;
if lastadm le 185 and lastadm gt 0 then prev_adm_180=1;
if nextadm le 90 and nextadm gt 0 then readm_90=1;
if region='1' then region_NE=1;
else if region='2' then region_NC=1;
else if region='3' then region_S=1;

else if region='4' then region_W=1;

else region_un=1;

if dstatus in ('01','06','50') then Dischg_home=1;

else if dstatus in

('02','03','04','05'/*,'51','61','71','72','99'*/) then

dischg_transfer=1;

else if dstatus in

('20','21','22','23','24','25','26','27','28','29') then

dischg_died=1;

else dischg_oth=1;

if age ge 0 and age le 17 then age_grp1=1;

else if age ge 18 and age le 34 then age_grp2=1;

else if age ge 35 and age le 44 then age_grp3=1;

else if age ge 45 and age le 54 then age_grp4=1;

else if age ge 55 and age le 64 then age_grp5=1;

else age_grp6=1;

if _cluster_=1 then cluster_Screening=1;

else if _cluster_=2 then cluster_Biopsy=1;

else if _cluster_=3 then cluster_MRI=1;

else if _cluster_=4 then cluster_Doppler=1;

else cluster_oth=1;

run;
```

Table 7.11. Hazard ratio for the mortality of lung cancer.

Analysis of Maximum Likelihood Estimates								
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
typ_oth	1	-12.51181	362.82221	0.0012	0.9725	0.000	0.000	2.52E303
typ_med	1	1.09393	0.17066	41.0861	<.0001	2.986	2.137	4.172
typ_surg	0	0	.	.	.	.	.	.
age_grp1	1	-12.75666	1520	0.0001	0.9933	0.000	0.000	.
age_grp2	1	-13.31160	401.87273	0.0011	0.9736	0.000	0.000	.
age_grp4	1	-0.20518	0.25797	0.6326	0.4264	0.815	0.491	1.350
age_grp5	1	-0.30263	0.24120	1.5742	0.2096	0.739	0.461	1.185
age_grp6	1	0.48064	0.38189	1.5840	0.2082	1.617	0.765	3.418
age_grp3	0	0	.	.	.	.	.	.
cluster_Biopsy	1	-0.33550	0.23919	1.9674	0.1607	0.715	0.447	1.143
cluster_MRI	1	-0.20162	0.17630	1.3078	0.2528	0.817	0.579	1.155
cluster_Doppler	1	0.71388	0.20039	12.6917	0.0004	2.042	1.379	3.024
cluster_Screening	0	0	.	.	.	.	.	.
prev_adm_180	1	-0.22023	0.16741	1.7305	0.1883	0.802	0.578	1.114

In the Cox proportional hazards model, the effects of the covariates are to act multiplicatively on the hazard of the survival time. From Output Table 7.11, the hazard ratio estimate for *prev_adm_180* is 0.802, meaning that the existence of previous admissions will shrink the hazard rate by 1-(0.802)=20%. For a CLASS variable parameter, the hazard ratio is the ratio of the hazard rates between the given category and the reference category. The hazard rate of

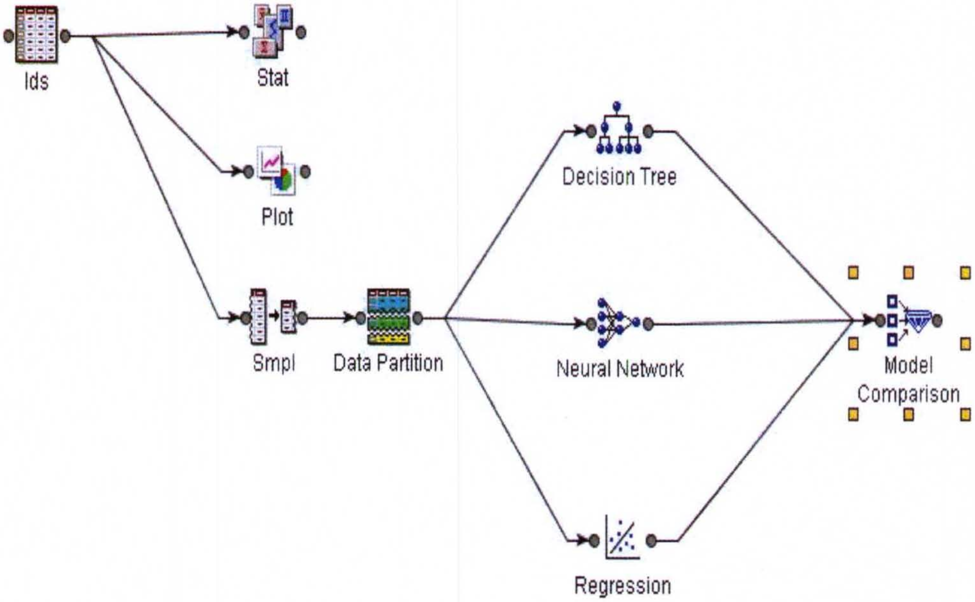
AdmitType=Medical is 299% that of AdmitType=Surgical, and the hazard rate of AdmitType=other is 0% that of AdmitType=Surgical. Similarly, the hazard rate of Cluster= Doppler is 200% that of Cluster=Screening; the hazard rate of Cluster= Biopsy is 72% that of Cluster=Screening and the hazard rate of Cluster= MRI is 82% that of Cluster=Screening.

From what we discussed above, the outcome of the Hazard Ratio analysis is consistent with the Kaplan-Meier method.

7.6. Predictive Modeling

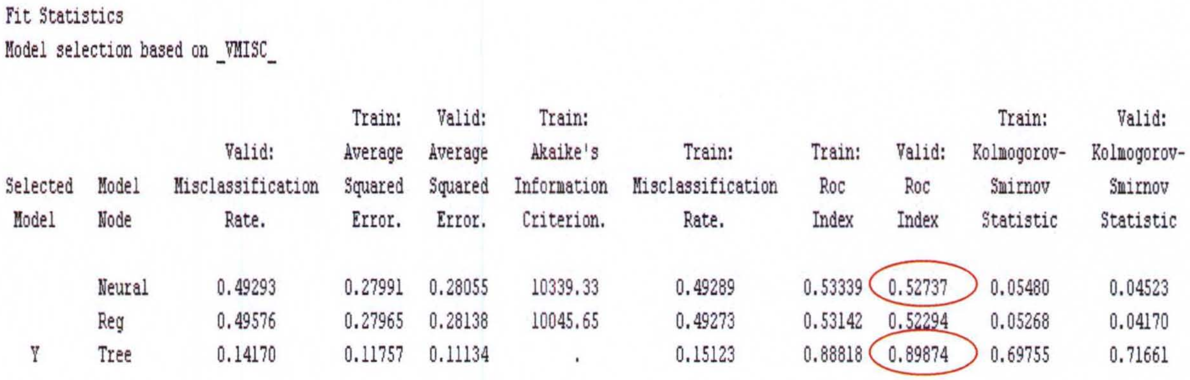
Next, we want to estimate the probability of the occurrence of lung cancer based on patient age, gender, clinical conditions, days of stay and total charge by using modeling techniques. Since lung cancer remains a rare occurrence, we use stratification as the sampling method and the sample proportion is 50/50. Figure 7.12 shows the predictive modeling process for Decision Tree, Regression and Neural Network in Enterprise Miner.

Figure 7.12. Predictive Modeling Process



Here, the model comparison node is used to compare the results of all the models to determine which model gives the most accurate or least costly results. Table 7.12 shows the model choice using the 50/50 proportion, stratified sampling and the misclassification rate.

Table 7.12. Model Choice with Misclassification Rate



```
Fit Statistics
Model selection based on _VMISC_

                                     Train:    Valid:     Train:                                                 Train:        Valid:
                      Valid:         Average   Average    Akaike's         Train:          Train:   Valid:   Kolmogorov-   Kolmogorov-
Selected  Model  Misclassification   Squared   Squared   Information  Misclassification    Roc      Roc        Smirnov       Smirnov
 Model    Node         Rate.         Error.    Error.    Criterion.         Rate.           Index    Index      Statistic     Statistic

          Neural      0.49293        0.27991   0.28055    10339.33        0.49289         0.53339  0.52737    0.05480       0.04523
          Reg         0.49576        0.27965   0.28138    10045.65        0.49273         0.53142  0.52294    0.05268       0.04170
   Y      Tree        0.14170        0.11757   0.11134        .           0.15123         0.88818  0.89874    0.69755       0.71661
```

Note that the Decision Tree is optimal with a 14.2% misclassification rate in the testing set. The ROC index for the Model Tree is 0.90 when it is 0.52 for the Neural Network and Regression models. Figures 7.13 and 7.14 give us the details of the Roc Curve and Decision Tree.

Figure 7.13. Roc Curve for model of Lung Cancer Occurrence.

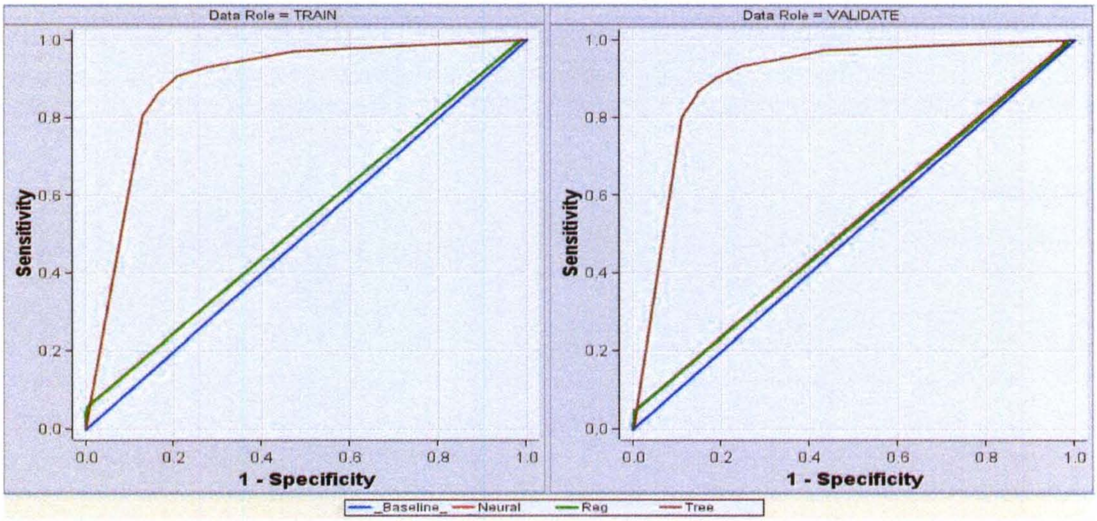
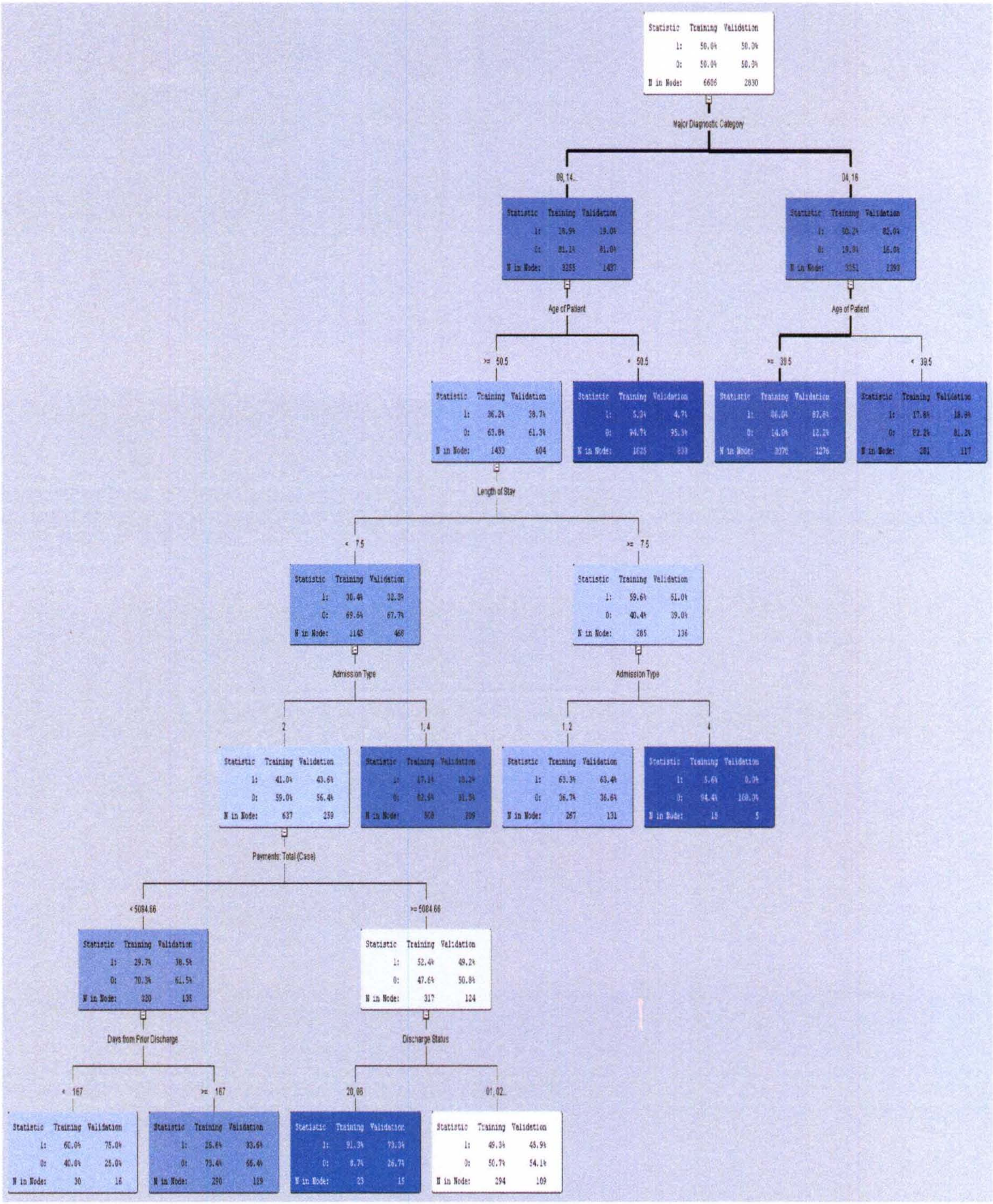


Figure 7.14. Decision Tree Results



Note that the Major Diagnostic Category is the first major predictor based on diagnosis with the Respiratory System or the Blood, Blood Forming Organs, and Immunological Disorders. The next split is based on the Age of the Patient at

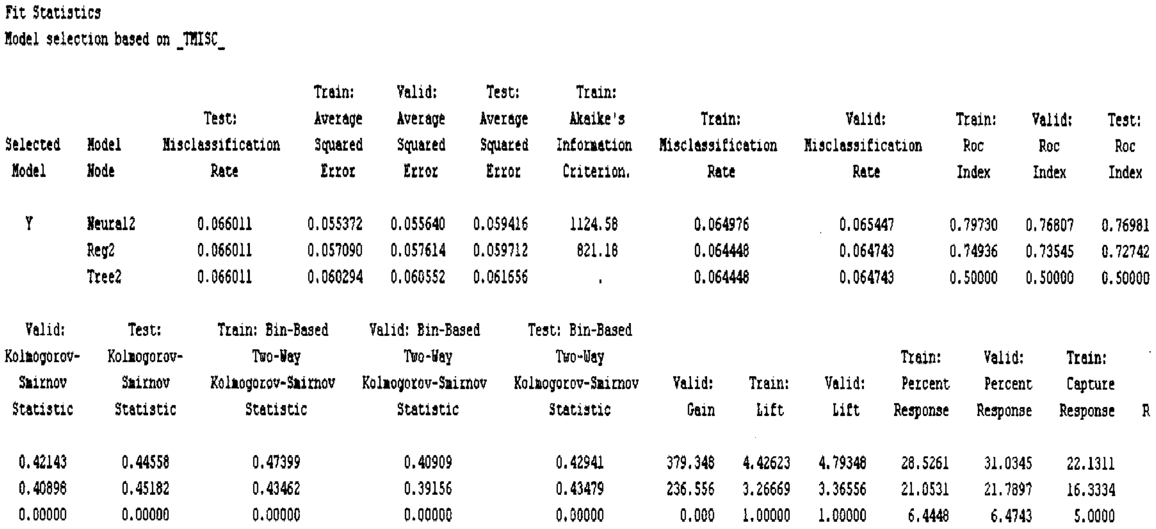
39.5 years old. With these two splits, we have reached 88% of the lung cancer cases with 1276 records. Here, a patient more than 40 years old is at higher risk of lung cancer compared to those less than 40 with the same diagnosis of Respiratory. The decision tree clearly shows that Major Diagnostic Category, Age of Patient and Length of Stay are the leading predictors of lung cancer diagnosis.

Next, we used several predictive models (Decision tree, Logistic regression and Neural Network) to analyze the mortality of lung cancer based on what we discussed above. Here, we list all the variables used in the model process as follows,

1. Length of stay
2. Age
3. Admit type
4. MDC code
5. Medication treatment Clusters
6. Previous Admission
7. Treatment procedure Clusters
8. Total Cost
9. Patient location
10. Diagnosis Clusters
11. Gender
12. Target variable: Death

First, we split the data into three sets with proportion of 40/30/30, training, validation and test sets. For the Neural network, there are three hidden units; multilayer preceptron architecture is used and the misclassification rate is selected as the model selection criteria. Also, the logistic regression model is used to compare with the Neural Network model. Again, the following table and figure show us the outcomes.

Table 7.13 Fit statistics for model selection

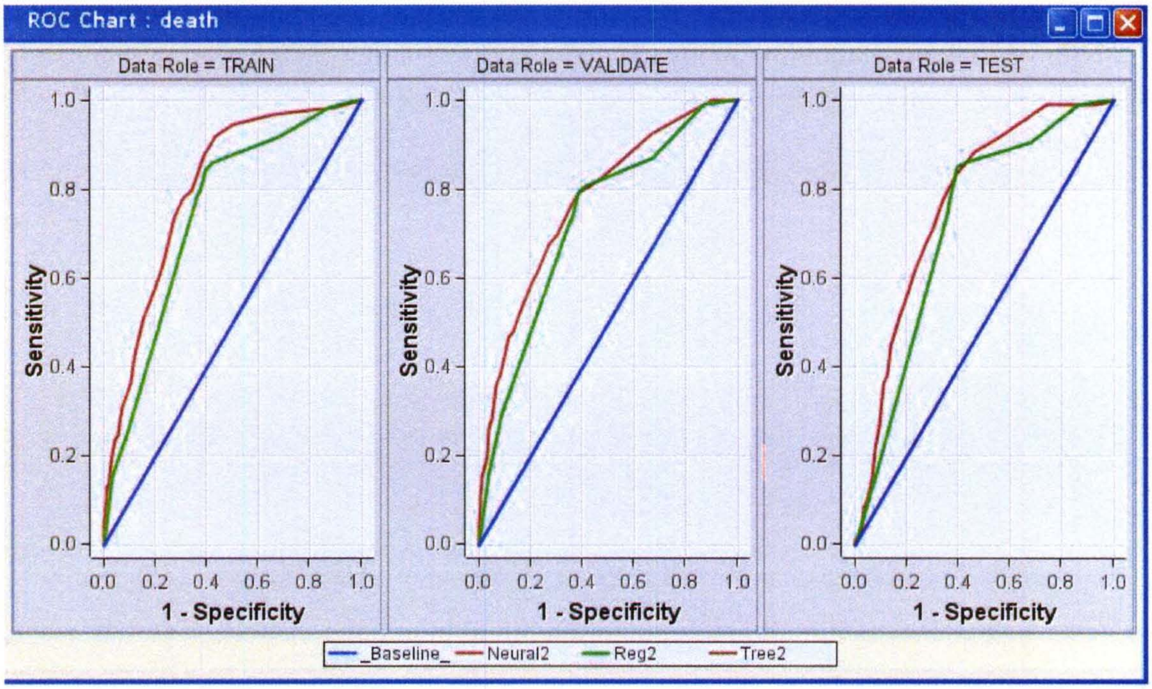


Fit Statistics
Model selection based on _TMISC_

Selected Model	Model Node	Test: Misclassification Rate	Train: Average Squared Error	Valid: Average Squared Error	Test: Average Squared Error	Train: Akaike's Information Criterion.	Train: Misclassification Rate	Valid: Misclassification Rate	Train: Roc Index	Valid: Roc Index	Test: Roc Index
Y	Neural2	0.066011	0.055372	0.055640	0.059416	1124.58	0.064976	0.065447	0.79730	0.76807	0.76981
	Reg2	0.066011	0.057090	0.057614	0.059712	821.18	0.064448	0.064743	0.74936	0.73545	0.72742
	Tree2	0.066011	0.060294	0.060552	0.061656	.	0.064448	0.064743	0.50000	0.50000	0.50000

Valid: Kolmogorov-Smirnov Statistic	Test: Kolmogorov-Smirnov Statistic	Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Test: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Valid: Gain	Train: Lift	Valid: Lift	Train: Percent Response	Valid: Percent Response	Train: Capture Response	R
0.42143	0.44558	0.47399	0.40909	0.42941	379.348	4.42623	4.79348	28.5261	31.0345	22.1311	
0.40898	0.45182	0.43462	0.39156	0.43479	236.556	3.26669	3.36556	21.0531	21.7897	16.3334	
0.00000	0.00000	0.00000	0.00000	0.00000	0.000	1.00000	1.00000	6.4448	6.4743	5.0000	

Figure 7.15 ROC curve for model selection



Note that the Neural Network won this time with a testing ROC of 77%. At the cutoff point of 0.07, we have an overall classification rate of 72% for the test data set (65 true positives out of 94 positive events). With the above 11 variables, the neural network model will generate a probability of mortality for each patient with lung cancer under different treatments.

7. 7 Summary

Relatively few models have been developed to estimate lung cancer risk. Previous lung cancer risk prediction models have tended to focus on smoking characteristics, sex, and age. In 2003, Peter B. Bach and Michael W. Kattan developed a lung cancer risk prediction model based on the Carotene and Retinol Efficacy Trial (CARET), a large, randomized trial of lung cancer prevention. This model used smoking history data of retinol and carotene in heavy smokers and asbestos-exposed individuals, which was applicable to smokers between 50 and 75 years of age, who are or were heavy smokers (10 – 60 cigarettes per day for 25 – 60 years). In 2007, to extend the work of Bach and Kattan and to include additional risk factors beyond smoking history and asbestos exposure, Margaret Spitz developed a risk prediction model for lung cancer using epidemiologic data for assessing lung cancer risk with a minimal number of risk predictors. In 2008, a logistic regression model was developed by Jennifer Beane and Paola Sebastiani using the biomarker, clinical factors, and these data combined were tested using the independent set of patients with non-diagnostic bronchoscopies. Most of these models are constructed using smoking history data and clinical data to predict the occurrence of lung cancer. Our models are constructed based on demographic data and clinical factors, such as treatment procedures and medication treatments, to identify the lung cancer and to determine the mortality under different treatments.

Based on text analysis on the diagnosis codes and KDE, the relationships between the variables and the target, Lung Cancer, were examined. With the *Kaplan-Meier method*, Age, Admission Type and Procedure clusters are used to identify the significant factors for the diagnosis of Lung Cancer. The Cox Hazard Proportion Model yields the detail of the relationships by the hazard ratio of the contribution of each variable to the target. Then, predictive modeling was used to investigate the rules between the variables and the target. Based on the tests and the ROC curve, the Decision tree is selected as the optimal model to predict the occurrence of lung cancer, and the Neural Network model is the optimal model to predict the death because of lung cancer with different treatments.

CHAPTER VIII CONCLUSION

8.1 Methods Comparison

Since the first risk predictive model for chronic disease was developed in 1976, many models have been developed to predict the occurrence of a disease or the mortality of patients. However, relatively few models have been developed to predict the risk of Lung Cancer in past years. Moreover, the previous predictive models for lung cancer focused on smoking characteristics, sex, and age.

As far as Lung Cancer is concerned, F L Rice and R Park analyzed the relation between mortality from Lung Cancer and cumulative crystalline silica exposure with Poisson regression and Cox's proportional hazards models. In their paper, several model forms for Poisson regression were evaluated: Log linear, Log square root, Log quadratic, Power, Linear relative rate, Shape (α) and Additive excess rate. The Cox's hazards models were applied as alternative methods to examine the relationship. They got very similar results from these methods, which show a significant exposure-response relation between the mortality from Lung Cancer and the silica exposure [40].

During 2002, F. D. K. Liddell and B.G. Armstrong analyzed the effects on Lung Cancer of cigarette smoking and exposure in Quebec Chrysotile miners and millers by using principal components analysis and linear and log-linear models. A standardized mortality ratio (SMR) was defined by the number of lung cancer deaths observed and the numbers of expected in the subdivision. The SMR is the dependent variable for the linear and log-linear models.

In 2003, a predictive model was developed by Peter B. Bach and Michael W. Kattan to estimate the risk of Lung Cancer using the data from the Carotene and Retinol Efficacy Trial (CARET), which is a large and randomized trial of lung cancer prevention [2]. In their research, smoking history data of retinol and carotene in heavy smokers and asbestos-exposed individuals were used to train and validate the model that was applicable to heavy smokers at the age of 50 or older only. Two 1-year models were developed with similar methods and predictors to predict the probability of occurrence of lung cancer and the mortality without having been diagnosed with lung cancer. The Cox proportional hazards regression models were used to examine the relationship between the occurrence of Lung Cancer and the predictors. However, Bach's lung cancer risk model was only focused on seniors with a smoking history and was not able to distinguish the different histological types of Lung Cancer.

Then, in 2007, to extend the work of Bach and Kattan, Margaret Spitz developed a risk prediction model for Lung Cancer using epidemiologic data for assessing lung cancer risk with a minimal number of risk predictors [41]. In her research, Spitz also collected patients' demographic information, medical history (including physical conditions), and family history of cancer besides smoking history. In the article, the multivariable logistic regression model with backward selection was used to estimate the risk of lung cancer. The classification and regression tree (CART) methods were used to evaluate the higher order interactions in the training sets. However, the limitation for Margaret's work is that the data were derived from a single large case control study; that is, the case patients were recruited from a single tertiary cancer center and the control group was not population based. Then, Spitz expanded her work by adding two markers of DNA repair capacity in 2008. Again, multivariable logistic models were constructed to estimate the individual risk of Lung Cancer. As a conclusion of this article, the biomarker assays improved the sensitivity of the models over epidemiologic and clinical data only.

Meantime, a novel data mining technique, Backward-Chaining Rule Induction (BCRI), was introduced by Mary E. Edgerton, Douglas H. Fisher for Gene Networks relevant to poor prognosis in Lung Caner. BCRI was embedded in a semi-supervised approach, C4.5, which is a method for learning decision trees. In their paper, C45W-BCRI (Wrapper-based implementation of BCRI with C4.5) was applied to generate the rules to estimate the long versus short survival periods for Lung Cancer. Seventeen rules were generated with gene interactions out of 19 molecular species, and 12 were associated with Neoplasia in general and 5 for Lung Cancer [42].

In 2008, three logistic regression models were developed by Jennifer Beane and Paola Sebastiani using the biomarker, clinical factors, and these combined data were tested using the independent set of patients with non-diagnostic bronchoscopies [43]. The combined clinicogenomic model had better performance and got higher specificity and positive predictive value compared with the other models.

Dursun Delen developed predictive models for survivability and variable explanation using Logistic Regression, Decision Trees, Artificial Neural Networks (ANN) and Support Vector Machines (SVMs) based on the data from the SEER Program of the National Cancer Institute during 2009. From their results, SVM models had the best performance, and the ANN models still performed better that other two modeling techniques [44].

As discussed above, several statistical model techniques were used to estimate the risk of Lung Cancer and smoking history was the most important predictor. Since the complexity of medical data, some information such as diagnosis, medication, or treatment procedures were not used to examine the relationship with the risk of Lung Cancer.

In this dissertation, we used Text mining to look for patterns and group patients with similar condition or similar treatments into the same clusters, creating new categorical variables for modeling.

Medical data used in this dissertation are from a collection of claims for each patient. Each record includes the patient identification (Name and/or medical record), demographic information, clinical information, and cost information. Some information such as medication, diagnosis, and treatments are in textual format; some such as age, length of stay and costs are in numerical format. All of these characteristics make medical data analysis a challenging task. Krzysztof J. Cios and G. William Moore gave the major uniqueness of medical data in their article [45].

Medication, diagnosis, and treatments constitute the highest source of information about the patient in the data. Unfortunately, they cannot be used in the analysis as they stand. Hence, preprocessing of these variables is necessary to make them understandable by the models. Patients that have similar symptoms and diagnoses should have similar treatments, and can be grouped together according to these treatments. Therefore, text analysis was applied to allow those variables to be used with other techniques.

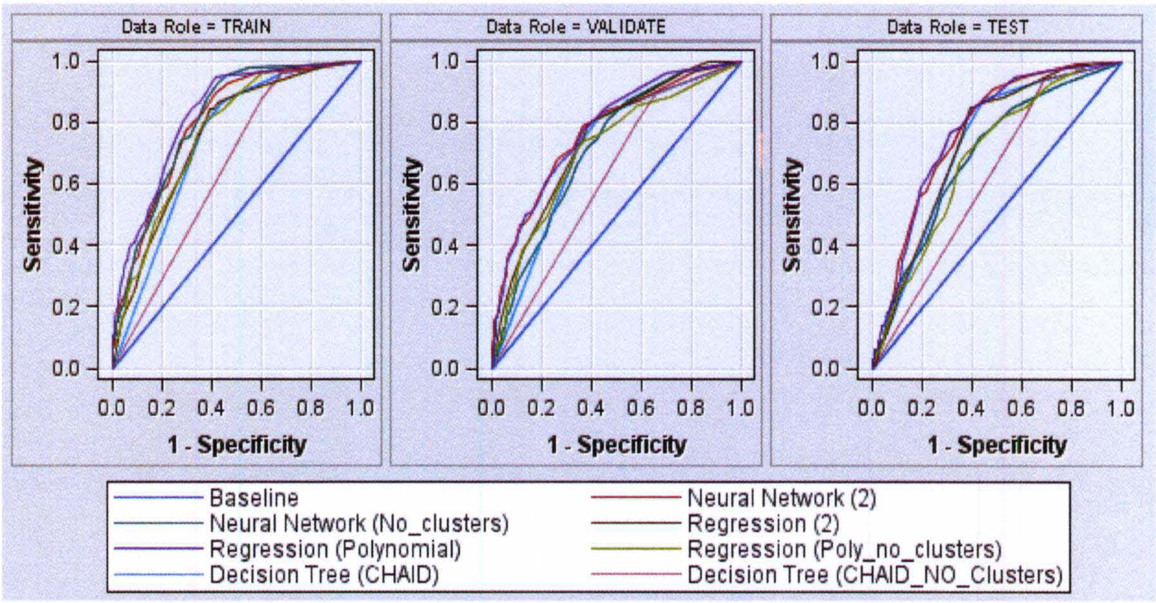
We applied Text Miner with Cluster analysis to identify the claims data for Lung Cancer and to determine the category of diagnosis, treatment procedures and medication treatments for those patients. Moreover, they were used to define severity level and treatment categories. Compared with using diagnosis codes directly, our method is more efficient and captures more useful information for further analysis. A decision tree was built to generate rules for identifying high risk lung cancer cases amongst the regular inpatient population. Decision trees provide interpretable decision rules and logic statements for a good understanding of the model. This model can then be used to predict the class to which an unseen object belongs by using an equivalent set of rules that are often easier to understand than the tree itself.

In order to analyze the mortality of Lung Cancer, we also found that survival analysis is appropriate to preprocess the data for the relationship between a predictor variable of interest and the time of an event. The proportional hazard model examined the effects of different treatment clusters using the hazard ratio and the proportional effect of a treatment cluster (treatment procedure or medication treatment) that may vary with time. Several statistical models were developed based on demographic data and clinical factors, such as treatment procedures and medication treatments, to determine the mortality under different treatments.

Among these techniques that were used for the analysis of the dataset, the Polynomial Regression Model, Neural Network, Regression Model and Decision

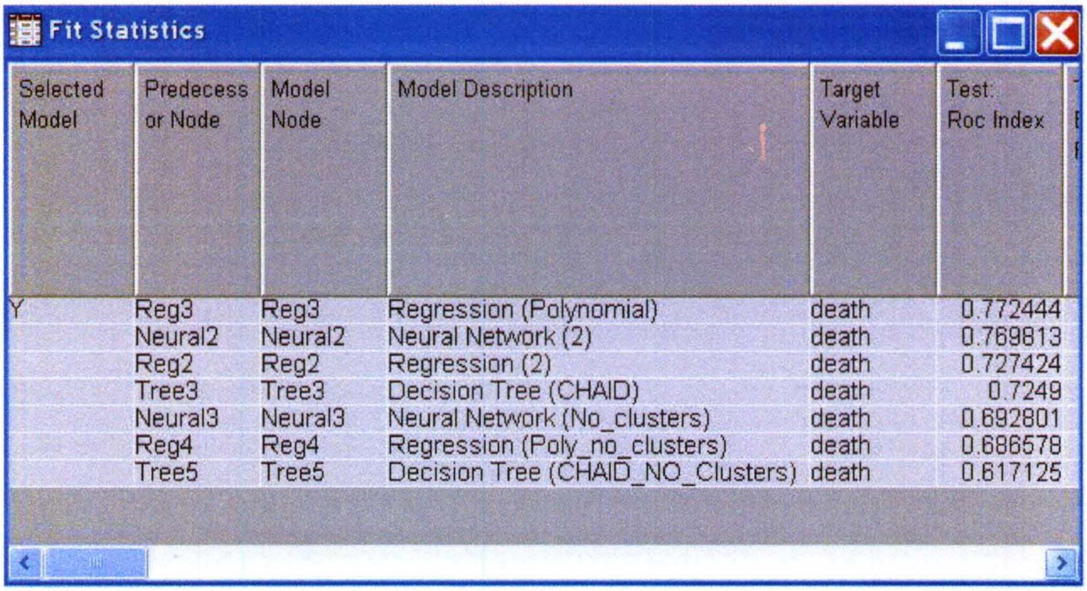
Tree with the CHAID algorithm provided insight into the data. ROC curves were used to compare the performance of the models with or without diagnosis, medication and treatments. The models with clusters of diagnosis, medication and treatments generated by Text mining had better performance than the models excluding those variables (Figure 8.1).

Figure 8.1 The comparison of ROC curves



The models with text mining clusters, polynomial regression (purple), Neural network (red) and Decision Tree (light blue), have better performance than others, which were the top category in the graph. The models without the cluster variables, polynomial regression (green), Neural network (dark blue) and Decision Tree (gray), were in the middle of the graph. From the test dataset, it was clear that there was a significant difference in the area under the curves. The following table showed the detail about the difference of the ROC index.

Table 8.1 The ROC index for the models



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Test: Roc Index
Y	Reg3	Reg3	Regression (Polynomial)	death	0.772444
	Neural2	Neural2	Neural Network (2)	death	0.769813
	Reg2	Reg2	Regression (2)	death	0.727424
	Tree3	Tree3	Decision Tree (CHAID)	death	0.7249
	Neural3	Neural3	Neural Network (No_clusters)	death	0.692801
	Reg4	Reg4	Regression (Poly_no_clusters)	death	0.686578
	Tree5	Tree5	Decision Tree (CHAID_NO_Clusters)	death	0.617125

The differences with regard to the cluster variables were about 9%, 7%, and 11% for Polynomial regression models, Neural Network and Decision trees with the CHAID algorithm, respectively. The information from diagnosis, medication and treatments improved the predictive capability of these models.

8.2 Conclusion

The main aim of this dissertation was to examine the relationships between the patient's conditions and the outcome of Lung cancer and to develop, evaluate and apply specific Lung cancer risk prediction models for the prediction of the cost and mortality of treatment, or developing rules for identifying the diagnosis of Cancer, which will help clinical decision-making or reduce cost.

Two large databases, the National Inpatient Sample for the years of 2000 to 2004 and the Thomson MedStat MarketScan data containing all patient claims for 40 million people followed for the years 2000-2001 were used to examine our aim.

The Kernel Density Estimation (KDE procedure) was used to examine the relationships between Lung cancer and Age, Length of Stay and Total Charges by using data visualization. With text analysis on the diagnosis codes and KDE, it shows that the malignant neoplasm of lobe, bronchus or lung is of higher risk and has a higher cost compared to other lung cancers.

Text mining and cluster analysis were used to reduce a large number of patient condition codes into cluster categories with different severity levels based on diagnosis codes or treatment procedure codes. These categories were used to examine the relationship to the costs and mortality because of cancer.

Based on the outcomes we discussed above, a Time series model and Logistic Model were applied. The ARIMA model with ordinary and dynamic regressors for the inflation rate was used to analyze the hospital's financial data. It provides the hospital with the ability to predict total charges of lung cancer based on previous costs. The ordinary and dynamic regressors model showed the effect of the length of stay and age on the predicted values of total charges.

The Logistic model was used to examine the relationship between death and conditions of patients with lung diseases. In this paper, we just focused on Age in years at admission, Admission type, Elective versus non-elective admission, Length of stay, Total charges, and Median household income for patient. By using the backward elimination method, we removed Admission type, Elective versus non-elective admission and Median household income for the patient, which are not statistically significant. Then we refit the data with Age, Los and Totchg effects that are all highly statistically significant.

With the *Kaplan-Meier method*, Age, Admission Type and Procedure clusters from the Thomson MedStat MarketScan data are significant for the diagnosis of Lung Cancer. The Cox Hazard Proportion Model gave us the detail of the relationships by hazard ratio, the contribution of each variable to the target. Then, predictive modeling was used to investigate the rules between the variables and the target. Based on the tests and ROC curve, Decision tree and Neural Network models are selected as the optimal model for occurrence and mortality of lung cancer, respectively.

Throughout our analysis, some significant factors for lung cancer are examined and diagnosis cluster, treatment procedure cluster and medication cluster are created based on the severity level of lung cancer. Specific lung cancer risk prediction models based on available variables were developed, evaluated and used to predict the cost and mortality for the patients and to build rules for identifying the diagnoses of Cancer, which will help the physician's clinical decision making (treatment procedure and medication), to identify high-risk individuals and to provide better care.

REFERENCE

1. Field, John, Lung Cancer Risk Models Come of Age, Cancer Prev Res 2008;1(4) September 2008.

2. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003 ; 95 : 470 – 478.

3. Andrew N. Freedman , Daniela Seminara , Mitchell H. Gail , Cancer Risk Prediction Models: A Workshop on Development, Evaluation, and Application, Journal of the National Cancer Institute, Vol. 97, No. 10, May 18, 2005.

4. Stratton, M.R., Campbell, P.J., The Cancer Genome, Nature 458, 719-24, 2009.

5. Anirban Mahapatra, Lung Cancer-Genomics and Personalized Medicine, American Chemical Society, 2010.

6. Agency for Healthcare Research and Quality, http://www.ahrq.gov.

7. MarketScan Research Database, Thomson MedStat; Ph.D. Dissertation Support Program. Cited; Available from: http://www.medstatmarketscan.com/.

8. Parkin, D. (2005). Max Global Cancer Statistics, 2002. *CA Cancer J Clin* 2005, 55, 74-108.

9. Thompson Cancer Survival Center (2008). Lung Cancer Diagnosis and Staging. From Thompson Cancer Survival Center web site http://www.thompsoncancer.com/tcsc-lungcancer-diagnosis.cfm.

10. M. Patricia Rivera, Frank Detterbeck. Diagnosis of Lung Cancer: The Guidelines. Chest, 2003;123;129-136.

11. Cancer Research UK, http://www.cancerhelp.org.uk/help/default.asp?page=6706#typestypes.

12. Thompson, JR, Lung Cancer Update: Diagnosis and Treatment. Pulmonary Associates, PSC. 2005.

13. Hopenhayn-Rich, Claudia (2001). Lung cancer in the commonwealth: A closer look at the data. *Lung Cancer Policy Brief 2001.* Volume 1, Issue 2.

14. ICD9 Website, http://icd9cm.chrisendres.com/.

15. Patricia Cerrito, Introduction to Data Mining: Using SAS Enterprise Miner. Department of Mathematics, 2007, University of Louisville.

16. George Fernández, Data Mining Using SAS Applications, chapman & Hall, 2003.

17. Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, IEEE Press, 2003.

18. Vipin Kumar, Text Mining Classification, Clustering, and Applications, Chapman & Hall, 2009.

19. SAS Enterprise Miner Help, http://support.sas.com/documentation/onlinedoc/miner/.

20. Michael E. Wall, Andreas Rechtsterner, Singular value decomposition and principal component analysis, Kluwer: Norwell, MA, 2003,pp: 91-109.

21. Tyler Smith, and Besa Smith, Survival Analysis And The Application Of Cox's Proportional Hazards Modeling Using SAS, SUGI, 2001.

22. SAS/STAT® User's Guide, Version 8,1999.

23. Mara Tableman, Jong Sung Kim. Survival Analysis Using S, Chapman & Hall, 2004.

24. Regina C. Elandt, Norman L. Johnson, Survival Models and Data Analysis, JOHN WILEY & SONS, 1980.

25. Viv Bewick, Liz Cheek, Survival analysis, *Critical Care* 2004, 8:389-394.

26. Margaret R. Spitz, Waun Ki Hong, A Risk Model for Prediction of Lung Cancer. J Natl Cancer Inst 2007; 99 : 715 – 726.

27. Peter Brockwell, Richard Davis, Time Series: theory and methods, Springer-Verlag, 1987.

28. Rafal Weron, Modeling and Forecasting Electricity Loads and Prices, JOHN WILEY & SONS, 2007.

29. Randall Matignon, Data Mining Using SAS Enterprise Miner, JOHN WILEY & SONS,2007 .

30. Paul Allison, Logistic regression using the SAS system, SAS, 2001.

31. David Hosmer, Stanley Lemesbow, Applied Logistic Regression, JOHN WILEY & SONS,2000.

32. Ewout W Steyerberg, Clinical Prediction Models, Springer,2009.

33. Leland Wilkinson, Tree Structured Data Analysis: AID, CHAID and CART, Sawtooth/SYSTAT Joint Software Conference 1992.

34. SAS 9.1 Help Menu, http://support.sas.com/onlinedoc/913/docMainpage.jsp.

35. Inflation Data Website. [InflationData.com is published by Financial Trend Forecaster®] [cited; Available from: http://inflationdata.com.

36. Anne Senter, Time Series Analysis, http://userwww.sfsu.edu/~efc/classes/biol710/timeseries/timeseries1.htm.

37. David M Adamson, Stella Chang, Health Research Data for the Real World: The MarketScan Databases, Thomson Healthcare, 2008.

38. US. Food and Drug Administration website, www.fda.org; used to concert NDC Numbers [cited; Available from: http://www.fda.gov/cder/ndc/database/Default.htm.

39. Eleuteri A, T.R., Milano L, et al, *Survival analysis and neural networks. Paper presented at 2003 Conference on Neural Networks; Portland, Oregon.* 2003.

40. F L Rice, R Park, Crystalline silica exposure and lung cancer mortality in diatomaceous earth industry workers: a quantitative risk assessment, Occup Environ Med, 2001; 58:38-45.

41. Margaret R. Spitz, Carol J. Etzel, An Expanded Risk Prediction Model for Lung Cancer. Cancer Prev Res 2008; 1(4) September 2008.

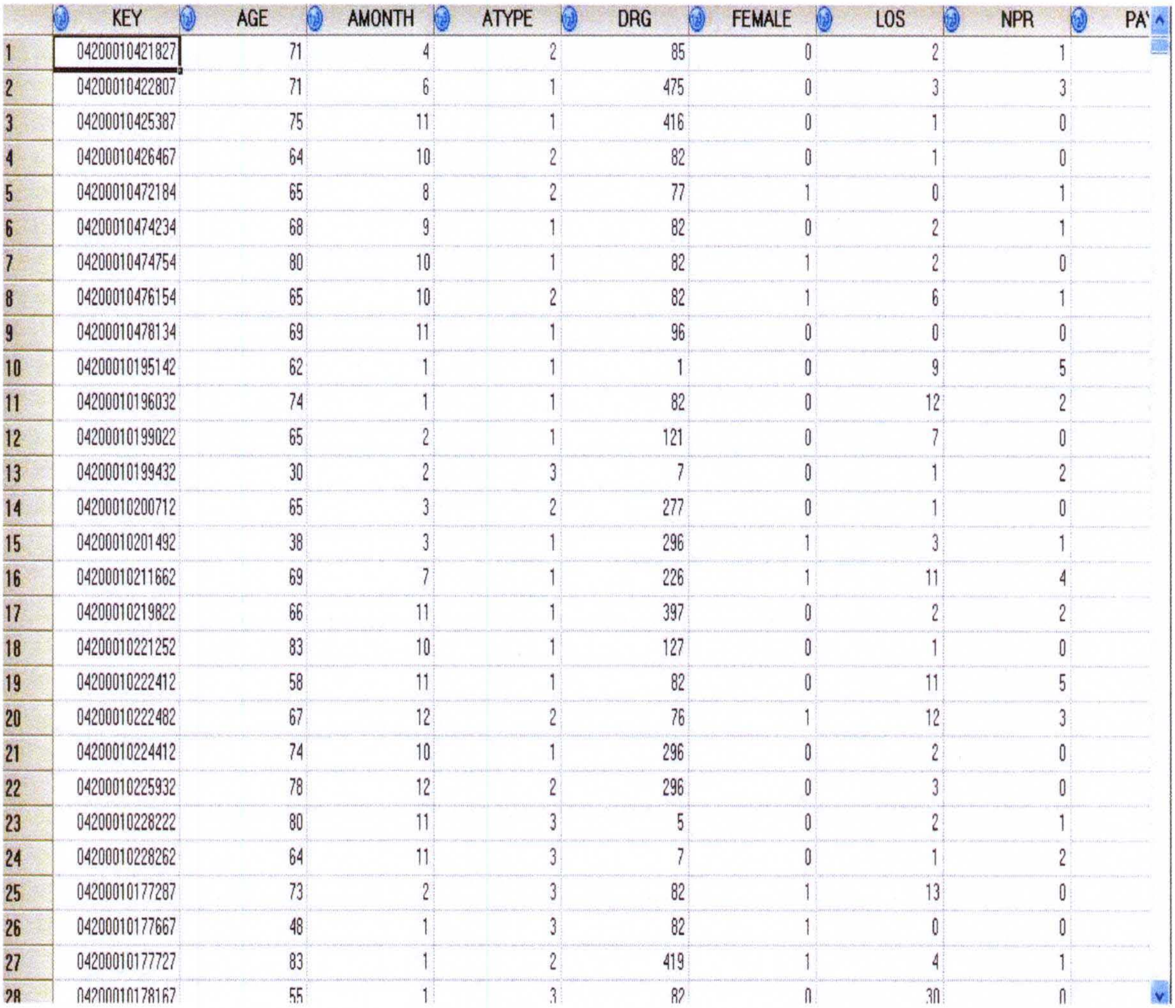
42. Mary E. Edgerton, Douglas H. Fisher, Data Mining for Gene Networks Relevant to Poor Prognosis in Lung Cancer Via Backward-Chaining Rule Induction, Cancer Informatics, 2007; 3:93-114.

43. Jennifer Beane, Paola Sebastiani, A prediction Model for Lung Cancer Diagnosis that Integrates Genomic and Clinical Features. Cancer Prev Res 2008; 1(1) June 2008.

44. Dursun Delen, Analysis of Cancer Data: a Data Mining Approach. Expert Systems, Feb 2009, Vol 26, No. 1.

45. Krzysztof J. Cios, G. William Moore, Uniqueness of Medical Data Mining, Artificial Intelligence in Medicine, 2002.

APPENDIX A. Sample Data Elements for National Inpatient Sample (NIS) Database (First 8 columns out of 129).

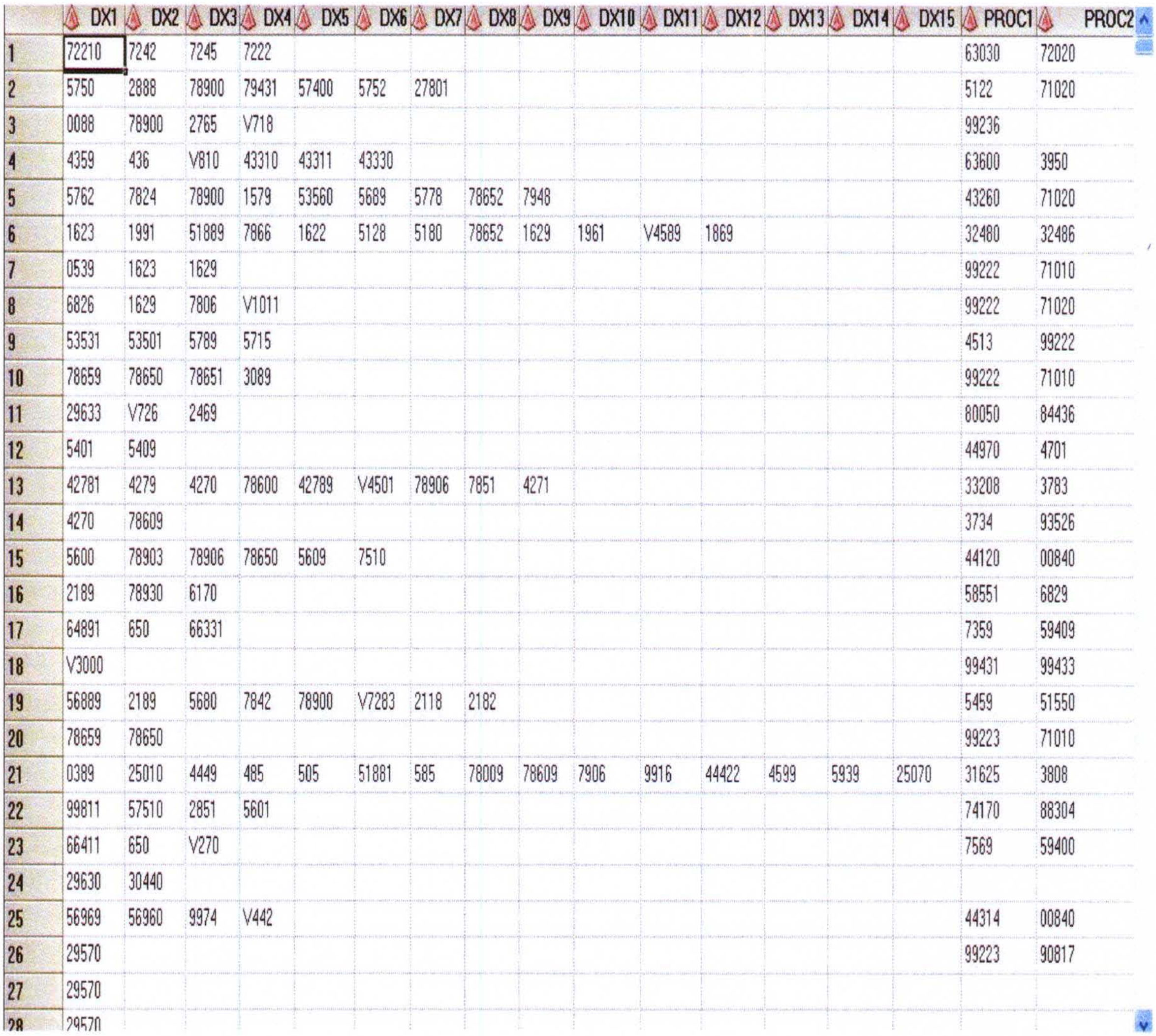


	KEY	AGE	AMONTH	ATYPE	DRG	FEMALE	LOS	NPR	PA'
1	04200010421827	71	4	2	85	0	2	1	
2	04200010422807	71	6	1	475	0	3	3	
3	04200010425387	75	11	1	416	0	1	0	
4	04200010426467	64	10	2	82	0	1	0	
5	04200010472184	65	8	2	77	1	0	1	
6	04200010474234	68	9	1	82	0	2	1	
7	04200010474754	80	10	1	82	1	2	0	
8	04200010476154	65	10	2	82	1	6	1	
9	04200010478134	69	11	1	96	0	0	0	
10	04200010195142	62	1	1	1	0	9	5	
11	04200010196032	74	1	1	82	0	12	2	
12	04200010199022	65	2	1	121	0	7	0	
13	04200010199432	30	2	3	7	0	1	2	
14	04200010200712	65	3	2	277	0	1	0	
15	04200010201492	38	3	1	296	1	3	1	
16	04200010211662	69	7	1	226	1	11	4	
17	04200010219822	66	11	1	397	0	2	2	
18	04200010221252	83	10	1	127	0	1	0	
19	04200010222412	58	11	1	82	0	11	5	
20	04200010222482	67	12	2	76	1	12	3	
21	04200010224412	74	10	1	296	0	2	0	
22	04200010225932	78	12	2	296	0	3	0	
23	04200010228222	80	11	3	5	0	2	1	
24	04200010228262	64	11	3	7	0	1	2	
25	04200010177287	73	2	3	82	1	13	0	
26	04200010177667	48	1	3	82	1	0	0	
27	04200010177727	83	1	2	419	1	4	1	
28	04200010178167	55	1	3	82	0	30	0	

APPENDIX B. List of data variables for Thomson MarketScan Commercial Database

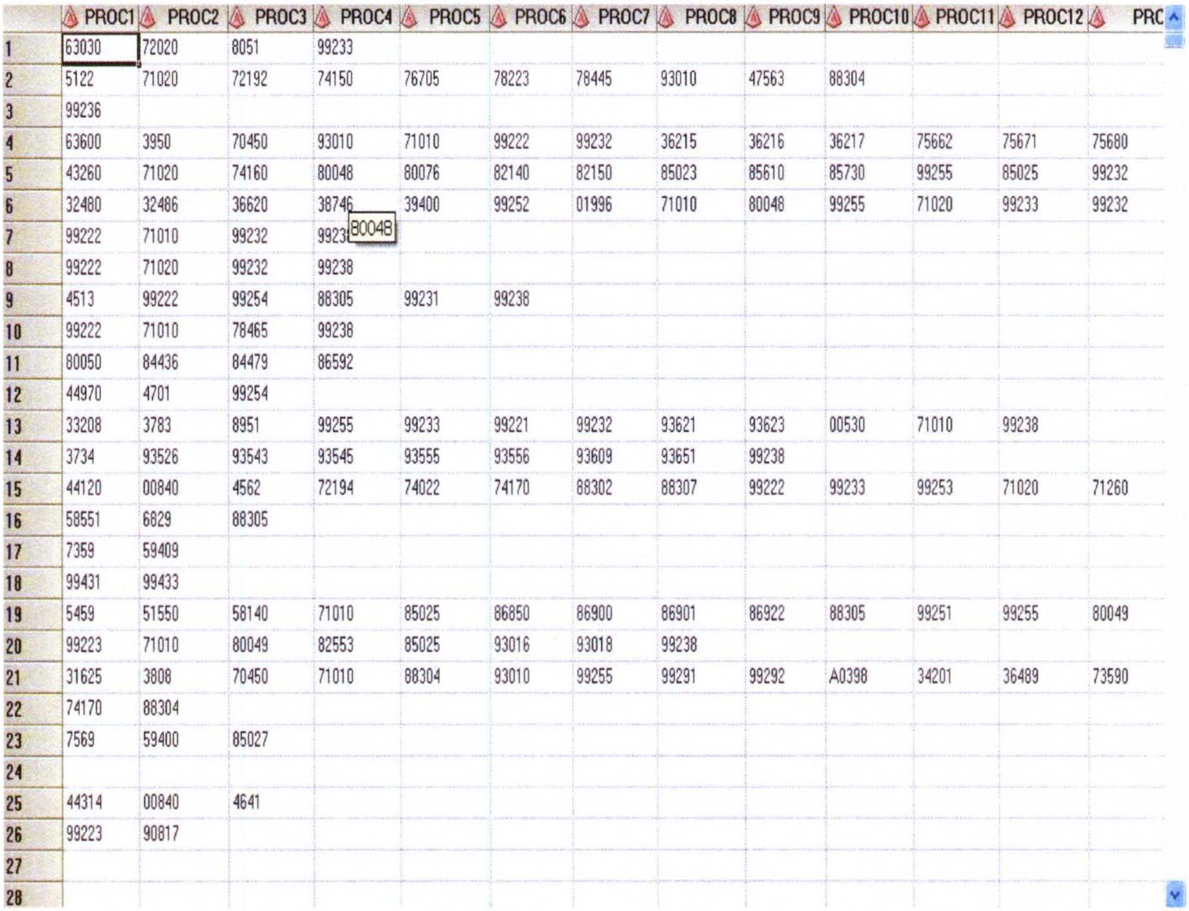
Demographic	Medical Information (Inpatient and Outpatient)	Health Plan Features	Financial Information	Drug Information	Enrollment Information
Patient ID	Admission date and type	Coordination of benefits amount	Total payments	Generic product ID	Date of enrollment
Age	Principal diagnosis code	Deductible amount	Net payments	Average wholesale price	Member days
Gender	Discharge status	Copayment amount	Payments to physician	Prescription drug payment	Date of disenrollment
Employment status and classification (hourly, etc.)	Major diagnostic category	Plan type	Payment to hospital	Therapeutic class	
Relationship of patient to beneficiary	Principal procedure code		Payments—total admission	Days supplied	
Geographic location (state, ZIP Code)	Secondary diagnosis codes (up to 14)			National drug code	
Industry	Secondary procedure codes (up to 14)			Refill number	
	DRG			Therapeutic group	
	Length of stay				
	Place of service				
	Provider ID				
	Quantity of services				

APPENDIX C. Sample Data Elements for Thomson MarketScan Commercial Database (15 Diagnosis columns out of 82).



	DX1	DX2	DX3	DX4	DX5	DX6	DX7	DX8	DX9	DX10	DX11	DX12	DX13	DX14	DX15	PROC1	PROC2
1	72210	7242	7245	7222												63030	72020
2	5750	2888	78900	79431	57400	5752	27801									5122	71020
3	0088	78900	2765	V718												99236	
4	4359	436	V810	43310	43311	43330										63600	3950
5	5762	7824	78900	1579	53560	5689	5778	78652	7948							43260	71020
6	1623	1991	51889	7866	1622	5128	5180	78652	1629	1961	V4589	1869				32480	32486
7	0539	1623	1629													99222	71010
8	6826	1629	7806	V1011												99222	71020
9	53531	53501	5789	5715												4513	99222
10	78659	78650	78651	3089												99222	71010
11	29633	V726	2469													80050	84436
12	5401	5409														44970	4701
13	42781	4279	4270	78600	42789	V4501	78906	7851	4271							33208	3783
14	4270	78609														3734	93526
15	5600	78903	78906	78650	5609	7510										44120	00840
16	2189	78930	6170													58551	6829
17	64891	650	66331													7359	59409
18	V3000															99431	99433
19	56889	2189	5680	7842	78900	V7283	2118	2182								5459	51550
20	78659	78650														99223	71010
21	0389	25010	4449	485	505	51881	585	78009	78609	7906	9916	44422	4599	5939	25070	31625	3808
22	99811	57510	2851	5601												74170	88304
23	66411	650	V270													7569	59400
24	29630	30440															
25	56969	56960	9974	V442												44314	00840
26	29570															99223	90817
27	29570																
28	29570																

APPENDIX D. Sample Data Elements for Thomson MarketScan Commercial Database (15 Procedure columns out of 82).



	PROC1	PROC2	PROC3	PROC4	PROC5	PROC6	PROC7	PROC8	PROC9	PROC10	PROC11	PROC12	PRC
1	63030	72020	8051	99233									
2	5122	71020	72192	74150	76705	78223	78445	93010	47563	88304			
3	99236												
4	63600	3950	70450	93010	71010	99222	99232	36215	36216	36217	75662	75671	75680
5	43260	71020	74160	80048	80076	82140	82150	85023	85610	85730	99255	85025	99232
6	32480	32486	36620	38746	39400	99252	01996	71010	80048	99255	71020	99233	99232
7	99222	71010	99232	9923 80048									
8	99222	71020	99232	99238									
9	4513	99222	99254	88305	99231	99238							
10	99222	71010	78465	99238									
11	80050	84436	84479	86592									
12	44970	4701	99254										
13	33208	3783	8951	99255	99233	99221	99232	93621	93623	00530	71010	99238	
14	3734	93526	93543	93545	93555	93556	93609	93651	99238				
15	44120	00840	4562	72194	74022	74170	88302	88307	99222	99233	99253	71020	71260
16	58551	6829	88305										
17	7359	59409											
18	99431	99433											
19	5459	51550	58140	71010	85025	86850	86900	86901	86922	88305	99251	99255	80049
20	99223	71010	80049	82553	85025	93016	93018	99238					
21	31625	3808	70450	71010	88304	93010	99255	99291	99292	A0398	34201	36489	73590
22	74170	88304											
23	7569	59400	85027										
24													
25	44314	00840	4641										
26	99223	90817											
27													
28													

CURRICULUM VITAE

Guoxin Tang

EDUCATION:

·2007-2010 Applied Mathematics (Data Mining and Statistics), University of Louisville, Ph.D.

·2005-2007 Applied Mathematics, University of Louisville, M.A.

·2001-2004 Operations Research, Tianjin University, M.S.

·1997-2001 Applied Mathematics, Tianjin University, B.S.

RESEARCH/WORK EXPERIENCE:

·2009-present Metrician, Clinical Outcomes and Data Acquisition ("CODA") in Humana. Inc. Develop statistical models and methodologies for inpatient readmission project to predict readmissions and provide expertise and assistance in implementing model results in ongoing business processes.

·2008 *Data Mining and Analysis of Lung Cancer for Thomson Medstat Marketscan data*. The purpose is to investigate the relationship between conditions of lung cancer patients and patient outcomes for those undergoing different treatment procedures. Several clusters were defined based on the procedures of treatment, kernel density estimation was use to make a comparison of age, length of stay and cost by clusters and predictive models were used to examine the patient events of lung cancer.

·2008 *Analysis and Forecasting Total Costs for Lung Cancer.* Lung Cancer data were extracted from the NIS by using CATX and RXMATCH functions; then, statistical models were established to examine the relationship between total charges and Age in years at admission, Diagnosis Related Group, Length of stay and Died during hospitalization. To forecast the costs for future, the appropriate ARIMA model was used with the inflation rate, Age and length of stay as dynamic regressors to model the data which was more reliable and gave a trend of costs with more certainties.

·2007 *Analysis and Forecasting for the S&P Stock Price Index Monthly Average Data with Inflation Rate.* To investigate and forecast the trend of the S&P 100 index to determine the future price movement with consideration of the inflation rate, ARIMA model with the first difference based on the inflation rate as dynamic regressor gave us more accurate predictions.

·2007 *Text and Data Mining to Investigate Expenditures of Prescribed medicines.* Text Miner and Concept links were used to examine combinations of medications used in relationship to patient therapeutics. We also used the linear model to examine the relationship between self cost and total cost, Medicare cost, Medicaid cost, private insurance.

·2006 *Data Mining to Investigate University Expectations of Work.* Course syllabi from representative departments were collected to investigate variability in course requirements. Text Miner divided the 107 syllabi into 10 clusters.

·2005-2008 Graduate Assistant, Mathematics Department, University of Louisville.

·2004-2005 Lecturer, Mathematics Department, Tianjin University of Commerce.

·2002-2004 The Study of Wavelet Transforms to Compress the Analytical data and Denoise (National Key Lab of Precision Measuring & Testing Techniques and Instruments).

COMPUTER SKILLS:

·SAS, SAS Enterprise Guide, Enterprise Miner, Text Miner. SAS is the main tool I used to analyze and model the data, with experience from 2006.

·R, programming for the methods of Linear Models and Classification.

·Microsoft Office

PRESENTATIONS AND HONORS:

·2008 Poster was selected to present at M2008 Data Mining Conference

·2008 Paper was selected to present at 2008 Midwest SAS Users Group Conference, Honorable Mention, SAS Ambassador Program

·2008 Poster winner at F2008 Business Forecasting Conference, Analysis and Forecasting Total Costs for Lung Cancer Data with Inflation Rate

·2008 Poster was selected to present at ISPOR conference

·2008 Paper was selected to present at SAS Global Forum 2007 Conference, Honorable Mention, SAS Ambassador Program

·2007 Poster was selected to present at SESUG conference (Southeast SAS User's Group)

·2007 Poster was selected to present at M2007 Data Mining Conference, Data Mining with Linear Model to Examine Expenditures of Prescribed medicines

·2007 Fellowship winner in the National Policy Institute

·2007 Poster winner at F2007 Business Forecasting Conference, Analysis and Forecasting for the S&P Stock Price Index Monthly Average: Data with Inflation Rate

·2006 Poster winner at M2006 Data Mining Conference, Data Mining to Investigate University Expectations of Work

·2004 Scholastic Honor: Excellent Student Cadre, Tianjin University

PUBLICATIONS:

·Tang, Guoxin, In Cases on Health Outcomes and Clinical Data Mining: Studies and Frameworks, Patricia Cerrito, editor, IGI Publishing 2010.

·Tang, Guoxin, *Data Mining and Analysis to Lung Disease Data*， MWSUG 2008 Proceedings.

·Tang, Guoxin, *Text and Data Mining to Investigate Expenditures of Prescribed medicines*， Global Forum Proceedings 2008.

·Tang, Guoxin, *Data Mining to Investigate University Expectations of Work*, SESUG Proceedings 2007.

·Liu, Zeyi, Tang, Guoxin, *Application of Wavelet Transform in Fundamental Study of Measurement of Blood Glucose concentration with Near-Infrared Spectroscopy*, Transactions Of Tianjin University, vol.37 No.6 Jun.2004.