5-2011

Statistical analysis and data mining of Medicare patients with diabetes.

Xiao Wang
University of Louisville

STATISTICAL ANALYSIS AND DATA MINING OF MEDICARE PATIENTS WITH DIABETES

By

Xiao Wang
B.A., Qingdao University, China, 2002
M.A., Shanghai University of Finance& Economics, China, 2006
M.A., University of Louisville, 2008

A Dissertation
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulllment of the Requirements
for the Degree of

Doctor of Philosophy

Department of Mathematics
University of Louisville
Louisville, Kentucky

May 2011

STATISTICAL ANALYSIS AND DATA MINING OF MEDICARE PATIENTS WITH DIABETES

By

Xiao Wang
B.A., Qingdao University, China, 2002
M.A., Shanghai University of Finance& Economics, China, 2006
M.A., University of Louisville, 2008

A Dissertation Approved On

March 25, 2011

by the following Dissertation Committee:

Dr. Patricia Cerrito, Committee Chair

Dr. Ryan Gill

Dr. Jiaxu Li

Dr. Adel Elmaghraby

Dr. Ibrahim Imam

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Dr. Patricia Cerrito for her guidance and help during my PhD study. I am grateful for her patience in leading me into a brand new area of data mining in the healthcare industry. What I have benefited from her are not only her contributions to the problems in this dissertation, but also her diligence, insights, and dedication to the intellectual and personal growth of her graduate students. I can never forget the hours that we spent together on our research projects. Working with her has been a true privilege.

I would like to thank all my committee members, Dr. Ryan Gill, Dr. Jiaxu Li, Dr. Adel Elmaghraby and Dr. Ibrahim Imam, for all their continual support, advice and the time they contributed in reviewing my work.

I specially thank my husband Gang Zhao for his tremendous help and support during my graduate study.

Above all, I am especially grateful to my loving parents, for their dedication, guidance, encouragement and support.

ABSTRACT

STATISTICAL ANALYSIS AND DATA MINING OF MEDICARE PATIENTS WITH DIABETES

Xiao Wang

March 25, 2011

The purpose of this dissertation is to find ways to decrease Medicare costs and to study health outcomes of diabetes patients as well as to investigate the influence of Medicare, part D since its introduction in 2006 using the CMS CCW (Chronic Condition Data Warehouse) Data and the MEPS (Medical Expenditure Panel Survey) data.

In this dissertation, we introduce pattern recognition analysis into the study of medical characteristics and demographic characteristics of the inpatients who have a higher readmission risk. We also broaden the cost-effectiveness analysis by including medical resources usage when investigating the effects of Medicare, part D. In addition, we apply several statistical linear models such as the generalized linear model and data mining techniques such as the neural network model to study the costs and outcomes of both inpatients and outpatients with diabetes in Medicare. Moreover, some descriptive statistics such as kernel density estimation and survival analysis are also employed. One important conclusion from these analyses is that only diseases and procedures, rather than age are key factors to inpatients' mortality rate. Another important discovery is that

at the influence of Medicare part D, insulin is the most efficient oral anti-diabetes drug treatment and that the drug usage in 2006 is not as stable as that in 2005. We also find that the patients who are discharged to home or hospice are more likely to re-enter the hospital after discharge within 30 days. Two – way interaction effect analysis demonstrates that diabetes complications interact with each other, which makes healthcare costs and health outcomes different between a case with one complication and a case with two complications. Accordingly, we propose some useful suggestions. For instance, as for how to decrease Medicare payments for outpatients with diabetes, we suggest that the patients should often monitor their blood glucose level. We also recommend that inpatients with diabetes should pay more attention to their kidney disease, and use prevention to avoid such diseases to decrease the costs.

TABLE OF CONTENTS

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURES

FIGURE **PAGE**

CHAPTER I

INTRODUCTION

As healthcare reform has passed and is now law, Medicare will come under more intense scrutiny in the healthcare industry. Diabetes is one of the most common chronic condition diseases and the 7^{th} leading cause of death in America. It is necessary to study the patients with diabetes in the Medicare insurance program. Although there are many researchers who study such patients, few of them investigate those patients from a statistical or data mining perspective. That is why we choose statistical analysis and data mining of the Medicare beneficiaries with diabetes as the topic of this dissertation.

The primary purpose of this research is to apply statistical knowledge and data mining algorithms to reduce the Medicare expenditures on diabetes beneficiaries, improve the health outcomes of patients as well as to investigate the influence of Medicare, part D since its introduction in 2006. In this chapter, we will briefly introduce the background of this research. First, we will present the objectives of this dissertation. Then we will review the previous research about diabetes patients within the Medicare population and describe our main contributions. Next, we will make some assumptions concerning the Medicare data. Then, we will introduce the basic concepts needed for this study, including the types of diabetes and their oral medication treatments, Medicare insurance and Medicare reform, the data sources and the medical codes. After we introduce the basic statistical methodologies used in this dissertation, we will introduce

the outline for the rest of the chapters.

1.1 Basic Objectives of Dissertation

There are several objectives in this dissertation:

- To see how to decrease Medicare reimbursements for both inpatients and outpatients. We try to find out the most important factors influencing Medicare payments through various statistical models and data mining algorithms such as the decision tree model; we also utilize multivariate analysis to find the directions of the relationships between the important predictors and the costs. In that way, we propose strategies to decrease Medicare payments.
- To see how to improve health care quality and outcomes. How to reduce the mortality rate of patients and how to decrease the length of hospitalizations are two concerns in this dissertation. We mainly employ various kinds of predictive models to analyze how variables such as age, gender, and procedures affect the outcomes so that we can provide methods to improve health outcomes.
- To investigate the influences of the Medicare drug plan since it became effective in 2006. We study the impacts through the usage of diabetes medications, the cost-effectiveness of the drugs and the medical resources utilization.

1.2 Literature Review and Main Contributions

1.2.1 Literature Review

In most cases, patients with diabetes in the Medicare population are mostly studied by medical methods, but occasionally, quantitative methods are also utilized.

Traditionally, statistical analysis is applied in three ways: cost analysis, outcomes research and the study of the impacts of the implementation of Medicare, Part D. For cost analysis, researchers only utilized simple statistical models to analyze the relationship between the predictors and the dependent variables, costs. For example, Li et al. [1] applied a 2-stage, least-squares model to estimate the effects of the predictor on Medicare expenditures. Bhattacharyya et al. [2] employed a general linear regression model to identify principal cost drivers among the identified cohort to the managed care system. Herrin et al. [3] applied a hierarchical generalized linear model to analyze the relationship between the costs and physicians and the general practice for Medicare diabetes patients.

For outcomes analysis, investigators employed basic statistical methods. For instance, Kuo et al. [4] utilized the two-tailed t-test statistics to analyze the trend of care practice and outcomes among the Medicare beneficiaries with diabetes. McBean et al. [5] applied a t-test when measuring the differences between the year 1999 and the year 2001 in diabetes care.

In terms of Medicare part D, only a few researchers have touched this topic so far and they also used one basic statistical model for data analysis. For instance, Karaca et al. [6] applied a multiple regression model to analyze how this drug plan affects the beneficiaries' out-of –pocket costs. Another research group represented by Schmittdiel et al. [7] used multiple logistic regression models to create adjusted percentages of diabetes patients across demographic and health plan characteristics responding to the survey questions about the implementation of the Medicare, Part D.

1.2.2 Main Contributions

Just as mentioned above, although some researchers performed simple statistical

analyses when they studied diabetes patients in Medicare, no one has yet investigated those using data mining techniques. In this dissertation, we introduce the data mining principles into the study of diabetes patients and make several contributions.

- Introduce pattern discovery into patients' demographic characteristics and diseases characteristics analysis.
- Apply supervised machine learning, decision trees and neural networks to analyze health outcomes.
- In cost effectiveness analysis, instead of comparing two drugs or two clinical trial methods, we compare differences of the two years for the same patients.
- Introduce the quantitative methods into readmission risk analysis for diabetes inpatients in Medicare.

1.3 Assumptions

Although we will make different hypotheses for different statistical models, we also make some general assumptions in the entire dissertation.

- In the dissertation, we are only concerned about type II diabetes patients in Medicare and when investigating the influence of Medicare, Part D, the research subjects are restricted to those who join in the drug plan.
- We do not consider the differences in severity of the diseases among the patients, and when we compare two years of cases, we assume that the diseases of the same patient will not become more severe in time.
- We also assume that there are no differences in the costs and treatments among different geographic regions.

- We do not consider the discount rate for the QALY (Quality Adjusted Life Year) for different years.
- When it refers to medications, we only consider the generic drugs and do not consider the brand name drugs.

1.4 Basic Concepts

1.4.1 Information about Diabetes, Its Types and Its Complications

Diabetes mellitus, or diabetes, is a group of diseases characterized by high blood glucose levels that result from defects in the body's ability to produce or use insulin. It is a devastating disease that greatly impacts long-term care and is the 7th greatest cause of death in the US. In recent years, diabetes has become a serious problem. According to the CDC (the Centers for Disease Control and Prevention) [8] 23.6 million children and adults had diabetes in the U.S.in 2007, and 12.2 million were over 60 years old.

There are mainly three types of diabetes [9]. Type I diabetes is usually called juvenile diabetes since it is common in children and young adults. The cause for Type I diabetes is that the body does not produce insulin. Type II diabetes, the most common form of diabetes, occurs either because the body does not produce enough insulin or the cells ignore insulin. Another type of diabetes is gestational diabetes, which is common to pregnant women. However, this dissertation will only focus on type II diabetes in persons 65 years of age and older.

Diabetes can lead to many complications such as heart disease and renal failure, high blood pressure and anemia. Statistics carried out by the American Diabetes Association [10]showed that the probability of people with diabetes have heart diseases is

twice that of people without diabetes. Diabetes is the primary cause of end-stage renal disease. The CDC [8] states that approximately 60% to 70% of those with diabetes have mild to severe forms of nervous system damage, and severe forms of diabetic nerve disease are a major contributing cause of lower-extremity amputations.

In addition, the co-morbidities often suffered by patients with diabetes can affect each other. For example, diabetes is the leading cause of renal failure. The National Institute of Diabetes and Digestive and Kidney Diseases study [11] showed that nearly 24 million people in the United States have diabetes; the Annual Data Report supported by the United States Renal Data System [12] illustrated that nearly 180,000 persons are suffering renal failure due to diabetes. Diabetic nephropathy likely contributes to the development of anemia in diabetes patients. Anemia often develops early in the course of chronic kidney disease in patients with diabetes and also contributes to the high incidence of cardiovascular disease observed in diabetic patients.

Each year, it takes a large amount of resources to treat diabetes and its complications, including organ dysfunctions and neurological disorders (Table1.1) [13] . According to the American Diabetes Association [10], a total of $174 billion was expended on the treatment of diabetes in 2007; among the total medical expenditures ($116 billion), 23% ($27 billion) was for diabetes care and 50 % ($58 billion) was for chronic diabetes-related complications. Therefore, it is fundamental to control diabetes. Among all the measures to control diabetes, blood glucose monitoring is the best.

The Diabetes Control and Complications Trial funded by the National Institutes of Health reported in 1993 [14] that intensive glucose control prevents or delays the eye, nerve and kidney complications of type I diabetes (as cited in Glucose Control Cuts Risk

of Heart Disease in Type I Diabetes); and DCCT/EDIC study illustrated that intensive glucose control lowers the risk of heart disease and stroke by about 50 % in people with type I diabetes. However, renal failure treatment is expensive, since the expenditures for the treatment of this disease account for 30 % of the costs of the treatment of diabetes, so it is essential to find and examine the factors that impact renal failure in order to reduce the total charges of diabetes treatment.

Table 1.1 Health Care Expenditures Attributed to Diabetes (in millions of dollars)

		Chronic Complications						
Setting	Diabetes	Neurological	Peripheral vascular	Cardio-vascular	Renal	Ophthalmic	General conditions	Total
Hospital	1,535	3,115	2,719	20,790	3,285	36	23,473	58,344
Physician's	2,899	382	382 279	1,004	323	899	3,830	9,897
Emergency	234	138	43	403	132	11	2,717	3,870
Hospital	842	75	135	317	87	130	1,321	2,985

Source: Table 12(Abridged) in Economic Costs of Diabetes in the U.S. in 2007

Type II diabetes can be treated by oral medications [15]. The sulfonylureas such as glyburide, glipizide, and glimepiride are all generic medications, which make them inexpensive drugs for the management of Type II diabetes mellitus. Another class of medication that has a similar mechanism of action to the sulfonylurea medications is the meglintinides, including repaglinide (prandin) and neteglinide (starlix). Another drug class is alpha glucosidase inhibitors, such as precose. Metformin, the only biguanide, is recommended as a mainstay in therapy in patients with type II diabetes. When the above medications become ineffective, insulin should be used alone or combined with other drugs.

1.4.2 Medicare and Medicare Reform Related to Part D Plan

Medicare is the voluntary health insurance for people of age 65 or older, under age 65 with certain disabilities, and any age with permanent kidney failure (called "End-Stage Renal Disease"). It is required for any senior citizen on social security. It basically consists of four parts, Part A (Hospital Insurance), Part B (Medical Insurance), Part C (Medicare Advantage Plans) and Part D (The optional prescription drug program). Part D uses competing private plans to provide beneficiaries access to appropriate drug therapies. As of January 2008 [16], almost 90 % of Medicare enrollees had their prescription drugs covered by the part D plan or other creditable sources. In this dissertation, we will not study Medicare Advantage Plans because they are run by commercial insurance companies instead of CMS (Centers for Medicare and Medicaid Services).

As the two acts, the Patient Protection and Affordable Care Act and the Health Care and Education Reconciliation Act of 2010 were signed into law; several changes with the Medicare insurance program have taken effect since 2010. One of the significant changes is about the Medicare, part D coverage gap or 'donut hole' [17]:

- In 2010, if the beneficiary's expenditures reached the donut hole and enter the gap, then s/he received a $250 rebate from Medicare.
- From 2011 to 2020, if the enrollee's expenditures are in the gap, s/he will get a 50% discount on the total cost of brand name drugs.
- Medicare will phase in additional discounts on the cost of both brand name and generic drugs.
- By 2010, the Medicare coverage gap would have been cancelled. Instead of paying 100 % of the costs during the gap, the beneficiaries only needed to

pay for 25 % of the costs.

The Medicare drug plan was always a controversial issue since its introduction. This time, the amendments to the plan make it again a topic of concern. Instead of studying of the influences of cancelling the gap, we will investigate the influences regarding the implementation of the drug plan due to limited data information.

1.4.3 Data Sources

In this research, we utilize two kinds of data sets. One is the CMS CCW (Chronic Condition Data Warehouse) data [18] for the year 2004 and the other is the MEPS (Medical Expenditure Panel Survey) data [19] for the years 2005 and 2006.

The CCW data are collected by CMS and provides researchers with Medicare beneficiary, claims, and assessment data linked by beneficiary across the continuum of care. Between the years 1999 to 2004, it covers a random 5% of the Medicare beneficiary population each year. In this research, we use three data files for the year 2004, outpatient_base_claims with 2,030,078 records and inpatients_base_claims with 244,299 items. Since both of them do not cover demography information, the dataset, beneficiary_summary_file is needed.

The MEPS data, collected by the Agency for Healthcare Research and Quality, is a class of survey data sets containing such information about medical services and employers across the United States. In order to compare the differences in 2005 and 2006, we harness various kinds of data for these two years. They contain such information as office-based visits, outpatient visits, inpatients, prescription drugs and the full year consolidation. However, such data have some disadvantages. For instance, time information is incomplete.

1.4.4 Medical Codes

Medical codes are very useful tools in medical billing and reimbursement. In this study, we will use HCPCS (Healthcare Common Procedure Coding System) codes [20], CPT (Current Procedural Terminology) codes [21] and ICD -9-CM (International Classification Diagnosis, Clinical Modification, 9th edition) codes [22].

HCPCS codes are utilized by CMS for explaining the claims for payments. There are two levels of HCPCS codes. Level I is comprised of CPT and Level II is a standardized coding system that is used primarily to identify products, supplies, and services not included in the CPT.

CPT codes are numbers used to represent medical procedures and services under public and private health insurance programs. CPT codes are developed, maintained and copyrighted by the AMA (American Medical Association). As the practice of health care changes, new codes are developed for new services, current codes may be revised, and old, unused codes are discarded.

ICD- 9-CM is an official system of codes used to stand for the diagnoses and procedures associated with hospital utilization. Those codes are overseen and modified by NCHS (the National Center for Health Statistics) and CMS. This system contains two kinds of codes; one code is for the diagnosis of diseases and the other is for diagnostic, surgical or therapeutic procedures. The common diagnostic ICD9-CM codes for diabetes are shown below [23].

250.00-250.03 Diabetes mellitus without mention of complication

250.10-250.13 Diabetes with ketoacidosis

250.20-250.23 Diabetes with hyperosmolarity

250.30-250.30	Diabetes with other coma
250.40 -250.43	Diabetes with renal manifestations
250.50 -250.53	Diabetes with ophthalmic manifestations
250.60-250.63	Diabetes with neurological manifestations
250.70-250.73	Diabetes with peripheral circulatory disorders
250.80-250.83	Diabetes with other specified manifestations
250.90- 250.93	Diabetes with unspecified complication

1.5 Basic Statistical Methodology

In this dissertation, several statistical methods and data mining algorithms are utilized. The descriptive statistics such as kernel density estimation are applied to study the distribution of the costs. The various kinds of statistical linear models such as the general linear model and the generalized linear model are used to examine the relationships between the predictors and the costs. Two-way interaction effect analysis is used to analyze the influence of one diabetes complication on another complication in costs and outcomes. The data mining algorithms such as the decision tree model and the neural network model are employed for health outcomes and health quality analyses. The survival model is utilized for our diabetes medication study. In addition, other theories such as cost-effectiveness analysis in health economics are also utilized in this dissertation.

1.6 Outline

The rest of the dissertation is organized as follows: Chapter II describes the data

processing. Chapters III- IV introduce statistical models and supervised / unsupervised machine learning. Chapters V- VI discuss Medicare costs and their influencing factors for both inpatients and outpatients with diabetes, outcomes and readmission risk factors of diabetic inpatients as well as 2-way interaction effects of diabetes complications on costs and outcomes. Chapter VII investigates how the Medicare, Part D program affects the usage and the cost effectiveness of diabetes medications. During the discussion of our research in chapters V - VII, several theories applied in the study such as survival analysis and cost-effectiveness analysis are also introduced. The last chapter summarizes the dissertation results and gives the conclusions.

CHAPTER II

DATA PREPARATION AND PROCESSING

Data preparation and processing is a key to successful analysis and accurate results. Before we develop a statistical model or perform data mining analysis, we need to preprocess the data to get them ready for study. In this chapter, we will briefly discuss what techniques are utilized for data processing; that is, data validation, merging data, data reduction, data transformation and some SAS functions used for data processing.

2.1 Data Validation

During the data validation (a.k.a. data cleaning), several things needed to be checked: whether the information such as the diagnosis code is correct and whether the variable type is correct. If some problem exists, then it needs to be dealt with by some measure.

2.1.1 Dealing with Missing Values

It is very common that there are missing values in a large-size data set or a survey data and they are needed to be examined in most cases. There are three ways to process missing values: elimination, imputation and substitution.

1. Elimination

Sometimes, if the missing values only account for a small percentage of the sample, then they can be deleted without any processing. However, this method may result in bias or inaccurate results if too many observations are eliminated.

2. Imputation

Imputation [24] is a prevailing approach of manipulating the missing values and there are several different methods for imputation according to the types of missing values. If the values are missing at random; that is to say, the probability of a missing value appearing in one variable is not related to the probability of existing missing values in another variable, then the values can be simply eliminated from the sample data. However, this method may result in unnecessary elimination. If the values are completely missing at random and can be judged in this way that the probability of missing values in one variable is unrelated to the value of the variable itself or to values of any other variable, then imputation is needed. One way to do this is MCMC (Markov Chain Monte Carlo), which creates multiple imputations by using simulations from a Bayesian prediction distribution for normal data.

3. Substitution

Substitution [25] is another approach of handling missing values. Sometimes, according to the characteristics and types of variables, the missing values can be replaced with some other values, such as the mode, the median, the mean, the maximum or the minimum of the variables.

4. Our Approach

In our research, only the third method works to some extent. When we perform

survival analysis about diabetes drug usage, considerable information about the day, the month and the year of the prescriptions in the MEPS data is missing. Therefore, we deal with missing values in this way: according to the rule of prescription of drugs, we set the missing values of the variable, DAY to the first day of the month and the missing values of the variable, YEAR to the year when the data were collected, but we eliminate an observation whose month information is missing.

2.1.2 SAS Functions for Data Cleaning

1. Characteristic Functions

In our study, we utilize SAS characteristic functions to remove trailing blanks before and after the nominal variables and to find certain letters in a string. The functions [26] that we apply include LEFT, TRIM and TRANSLATE. For example, consider what we do using the prescription drugs data.

```
/*To replace  '_' with a blank in the names, removes the
trailing blanks from theright-hand side of a variable value
and left justifies the variable value*/
NRXNAME=TRANSLATE(LEFT(TRIM(NRXNAME)),'_',' ');
LEFT (TRIM(CONCAT[I])); END; RUN;
```

2. Date Functions

We utilized the MDY function to return a date value from the numeric values for month, day and year into a SAS date value. For instance, in our project about prescription drugs usage, we use the following code to combine the three variables, DAY, MONTH and YEAR into one variable, DATE.

```
DATE = MDY (RXMM, RXDD, RXYY);
```

2.2 Merging Data

None of our data contains all the information we need; therefore, we have to merge different data files into one file before our analysis. The tools for merging data are SAS Enterprise Guide, the Base SAS data step and SAS SQL.

2.2.1 SAS Enterprise Guide

This is a user-interface-design SAS module in which we only need to click some buttons, and then we can merge the different data sets together. For example, in the project to discuss how to reduce the Medicare reimbursements for outpatients, we click Filter and Query->Add table to join columns from revenue and beneficiary data sets to generate another new data set containing beneficiary ID, HCPCS codes, total charges, and so on.

2.2.2 Base SAS Data Step

Base SAS is the most commonly utilized method of merging data among the SAS users. We employ it for data combination in several analyses. For instance, in analyzing the outcomes of diabetes outpatients, we used the following SAS code:

```
PROC SORT  DATA=SASUSER.IPCLUS; BY_CLUSTER_ ;
PROC SORT DATA= SASUSER.IPTCHDEM; BY _CLUSTER_;
DATA SASUSER.IPKDETCHDEM;
MERGE SASUSER.IPCLUSTER SASUSER.IPTCHDEM; BY _CLUSTER_;
```

2.2.3 SAS SQL

SAS SQL has some advantages over the base SAS data step; for one thing, before merging the data by their common variables, we do not need to sort each data set by their common variables. For another, the names of the common variables in different data sets

are not required to be the same. Consider a cost-effectiveness analysis of the diabetes medications; for example, we can apply an SQL conditional inner join to merge the data.

```
/*Combine the life table and 2006 Medicare part D
beneficiary table */
PROC SQL;
CREATE TABLE SASUSER.LE06 AS
SELECT *
FROM SASUSER.LIFETABLE1 AS LT,
SASUSER.BCHWLQ06 AS BC
WHERE LT.AGE=BC.AGE06X;
QUIT;
```

2.3 Data Reduction

After we get the different data sets into one data set, we need to reduce the data size, if possible. There are several approaches: (1) sampling and partitioning (2) principal component analysis (3) factor analysis (4) observations or variables selection.

2.3.1 Sampling and Partitioning

1. Sampling

There are mainly three types of data sampling methods [27]: (1) simple random sampling (2) stratified random sampling (3) cluster sampling.

(1) Simple random sampling: A sample is selected in such a way that every possible sample of the same size is equally likely to be chosen. We can realize this through clicking the Random Sampling button in SAS Enterprise Guide by choosing either the number of observations or a percentage of the sample. We can also perform random sampling through the SAS SURVEYSELECT procedure.

(2) Stratified random sampling: This sample can be obtained by separating the population into mutually exclusive strata, and then drawing simple random samples from each stratum. When we discuss the relationships between other diseases and renal failure, we need to use this method to guarantee that all the rare occurrence events of renal failure are selected in the large data set. We realize this through setting the sample node in Enterprise Miner in this way [28]: set sample method to stratify, stratified criterion to level based, and level selection to rarest level.

(3) Cluster sampling: This can be obtained by dividing the data into several groups or clusters of elements. When we compare the differences between the different diagnosis procedures in Medicare payments, we first cluster the procedures into several groups.

2. Partitioning

Before we build a predictive model in SAS Enterprise Miner, we often use the Partition node to divide the sample into three smaller data sets: training, validation and testing [29]. The training set is used to build a model. The validation data set is utilized to ensure a model of good fit while the testing data set is applied for comparison to find an optimal model.

2.3.2 Principal Component Analysis and Factor Analysis

Although principal component analysis and factor analysis are very popular approaches, we apply neither of them in our analyses due to a very large data size. Instead, we utilize sampling and clustering to reduce the sample size.

2.3.3 Observations or Variables Selection

Sometimes, not all the observations meet our requirements, or some variables

have nothing to do with our predicted targets; therefore, we need to select observations or variables. One method is to apply the KEEP or DROP statements and the following SAS code demonstrates how to use them:

```
DATA SASUSER.IPCLUS(KEEP=_CLUSTER_  _FREQ_  _RMSSTD_
CLUS_DESC);
SET  EMST.TEXT_CLUSTER;
```

In order to avoid duplicate observations, we can apply NODUPKEY:

```
PROC SORT DATA=SASUSER.COM06 OUT=SASUSER.NREP06 NODUPKEY;
BY DUPERSID DATE DATE1 DATE2; RUN;
```

Another method to select observations (a.k.a. rows in SQL) and variables (a.k.a. columns in SQL) is SQL conditional selection:

```
/*To sort out the diabetes patients*/
PROC SQL;
 CREATE TABLE SASUSER.OBDIA05 AS
 SELECT t1.DUPERSID, t1.OBICD1X, t1.OBICD2X, t1.OBICD3X,
 t1.OBICD4X
 FROM SASUSER.FILTER_FOR_QUERY_FOR_FILTER_FOR_ AS t1
 WHERE t1.OBICD1X = '250' OR t1.OBICD2X = '250' OR
 t1.OBICD3X =   '250' OR t1.OBICD4X = '250'; QUIT;
```

2.4 Data Transformation

The dominant method of transformations is the Box–Cox transformation [25], which attempts to transform a continuous variable into an almost normal distribution. This is achieved by mapping the values using the following set of transformations:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} when \lambda \neq 0 \\ \log(y_i) when \lambda = 0 \end{cases} \tag{2.1}$$

where λ is the transformation parameter, and y_i is a continuous variable that needs to be transformed. What we use data transformation for is to define dummy variables. Sometimes, or defining procedures to new variables, we use 0-1 indicator functions and the SAS code is shown below to get the new data shown in Figure 2.1.

```
IF(Recode_ICD9 EQ:'25000')
THEN R25000=1;
ELSE R25000=0;
```

CLM_ID	Recode_ICD9	COUNT	PERCENT	R25000	R4019	R585	RV5861	R2724	R42731
WWWWE	other	16	100	0	0	0	0	0	
WWWWE	42731	1	6.25	0	0	0	0	0	
WWWWE	V5861	1	6.25	0	0	0	1	0	
WWWWE	other	14	87.5	0	0	0	0	0	
WWWWE	42731	1	6.25	0	0	0	0	0	
WWWWE	V5861	1	6.25	0	0	0	1	0	
WWWWE	other	14	87.5	0	0	0	0	0	
WWWWE	other	16	100	0	0	0	0	0	
WWWWE	2859	1	6.25	0	0	0	0	0	
WWWWE	5990	1	6.25	0	0	0	0	0	
WWWWE	other	14	87.5	0	0	0	0	0	

Figure2. 1. Newly-defined Variables

2.5 Other SAS Functions

Besides what was discussed above, we also apply some base SAS functions to process the data sets. For instance, when we want to define a string containing all possible diagnosis codes, we use the CATX statement, which concatenates character strings, removes leading and trailing blanks, and inserts separators. The code is [26]:

```
DIAGNOSIS=CATX('',ICD9_DGNS_CD1,ICD9_DGNS_CD2,ICD9_DGNS_CD3
,ICD9_DGNS_CD4,ICD9_DGNS_CD5,ICD9_DGNS_CD6,ICD9_DGNS_CD7,IC
D9_DGNS_CD8,ICD9_DGNS_CD9,ICD9_DGNS_CD10,ICD9_DGNS_CD11,ICD
```

```
9_DGNS_CD12,ICD9_DGNS_CD13,ICD9_DGNS_CD14,ICD9_DGNS_CD15,IC
D9_DGNS_CD16);
```

When we want to convert all the values of a variable into observations, we utilize the TRANSPOSE function and the code is shown below [26]:

```
PROC TRANSPOSE DATA=SASUSER.SORTMR06 OUT=SASUSER.TRANMR06
PREFIX=MED_; VAR NRXNAME; BY DUPERSID; RUN;
```

If we want to calculate the differences between different dates, we apply the DATDIF functions:

```
DAYS=DATDIF (SDATE, EDATE,'ACT/ACT');
```

In this chapter, we briefly discussed our approaches of processing the data and we will elaborate them in detail in the later application chapters. The process of data preparation contains more contents than what we introduced above. For example, it also covers exploratory data analysis such as the sample mean analysis or a frequency count study. We also utilize SAS to process the data during our study. Other preprocessing will be discussed in later chapters.

CHAPTER III

STATISTICAL LINEAR MODELS

Statistical models and the statistical methods associated with them are versatile and robust. There are mainly two kinds of statistical models, one is the linear model, which is simple and widely used and the other is the non-linear model. In spite of the availability of highly innovative tools in statistics, the main tool of researchers remains the linear model, which involves the simplest and seemingly most restrictive statistical properties of independence, normality, constancy of variance and linearity. It can be divided into several subgroups. (1) the general linear model, including the general linear univariate model (ANOVA), the general linear multivariate model (MANOVA), the regression model; (2) the generalized linear model, including the generalized univariate/multivariate model; (3) the linear mixed model. The general linear model and the logistic regression model can be thought of as special cases of the generalized model.

3.1 The General Linear Model

The general linear models are a class of linear models, and they can be represented by the general linear regression model and ANOVA (Analysis of Variance). As for the regression model, there are three subgroups, the simple linear regression model, the multiple linear regression model and the logistic regression model.

3.1.1 The Multiple Linear Regression Model

The regression model [30] is the oldest and most used model. It is also the most understood model in terms of performance, mathematics, and diagnostic measures for model quality and goodness of fit. The regression model has been applied to a very wide range of problems in healthcare, finance or medical fields.

1. Assumptions

Understanding of the assumptions of a model is the key to successfully build a model. The following are the assumptions for the multiple linear model:

- The relationship between the response variables and the predictors is linear.
- The response vector $y_{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ is mutually independent, and the variance of each element of the vector is σ^2, which is fixed and unknown.
- The parameter vector $\beta_{p\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ is fixed and unknown.
- The elements of the error vector $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ have the characteristics: (1) The elements of the vector are independent from each other and identically distributed. (2) The expectation of each element: $E(\varepsilon_i) = 0$; (3) The variance of each element, $\text{var}(\varepsilon_i) = \sigma^2$, where $i=1,2,...n$; (4) $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$.

2. Expressions

The matrix formulation of the linear multiple regression model is shown below

$$y = X\beta + \varepsilon \tag{3.1}$$

where y, β, ε are as mentioned above, and the vector $X = \begin{pmatrix} I & \vdots & x_1 & \vdots & \cdots & \vdots & x_{p-1} \end{pmatrix}$ is the design matrix and $I = (1,1,\ldots 1)^T$.

3. Model Development

In developing the model, predictors and the response variables are chosen and the choice of algorithms to be applied is also considered. However, model development belongs to the application of the model; we will discuss it in detail when presenting our analyses. In the following model theories discussion, we will also postpone a discussion of our analyses.

4. Assumption Diagnostics

The examination of basic assumptions is a fundamental procedure in building a model and it is even more important than developing models themselves. The first thing we need to do is to examine whether the model meets the assumptions. If a model does not meet the basic requirements, then the model building is a failure. There are several criteria that are used to check whether the assumptions are met.

I. Independence Test

The primary method of testing the independence of the response variables is the Durbin-Watson statistic [31]. The test statistic can be written as:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{3.2}$$

In the above expression, $e_i = y_i - \hat{y}_i$ is the residual of individual i, which is the difference of the observed and predicted values of the response variable for individual i. The value of d is always between 0 and 4. If $d = 2$, it indicates that the model meets the independence requirement and it also indicates that no autocorrelation exists.

II. Normality Test

There are three common measures to examine whether the residuals follow normal distributions: (1) S-W (Shapiro – Wilk) test, (2) K-S (Kolmogorov – Smirnov) test, (3) A-D (Anderson- Darling) test.

(1) S – W test [32]: The test can be conducted through a W statistic and it can be calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{n} \alpha_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.3}$$

In the above expression, $x_1, x_2 \ldots x_n$ represent a random sample that follows a normal distribution; $\bar{x}$ stands for the mean of the sample; $x_{(1)}, \ldots x_{(n)}$ are an ordered sample; $\alpha_1, \ldots \alpha_n$ are constants generated from the means, variances and co-variances of the order statistics of a sample of size n from a normal distribution; to be specific, they can be calculated in this way:

$$(\alpha_1,\cdots,\alpha_n)=\frac{E^T V^{-1}}{\sqrt{E^T V^{-1} V^{-1} E}} \tag{3.4}$$

where $E=(E_1,\ldots E_n)^T$ and the E_i's are the expected values of the order statistics of the random sample and V is the covariance matrix of the order statistics. Whether the sample meets normality can be judged by the value of W. The value of W is between 0 and 1. Small values of W indicate that the sample does not follow a normal distribution, while the value of close to 1 means normality.

(2) K-S test [33]: This test can be realized through calculating a D-statistic and the process is shown below.

Test hypothesis: H_0: The sample follows a normal distribution

H_a: The sample does not follow a normal distribution

The D-statistic is defined as follows:

$$D_n=\sup_x\left|F_n(x)-F(x)\right| \tag{3.5}$$

in which D_n is the largest vertical distance between the distribution function $F(x)$ and the empirical distribution function $F_n(x)$; the empirical distribution function is defined as:

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n} I_{X_i\le x} \tag{3.6}$$

where the $x_{(i)}$'s are ordered statistics, $I_{x_i\le x}$ is an indicator function, and its expression is shown below:

$$I_{x_i\le x}=\begin{cases}1 & if \quad x_i\le x\\ 0 & otherwise\end{cases} \tag{3.7}$$

The criterion of judging whether the sample follows a normal distribution is based on the

value of the D-statistic. If the value of D is greater than the critical value, then the sample does not follow a normal distribution.

(3) A- D test [32]: In the Anderson- Darling test, the statistic A is calculated and its expression is shown below:

$$A^2 = -n - S \tag{3.8}$$

where F is the cumulative distribution function and the variable S is defined in this way:

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} [\log F(y_i) + \log(1 - F(y_{n+1-i}))] \tag{3.9}$$

If the value of A is smaller than the critical value, then the sample follows a normal distribution.

If the model's requirements are met, then the fitness of the model is another vital thing needed to be checked.

III. Autocorrelation Diagnostics

In statistics, the autocorrelation [34] describes the correlation between values of a random process at different points in time and it can be defined as in (3.10). Autocorrelation of the error terms often occur in time series data.

$$R(i,j) = \frac{E(x_i - \mu_i)(x_j - \mu_j)}{\sigma_i \sigma_j} \tag{3.10}$$

In the above expression, if $R(i,j)$ is well defined, then its values should lie in [-1,1]. x, μ, σ represent the value, the mean and the standard deviance, and i, j stand for two different time points. The measure to examine the existence of the autocorrelation is the Durbin–Watson statistic [31]. The basic rule is to compare the values of d to the lower level $d_{L,\alpha}$ and the upper level $d_{U,\alpha}$ of the critical values at the significance α.

IV. Multicollinearity Diagnostics

Multicollinearity exists when the samples are not independent from each other. The statistic, VIF (Variance Inflation Factor) [35], can be employed to test for it. The variable VIF can be defined in (3.11).

$$VIF = \frac{1}{1-R^2} \tag{3.11}$$

where R^2 is the model variance, defined in (3.12), in which n is the sample size (a.k.a. the number of observations), k is the number of predictors (a.k.a. the number of coefficients) and the expression of $1\text{-}R^2$ is tolerance.

$$R^2 = \frac{\dfrac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}}{\dfrac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-k+1}} \tag{3.12}$$

A high value of VIF indicates the existence of multicollinearity and a value greater than 10 means a serious problem.

V. Outlier Diagnostic

Outlier detection is also an important step to test whether the mode is good or not. There are mainly four statistics [35] : (1) Leverage (2) Studentized deleted residual (3) Dffits-statistic (4) Cook' D statistic.

(1) Leverage: It is called the hat diagonal, which is used to detect outliers among the predictor variables. If we write a prediction model into matrix form as:

$$\hat{y} = X^T \hat{\beta} \tag{3.13}$$

and if we use the least squares approach to get the minimizer with respect to coefficients

β, we get $\hat{\beta}$ and then we get the expression of the hat matrix H defined in (3.14).

$$H = X(X^T X)^{-1} X^T \tag{3.14}$$

The leverage, h_{ii}, is the i^{th} diagonal element of the hat matrix. As a rule of thumb, any observation with an h_{ii} that meets the following inequality (3.15) is considered as a leverage point and it has the potential to change the model. For a small sample, the criterion is changed to an inequality (3.16).

$$h_{ii} > 2\frac{k+1}{n} \tag{3.15}$$

$$h_{ii} > 3\frac{k+1}{n} \tag{3.16}$$

In (3.15) and (3.16), n is the number of response variables and $k+1$ is the number of coefficients, $\beta_0 \beta_1 \ldots \beta_k$. The observation with the largest h_{ii} can be said to have the most extreme predictor variables, while the observation with the smallest h_{ii} values might be said to be the most typical.

(2) Studentized deleted residual: This residual is another useful tool to detect extreme values of the observations and it can be calcuated as the same way for the standardized residual (a.k.a. studentized residual), except without considering the i^{th} observation. The expression for the residual t_i can be written as:

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \tag{3.17}$$

where e_i is the residual for the i^{th} observation; $MSE_{(i)}$ is the mean squared error for the regression model given that the i^{th} observation is left out; h_{ii} is a leverage. If the value t_i of an observation is greater than 2, then the observation is probably an outlier.

(3) Dffits statistic [36]: Although some potential outliers may be detected by the above two criteria, they may not affect the model. On the contrary, the DFFITs statistic can find the outliers that actually have influence on the model. The statistic can be expressed as:

$$DFF_i = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{\frac{1}{2}} \tag{3.18}$$

The observation whose DFF_i meets the following inequality will be thought to be influential.

$$DFF_i > 2\sqrt{\frac{k+1}{n}} \tag{3.19}$$

Here, k and n have the same meaning as the above.

(4) Cook' D statistic: This is the most often used criterion for outlier detection. Its expression is shown below and the signs in it have the same meaning as above.

$$D_i = \left(\frac{1}{k+1} \right) \left(\frac{h_{ii}}{1 - h_{ii}} \right) \left(\frac{e_i^2}{MSE(1 - h_{ii})} \right) \tag{3.20}$$

Typically, if the value of D_i is greater than 2, the observation should be investigated.

Through I - V, we discussed methods and criteria to examine whether the assumptions of a model are met and we will utilize them for model diagnostics before developing our models.

5. Model Evaluation [35]

I. Significance Test

R^2 and Adjusted-R^2 are used to study the magnitude of effects. If both the predictors and the response variables are continuous variables, then R^2 will be a good measure to demonstrate that the proportion of the variation in the dependent variable is

accounted for by the explanatory variables. The rule is that the value of R^2 varies from 0 to 1, and the higher the value, the greater the effect. The definition of R^2 is displayed in (3.12). Unlike R^2, the adjusted R^2 (3.21) increases only if the new term improves the model more than would be expected by chance and it can be defined as:

$$R^2_{adjusted} = 1-(1-R^2)\frac{n-1}{n-k-1} \tag{3.21}$$

The F- test is often used for a model significance check (a.k.a. overall fit of the model) and it can be calculated as:

$$F_{(k,n-k-1)} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} \tag{3.22}$$

II. Goodness Test

One useful tool to check whether the model is good is the root MSE [37] and it can be defined in this way:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}} \tag{3.23}$$

Since the *RMSE* is a frequently-used measure of the differences between values predicted by a model and the values actually observed from the parameter being estimated, the smaller the value, the better the model.

3.1.2 Linear Logistic Regression Model

The linear logistic regression model is a special case of the general linear regression model used when the response variable is dichotomous. Most of its characteristics are similar to those of the general regression model; therefore, we will

focus on its three unique parts in this section: analysis of the results, rare event and model selection.

1. Expression

The logistic regression model (a.k.a. logit model) [38] is used for prediction of the probability of occurrence of an event by fitting data to a logit function:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta^T X \tag{3.24}$$

where p is the probability and the odds ratio is expressed as:

$$odd = \frac{p}{1-p} \tag{3.25}$$

2 . Rare Event and Oversample

Rare event: If a target variable appears in a fraction of less than 10% in a large data set, then it is a rare occurrence event. For instance, fraud rate and mortality rate are two rare events with two values, 0 and 1.

Although the logistic regression model is a very useful model, there arise some problems when the model is applied in rare event data. Gary King et al.(2001) [39] demonstrated that the logistic regression model could greatly underestimate the probability of a rare occurrence event and they suggested that all the rare target variables should be included in the model to fix this problem.

What Dr. King suggested is an approach of oversampling, in which a data set is stratified by the levels of a rare event variable and assigned different sample weights to different levels. We can compensate for oversampling in two ways, one method is to use the sample node in SAS Enterprise Miner as we mentioned in chapter 2 and another

method is to employ the surveyselect procedure in SAS/BASE, which we will discuss in chapter 6.

3. Model Selection

When developing the model, effect selection is often considered. The purpose of the selection is to choose the predictors that are significant to the response variable and the criteria for each selection vary from method to method. There are 3 main frequently-utilized methods: (1) forward selection, (2) backward elimination, (3) stepwise selection.

(1) Forward selection: In this method, at the beginning, only the intercepts and the first n explanatory effects are put into the model; then, another new effect significant at the level α is input into the model. The process will continue until all the remaining effects that are significant outside the model are imported. The criterion for the selected entry is the score χ^2 statistic defined as:

$$T_{score} = l^T(\hat{\beta}_{H_0})(-H(\hat{\beta}_{H_0}))^{-1} l(\hat{\beta}_{H_0}) \tag{3.26}$$

where the H_0: hypothesis is defined as follows; $l = \frac{\partial L(\beta)}{\partial \beta}$; L is the log likelihood function; $\hat{\beta}_{H_0}$ is the maximum likelihood estimation of the coefficient vector β under H_0: and H^S is the hessian matrix defined in the form:

$$H^S = \begin{pmatrix} \frac{\partial^2 f(\beta)}{\partial \beta_1^2} & \frac{\partial^2 f(\beta)}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 f(\beta)}{\partial \beta_1 \partial \beta_k} \\ \frac{\partial^2 f(\beta)}{\partial \beta_1 \partial \beta_2} & \frac{\partial^2 f(\beta)}{\partial \beta_2^2} & \cdots & \frac{\partial^2 f(\beta)}{\partial \beta_2 \partial \beta_k} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f(\beta)}{\partial \beta_k \partial \beta_1} & \frac{\partial^2 f(\beta)}{\partial \beta n \partial \beta_2} & \cdots & \frac{\partial^2 f(\beta)}{\partial \beta_k^2} \end{pmatrix} \tag{3.27}$$

In the above expression, f is a real-value function: $f(x_{1,\ldots,}x_n)$.

(2)Backward elimination: In this method, at the beginning, all the effects are fitted into the model; then, at each step, the least significant effect among those staying in the model will be removed. The process will not stop until all the remaining effects are significant. In this step, the results of the Wald test for individual parameters are examined.

(3) Stepwise selection: The process of this method is a combination of the above two methods. At the beginning, only intercepts and n effects are in the model; then, another significant effect will enter into the model in each step. However, during the process, if some effect in the model becomes not significant, then it will be eliminated from the model. The process will continue until no new effect can be input into the model or no existing effect can be eliminated from the model.

4. Model Convergence

It is important to check whether the convergence criterion is met. The default criterion by SAS is the relative Hessian convergence criterion with tolerance number [40] and it can be expressed as follows:

$$\frac{\frac{\partial^2 f(\beta)}{\partial \beta_k^2}(H^s)_k^{-1}\frac{\partial f(\beta)}{\partial \beta_k}}{|f_k(\beta)|} \leq number \tag{3.28}$$

where $f(\beta)$ is the function with the k dimensional vector: $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$; $\frac{\partial f(\beta)}{\partial f_k(\beta)}$ is the gradient (a.k.a. first derivative) of f_k (the objective function at iteration k) ; and H_k is the Hessian (a.k.a. second derivative) of the objective function at iteration k. The expression of the Hessian for the function $f(\beta)$ is similar to the one shown in (3.27).

5. Output Analysis

I. Model Fit Statistic

Three criteria can be applied to examine the model fitness [40] (1) -2 Log L (Log Likelihood), (2) AIC (Akaike Information), (3) SC (Schwarz Bayesian Information).

(1) -2 Log L: This is the most popular criterion used in hypothesis tests for nested models. In this case, ω_m means the weight, f_m means the frequency, $\hat{p}_m$ indicates the estimated probability of the event, n is the total number, m is the number of successful events; then for the m^{th} event, the expression of the log likelihood for the binary target is:

$$LogL = \sum_{i=1}^{m} \frac{w_m}{\sigma^2} f_m [n \log(\hat{p}_m) + (n-m)\log(1-\hat{p}_m)] \tag{3.29}$$

(2) AIC: It is used for the comparison of non- nested models on the same sample and it can be calculated as:

$$\text{AIC} = -2\ \text{LogL} + 2\ [(k-1) + s] \tag{3.30}$$

where k means the number of levels of the response variable and s stands for the number of input variables.

(3)SC: It is another adjusted form of -2 Log L. The form of SC is shown below and the parameters have the same meaning as before:

$$\text{SC} = -2\ \text{Log L} + ((k-1) + s) * \log(\Sigma f_i) \tag{3.31}$$

II. Analysis of Maximum Likelihood Estimates

There are two items in the output worthy of notice [40]:

Estimate: It explains how the input variables affect the dependent variable, given that the other predictors in the model are held constant.

95% Wald Confidence Limits: This is the Wald Confidence Interval (CI) of individual odds ratio, given the other predictors is in the model. For a given predictor variable with a

level of 95% confidence, we say that we are 95% confident that upon repeated trials, 95% of the CI's would include the "true" population odds ratio.

3.2 ANOVA

Analysis of Variance is another popular tool among researchers. There are several ways to divide ANOVA into different types. According to the structure of the data, it can be divided into balanced and unbalanced ANOVA. According to the number of dependent variables, there are ANOVA and MANOVA. According to the number of predictors, there are one-way ANOVA (a.k.a. one-factor ANOVA), two-way ANOVA (two-factor ANOVA) and multi-factor ANOVA. In our analyses, we only study one response variable each time; therefore, we will not consider MANOVA, in which there are two or more dependent variables. ANOVA can be thought of as a special case of a linear model, and hence, it has many characteristics that a general linear model has. Besides, with only a slight exaggeration, a multi-factor analysis is very similar to a two-way ANOVA. Therefore, in this section, we will focus on the special characteristics of a one-way ANOVA and two-way ANOVA.

3.2.1 One-way ANOVA [41]

There are two ways of parameterizing ANOVA; one is the cell means model while the other is the factor effects model. However, as the former is not as robust as the latter one; we only explain the factor effects model in this section.

1. Assumption

The basic purpose of a one-factor ANOVA is to compare whether there is a

difference between the means of the different groups. The assumptions of one-factor ANOVA and two-factor ANOVA are the same:

- The population distributions should be normal, and have equal means.
- Variances across all of the levels should be equal.

2. Model Expression

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \tag{3.32}$$

where the notations related will be explained below:

N: total number of observations

$n_i = 1,2,\ldots r$: number of levels of factor x

$n_j = 1,2,\ldots s$: number of observations at each level i

the mean for level i : $\mu_i = \dfrac{\sum_{j=1}^{n_j} Y_{i,j}}{n_j}$ (3.33)

the overall mean μ : $\mu = \dfrac{\sum_{i=1}^{r}\sum_{j=1}^{n_j} Y_{i,j}}{N}$ (3.34)

the difference between the mean of the sample and the mean of x at level i,

$\alpha_i = \mu - \mu_i$.

3. Hypothesis

$H_0: \mu_1 = \mu_2 = \cdots = \mu_i$ H_α: not all the μ_i are equal

4. Logic of ANOVA

There are two important parameters in this analysis. One is the sum of squares, and the other is the F test. In this part, the notations have the same meanings as before.

(1) Sum of squares: The total sum of squares of the model can be divided into the sum of the among-group sum of squares and the within-group sum of squares:

Among- group sum of squares: $$SSG = n_j \sum_{i=1}^{n_j} \alpha^2{}_i \tag{3.35}$$

Within- group sum of squares: $$SSE = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \left(Y_{i,j} - (\alpha_i + \mu)\right)^2 \tag{3.36}$$

Total sum of squares: $$TSE = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} (Y_{i,j} - \mu)^2 = SSG + SSE \tag{3.37}$$

(2) *F* test: The F-test is used for comparisons of the components of the total deviation and it can be defined as the ratio MSG/MSE (defined in 3.38 – 3.40) where the *p* value is the area of the tail outside the value given by the ratio. If it is small, then the significance level is larger than the p-value and so we can reject the null hypothesis; it indicates that the predictors are significant. A large F value indicates that there is more difference between groups than within groups.

$$MSE = \frac{1}{n_i} \sum_{i}^{n_i} [\frac{1}{n_j - 1} \sum_{j=1}^{n_j} (Y_{ij} - (\alpha_i + \mu))^2] \tag{3.38}$$

$$MSG = n_j [\frac{1}{n_i - 1} \sum_{i=1}^{n_i} (\alpha_i - \mu)^2] \tag{3.39}$$

$$F = \frac{MSG}{MSE} \tag{3.40}$$

3.2.2 Two-way ANOVA [42]

A multi-factor ANOVA can be represented by the two-factor ANOVA; we will discuss the two-way analysis for simplicity.

1. Hypothesis

Main effect of factor A : $H_0: \mu_1 = \mu_2 = \cdots = \mu_i$ where $\alpha_i = \mu - \mu_i$

H_α: not all the μ_i are equal

Main effect of factor B: $H_0: \mu_1 = \mu_2 = \cdots = \mu_j$ where $\beta_j = \mu - \mu_j$

H_α: not all the μ_j are equal

Interaction effect: $H_0: \mu_{(1,1)} = \mu_{(1,2)} = \cdots = \mu_{(i,j)}$ where $\gamma_{i,j} = \mu - \mu_j$

H_α: not all the $\mu_{(i,j)}$ are equal

2. Expression

The expression for two-factor ANOVA is:

$$\mu_{(i,j)} = \mu + \alpha_i + \beta_j + \gamma_{(i,j)} \tag{3.41}$$

3. Sums of Squares

There are at least four types of sums of squares [40].

Type I sums of squares: This type of sums of squares is also called sequential sums of squares. It can be computed as the decrease in the error sum of squares when the effect is added to a model. Type I sums of squares are appropriate for balanced analyses of variance in which the effects are specified in proper order and for trend analysis where the powers for the quantitative factor are ordered from lowest to highest in the model statement.

Type II sums of squares: Type II sums of squares can also be calculated by comparing the error sums of squares for subset models. It is the reduction in the SSE due to adding the effect to a model that contains all other effects except those being tested.

Type III sums of squares: Type III sums of squares is also referred to as the partial sums of squares. Because they do not depend upon the order in which effects are

specified in the model, this type is more popular and useful than type I sums of squares. They can also be used for unbalanced designs.

Type IV sums of squares: Type IV sums of squares are used for designs with missing cells. The results are not unique.

3.3 The Generalized Linear Model

The generalized linear models [43] are a class of linear models that includes the Poisson regression model and the gamma model, etc. The model uses a nonlinear link function to describe how the mean of a population is related to a linear predictor and allows the dependent variable to follow any distribution belonging to the exponential family of distributions (including many common distributions such as a normal distribution, an exponential distribution, a gamma distribution, etc.). With the introduction of GEEs (the generalized estimating equations) by Liang and Zeger [44], the correlated data also can be fit into a generalized linear model. Therefore, the generalized linear model can be used in more cases than the traditional linear model.

3.3.1 Assumption

The generalized linear model still assumes that the relationship between the predictors and the response variable is linear. However, it has its own different assumptions from the other linear models.

- The distribution of the dependent variable is not necessarily a normal distribution.
- The variance of the sample is constant for all observations.
- The relationship between the mean of a sample and an input variable can be linked through a nonlinear link function.

3.3.2 Expression

The linear component of this model is the same as the general linear model. If we use y_i to stand for an element of the response vector, μ_i for the mean of y_i, τ for the link function, which describes how μ_i is related to y_i, $V(\mu_i)$ is a variance function with respect to μ_i, and σ is the variance of y_i, then the expression for the model can be described as follows:

Linear components: $$y_i = x_i^T \beta \tag{3.42}$$

Link function: $$\tau(\mu_i) = x_i^T \beta \tag{3.43}$$

Variance function: $$\sigma(y_i) = \frac{\phi V(\mu_i)}{w_i} \tag{3.44}$$

3.3.3 Link Function for Different Distributions

The link functions vary from distribution to distribution. The differences among these link functions are the range and scale of the probabilities they produce.

1. The General Linear Model

Distribution: normal: $$f(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma} \right)^2 \right] \quad for(-\infty,+\infty) \tag{3.45}$$

Link function: identity: $$\gamma(\mu) = \mu \tag{3.46}$$

2. Logistic Regression

Distribution: binomial: $$f(r) = \binom{n}{r} \mu^r (1-\mu)^{n-r}, r = 0,1,2,...n \tag{3.47}$$

Link function: $$\log it \gamma(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \tag{3.48}$$

3. Poisson Regression in Log-linear Model

Distribution: Poisson: $$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0,1,2... \quad (3.49)$$

Link function: $$\log \gamma(\mu) = \log(\mu) \quad (3.50)$$

4. Gamma Model with Log Link

Distribution: gamma $$f(y) = \frac{1}{\Gamma(\nu) y}\left(\frac{y\nu}{\mu}\right)^{\nu} \exp\left(-\frac{y\nu}{\mu}\right), y \in (0,\infty) \quad (3.51)$$

Link function: $$\log \gamma(\mu) = \log(\mu) \quad (3.52)$$

3.3.4 Output Analysis [38, 41]

1. Goodness of Fit

When evaluating the adequacy of the generalized liner model, one criterion, the deviance of the model is often used. The rule is that the deviance is compared with its asymptotic χ^2 with the same degree of freedom distribution to find the *p*-value; if the *p*-value lies in a certain allowable range, then the model fits the sample well.

2. Type I analysis & Type III analysis

One fundamental application of the generalized linear model is to find out all the prominent predictors to the dependent variable. The importance of the variables can be measured by χ^2, which presents the difference in deviances of fitted log likelihoods between successive models. In other words, if the value of χ^2 for a variable is larger than those of the other variables, then this variable is the most important variable to the dependent variable.

I. Type I Analysis

One special property of type I analysis is that its results depends on the order by which the independent variables are input into the model. The analysis can be run in this way such that at the beginning, only an intercept term is input into the model. In subsequent steps, one of the additional effects enters into the model. During each step, a likelihood ratio statistic is computed between successive pair of models. *P*-values are calculated based on the asymptotic distributions of the likelihood ratio statistics. By comparing the corresponding p-value for each variable to a significance level, we can judge whether a variable is significant in the model.

II. Type III Analysis

A Type III analysis does not depend on the order in which the terms for the model are specified. This analysis consists of specifying a model and calculating likelihood ratio statistics for type III contrasts for each term in the model. Under this analysis, a maximum likelihood statistic is calculated through constrained optimization under the condition that the Type III function of the parameters is equal to 0 and can be defined as below:

$$LR = 2[L(\hat{\beta}) - L(\tilde{\beta})] \tag{3.53}$$

in which $\hat{\beta}$ is the unconstrained estimate, $\tilde{\beta}$ is the constrained parameter estimate, with an asymptotic χ^2 distribution under the condition that the Type III function of the parameters is equal to 0, with degrees of freedom equal to the number of parameters associated with the effect. However, this analysis is time-consuming; it is recommended that the researcher reduce the sample size before applying it.

3.3.5 The Poisson Regression Model

The Poisson regression model is another common linear model with a count variable as a response variable and it assumes that the dependent variable follows a Poisson distribution.

1. Expression[45]

The frequently- used expression for the Poisson regression model is:

$$\log(E(Y)) = \log(t) + \beta^T X \tag{3.54}$$

where log(t) is an offset.

2. Adequacy of the model

After developing a Poisson regression model, adequacy of the model is the first thing to check since overdispersion [46]often occurs in a Poisson regression model. Overdispersion occurs when there is more variability than a model is expected to have. Typically, if the deviance or Pearson's χ^2, divided by the degrees of freedom, is greater than 1, then the model may be overdispersed.

One approach of adjusting for overdispersion [47] derives from the theory of quasilikelihood and the basic idea is as follows:

- Import a scale parameter Φ, then $E(y) = \mu$ and $Var\ (y) = \Phi\mu$; if $\Phi > 1$, then the model is overdispersed.
- $\hat{\Phi} = \frac{Pearson\chi^2}{N - p}$ [45], where N is the number of sample cases and p is the number of parameters.
- Φ is unknown, and therefore, under this modification, the frequently-used method

called the Fisher-scoring procedure for the estimated covariance matrix is changed from $(X^T WT)^{-1}$ (standard error) to $\Phi(X^T WT)^{-1}$, where W= *diag* $(w_1, w_{2,} \ldots w_n)$.

This adjustment will be further discussed later in chapter 6.

3.4 The Generalized Linear Mixed Model

The generalized linear mixed model [48] is a statistical model that extends the generalized linear model by incorporating a normally distributed random effect [49], which is an effect whose levels are assumed to be selected randomly.

3.4.1 Assumptions

- The relationship between the exploratory variables and the response variable is linear. Otherwise, a non-linear mixed model should be applied.
- The random effect should follow a normal distribution; otherwise, a hierarchical linear mixed model will be employed.
- The model cannot be used when the data are correlated.

3.4.2 Expression

The generalized linear mixed model can be defined in this way:

$$y = X\beta + Z\gamma + \varepsilon \tag{3.55}$$

where X and Z are known design matrices, the error $\varepsilon \sim N(0,R)$; the random effect $\gamma \sim N(0,G)$, where G, R are variance matrices. The model has the following properties:

- $\eta = X\beta + Z\gamma$ (3.56)

  is a linear predictor with combined fixed effects and random effects.

- $E(y|\gamma) = \tau^{-1}(X\beta + Z\gamma) = \mu$ (3.57)

  $\tau^{-1}(.)$ is an inverse link function and its selection is typically based on the error distribution.

- $Var(\gamma) = G$ (3.58)

  The variance function is used to model non-systematic variability.

- $Var(y|\gamma) = \sqrt{D}R\sqrt{D}$ (3.59)

where D is a diagonal matrix containing the variance functions and R means "R-side" random effect and $Var(y/\gamma)$ is the variance matrix in a model with only R-side random components.

3.4.3 Estimation Method

In the generalized linear mixed model, most estimation approaches also rest on some likelihood principle. To obtain maximum likelihood estimates, we should maximize the marginal likelihood shown below:

$$l(\beta, y) = \int f(y|\gamma)p(\gamma)d\gamma \quad (3.60)$$

3.4.4 Output Analysis

1. Assumptions

The generalized linear mixed model is a very complicated model and its output analysis varies in different sample data. In our later research, we will utilize it for a binary response variable, and hence we will focus on the logistic regression with random effects. In this case, there are n $(=s+t)$ groups of subjects, two properties (P_1, P_2) and one event y. Suppose that s subjects are randomly selected to have property 1and t subjects

have property 2. The purpose of the model is to compare the occurrence of the event for the two properties among different groups. Therefore, the event that a subject is chosen to have property 1 or 2 can be counted as fixed effects and the group effects are random effects. We define the following terms:

- n_{i1} and n_{i2} represent the number of subjects reported to have property 1 or 2 and i stands for a group.
- p_{i1} and p_{i2} stand for the probability of the event occurring to property 1 or 2.
- γ_i is a random effect such that $y_{i1}|\gamma_i \sim B(n_{i1}, p_{i1})$; $y_{i2}|\gamma_i \sim B(n_{i2}, p_{i2})$; $\gamma_i \sim N(0, \sigma^2)$

and $$\log(\frac{p_{i1}}{1-p_{i1}}) = \beta_0 + \beta_1 + \gamma_i \qquad (3.61)$$

and $$\log(\frac{p_{i2}}{1-p_{i2}}) = \beta_0 + \beta_2 + \gamma_i \quad . \qquad (3.62)$$

2. Fitness Evaluation

Whether the model fits the sample can be judged through the fit statistics table, which has three criteria, pseudo-likelihood, generalized χ^2 and generalized χ^2 divided by its degrees of freedom. We consider the measure, $\chi 2$ / DF. If its value is very close to 1, then it indicates that the model fits the sample pretty well.

3. Random Effects Prediction

In order to predict the probability of the random effects, two methods are utilized; one uses prediction in the whole model, the other only rests on the fixed effects. In this procedure, two kinds of predictions of probabilities are generated for each observation (take one observation with property 1 for example) and they are shown below,

$$E(\hat{y}_{i1}|\gamma_i) = \frac{1}{1+\exp\{-\hat{\beta}_o - \hat{\beta}_1 - \hat{\gamma}_i\}} \tag{3.63}$$

$$E(\hat{y}_{i1}) = \frac{1}{1+\exp\{-\hat{\beta}_o - \hat{\beta}_1\}} \cdot \tag{3.64}$$

3.5 Comments

Each linear model discussed here has its own unique characteristics and has different applications. The logistic regression model is by far the most popular classification model, extending the techniques of the multiple regression model to research situations in which the outcome variable is categorical, especially for the target, which is a binary variable. However, for a rare occurrence of mortality, adjustments need to be made to the model. Analysis of Variance is often applied to compare the means of different sample data, no matter whether the sample is balanced or unbalanced. It is widely employed in sociology and psychology and so on. The generalized linear model extends the applications of the general linear model to a dependent variable following a non-normal distribution such as a gamma distribution or a Poison distribution as well as to the correlated data. Therefore, it is a popular research tool in the cost analysis, which often follows a gamma distribution and longitudinal analysis in the healthcare industry. The generalized linear mixed model can be applied for dichotomous, ordinal and nominal outcomes as well as ranked data. Therefore, it can be employed for assessing the trends in disease rates, modeling counts or predicting the probability of occurrence in time series.

In this chapter, we briefly introduced the theories and methods related to our analyses, most of which are utilized to study one dependent variable. So far, we have only discussed cases when there exist linear relationships between the predictors and the

response variables. What if the relationship between the exploratory variables and the dependent variable is non-linear, or what if there is no specific predicted variable? In those cases, we will employ supervised/unsupervised machine learning, which we will introduce in the next chapter.

CHAPTER IV

UNSUPERVISED /SUPERVISED MACHINE LEARNING

Machine learning [50] is a subfield of data mining, and it was conceived in the early 1960's with the clear objective to design and develop algorithms and techniques that implement various types of learning, mechanisms capable of inducing knowledge from examples of data. Machine learning is widely applied to medical diagnosis, bio-informatics and object recognition in computer vision and so on. According to the causal structure of the model, the learning can be divided into two types: I. unsupervised learning, II. supervised learning. Unsupervised machine learning includes:

- Cluster analysis (e.g. means, hierarchical algorithms)
- Association rules (a.k.a. market basket analysis)
- Collaborative filtering

Supervised machine learning contains:

- Classification trees (e.g. decision trees)
- Regression analysis (including the logistic regression)
- Neural networks
- *K-nearest* neighbors
- Rule induction
- Support vector machine

We have already discussed some of the algorithms listed above, such as the regression analysis in chapter III, and some of these algorithms we will not use in our

research. Therefore, in this section, we will elaborate on cluster analysis, association rule analysis, the decision tree model, the neural network model and rule induction.

4.1 Unsupervised Machine Learning

In unsupervised learning situations, all observations are assumed to be caused by latent variables and there is no distinction between the independent and the dependent variables. Although it does not have a target variable, it does have some purposes. Unsupervised learning is primarily composed of two techniques: cluster analysis and market basket analysis, which are often utilized in consumer market analysis or patients' medical conditions analysis.

4.1.1 Cluster Analysis

The purpose of clustering techniques is to detect similar subgroups among a large collection of cases and to assign the homogeneous observations into one cluster. The clusters are assigned a sequential sequence number to identify them in results reports. In most cases, it is performed to reduce sample size and to prepare for another analysis. The primary algorithms of clustering are *k*-means clustering, EM (Expectation Maximization) clustering and hierarchical clustering.

1. K –means Clustering

The k –means algorithm [24] is an old and simple method applied in cluster analysis. The process can be operated in this way: first, a fixed number of clusters, k is given; then, observations are assigned to those clusters so that the means across clusters are as terms different from each other as possible. The difference between observations is

measured in a distance measure such as Euclidean or Squared Euclidean.

2. EM Clustering

Expectation Maximization [51] is another algorithm for clustering and its goal is to find the most likely set of clusters for the observations. The basis of this approach is a body of statistical theory, called finite mixture, in which a set of probability distributions represent k clusters. The technique consists of two steps, E-step (estimation) and M-step (maximization).

E-step: to compute the conditional expectation:

$$Q(\theta;\theta^{(t)}) = E_{\theta^{(t)}}(\log L(\theta, y)|\, y_{obs}) \tag{4.1}$$

M-step: to find the θ , which maximizes the expectation:

$$\hat{\theta}^{(t+1)} = \arg\max_{\theta} Q(\theta|\hat{\theta}^{(t)}) \tag{4.2}$$

These two steps are repeated until

$$\left|L(\theta^{(t+1)}) - L(\theta^{(t)})\right| < \varepsilon \tag{4.3}$$

where L (θ, y) is a likelihood function, $Q(\theta,\theta^{(t)})$ is a conditional expectation, θ is an unknown parameter and $\hat{\theta}^{(t)}$ is the estimate of the unknown parameter at iteration t >0, and ε is a very small positive number.

3. Hierarchical Clustering

Hierarchical clustering [52] is another algorithm for clustering, in which Ward's method is utilized. We first defined several terminologies, error sum of squares, total sum of squares, τ and R^2 and they are shown in Equations (4.4) – (4.6).

Error Sum of Squares: $$\delta = \sum_i \sum_j \sum_k \left| X_{ijk} - \bar{x}_{i.k} \right|^2 \quad (4.4)$$

Total Sum of Squares: $$\tau = \sum_i \sum_j \sum_k \left| X_{ijk} - \bar{x}_{.k} \right|^2 \quad (4.5)$$

R^2: $$R^2 = \frac{\tau - \delta}{\tau} \quad . \quad (4.6)$$

In the above equations, x_{ijk} denotes the value for the variable k in observation j belonging to cluster i; $\bar{x}_{ik}$ stands for the cluster mean for that variable; $\bar{x}_{.k}$ denotes the grand mean for that variable. Using Ward's method, the like units should be clustered together in each step; therefore, the error should be computed and minimized (or if the selection criterion is R^2 instead, then it should be maximized) in each step. At the end of the process, a single large cluster of size n will be formed. The process starts with all sample units in n clusters of size 1. Then, the clusters or observations are combined in such a way to minimize the error or maximize R^2. In that way, n-1 clusters are formed in the first step, n-2 clusters are generated in the second step, and this process will not stop until all sample units are combined into one cluster.

4. Comments

We apply the last two methods of clustering. The merit of clustering analysis is that it is simple and easy to perform. It provides a quick way to explore the data structure without providing an explanation or interpretation, especially if the objects are classified into many groups. However, it cannot be used to predict either the relationships between different groups or the relationships between the predictors and the predicted variables.

4.1.2 Association Rule Analysis

Association rule analysis [53], often referred to as market basket analysis, is used to investigate relationships or associations between specific values of categorical variables in large data sets. This technique can be used to uncover hidden patterns in large data sets, and hence, it is especially well suited for data mining.

1. Basic Rule

The goal of association rule analysis is to find subsets of variable values s_1, $s_{2...}s_n$ such that the probability of each of the variables simultaneously assumes that a value within their respective subsets,

$$P\left(\bigcap_{i=1}^{n}(x_i \in s_i)\right) \tag{4.7}$$

is relatively large. If we apply dummy variables, and X_i stands for a variable, x_{ij} is the j^{th} observation of the variable X_i , i means a set of all items associated with the variable X_i, $S(I)$ is the support of I, and t is the lower support bound; then the formulation of the rule is:

$$P(\bigcap_{i\in I}(X_i = 1)) = P(\prod_{i\in I} X_i = 1) \tag{4.8}$$

and the purpose is to find all the item sets I_m that meet the following condition:

$$\{I_m | S(I_m) > t\} \tag{4.9}$$

2. Confidence, Support and Lift

According to the rules, a high support item returned can be partitioned into two subsets, A and B such that

$$\begin{cases} A \cup B = I \\ A \cap B = \phi \end{cases} \tag{4.10}$$

If A is the antecedent, and B is the consequence, then the confidence can be defined in this way: given that event A occurs, the probability that B also happens is equal to

$$P(B/A) = \frac{\#of(A \Rightarrow B)}{\#A} \tag{4.11}$$

The support is defined as the ratio of the number of both events, A and B, appearing together compared to the number of total transactions:

$$P(A \cap B) = \frac{\#of(A \cap B)}{\#ofall} \tag{4.12}$$

The lift can be written as the ratio of confidence and its expected confidence:

$$P(L) = \frac{P(B|A)}{E(B|A)} \tag{4.13}$$

When studying the associations between items, the confidence, the support and the lift should be considered. The basic rule is that the higher the value of these three measures, the stronger the relations.

3. Application

Association rule analysis is one of the popular marketing strategies; many retailers utilize it to find potential customers and increase their response rate when they mail their product catalogs to their targeted customers. In our research, we apply it to medical procedures. For instance, we explore whether the patients receive procedures related to heart disease along with what additional procedures and treatments are also given to them.

4.2 Supervised Machine Learning

Although unsupervised learning has some merits, it cannot be utilized for accurate prediction. Therefore, supervised learning algorithms are frequently used. In our research, we often employ the decision tree model, the neural network model, and the rule induction model. We then compare these models to find the optimal one. In this section, we will discuss these models.

4.2.1 Decision Tree Model

We can apply three algorithms to the decision tree model in SAS EM (Enterprise Miner 6.2): (1) the default tree methodology in EM, (2) CHAID (Chi-square automatic interaction detection), and (3) CART (Classification and regression tree).

1. The Default Algorithm

We first discuss the default methodology [52] for the decision tree model in SAS EM; it is a little different from the other two algorithms. The main idea of this approach can be explained in this way:

- The split is based on some measures, either a node impurity measure or χ^2 test; the F test criteria, i.e. the p- value can also be utilized.
- If the node has many observations, then a sample is used for the split search.
- If the target is binary, nominal or ordinal, then the sample should be as balanced as possible.
- If after consolidation, the number of possible splits is greater than the number specified, then a heuristic search is used.
- At the beginning of the heuristic algorithm, each consolidated group of

observations is assigned to a different branch; then at each step, the two branches are merged; the process will not stop until no group can be assigned.

2. CHAID Algorithm

The algorithm, CHAID [54], applied in decision tree model building, was originally proposed by Kass in 1980. Although this method can be utilized for both a continuous dependent variable and a categorical target, here, we discuss the latter .The method consists of three steps: merging, splitting and stopping.

Merging step: In this step, a significant value α_{merge} should be defined at the beginning; then, by the equations shown in (4.14) – (4.16), the variable X^2 and the p-value can be computed. If the significance (i.e., p-value) for a given pair of categories is larger than α_{merge}, then it will merge the respective predictor categories; otherwise, the adjusted p-value will be computed using the equation displayed in (4.17).

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(\eta_{ij} - \varpi_{ij})^2}{\varpi_{ij}} \tag{4.14}$$

where η_{ij} is the observed cell frequency, ϖ_{ij} is the estimated expected cell frequency for cell ($x_n = i$, $y_n = j$) and there are i categories of predictor X and j categories of the dependent variable Y.

$$\varpi_{ij} = \frac{\sum_{k=1}^{I}\omega_{i,k}\sum_{k=1}^{J}\omega_{k,j}}{N} \tag{4.15}$$

$$p = P_r(\chi_d^2 > X^2) \tag{4.16}$$

where χ_d follows a χ^2 distribution with degrees of freedom $d = (J- 1)(I- 1)$. The adjusted p-value is calculated as the p-value multiplied by β (defined in Equation (4.18)). Suppose that the total number categories of X *is* I, and after several merges, r classes are left. If $2 \leq r < I$, then

$$p_{adjusted} = p * \beta \tag{4.17}$$

$$\beta = \sum_{k=0}^{r-1} (-1)^k \frac{(r-k)^I}{k!(r-k)!} \tag{4.18}$$

Splitting step: In this step, the adjusted p-value of each predictor will be compared with the user-defined α_{split}. If the adjusted p-value is not greater than the α_{split}, then its corresponding predictor is used for the split; otherwise, the node is the terminal one.

Stopping step: The splitting step will continue until no more splits can be performed. During development, a tree with a binary target is generated by repeatedly using the above three steps on each node starting from the root node using the CHAID algorithm.

3. CART Algorithm

The Classification and regression tree methodology [24] was first introduced in 1984 by UC Berkeley and Stanford researchers Leo Breiman et al. This algorithm can be implemented in two cases: the classification tree is used for a categorical predicted variable while the regression tree is used for a continuous predicted variable. The primary techniques included in this algorithm are how to select splits and how to prune a tree [49].

(1) Selecting Splits: The rule of thumb is that the split at each node will improve the predictive accuracy to the greatest extent. For classification trees, there are several impurity measures, including χ^2, G^2 or the Gini index (shown in (4.19)), among which the

Gini index is the most frequently used criterion,

$$Gini(S) = \sum_{i \neq j} p_i p_j = 1 - \sum_{j=1}^{n} p_j^2 \tag{4.19}$$

where S is a set containing n classes, and p_j is the relative frequency of class j in S. For regression trees, a least squares deviation is most frequently used. If $N_w(S)$ represents the weighted number of cases in node S, w_i is the value of the weighting variable for case i, f_i is the value of the frequency variable, y_i represents the value of the dependent variable, and $\bar{y}(S)$ is the weighted mean for node S, then a least squares deviation can be computed as:

$$LSD = \frac{1}{N_w(S)} \sum_{i \in S} w_i f_i (y_i - \bar{y}(S))^2 \tag{4.20}$$

(2) Pruning Trees: For the CART algorithm, it is more important to prune back to find the optimal tree than to find when to stop splitting. A tool called V-fold cross validation can be used to prune a tree, and the process can be addressed in the following way:

- Partition the entire data set into V folders.
- Train V folders on different combinations of V-1 folds and estimate the error for the fold that is left out of the tree at each time.
- Estimate tree accuracy based on the error measurements.
- Find the design parameters to minimize the error.
- Refit the tree using all of the data, the chosen parameters.

4. Comments

The CHAID algorithm can build non-binary trees, which makes it popular among market researchers while the CART algorithm is always building binary trees. One of the

characteristics of the decision tree model in SAS is that it can demonstrate the importance of the variables. A rule of thumb is that the level where a variable lies indicates its importance. The higher the level, the more vital the variable. In addition, compared to other methodologies, the decision tree model has three primary advantages. First, it is non-parametric and non linear and hence it does not require specifications of a data distribution. Moreover, it can automatically group the missing values into one category without preprocessing them. In addition, its output is easy to understand and interpret. Therefore, the decision tree model is a very popular tool for decision makers.

4.2.2 Neural Network Model

The neural network model [55], often simply called Neural Nets, originated from early understandings of the structure and function of the human brain. The type of the model that we utilize is MLP (Multilayer perceptron), and the algorithm is back propagation. In this section, we will focus on this algorithm and this type of neural network.

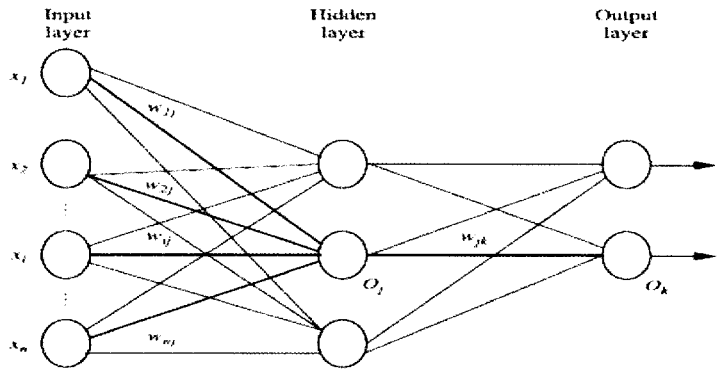


Figure 4.1. A Feed-forward Neural Networks

1. Basic Rule

As figure 4.1 [55] demonstrates, the structure of an MLP consists of neurons organized in three layers: input layer, hidden layer and output layer. Weights W_{ij} are assigned to each connection between the input neuron and the middle neuron, and between the middle neuron and the output neurons. With nonlinear transfer functions, hidden neurons can process complex information received from input neurons and then send processed information to output layer for further processing to generate outputs. The whole process of MLP can be described in the following way:

Step 1: The inputs x_i are combined together to form a weighted sum of inputs and the weights w_i of connecting links.

Step 2: A transformation is performed to convert the sum to an output via a transfer function f, and a transformation can be expressed in Equation (4.21). Among all transfer functions, the logistic function is the most popular one.

$$F = f(\sum_i w_i x_i) \tag{4.21}$$

Step 3: Network output values are calculated and compared to the target values and the weights of the connections are changed to produce a better approximation to the desired output. The way of adjusting the weights is demonstrated in Equations (4.22).

$$W_{ij}^{new} = W_{ij}^{old} + \Delta w_{ij} = W_{ij}^{old} + (-\eta \frac{\partial E}{\partial_{ij}}) \tag{4.22}$$

in which E is the objective function and η is the learning rate which controls the size of the gradient descent step.

Step 4: The above step will repeat until the output is ideal.

2. Comments

Besides MLP, there are several other types of neural networks such as auto neural networks, linear networks, Bayesian networks, etc. A neural network can be thought of as a complicated combination of regression models, but it is a powerful classifier. It can handle problems with many parameters, and it tends to fit the training data well and thus has low bias. Therefore, in most cases, this model outperforms other models. If it is the optimal model, we should use another tool to express the information in an easily-understood way.

4.2.3 The Other Models

1. Rule Induction

Rule induction [29] is another tree-based algorithm for classification. It is mainly utilized to improve the classification of rare events. Usually rules are expressions of the form:

if (*attribute*-1, *value*-1) *and* (*attribute*-2, *value*-2) *and* $\cdots$ *and* (*attribute*$-n$, *value*$-n$) *then* (*decision, value*).

Some rule induction systems induce more complex rules, in which values of attributes may be expressed by the negation of some values or by a value subset of the attribute domain.

2. Memory-Based Reasoning

MBR (Memory- Based Reasoning) [52] tries to mimic human behavior in an automatic way. MBR needs a distance measure to assign the dissimilarity of two observations and a combination function to combine the results from the neighboring points to achieve an answer. The distance measure is the k-nearest neighbor algorithm, in

which the *k*- nearest neighbors are determined by the Euclidean distance between an observation and the probe. By MBR means, it is easier to generate examples than to generate rules, and hence this method is very attractive.

4.3 Data Mining Model Comparison

After developing various models, several measures [56] can be used to find the optimal model, including the AIC (Akaike Information Criterion), the BIC criterion (Bayesian Information Criterion), the ROC (Receiver Operating Characteristic) curve, the Lift Chart and the misclassification rate, etc. We have already discussed the first two criteria in Chapter III, and hence we will introduce the other measures in this section.

4.3.1 ROC Curve

The ROC curve is a graph (shown in Figure 4.2, part of Figure 6.7) that measures the predictive accuracy of a model. In a Cartesian plane, the x-axis represents the false positive value (a.k.a. 1-specificity) and the y-axis represents the sensitivity value. Each point in the curve corresponds to a particular cut-off. The model where the ROC curve is leftmost is the best model among the different models.

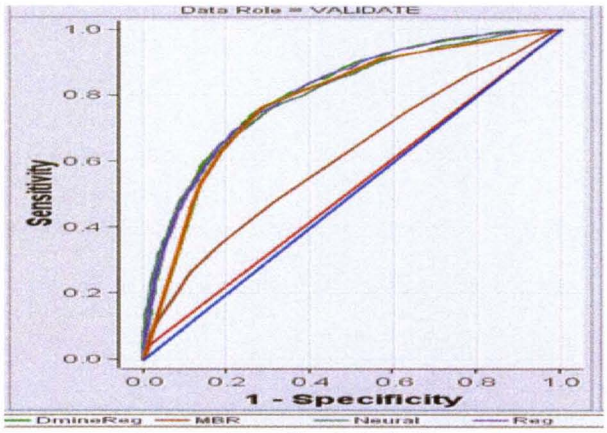


Figure 4.2. The ROC Curve

4.3.2 Lift Chart

A lift chart (displayed in figure 4.3[56]) is another graph criterion of measuring the predictive accuracy of models. The scored data set is sorted by the probabilities of the target event in descending order; observations are then grouped into deciles. For each decile, a lift chart can calculate the ratio between the result obtained with a model and the result obtained without a model that is based on randomly selected records. In the chart, it is represented by the horizontal base line. Lift charts show the percentage of positive response or the lift value on the vertical axis. If the distance between a curve and the base line is the greatest, then the model to which the curve corresponds is the best one.

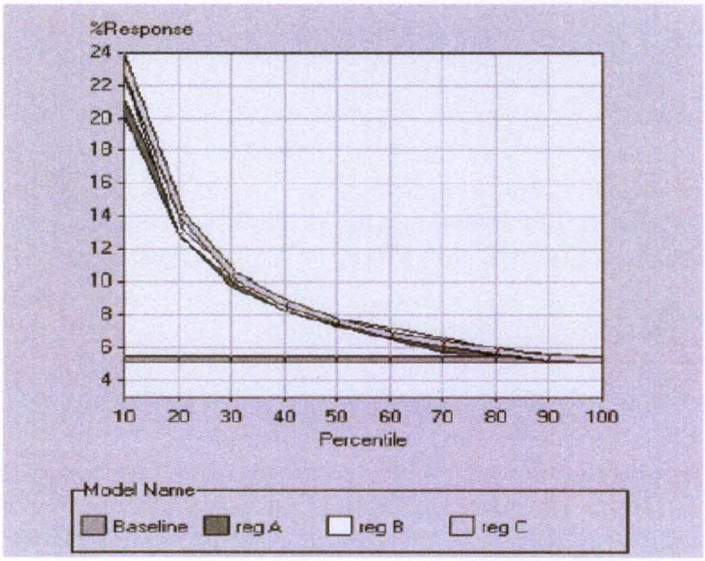


Figure 4.3. The Lift Chart

4.3.3 Misclassification Rate

When comparing the models, we often utilized the misclassification rate. We divide the sample into decision regions R_m, for each class C_k. Misclassification occurs if the input vector x belonging to C_i is assigned to C_j, where $i \neq j$, the

misclassification rate can be computed as:

$$p_w = \sum \int_{R_{i_{ir}}} p(x, C_j) dx \tag{4.23}$$

The optimal model should be the one whose misclassification rate is the minimum.

Up to this point, we have discussed the basic data processing methodologies, basic concepts and the primary theories or algorithms that we will often use in our research. As we move into the chapters ahead, we will become immersed in studying diabetes outpatients and inpatients in the Medicare population without further discussing these theories and algorithms.

CHAPTER V

COST ANALYSIS OF MEDICARE OUTPATIENTS WITH DIABETES

Outpatients with diabetes use considerable Medicare resources, and hence it is essential to take measures to reduce the costs. The primary purpose of this chapter is to address this problem from prevention and medical services perspectives. We will first introduce some knowledge about descriptive statistics and correlation analysis; then we propose methods to decrease Medicare payments through analyses of claims data and data from revenue centers for outpatients.

5.1 Descriptive Statistics

Descriptive statistics [57] are used to quantitatively describe the basic features of the data in a study and they include four types: (1) measures of central tendency including the arithmetic mean, median and mode, etc.; (2) measures of dispersion such as the standard deviance; (3) measures of association such as odds ratio and correlations coefficients; (4) a non-parametric analysis such as kernel density estimation. In this section, we will talk about the last two types.

5.1.1 Kernel Density Estimation

In statistics, kernel density estimation [57] is a non-parametric way of estimating the probability density function of a random variable. For our research, we use the

univariate case. The kernel estimator for the univariate case can be defined as:

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}) \tag{5.1}$$

where $X_1, X_2, \ldots X_n$ are independent and identically distributed random variables; f is the density function; K stands for some known density function and h is a smoothing parameter called the bandwidth. As an illustration, given some data about a sample of a population, kernel density estimation makes it possible to extrapolate the data to the entire population. To better display the distribution, it is important to make sure that the smoothness of the graph is reasonable, which is decided by the bandwidth. One simple method to choose the optimal bandwidth is to find the minimum of the asymptotic mean integrated squared error (AMISE, defined below) with respect to h,

$$\text{AMISE} = \frac{R(K)}{nh} + \frac{1}{4}\sigma_k^4 h^4 R(f''(x)) \tag{5.2}$$

where $R(K) = \int K^2(x)dx$.

5.1.2 Pearson Correlation Analysis

Correlation analysis is a common method of studying the relations between two or more variables. The typical measure for the linear relationship analysis is the Pearson correlation coefficient. The Pearson Product-Moment Correlation coefficient r [58], simply called the Pearson coefficient gives information about the degree and the directions of how the two variables are related. Its computation is based on covariance and its value range is [-1, 1]. The values ± 1 mean that there exists a perfect positive or negative linear relationship between the two variables; the intervals of [-1, - 0.75) or [0.75, 1) indicate a high degree of correlation; the intervals of [-.25, 0) or (0, 0.25]

indicate a low degree of correlation; the value 0 means that there is no predictability. When it comes to the sample analysis, the sample coefficient R is seldom utilized; instead, R^2 is frequently utilized to display the relationship between the dependent variable and the predictors. The formula for r and R are shown below.

$$r = \frac{\operatorname{cov}(x, y)}{\sqrt{\sigma_x{}^2 \sigma_y{}^2}} = \frac{E((x-\mu_x)(y-\mu_y))}{\sqrt{E(x-\mu_x)^2 (y-\mu_y)^2}} \tag{5.3}$$

$$R = \frac{\sum_i ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{5.4}$$

In the above two formulas, *cov (x,y)* denotes the covariance; $\mu_x, \mu_y, \sigma_x{}^2, \sigma_y{}^2$ respectively represent the expectations and the variances of the variables *x*, *y*; $\bar{x}, \bar{y}$ denote the sample mean of *x* and *y*.

5.2 Cost Analysis using Data from Revenue Center

In this section, we use data from the revenue center and demographic data for diabetes outpatients in Medicare from the CMS CCW data [18] to find ways to decrease the costs through the general linear model and the generalized linear model as well as Pearson correlation analysis.

5.2.1 Kernel Density Estimation

After using Filter and Query in SAS EG (Enterprise Guide) 4.1 to select the necessary variables and combine two data sets together, we perform KDE (kernel density

estimation) to see the differences in the costs among different races. The SAS code and the results are shown below:

```
PROC SORT DATA=SASUSER.RANSAMPLE   OUT=SASUSER.SRANSAM;
BY BENE_RACE_CD  BENE_SEX_IDENT_CD;
PROC KDE DATA=SASUSER.SRANSAM;
UNIVAR REV_CNTR_TOT_CHRG_AMT/ GRIDL=0 GRIDU=1500
METHOD=SNR OUT=SASUSER.KDECHAR; RUN;
```

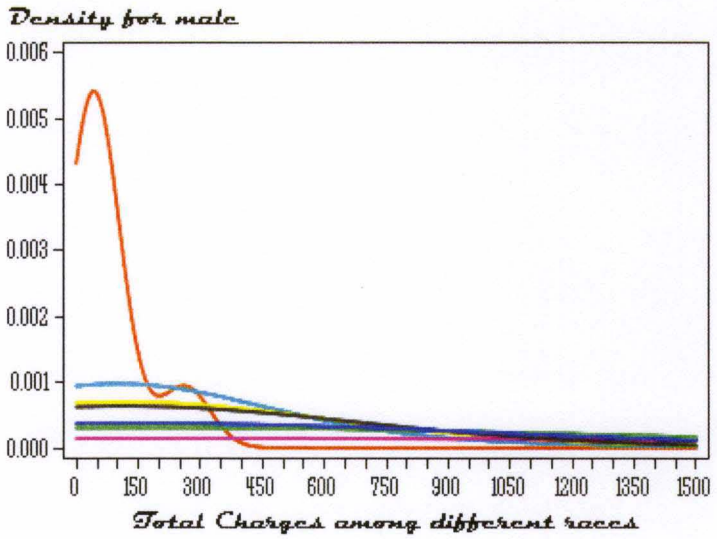
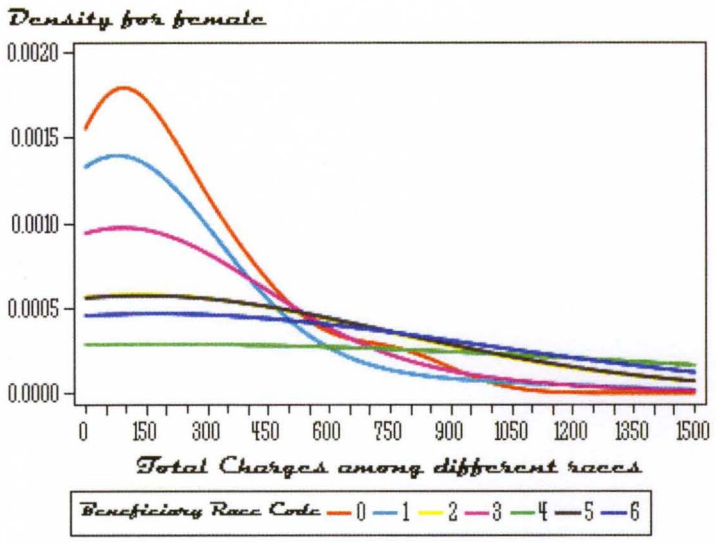


Figure5.1. KDE of Costs among Different Races (Male)



Beneficiary race code:0: Unknown 1: White 2: Black 3: Other 4: Asian 5: Hispanic 6: North American Native

Figure 5.2. KDE of Costs among Different Races (Female)

Figures 5.1 & 5.2 visualize the distributions of the costs; the densities of costs for males and females follow gamma distributions and all of them approach zero after 1500 dollars. The costs of the whites have a higher probability of being lower compared to all other races. There exist differences between the males and the females; the densities arrive at the peak when the costs reach 50 dollars for males and 100 dollars for the females.

5.2.2 Statistical Model Analysis

We want to find the reasons for Medicare payments from diagnosis information utilizing the linear statistical models. Prior to analysis, we used the One-Way Frequency in SAS EG to find the top 20 procedures displayed in Table 5.1[23].

Table 5.1 Top 20 HCPCS Codes

1	82962	Glucose, blood by glucose monitoring device(s) cleared by the FDA specifically for home use
2	G0001	Routine venipuncture for collection of specimen(s)
3	97110	Therapeutic procedure, therapeutic exercises to develop strength and endurance, range of motion and flexibility
4	85025	Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count
5	85610	Prothrombin time;
6	80048	Basic metabolic panel
7	80053	Comprehensive metabolic panel
8	83036	Hemoglobin; glycosylated (A1C)
9	90999	Unlisted dialysis procedure, inpatient or outpatient
10	97530	Therapeutic activities, direct (one-on-one) patient contact by the provider (use of dynamic activities to improve functional performance),
11	Q4055	Injection,
12	80061	Lipid panel
13	97116	Therapeutic procedures ;gait training (includes stair climbing)
14	97112	Therapeutic procedure,; neuromuscular reeducation of movement, balance, coordination, kinesthetic sense, posture, and/or proprioception for sitting and/or standing activities
15	99212	Office or other outpatient visit for the evaluation and management of an established patient; Physicians typically spend 10 minutes face-to-face with the patient and/or family.
16	93005	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report
17	99213	Office or other outpatient visit for the evaluation and management of an established patient physicians typically spend 15 minutes face-to-face with the patient and/or family.
18	A4657	Syringe, with or without needle, each
19	84443	Thyroid stimulating hormone (TSH)
20	71020	Radiologic examination, chest, two views, frontal and lateral;

Some of extracted 20 HCPCS codes are regular diabetes examination procedure such as A1C test (HCPCS: 83036) and some of them are procedures needed by diabetes complications such as an Electrocardiogram (HCPCS: 93005). The 20 codes are recorded as binary 0-1 indicator functions in 20 columns with SAS coding. We can use

Analyze->ANOVA->Linear Model with costs as the dependent variables; the newly-generated 20 indicator variables were used as the class variables in the model. The results are given below:

Table 5.2 Overall the General Linear Model Information

R-Square	Coeff Var	Root MSE	REV_CNTR_TOT_CHRG_AMT Mean
0.033549	434.3811	1737.414	399.9746

The r-square value is 3.3%, which means that 3% of the variability of the costs can be explained by the above 20 factors. Although this is a small number, it gives us an idea of what is significant. Table 5.3 shows that the 20 variables are significant to the costs. However, since the distributions of the costs are gamma distributions, we should also consider the generalized linear model with a gamma distribution.

Table 5.3 Type III Sum of Squares

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Blood glucose monitoring	1	144951012755	144951012755	48019.2	<.0001
Routine venipuncture	1	135540500264	135540500264	44901.7	<.0001
Therapeutic exercises to develop strength and endurance	1	76340503128	76340503128	25290.0	<.0001
Blood count	1	60850533348	60850533348	20158.5	<.0001
Prothrombin time	1	57263486670	57263486670	18970.2	<.0001
Basic metabolic panel	1	37054519078	37054519078	12275.4	<.0001
Comprehensive metabolic panel	1	30419948211	30419948211	10077.5	<.0001
Hemoglobin; glycosylated	1	37500526373	37500526373	12423.1	<.0001
Unlisted dialysis procedure,	1	417296629150	417296629150	138241	<.0001
Therapeutic activities, direct patient contact by the	1	34114806835	34114806835	11301.5	<.0001
Injection	1	169829929901	169829929901	56261.0	<.0001
Lipid panel	1	26676907909	26676907909	8837.49	<.0001
Therapeutic procedure, gait training	1	29359815777	29359815777	9726.28	<.0001
Neuromuscular reeducation of movement, balance	1	24582913341	24582913341	8143.79	<.0001
office or other outpatient visit, typically 10 minutes	1	20034618172	20034618172	6637.04	<.0001
Electrocardiogram, (EGG)	1	13816375074	13816375074	4577.07	<.0001
Office or other outpatient visit, typically15minutes	1	17996085232	17996085232	5961.72	<.0001
Syringe,	1	20096703187	20096703187	6657.61	<.0001
Thyroid stimulating hormone	1	15385539003	15385539003	5096.90	<.0001
Radiologic examination	1	8767439243.1	8767439243.1	2904.47	<.0001

Table5.4 Criteria for Assessing Goodness of Fit

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	9964	22374.1763	2.2455
Scaled Deviance	9964	12474.8191	1.2520
Pearson Chi-Square	9964	130667.2143	13.1139
Scaled Pearson X2	9964	72854.0725	7.3117
Log Likelihood		-63303.4243	

In Table 5.4, the value of deviance divided by degree is 2.25.The value shows the adequacy of this model, which means that the model fits the data reasonably well. In Table 5.5, all the *p* values are smaller than 0.0001, indicating that all 20 variables are statistically significant. The χ^2 value of 1070.13 for blood glucose monitoring represents twice the difference in log likelihoods between fitting a model with only an intercept and a model with an intercept and blood glucose monitoring. Similarly, every χ^2 value for each variable represents the differences in log likelihoods between successive models. The output shows that the χ^2 values for venipuncture and blood glucose monitoring are the highest among all the χ^2 values; therefore, we conclude that blood glucose monitoring and venipuncture have more important effects on the costs than any other treatments do.

Table 5.5 Type I analysis

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	-131538.88			
Blood glucose monitoring	-130468.75	1	1070.13	<.0001
Routine venipuncture	-129441.88	1	1026.87	<.0001
Therapeutic exercises	-129197.95	1	243.93	<.0001
Blood count	-128915.08	1	282.87	<.0001
Prothrombin time	-128568.73	1	346.35	<.0001
Basic metabolic panel	-128424.85	1	143.88	<.0001
Comprehensive metabolic panel	-128315.64	1	109.21	<.0001
Hemoglobin; glycosylated	-128081.87	1	233.78	<.0001
Unlisted dialysis procedure	-127823.38	1	258.49	<.0001
Therapeutic activities, direct patients contact by the	-127633.20	1	190.18	<.0001
Injection	-127565.57	1	67.63	<.0001
Lipid panel	-127459.23	1	106.34	<.0001

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Therapeutic procedure, gait training	-127263.69	1	195.55	<.0001
Neuromuscular reeducation of movement, balance	-127112.20	1	151.48	<.0001
Office or other outpatient visit, typically 10 minutes	-127004.16	1	108.04	<.0001
Electrocardiogram, EGG	-126952.54	1	51.62	<.0001
Office or other outpatient visit, typically 15 minutes	-126855.49	1	97.05	<.0001
Syringe	-126711.41	1	144.09	<.0001
Thyroid stimulating hormone	-126640.03	1	71.37	<.0001
Radiologic examination	-126606.85	1	33.18	<.0001

Table 5.6 Pearson Correlation

Pearson Correlation Coefficients, N = 12158258	
	coefficients
Blood glucose monitoring	-0.05110 (<.0001)
Routinevenipunctureforcollectionspecimen(s)	-0.04919 (<.0001)

In Table 5.6, the negative correlation coefficients between the total charges and blood glucose monitoring and routine venipuncture indicate that if we increase the monitoring and routine venipuncture, we can decrease the total charges.

5.3 Cost Analysis Using Claims Data

5.3.1 Newly-generated Predictors

In this section, we will analyze the relationships between the diagnosis and the costs of outpatients with diabetes using the Medicare claims data from the CMS CCW data warehouse [18]. In these data, diagnosed diseases or procedures appear in more than 2 columns; therefore, we use the method in [59] to get them into one column. First, we use Filter and Query to select Claim ID and ICD9_DGNS_CD1 to generate a new data set that contains two columns, and change the name of ICD9_DGNS_CD1 to ICD9. It is the same for the left nine ICD9_DGNS_CDn and six ICD9_PRCDR_CD n to generate 16 Tables. Then, we open one Table, and use Data−>Append Table; after that, the ICD9

Table (Figure 5.3) is generated. This is followed by Describe->One-way Frequencies to find the top 20 ICD9 codes as displayed in Table 5.7.

	CLM_ID	ICD9
1	vvvvvvvRRRVE	V5883
2	vvvvvvvRRRVE	V5883
3	vvvvvvvRRRVE	7038
4	vvvvvvvRRRVt	71947
5	vvvvvvvRRRVtp	5130
6	vvvvvvvRRRVB	4010
7	vvvvvvvRRRVB	25000
8	vvvvvvvRRRVB	25000
9	vvvvvvvRRRVB	7859
10	vvvvvvvRRRV7	25000
11	vvvvvvvRRRVp	29620
12	vvvvvvvRRRVp	V825
13	vvvvvvvRRRVp	25000
14	vvvvvvvRRRVp	25001
15	vvvvvvvRRRVV	V5861

Figure 5.3. ICD9 Table

Table 5.7 Top 20 ICD9 Diagnosis Codes

ICD9	Diagnosis	Frequency	Percent
25000	Diabetes mellitus without complication type ii or unspecified type not stated uncontrolled	2264	8.90
4019	Unspecified essential hypertension	1351	5.31
585	Chronic kidney disease (ckd)	629	2.47
V5861	Long-term (current) use of anticoagulants	509	2.00
2724	Other and unspecified hyperlipidemia	508	2.00
42731	Atrial fibrillation	469	1.84
2859	Anemia unspecified	452	1.78
4280	Congestive heart failure unspecified	429	1.69
V5869	Long-term (current) use of other medications	402	1.58
28521	Anemia in chronic kidney disease	357	1.40
4011	Benign essential hypertension	339	1.33
2720	Pure hypercholesterolemia	314	1.23
41400	Coronary atherosclerosis of unspecified type of vessel native or graft	296	1.16
25001	Diabetes mellitus without complication type i not stated as uncontrolled	244	0.96
5990	Urinary tract infection site not specified	232	0.91
2809	Iron deficiency anemia unspecified	230	0.90
2449	Unspecified acquired hypothyroidism	225	0.88
496	Chronic airway obstruction not elsewhere classified	210	0.83
25002	Diabetes mellitus without complication type ii or unspecified type	203	0.80
78079	Other malaise and fatigue	194	0.76

Next, we process the data to get the data (Figure 5.4) ready for analysis.

CLM_ID	CLM_TOT_CHRG_AMT	R25000_Max	R4019_Max	R585_Max	RV5861_Max	R2724_Max	R4280_Max	R4273
WWWWvEWWvv	262.00	0	0	1	0	0	0	
WWWWvEWWvb	327.42	0	1	0	0	0	0	
WWWWvEWWv	2055.00	0	0	0	0	0	0	
WWWWvEWWb	288.94	0	0	0	0	0	0	
WWWWvEWWb	163.00	0	0	0	0	0	0	
WWWWvEWWb	384.50	0	0	0	0	0	0	
WWWWvEWWO	200.00	0	0	0	0	0	0	
WWWWvEWWO	140.00	0	0	0	0	0	0	
WWWWvEWWR	410.00	0	0	0	0	0	0	
WWWWvEWWE	2446.00	0	0	0	0	0	0	

Figure 5.4. Newly-generated Data for GLM

5.3.2 Statistical Model Analysis

After the assumptions diagnostics, the generalized linear model is developed, with the costs as the predicted variable and the newly-generated variables as the input variables. The results are given in the following Tables.

Table 5.8 Overall Mode Information

Model Information		
Data Set	WORK.SORTTEMPTAB LESORTED	
Distribution	Gamma	
Link Function	Log	
Dependent Variable	CLM_TOT_CHRG_AMT	Claim Total Charge Amount

Table 5.8 gives information about distribution of the response variable, the link function. In Table 5.9, the value of deviance divided by degree is 2.32, and the scaled deviance divided by degree is 1.26. These two parameters demonstrate the adequacy of this model, which means that the generalized linear model fits the data reasonably well.

Table 5.9 Criteria for Assessing Goodness Fit

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	9980	23196.7333	2.3243
Scaled Deviance	9980	12550.5608	1.2576
Pearson Chi-Square	9980	79786.4792	7.9946
Scaled Pearson X2	9980	43168.3655	4.3255
Log Likelihood		-75829.4983	

Table 5.10 Type I Analysis

LR Statistics For Type 1 Analysis					
Source		2*LogLike-	DF	Chi-Square	Pr
Intercept		-154184.45			
R25000_Max	Diabetes mellitus without complication type II or unspecified type uncontrolled	-154160.76	1	23.68	<.0001
R4019_Max	Unspecified essential hypertension	-154078.23	1	82.53	<.0001
R585_Max	Chronic kidney disease (ckd)	-152265.81	1	1812.42	<.0001
RV5861_Max	Long-term (current) use of anticoagulants	-152018.52	1	247.29	<.0001
R2724_Max	Other and unspecified hyperlipidemia	-151969.30	1	49.21	<.0001
R4280_Max	Congestive heart failure unspecified	-151953.57	1	15.74	<.0001
R42731_Max	Atrial fibrillation	-151889.81	1	63.76	<.0001
RV5869_Max	Long-term (current) use of other medications	-151886.98	1	2.82	0.0929
R2859_Max	Anemia unspecified	-151880.30	1	6.69	0.0097
R4011_Max	Benign essential hypertension	-151782.39	1	97.91	<.0001
R28521_Max	Anemia in chronic kidney disease	-151737.71	1	44.68	<.0001
R2720_Max	Pure hypercholesterolemia	-151736.64	1	1.07	0.3014
R41400_Max	Coronary atherosclerosis of unspecified type of vessel native or graft	-151731.99	1	4.65	0.0311
R5990_Max	Urinary tract infection site not specified	-151731.98	1	0.01	0.9287
R25001_Max	Diabetes mellitus without complication type i not stated as uncontrolled	-151729.57	1	2.41	0.1203
R78079_Max	Other malaise and fatigue	-151729.57	0	0.00	.
R2809_Max	Iron deficiency anemia unspecified	-151708.85	1	20.72	<.0001

LR Statistics For Type 1 Analysis					
Source		2*LogLike-	DF	Chi-Square	Pr
R2449_Max	Unspecified acquired hypothyroidism	-151707.74	1	1.11	0.2922
R496_Max	Chronic airway obstruction not elsewhere classified	-151698.50	1	9.24	0.0024
R25002_Max	Diabetes mellitus without complication type ii or unspecified type uncontrolled	-151659.00	1	39.50	<.0001

Table 5.11 Variables Significant to the Model

R25000_Max	Diabetes mellitus without complication type II or unspecified type uncontrolled
R4019_Max	Unspecified essential hypertension
R585_Max	Chronic kidney disease (ckd)
RV5861_Max	Long-term (current) use of anticoagulants
R2724_Max	Other and unspecified hyperlipidemia
R4280_Max	Congestive heart failure unspecified
R42731_Max	Atrial fibrillation
R2859_Max	Anemia unspecified
R4011_Max	Benign essential hypertension
R28521_Max	Anemia in chronic kidney disease
R41400_Max	Coronary atherosclerosis of unspecified type of vessel native or graft
R78079_Max	Other malaise and fatigue
R2809_Max	Iron deficiency anemia unspecified
R496_Max	Chronic airway obstruction not elsewhere classified
R25002_Max	Diabetes mellitus without complication type ii or unspecified type uncontrolled

Type I analysis shows that the following factors (shown in Table 5.11) are statistically significant to the model. Since the χ^2 value for chronic kidney disease ranks the first among all the values (shown in Table 5.10), kidney disease is the most important variable related to the costs.

Since chronic kidney disease has the most important effect on the costs and chronic kidney disease counts for a large percentage of renal failure, the logistic regression in Enterprise Miner will be used to analyze the relationships between renal failure and the other diseases vital to the costs. Before that, a binary variable, Renal is generated using 0-1 indicator functions. The following steps are conducted in Enterprise Miner. First, R (Renal Failure) is set as the target and the variables are set on binary level. Then, a sample node is added and connected to the data; and the Stratify Method is used, choosing Level Based as the stratified criterion and Rarest Level in the Level Based options. The results are as shown in Tables 5.12 – 5.15.

Table 5.12 Misclassification Rate

TARGET	Fit	Statistic Label	Train	Validation	Test
R_Max		Misclassification Rate	0.38596491	NaN	NaN

Table 5.13 Type 3 Analysis of Effects

Effect	DF	Wald ChiSq	Pr > ChiSq
Unspecified acquired hypothyroidism	1	0.028	0.866
Diabetes mellitus without complication type ii or unspecified type not stated as uncontrolled	1	1.131	0.288
Diabetes mellitus without complication type i not stated as uncontrolled	1	0.360	0.548
Diabetes mellitus without complication type ii or unspecified type uncontrolled	1	0.027	0.868
Anemia unspecified	1	4.708	0.030
Benign essential hypertension	1	0.264	0.607
Coronary atherosclerosis of unspecified type of vessel native or graft	1	0.012	0.910
Atrial fibrillation	1	0.151	0.697
Congestive heart failure unspecified	1	0.226	0.634
Chronic airway obstruction not elsewhere classified	1	0.562	0.453
Chronic kidney disease (ckd)	1	0.428	0.513
Urinary tract infection site not specified	1	2.596	0.107
Long-term (current) use of anticoagulants	1	0.00	0.929
Long-term (current) use of other medications	1	0.015	0.902

The misclassification rate in Table 5.12 is 0.386, which is accepted. The results in Table 5.13 indicate that only unspecified anemia is significant to renal failure as only its p-value in the Type 3 analysis is less than 0.05 out of the 15 diseases that are considered for renal failure. The odds ratio for Anemia in Table 5.14 shows that if the diagnosis is not anemia, then the chance that it is related to renal failure is only 9 % of the probability that it is Anemia. Hence, Anemia has the most prominent relationship to renal failure.

Table 5.14 Odds Ratio Estimates

Effect	Diagnosis	Point Estimat
R2449_Max 0 vs 1	Unspecified acquired hypothyroidism	0.745
R25000_Max 0 vs1	Diabetes mellitus without complication type ii or unspecified	2.057
R25001_Max 0 vs1	Diabetes mellitus without complication type i not stated as	0.602
R25002_Max 0 vs1	Diabetes mellitus without complication type ii or unspecified	0.840
R2859_Max 0 vs 1	Anemia unspecified	0.093
R4011_Max 0 vs 1	Benign essential hypertension	0.562
R41400_Max 0 vs1	Coronary atherosclerosis of unspecified type of vessel native or	1.145
R42731_Max 0 vs1	Atrial fibrillation	1.551
R4280_Max 0 vs 1	Congestive heart failure unspecified	0.704
R496_Max 0 vs 1	Chronic airway obstruction not elsewhere classified	0.402
R585_Max 0 vs 1	Chronic kidney disease (ckd)	0.626
R5990_Max 0 vs 1	Urinary tract infection site not specified	0.121
RV5861_Max 0 vs1	Long-term (current) use of anticoagulants	1.139
RV5869_Max 0 vs1	Long-term (current) use of other medications	0.794

Table 5.15 Event Classification Table

Data Role	Target	False Negative	True Negative	False Positive	True Positive
Train	R_Max	32	45	12	25

Table 5.15 shows that the false negative rate is 32 %, which means that it is 32% likely for the model to predict that renal failure is not occurring when it in fact is; the false positive rate is 12, which means that it is only 12 % likely to predict the diagnosis of renal failure when it actually does not occur. Since a false negative is more critical than a false positive, the model is fairly good.

Just as the results show, the model probably needs improving; hence, it is necessary to utilize another model to analyze the relationship. Next, we utilized the generalized linear mixed model, in which the GLIMMIX procedure in SAS can be used, selecting Renal as the response variable, which is a binary variable; selecting Ane (Anemia), Hea (Heart disease) and Unctrl (Uncontrolled diabetes) as the classification variables; analyzing Ane and Hea in the fixed effects while analyzing Unctrl in the random effects. The least-squares means for Ane and Hea are also used. The following SAS code is used.

```
PROC GLIMIX DATA=SASUSER.RENAL;
CLASS HEA ANE UNCTRL;
MODEL RENAL=ANE HEA/DIST=BINARY LINK=LOGIT;
LSMEANS ANE HEA/;
RANDOM UNCTRL; RUN;
```

The model results of this analysis are shown in Table 5.16. In Table 5.17, the first two measures indicate that the model is statistically significant; and the third one demonstrates that the model fits the dataset very well.

Table 5.16 Overall Information

Model Information	
Data Set	SASUSER.RMIXW
Response Variable	Renal
Response Distribution	Binary
Link Function	Logit
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Residual PL
Degrees of Freedom Method	Containment

Table 5.17 Fit Statistics

Fit Statistics	
-2 Res Log Pseudo-Likelihood	80738.80
Generalized Chi-Square	10055.61
Gener. Chi-Square / DF	1.01

Covariance Parameter Estimates		
Cov Parm	**Estimate**	**Standard Error**
Unctrl	8.21E-22	.

Table 5.18 Type 3 Analysis for Fixed Effects

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Ane	1	9996	3.68	0.0551
Hea	1	9996	6.77	0.0093

Table 5.19 Least –square Means Analysis

Ane Least Squares Means					
Ane	**Estimate**	**Standard**	**DF**	**t Value**	**Pr > |t|**
0	4.7986	0.2116	999	22.67	<.0001
1	4.0978	0.3677	999	11.14	<.0001
Hea Least Squares Means					
Hea	**Estimate**	**Standard**	**DF**	**t Value**	**Pr > |t|**
0	4.9771	0.1895	999	26.26	<.0001
1	3.9192	0.4000	999	9.80	<.0001

Table 5.17 lists the covariance parameter estimates for a random variable. The variance for Unctrl is rather small and so the variable is significant to the model. The output in the Type 3 analysis (Table5.18) shows that both anemia and heart disease are significant to the model. Table 5.19 lists the information about the least-square means. The output shows that the estimate mean for anemia is 4.0978 and the standard error is 0.3677, while the value for heart disease is 3.9192 and its standard error is 0.4. Therefore, the two diseases are almost equally important to renal failure.

5.4 Conclusion

After this analysis, we can draw the conclusions that in order to decrease Medicare costs, we should often monitor blood glucose level of diabetes outpatients and we also should emphasize diabetes prevention. In addition, chronic kidney disease affects the charges the most. Finally, we should monitor heart disease and anemia of the diabetic outpatients with renal failure.

In this chapter, we focus on outpatients with diabetes in the Medicare population; in next chapter, we will study inpatients with diabetes.

CHAPTER VI

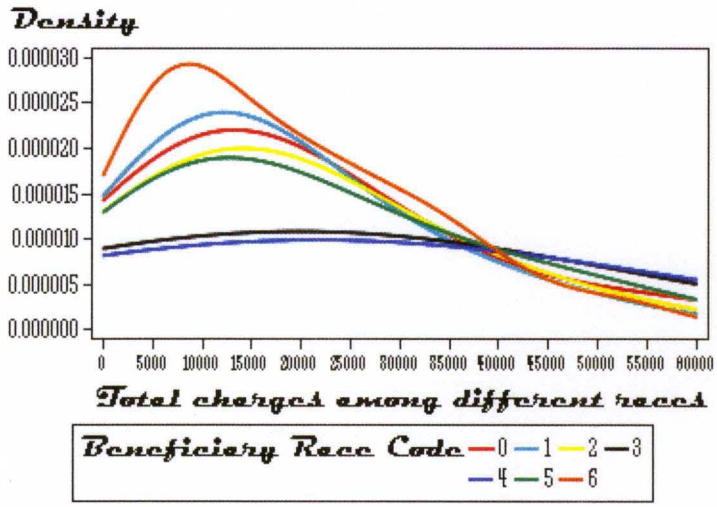
COST ANALYSIS AND OUTCOMES RESEARCH FOR MEDICARE INPATIENTS WITH DIABETES

In the previous chapter, we conducted a cost analysis of diabetes outpatients in Medicare. The analyses of inpatients encompass more than those of outpatients. In this chapter, based on the inpatient claims data and demographic data from the CMS CCW data[18], we will start a cost analysis among different races or different complications of diabetes; then, we will perform association rule analysis of different procedures. Next, we will use various kinds of supervised learning algorithms to study the outcomes and a readmission analysis of inpatients and we will end up with two-way interaction effects analysis of diabetes complications.

6.1 Cost Analysis

6.1.1 Inpatient Costs among Different Races

At the beginning, we study the general cost distributions among different races using kernel density estimation (Figure 6.1). Before the value of 40,000 dollars occurs, North American natives > Whites> Blacks> Hispanics in terms of densities for costs, and all of them spend more than Asians; while after that point, Asians cost more than the other races.



0:unkown ;1: white; 2: black; 3: other; 4: Asian; 5: Hispanic; 6: north American native

Figure 6.1. KDE of Total Charges among Different Races

6.1.2 Costs among Different Diagnosed Diseases

In this section, we want to see how different groups of diagnoses affect the costs. Before that, we define a string containing all possible diagnosis codes using the CATX statement in SAS.

Clusters

# ▲	Descriptive Terms	Freq	Percentage	RMS Std.
1	27800, 29570, 25000, 3051, 2724	327	0.0327	0.0994078...
2	412, 41401, v4581, 41400, v4582	1414	0.1414	0.1202243...
3	4139, 42789, 2948, 41401, 2720	539	0.0539	0.1212629...
4	4019, 25000, 2449, 71590, 311	1949	0.1949	0.1284052...
5	25060, 36201, 25050, 3572, 2724	362	0.0362	0.0988086...
6	5849, 4280, 49121, 40391, 486	1828	0.1828	0.1266924...
7	4240, 25001, 4280, 4254, 42731	1276	0.1276	0.1260251...
8	3310, 5990, 2859, 2765, 486	1515	0.1515	0.1238352...
9	25000, 53081, 2724, 4019, 2720	790	0.079	0.1176632...

Figure 6.2. Clusters of Diagnoses

After conducting clustering analysis with the E-M algorithm using SAS Text Miner, we get the results displayed in Figure 6.2. To view how the groups of diagnoses affect the total charges, we perform kernel density estimation.

Table 6.1 Translations for the Clusters

Cluster number	Diagnoses	Cluster label
1	Unspecified Obesity, Schizoaffective disorder, Diabetes mellitus without mention of complication, Tobacco use disorder, Other and unspecified hyperlipidemia	Diabetes
2	Old myocardial infarction, Of native coronary artery, Aortocoronary bypass status, Of unspecified type of vessel or native or graft, Percutaneous transluminal coronary angioplasty status	Heart disease
3	Other and unspecified angina pectoris, Other specified cardiac dysrhythmias, Other persistent mental disorders due to conditions classified elsewhere, Of native coronary artery, Pure hypercholesterolemia	Heart disease vascular disease
4	Unspecified Essential hypertension, Diabetes mellitus without mention of complication, Unspecified hypothyroidism, Osteoarthrosis which unspecified whether generalized or localized, Depressive disorder	vascular disease Diabetes
5	Diabetes with neurological manifestations, Background diabetic retinopathy, Diabetes with ophthalmic manifestations, Diabetes with ophthalmic manifestations, Other and unspecified hyperlipidemia	Ophthalmic disease Neurological disorder
6	Unspecified Acute renal failure, unspecified Congestive heart failure, Obstructive chronic bronchitis with exacerbation, Unspecified Hypertensive chronic kidney disease, Pneumonia	Heart disease Kidneydisease
7	Mitral valve disorders, Diabetes mellitus without mention of complication, unspecified Congestive heart failure, Other primary cardiomyopathies, Atrial fibrillation,	Diabetes Heart disease
8	Alzheimer's disease, Urinary tract infection, unspecified Anemia, Volume depletion, Pneumonia	Others
9	Diabetes mellitus without mention of complication, Esophageal reflux, Other and unspecified hyperlipidemia, Unspecified Essential hypertension, Pure hypercholesterolemia	Diabetes vascular disease

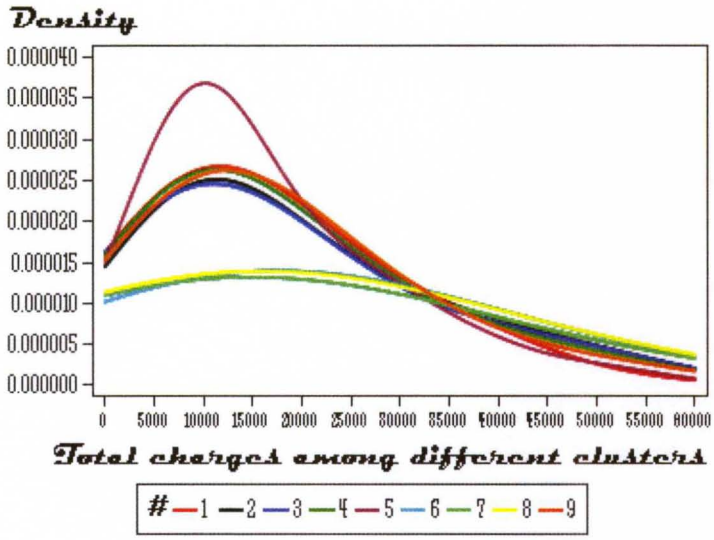


Figure 6.3. KDE of Total Charges for Diabetic Inpatients by Clusters (Male)

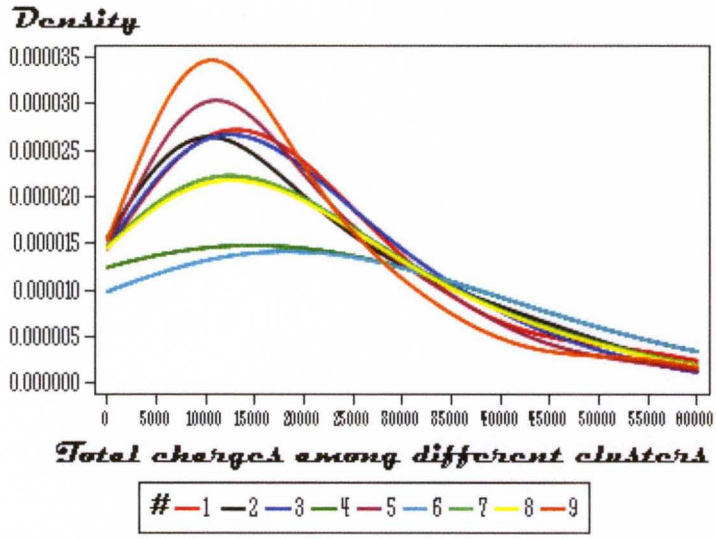


Figure 6.4. KDE of Total Charges for Diabetic Inpatients by Clusters (Female)

Figures 6.3 and 6.4 demonstrate that the distributions of the costs for male inpatients are different from the ones for females. The clusters yield the relationships in terms of ordering. For males, before the first cutpoint occurs at 19,200 dollars, in terms of density, cluster #5 is much greater than the other clusters; between the cutpoints of 19,200 dollars and 33,000 dollars, the ordering of estimated density is 1, 4, 5, 9>2, 3>6, 7,

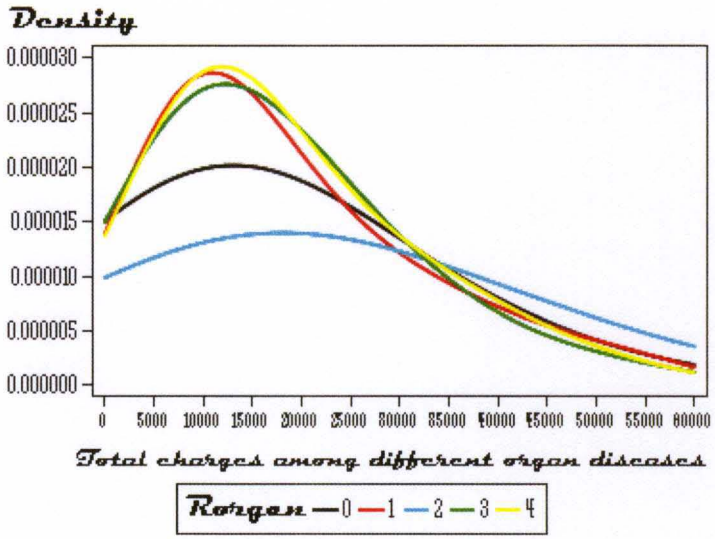
8. After 33,000 dollars, there are no differences among all clusters. The graph for female inpatients shows the costs in terms of ordering is 9>5>2>1,3>7,8>4,6 before the first cutpoint of 10,650 occurs.

6.1.3 Cost Distributions among Organ Diseases

The previous analysis indicates that diabetes has many complications such as heart disease or renal failure. We will demonstrate the expenditures on these organ diseases shown in Figure 6.4. Before that, we need to generate a new variable, ORGAN with the following SAS code:

```
DATA SASUSER.IPORGAN (KEEP=CLM_CD CLM_TOTO_CHRG_AMT
DIAGNOSES HEA  KID  OCULO NEU); SET SASUSER.IPCLAIMDEMO;
HEA=0; KID=0; OCULO=0; NEU=0;
DIAGNOSES= CATX('',ICD9_DGNS_CD1, …ICD9_DGNS_CD16);
IF(RXMATCH('4280',DIAGNOSES)>0)THEN HEA=1;
IF(RXMATCH('4254',DIAGNOSES)>0)THEN HEA=1;…
DATA SASUSER.ORGAN; SET SASUSER.IPORGAN;
IF HEA=1 THEN ORGAN=1; IF KID=1 THEN ORGAN=2; IF OCULO=1
THEN ORGAN=3; IF NEU=1 THEN ORGAN=4; RUN;
```

The graphs in Figure 6.5 indicate that before the costs reach the value of 9,900 dollars, the cost with heart disease, the cost with ophthalmic diseases and the cost with neurological disorders have almost the same probability, which is much higher than the cost without any of the organ diseases; the probability for the cost with kidney disease is the smallest. However, after the cutpoint at 34,350 dollars, the density of the cost with kidney disease is higher than any other densities. It shows that kidney disease has the highest probability of the highest cost while neurological disorders have the highest probability of lowest cost with heart disease and ophthalmic diseases having similar probabilities as neurological disorders.



0: None of the organ diseases; 1:Heart diseases; 2:Kidney diseases;

3: ophthalmic diseases, 4: Neurological disorders;

Figure 6.5. Costs by Different Organ Diseases

6.2 Outcomes Research

6.2.1 Association Rule Analysis of Procedures

Before we study the outcomes of inpatients, we conduct market basket analysis of various procedures and the results are displayed below. Figure 6.6 shows all the major connections between different procedures. The procedures shown in table 6.2 are important, since all of the rectangular boxes representing those procedures are bigger than the others. Among the procedures, five of them are used for cardiac disease and one is related to hematic disease, which form 6 centers of the diagram; they are marked with an asterisk, '*' in Table 6.2.

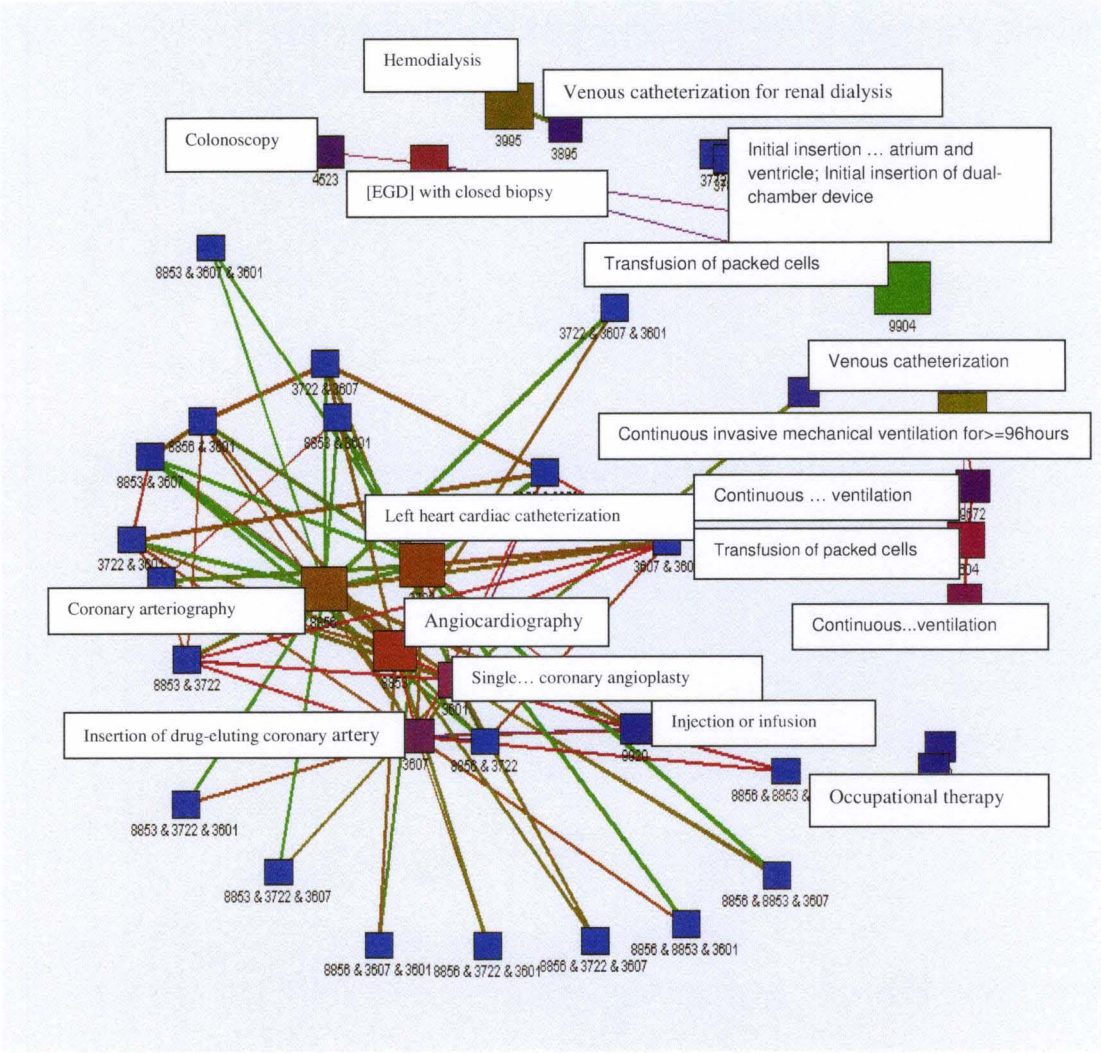


Figure 6.6. Associations of Procedures

Table 6.2 Translations for Important Procedures

Code	Procedure
3601*	Single vessel percutaneous transluminal coronary angioplasty
3607*	Insertion of drug-eluting coronary artery stent(s)
3722*	Left heart cardiac catheterization
3772	Initial insertion of transvenous leads [electrodes] into atrium and
3783	Initial insertion of dual-chamber device
3893	Venous catheterization, not elsewhere classified
3895	Venous catheterization for renal dialysis
3995	Hemodialysis
4516	Esophagogastroduodenoscopy [EGD] with closed biopsy
4523	Colonoscopy

Code	Procedure
8853*	Angiocardiography of left heart structures
8856*	Coronary arteriography using two catheters
9339	Other physical therapy
9383	Occupational therapy
9604	Insertion of endotracheal tube
9671	Continuous invasive mechanical ventilation for less than 96 consecutive
9672	Continuous invasive mechanical ventilation for 96 consecutive hours or
9904	Transfusion of packed cells
9920	Injection or infusion of platelet inhibitor

Table 6.3 Confidence and Lift for Rules

Rules	Confidence (%)	Lift
Initial insertion ... atrium and ventricle==> Initial insertion of dual-chamber device	100	73.22
Initial insertion of dual-chamber device==> Initial insertion ... atrium and ventricle	97.28	73.22
Angiocardiography of left heart structures ==> Coronary arteriography	94.17	9.38
Left heart cardiac catheterization==> Coronary arteriography	89.92	8.96
Angiocardiography of left heart structures==> Left heart cardiac catheterization	89.24	9.56
Coronary arteriography==> Left heart cardiac catheterization	83.64	8.96
Hem dialysis== > Venous catheterization for renal dialysis	79.61	7.41
Insertion of drug-eluting coronary artery stent(s)==> Single... coronary angioplasty	78.68	18.37
Left heart cardiac catheterization==> Angiocardiography	78.66	9.56
Coronary arteriography==> Angiocardiography of left heart structures	77.21	9.38
Insertion of drug-eluting coronary artery stent(s)==> Coronary arteriography	77.16	7.08
Occupational therapy==> Other physical therapy	77.08	43.06
Other physical therapy==> Occupational therapy	76.29	43.06
Continuous invasive mechanical ventilation==> Colonoscopy	74.27	14.74
Injection or infusion==> Coronary arteriography	73.27	7.28
Single ...coronary angioplasty==> Coronary arteriography	72.84	7.26
Insertion of drug-eluting coronary artery stent(s)==> Left heart cardiac catheterization	71.07	7.61
Single ...coronary angioplasty==> Left heart cardiac catheterization	68.10	7.29
Single ...coronary angioplasty==> Insertion of drug-eluting coronary artery stent(s)	66.81	18.37
Injection or infusion==> Left heart cardiac catheterization	64.36	6.89
Continuous invasive mechanical ventilation more than 96 hour==> Insertion of endotracheal tube	63.25	12.55
Continuous invasive mechanical ventilation more than 96 hour==>Continuous invasive mechanical ventilation	56.04	14.74
Single ...coronary angioplasty==> Angiocardiography of left heart structures	55.60	6.75
Insertion of drug-eluting coronary artery stent(s)==> Angiocardiography	55.33	6.72
Insertion of endotracheal tube==> Continuous invasive mechanical ventilation more than 96 hour	38.46	12.55

Table 6.3 just shows the important and meaningful rules; the combination cases are not considered. The initial insertion of transvenous leads [electrodes] into the atrium and ventricle will be used given that the procedure, initial insertion of dual-chamber device will be subsequently used. The lift value for this rule is 73.22, which indicates that the association between these two separate procedures is strong. For the same reason, the relationship between occupational therapy and other physical therapy is also strong. The confidence values are higher for the other rules in the table, which indicate that it is very likely that subsequent procedures will be used if the antecedent procedure is used, since all the left confident values are high.

6.2.2 Mortality Prediction of Diabetes Inpatients

One of dominant issues in outcomes research is about mortality. In this section, we use various kinds of supervised learning approaches shown in Figure 6.7 with mortality as the targeted variable; Age, Gender, Diagnosis Procedures and Diagnosis Procedures are identified as the input variables. To predict mortality, we use the regression model, the Dmine Regression model, the Neural Network model, the Auto Neural model, the Decision Tree model, the MBR model, the Rule Induction model and the Model Comparison model; and these nodes are shown in Figure 6.7. Table 6.4 shows that the Model Comparison node identifies the decision tree as the optimal model based on misclassification rates.

The ROC maps in Figure 6.8 show that for all three data subsets: Train, Validate and Test, there are no big differences in accuracy among the various models. According to the lift curves, we can find the patients at highest risk of dying. Figure 6.9 demonstrates that except for the MBR node, there are no differences among the other

nodes in terms of the prediction of mortality. In the train set, validate set and test set, 40 % of beneficiary records have a higher level of prediction than just chance.

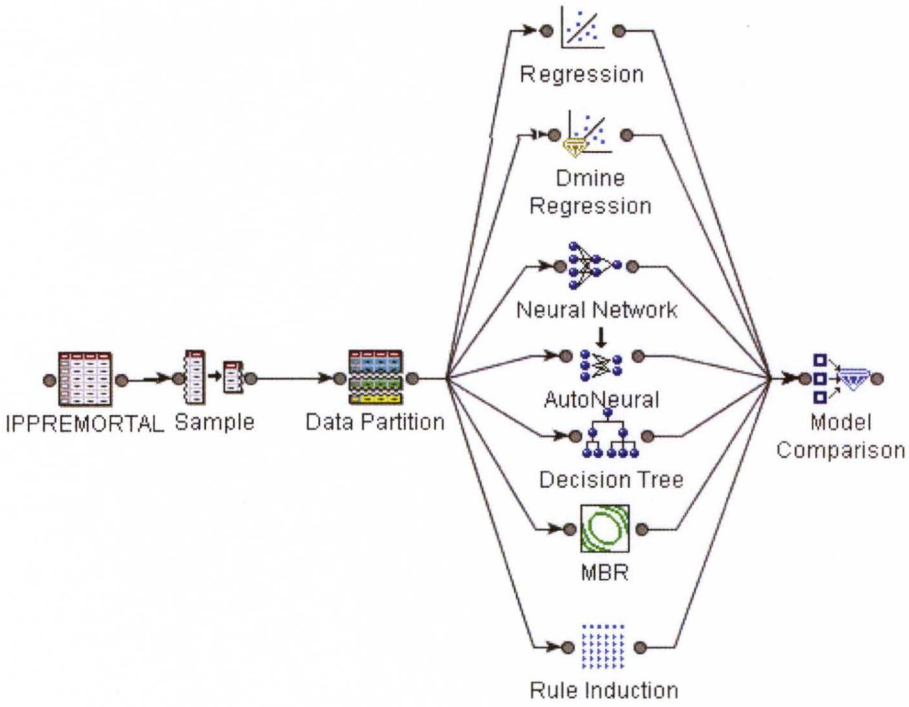


Figure 6.7. Predictive Models Diagram

Table 6.4 Fit Statistics of the Comparison Model Targeting at Mortality

Selected Model	Model Node	Train: Akaike's Information Criterion.	Train: Average Squared Error.	Average Squared Error.	Average Squared Error.	Train: Mis classification Rate.	Valid:Mis classification Rate.	Test Misclassification reate
	AutoNeural	12427.21	0.26	0.26	0.26	0.5	0.5	0.5
	DmineReg	NaN	0.18	0.18	0.18	0.26	0.26	0.26
	MBR	-7091.33	0.22	0.25	0.25	0.36	0.43	0.43
	Neural	9159.41	0.18	0.18	0.18	0.27	0.27	0.27
	Reg	9083.04	0.18	0.18	0.18	0.27	0.27	0.26
	Rule	NaN	NaN	NaN	NaN	0.26	0.26	0.26
Y	Tree	NaN	0.18	0.19	0.18	0.26	0.26	0.26

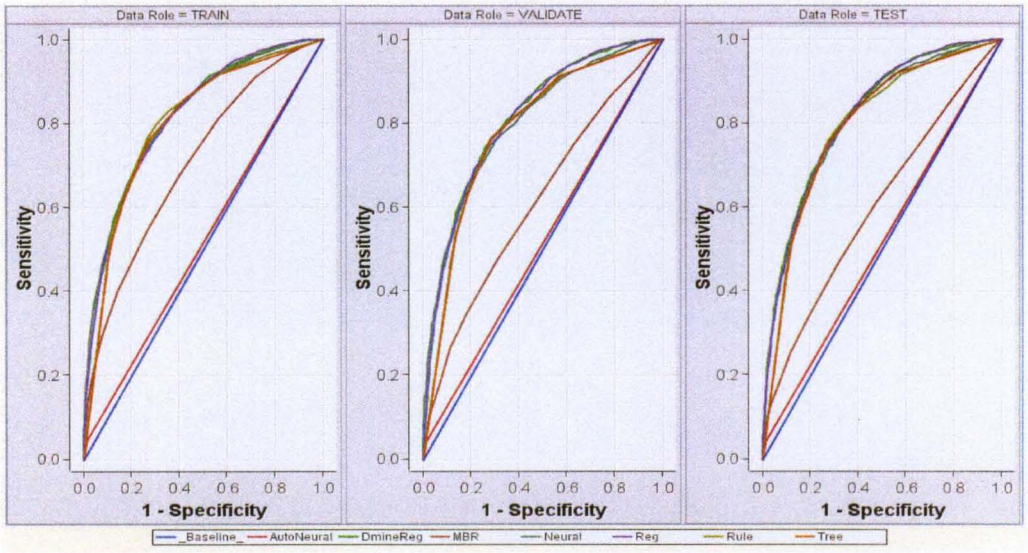


Figure 6.8. ROC Chart for Mortality Prediction

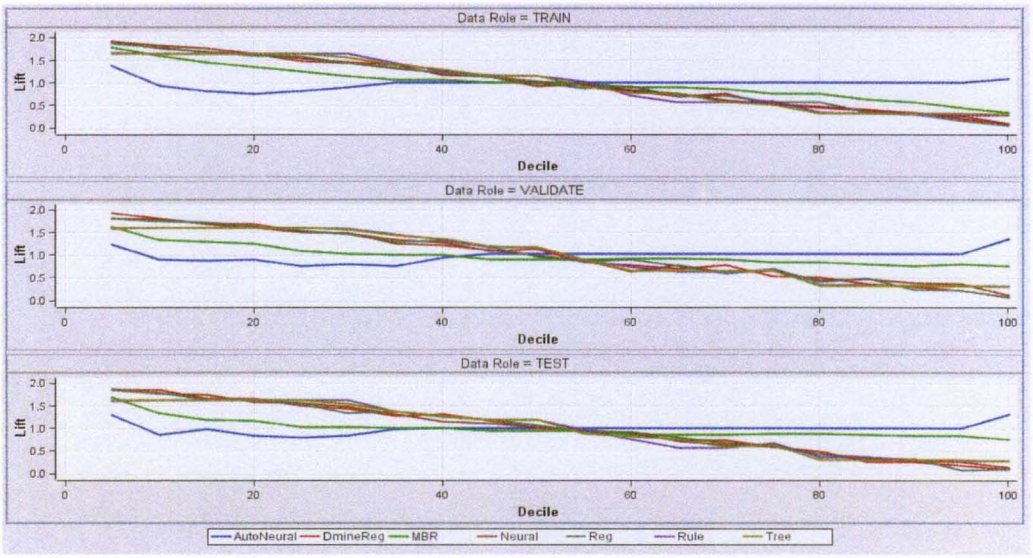


Figure 6.9. Lift Curve for Predictive Model for Mortality

Table 6.5 Variable Importance in Tree Targeting at Mortality

NAME	IMPORTANCE	VIMPORTANCE	RATIO
PROCLUSTER	1.0000	1.0000	1.0000
DIACLUSTER	0.7254	0.6807	0.9384
Age	0.3845	0.3993	1.0386
Utilization Day Count	0.3364	0.3464	1.0297
Total Charge	0.2466	0.2528	1.0252

Results in Table 6.5 demonstrate that the order of importance in the levels is procedures > diagnoses >age > days of staying in the hospital> total charges. The tree diagram in Figure 6.10 displays how the input variables affect mortality; the first segment is divided on the procedure cluster, indicating that procedures are essential to mortality; the next split is based upon the diagnosis cluster. The following split criteria vary from the left side to the right side. Age has no relation to mortality related to the procedure cluster #5 (endotracheal tube and catheterization) and #7 (Some heart operations); for the procedure cluster 1 and cluster3, age is also an important factor. Before the age of 82.5, both total charges and utilization day count should be considered, while after that, only utilization day count should be focused on.

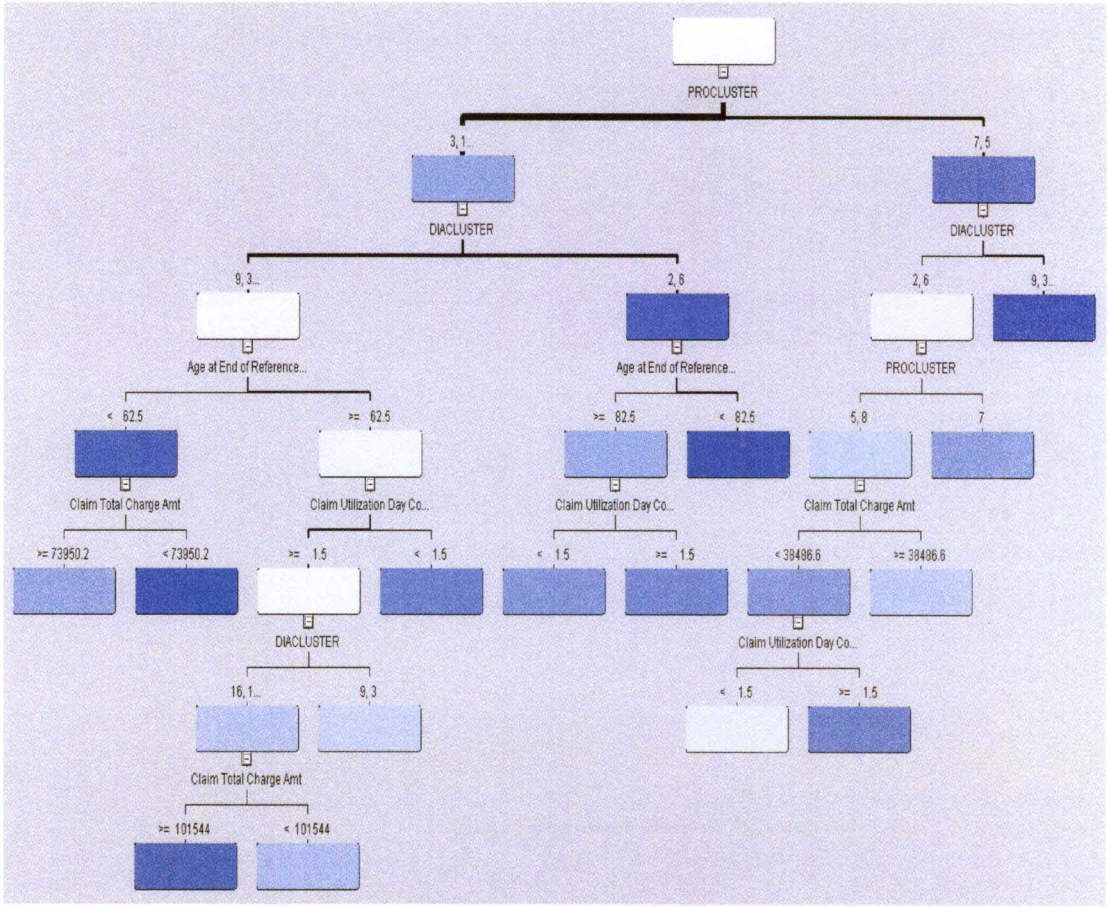


Figure 6.10. Tree Diagram Aiming at Mortality

6.3 Readmission Analysis

Under the current economic circumstances, medical care providers have become concerned about readmission rates. In this section, we attempt to find the diabetes patients who have higher risks of readmission using Medicare inpatient claims data from the CMS CCW data [18]. We first introduce the definition of readmission by CMS; then, we use SAS SQL to get the inpatients who re-enter the hospital after discharge within 30 days and the most frequent diseases and procedures for these patients; before we perform supervised learning, we conduct pattern recognition analysis.

6.3.1 Medicare Readmission

When defining readmission, different researchers give different definitions. CMS defines readmission as the re-entry to the hospital after discharge within 30 days and considers it as an important measure of poor quality health care. One report from the Atlantic Information Services, Inc. [60] demonstrates that almost 18% of Medicare patients are readmitted within 30 days of discharge and 13% of the readmissions ($12 billion worth) can be avoided if the correct treatments are provided. Medical providers are currently facing pressure to reduce the readmission rate while improving care quality. It is with this motivation that we evaluate the readmission risk of diabetic patients. We study three kinds of risk factors influencing readmission rates, (1) patient demographic information including age, gender, and race; (2) patient disease characteristics such as common co-morbidities and procedures for diabetes; (3) medical resources usage such as length of hospitalization and discharge locations.

6.3.2 Data Processing to Find Readmission Inpatients

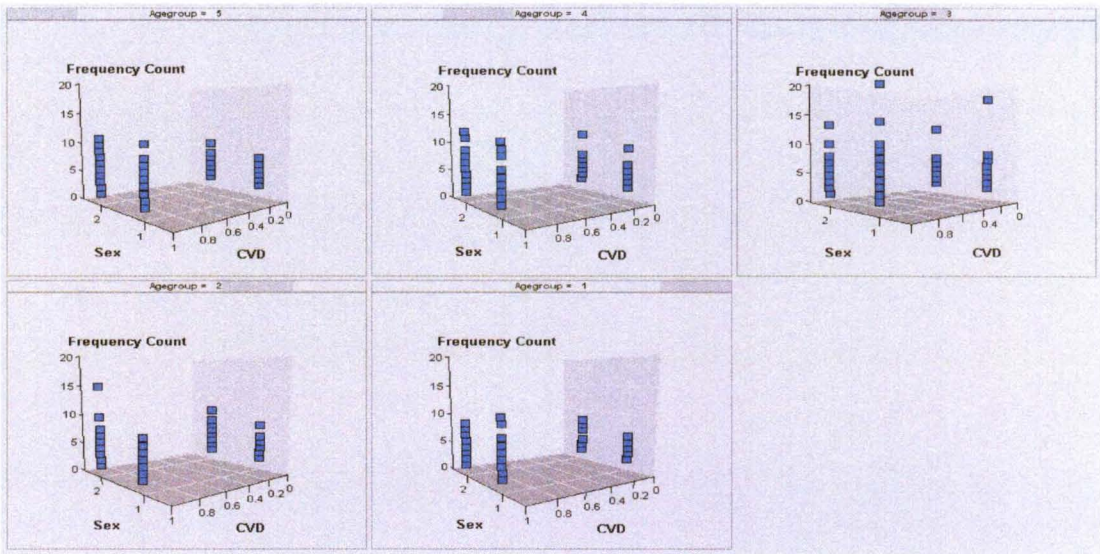
In the process of conditionally inner joining the two tables, observations are

grouped by patient ID and dates; each group having the property that the differences between the discharge date and the readmission date is less than 30 days. The code is:

```
PROC SQL; CREATE TABLE SASUSER.AD AS
SELECT T1.BENE_ID, T1.CLM_ADMSN_DT,T1.NCH_BENE_DSCHRG_DT,
DATDIF(T1.NCH_BENE_DSCHRG_DT,MIN(T2.CLM_ADMSN_DT),'ACT/ACT')
AS DATEDIF, MIN(T2.CLM_ADMSN_DT) FORMAT=DATE9. AS MINAD
FROM SASUSER.QRDM2010 AS T1
INNER JOIN SASUSER.QRDM2010 AS T2
ON T1.BENE_ID=T2.BENE_ID AND
T1.CLM_ADMSN_DT<T2.CLM_ADMSN_DT AND
T2.CLM_ADMSN_DT>T1.NCH_BENE_DSCHRG_DT
WHERE T1.NCH_BENE_DSCHRG_DT IS NOT NULL AND T1.CLM_ADMSN_DT
IS NOT NULL  GROUP BY
T1.BENE_ID,T1.CLM_ADMSN_DT,T1.NCH_BENE_DSCHRG_DT
HAVING DATDIF (T1.NCH_BENE_DSCHRG_DT, MIN(T2.CLM_ADMSN_DT)
,'ACT/ACT') <=30 ORDER BY DATEDIF; QUIT;
```

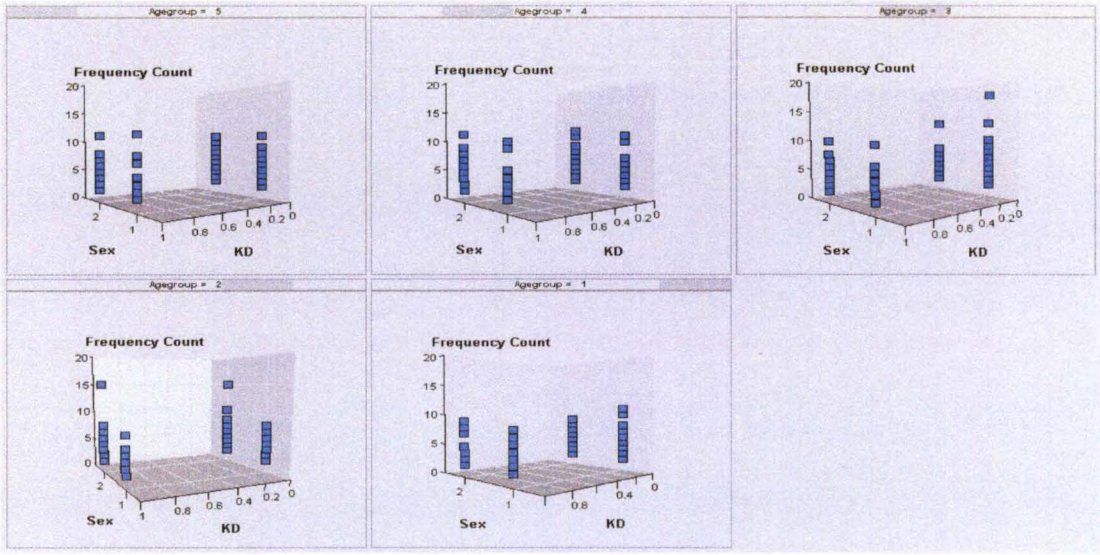
6.3.3 Pattern Recognition

We first conduct pattern recognition analysis using the cluster node in Enterprise Miner with the hierarchical clustering algorithm.



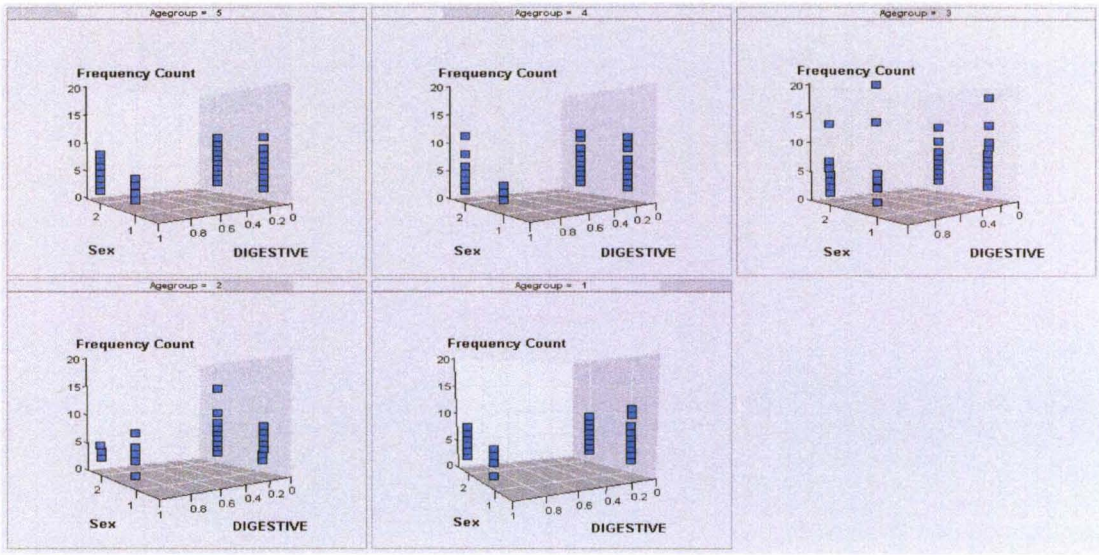
Agegroup 1:age 65-69; 2:70-74; 3: 75-79; 4: 80-84; 5:>=85 **Gender** 1: Male ; 2 : Female
CVD 0: No; 1: Yes

Figure 6.11. Demographic Characteristics of Patients with Cardiovascular Disease



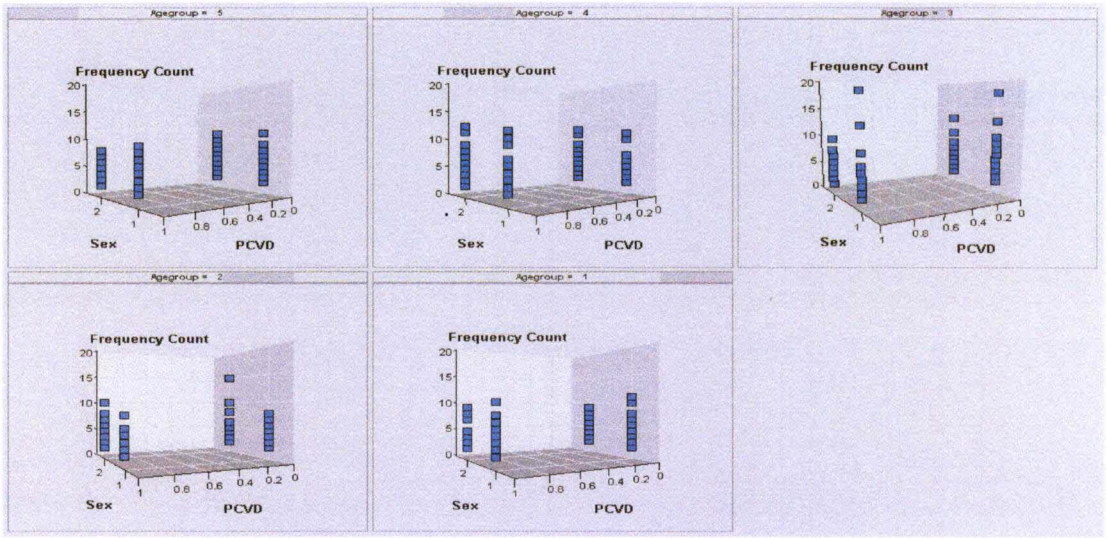
KD 0: No; 1: Yes

Figure 6.12. Demographic Characteristics of Patients with Kidney Disease



DIGESTIVE 0: No ; 1: Yes

Figure 6.13. Demographic Characteristics of Patients with Digestive Disorder



PCVD 0: No ; 1: Yes

Figure 6.14. Demographic Characteristics of Patients Having Cardiovascular Procedure

Some results shown in Figures 6.11 – 6.14 give us a rough description of the inpatients. The patients are segmented by 5 age groups. In Figure 6.11, the x-axis, the y-axis and the z- axis stand for CVD (Cardiovascular Disease), gender and admission frequency respectively. It is the same for the other graphs. Generally, as Figures 6.11-6.12 and 6.14 demonstrate, there is no obvious difference in frequency of hospitalization in different age groups for those patients who have cardiovascular disease or renal disease or have procedures related to these diseases. However, the patients who have heart disease are more likely to be admitted to the hospital than those who do not and this difference can be seen in the axis representing CVD in Figure 6.11. Figure 6.14 indicates that the male patients who receive heart procedures have a higher possibility of readmission compared to female patients. When comparing Figures 6.11- 6.12 to Figure 6.13, we can also conclude that in most cases, the patients with cardiovascular disease or kidney disease have a higher risk of readmission after discharge compared

to those with other diseases such as digestive disorder.

6.3.4 Supervised Learning with Readmission as the Target

In the following analysis, we need a more accurate estimate of the readmission risk; then, we will perform supervised learning. First, we combine the patients who readmitted to the hospital with the other inpatients and define a new variable, RA for readmission risk.

Next, we utilize the code displayed below to find out the top 20 most frequent procedure codes for the patients readmitted and apply a similar method to get the top 20 diagnosis disease codes. The most common co-morbidities are heart disease, kidney disease, digestive disorder and respiratory disease and the most frequent procedures are those related to the above diseases. Therefore, we apply the CATX, 0-1 indicator and RXMATCH functions to define the new predictors such as CVD (Cardiovascular disease) and PKD (Procedures related to kidney diseases) with the SAS code:

```
/*put all procedures into one column*/
PROC SQL;CREATE TABLE SASUSER.PRO04 AS
SELECT BENE_ID, ICD9_PRCDR_CD1 AS PRO FROM SASUSER.RD04
WHERE ICD9_PRCDR_CD1 IS NOT NULL   UNION
SELECT BENE_ID, ICD9_PRCDR_CD2 AS PRO FROM SASUSR.RD04
WHERE ICD9_PRCDR_CD2 IS NOT NULL...
  /*find the freq of procedures*/
CREATE TABLE SASUSER.COUNT AS
SELECT PRO, COUNT(*) AS C FROM SASUSER.PRO04
GROUP BY PRO  ORDER BY C DESCEDING; QUIT;
```

For a precise prediction, we utilize the variable selection node in SAS Enterprise Miner 6.2 to select the prominent input variables by R^2 shown in Figure 6.15. Considering the large data sample, we choose the variables with R^2 greater than 0.1 and

they are patient discharge location, Medicare status, diabetes, kidney disease, length of stay in the hospital, respiratory disease and procedures related to kidney dysfunctions.

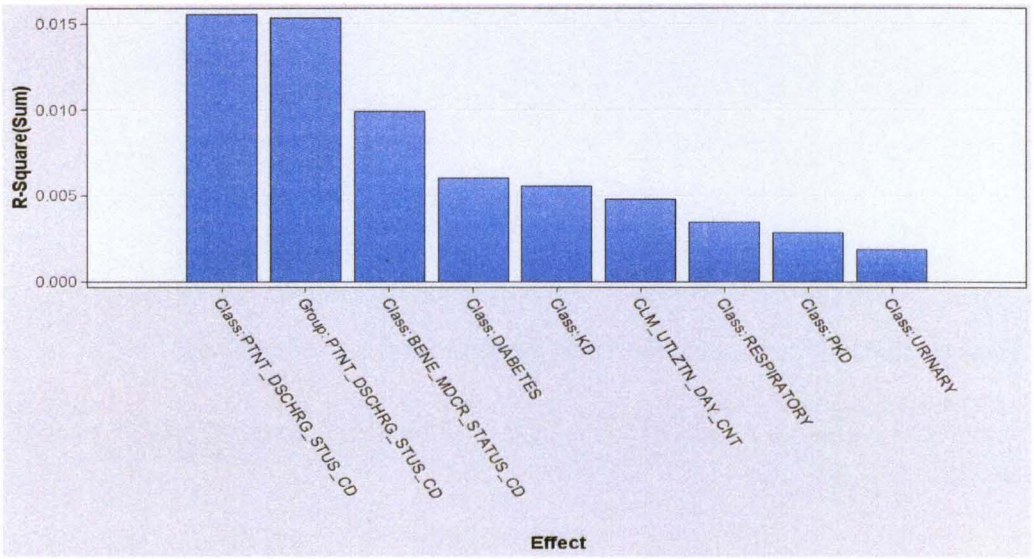


Figure 6.15. Selected Variables by R-square

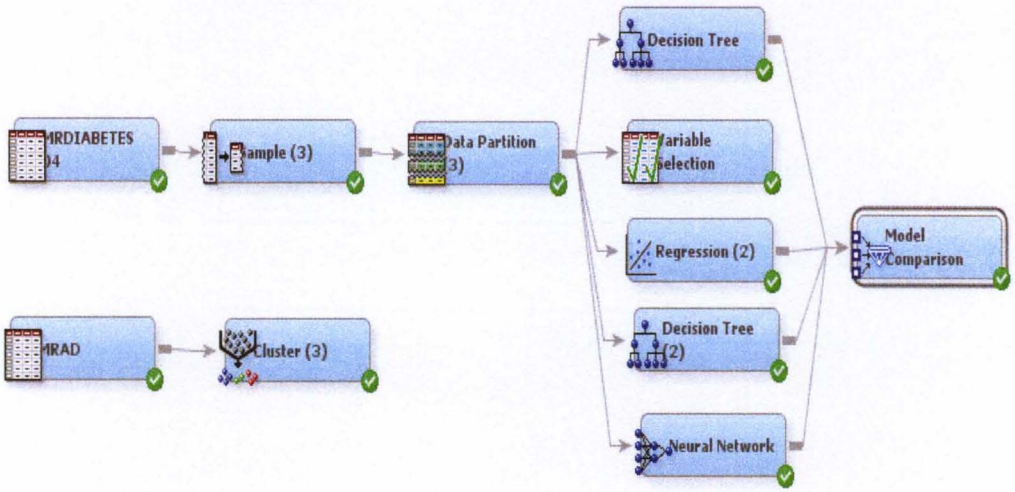


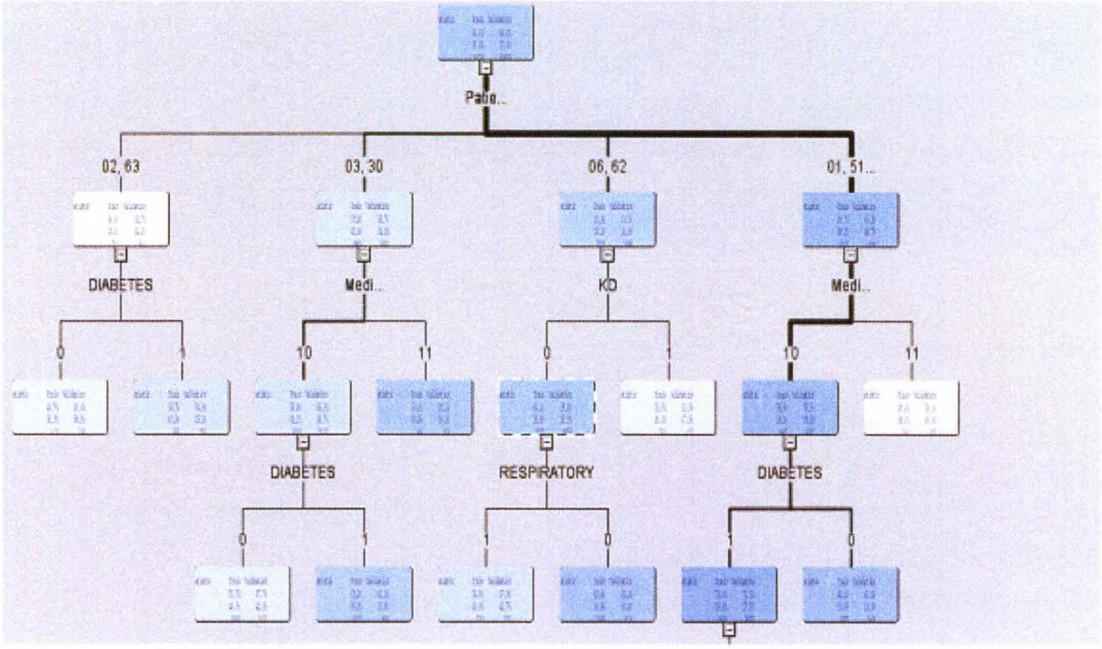
Figure 6.16. Various Predictive Models

Now, we can use different models such as the decision tree models (with the

default method or with the CHAID algorithm), the logistic regression model and the neural network model to predict readmission risk. Figure 6.16 shows the process flow.

Table 6.6 Important Variables to Readmission Estimation

Variable name	Number of splitting rules	Importance
Discharge places	1	1
Medciare status	2	0.74
Diabetes	3	0.48
Length of hospitalization	1	0.41
Kidney disease	1	0.33
Respiratory disease	1	0.22
Procedures related to kidney diseases	0	0



Discharge place code 01:home; 02: short term general hospital; 03:skilled nursing facility; 06:home health service; 30: still patient; 51 hospice - certified medical facility ;62:inpatient rehab facility; 63:Medicare certified long term care hospital.

Figure 6.17. Tree Diagram with CHAID for Readmission

The model comparison node chooses the decision model with the CHAID algorithm as the optimal model. The outputs shown in Table 6.6 demonstrate that patient discharge places, patient status (old or old with end stage renal failure), LOS (length of stay in the hospital), kidney disease and respiratory disease are key factors to predict readmission rate. In the above tree diagram (shown in Figure 6.17), such organizations as a short term hospital, a skilled nursing facility and a Medicare-certified long term care hospital can provide professional medical services while the other organizations do not. The diagram indicates that the inpatients who are discharged to home or hospice have a higher readmission rate than other patients.

6.4 Two – way Interaction Effects of Diabetes Complications

In the previous sections, we discussed the key factors to costs and outcomes without considering the interactions between each factor. In this section, we will analyze the two-way interaction effects of diabetes complications such as cardiovascular disease and hypertension on Medicare expenditures and health outcomes such as frequency and length of hospitalization. For two factors, A and B, two – way interaction effect analysis reflects whether a level of effect of factor A is influenced by a level of effect of factor B. An interaction plot can simply and vividly demonstrate the interaction, but in most cases, these interactions need statistical demonstration.

6.4.1 Data Processing

In order to get the data ready for our analysis, we process the data in the following steps.

- Conditionally inner join the two data sets, inpatient claims data and beneficiary

demography data by beneficiary ID.

- Use random sampling to reduce the data size from 244,299 records to 40,000 claims.
- Use the proc means procedure to get the average costs, average days in the hospital and total times of entering the hospital for each beneficiary ID (Each patient in Medicare has his/her own unique beneficiary ID).
- Utilize the CATX function, RXMATCH function and 0-1 indicator functions to generate binary variables, common diabetes complications and they are Hea (Heart disease), Kid (Kidney disease), Eye (Eye disease), Neu (Neurologic disease) and Hyper (Hypertension).
- When we develop the logistic regression model to predict mortality, we apply oversample to guarantee that all the observations of mortality are included in the sample and we use the following SAS code:

```
PROC SORT DATA=SASUSER.INTERACTIONNEW OUT=SASUSER.LOGISTIC;
BY STATUS; RUN;
PROC SURVEYSELECT DATA=SASUSER.LOGISTIC
OUT=SASUSER.OVERSAMPLE
SEED = 39585784  SAMPSIZE = 1256;
STRATA STATUS; RUN;
```

Now we get what we want and we can perform our analysis in the following sections.

6.4.2 Two – way Interaction Effects of Diabetes Complications on Cost

We start our study with cost analysis. In order to examine whether there exist significant two-way interaction effects between diabetes complications, we firstly plot interaction effects on Medicare payments. We use the following SAS code and the

similar codes for the other interaction effects plots .The variables are explained in Table 6.7.

```
PROC SORT DATA=SASUSER.INTERACTIONNEW OUT=SASUSER.NEW1;
BY HEA_MAX HYPER_MAX; RUN;
PROC MEANS DATA=SASUSER.NEW1;
VAR CLM_PMT_AMT_MEAN; BY HEA_MAX  HYPER_MAX;
OUTPUT OUT=SASUSER.NEW2 MEAN=CLM_PMT_AMT_MEAN;RUN;
SYMBOL VALUE=DOT I=JOIN;PROC GPLOT DATA=SASUSER.NEW2;
PLOT CLM_PMT_AMT_MEAN * HEA_MAX = HYPER_MAX;RUN;
```

Table 6.7 Explanations for Variables

Variable	Explanation
CLM_PMT_AMT_MEAN	Average payments for each patient
HEA_MAX	Heart disease
HYPER_MAX	Hypertension
OCULO_MAX	Eye disease
NEU_MAX	Neurological disease

An interaction effect plot gives a general estimation whether an interaction effect is significant and how the two different factors influence each other. The rule [61] is that if the two lines (representing the two levels of one factor) do not cross and are not parallel to each other, then the effect is significant. The difference of two levels of one factor indicates the main effect of the other factor. If the two lines interact with each other, the interaction effect is still significant, but the main effect should not be considered.

To be specific, let us look at the plots in Figure 6.18. These plots demonstrate some interactions between different diseases with Medicare payments as a response variable and they are significant according to the stated rule. Consider kidney disease, for

example; its interaction effect with a neurological disorder is significant and for diabetes inpatients with kidney disease, there is a big difference on their costs between the case when they have a neurological disorder and the case when they do not.

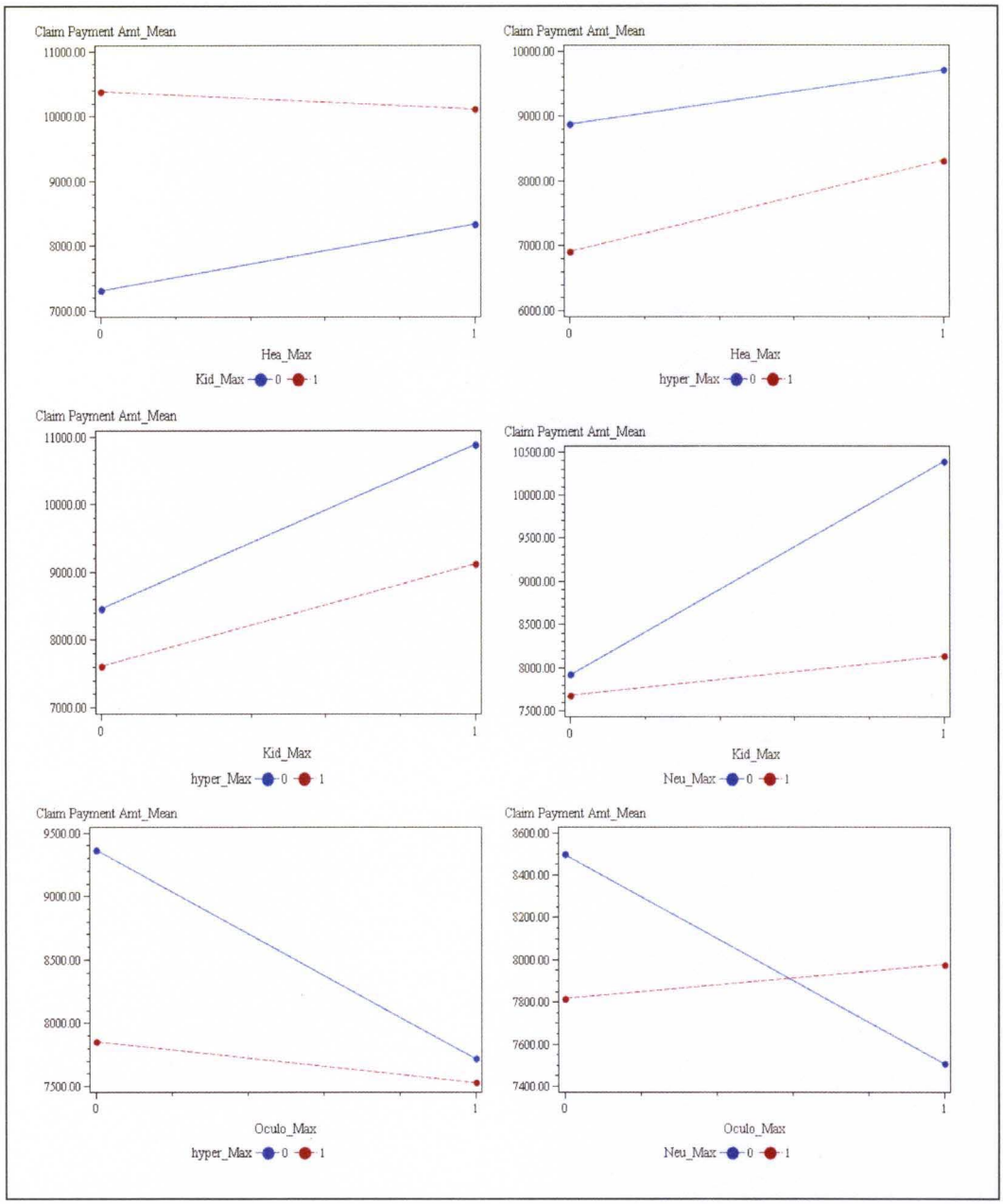


Figure 6.18. Interaction Effects on Costs Plots

However, the above plots just provide a basic idea; we need to use statistical models to further prove the significance of the interaction effects. We input the variables listed in Table 6.7 and their two-way interaction as predictors into the generalized linear model with a gamma distribution and use Medicare payments as the response variable, then we get the type III analysis results shown in Table 6.8. The table shows that without considering main effects, only these interaction effects are significant in the model of Medicare payments and they are: interaction between heart disease and kidney disease, interaction between heart disease and hypertension, interaction between kidney disease and neurological disease, interaction between kidney disease and hypertension, interaction between eye disease and neurological disease, and interaction between eye disease and hypertension since all of their p-values are smaller than the significance level 0.05.

Table 6.8 Type 3 Analysis for Interaction Effects(Cost)

Wald Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
Hea_Max*Kid_Max	1	22.64	<.0001
Hea_Max*hyper_Max	1	25.86	<.0001
Kid_Max*Neu_Max	1	17.96	<.0001
Kid_Max*hyper_Max	1	10.66	0.0011
Oculo_Max*Neu_Max	1	7.76	0.0053
Oculo_Max*hyper_Max	1	12.11	0.0005

6.4.3 Two – way Interaction Effects on Length of Hospitalization

In this section, we discuss the influence on length of stay (LOS) in the hospital. From this section on, we will directly investigate the effects with statistical models without demonstrating interaction plots. Because LOS also follows a gamma distribution

and has similar properties to Medicare payments, we use the generalized model with a gamma distribution. Table 6.9 displays the significant interaction effects. It demonstrates that for patients with other complications, the fact that they have kidney disease or not makes their LOS different. This conclusion is also true for hypertension.

Table 6.9 Type 3 Analysis for Interaction Effects (LOS)

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
Hea_Max*hyper_Max	1	38.37	<.0001
Kid_Max*Oculo_Max	1	3.99	0.0458
Kid_Max*Neu_Max	1	11.03	0.0009
Kid_Max*hyper_Max	1	3.88	0.0487
Neu_Max*hyper_Max	1	8.40	0.0038
Oculo_Max*hyper_Max	1	10.55	0.0012

6.4.4 Two – way Interaction Effects on Frequency of Hospitalization

In this section we will discuss the interaction effects on frequency of hospitalization, i.e., the average times of hospital stay for each patient. Since the frequency is a count variable, we assume that it follows a Poisson distribution. Therefore, we apply a Poisson regression model for our analysis. In order to adjust the over-fitting of this model, we set "scale = Pearson" and the SAS code is shown below.

```
PROC GENMOD DATA=SASUSER.INTERACTIONNEW;
CLASS Hea_Max Kid_Max Oculo_Max Neu_Max hyper_Max;
MODEL COUNT= Hea_Max*Kid_Max … hyper_Max/
LINK=LOG
DIST=POISSON
SCALE=PEARSON
TYPE3;
LSMEANS Hea_Max*Kid_Max …/ ALPHA=0.05; RUN;
```

The scaled deviance and the scaled Pearson χ^2 (Table 6.10) indicates that the model fit the data relatively well. Table 6.11 (Only significant interaction effects left in this table) demonstrates that for the patients with other diseases, there is a big difference in the frequency of being admitted to the hospital between the patients who have heart disease and those who do not.

Table 6.10 Goodness Fit of Poisson Regression model

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	34E3	4517.0538	0.1327
Scaled Deviance	34E3	26557.3993	0.7803
Pearson Chi-Square	34E3	5788.5522	0.1701
Scaled Pearson X2	34E3	34033.0000	1.0000

Table 6.11 Type 3 Analysis for Interaction Effects (Frequency)

LR Statistics For Type 3 Analysis						
Source	**Num DF**	**Den DF**	**F Value**	**Pr > F**	**Chi-Square**	**Pr > ChiSq**
Hea_Max*Kid_Max	1	34033	34.45	<.0001	34.45	<.0001
Hea_Max*Oculo_Max	1	34033	7.79	0.0052	7.79	0.0052
Hea_Max*Neu_Max	1	34033	6.99	0.0082	6.99	0.0082
Hea_Max*hyper_Max	1	34033	50.57	<.0001	50.57	<.0001
Kid_Max*Neu_Max	1	34033	24.67	<.0001	24.67	<.0001
Kid_Max*hyper_Max	1	34033	121.91	<.0001	121.91	<.0001

6.4.5 Two – way Interaction Effects on Mortality

In the last section of two-way interaction effects, we will analyze the effects on the mortality of inpatients with diabetes. Mortality is a rare occurrence event because the patients who are dead only account for 4.5% in the whole inpatient population. Hence, in order to get an objective result, we need to oversample those patients. We stratify

STATUS into two classes, dead or alive; then, we sort the data by STATUS and perform stratified sampling, setting sample size to be 1256 (the total number of the deaths in a sample with a size of 40,000 claims) and selecting the same number of patients alive. The code is shown below.

```
PROC SORT DATA=SASUSER.INTERACTIONNEW OUT=SASUSER.LOGISTIC;
BY STATUS; RUN;
PROC SURVEYSELECT DATA=SASUSER.LOGISTIC
OUT=SASUSER.OVERSAMPLE
SEED = 39585784  SAMPSIZE = 1256;
STRATA STATUS; RUN;
```

Once we get the data ready, we perform logistic regression analysis with the binary variable, STATUS as the dependent variable and the stepwise selection method to choose the best model. The selection process only proceeds to the second step and then stops, since no effect in the model can be removed and no additional new effect meets the requirement of entering the model. The R-square of this model is 13 %; considering the large data size, it is a reasonable fit, but it still means that 87% of the variability in the outcome variable remains unaccounted for. Table 6.13 demonstrates that the two –way interaction effects between heart disease and kidney disease or the interaction effects between neurological disease and hypertension are significant in the model.

Table 6.12 R-square for Logistic Regression Model

R-Square	0.1275	**Max-rescaled R-Square**	0.1700

Table 6.13 Type 3 Analysis for Interaction Effects (Mortality)

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Hea_Max*Kid_Max	1	14.9807	0.0001
Neu_Max*hyper_Max	1	7.9346	0.0048

6.5 Conclusion

Cost analysis of inpatients shows that many organ diseases and neurological disorders indeed have decisive effects on the costs of inpatients with diabetes: heart disease, eye disease and nervous system diseases raise the inpatients costs. The results demonstrate that under a threshold amount of costs, kidney disease does not impact the expenditures of inpatients as much as the other organ diseases do; however, as the costs increase, the effects of kidney disease become more and more important. Hence, all inpatients with diabetes should pay more attention to kidney disease, and use prevention to avoid such diseases to decrease the costs.

Outcomes research demonstrates that among the various procedures, the ones utilized for cardiac disease treatments are related to many different procedures. Association analysis also shows that hemodialysis is strongly related to venous catheterization for renal dialysis. Another discovery is that the procedures and the diagnoses are decisive to predicting mortality, which is contrary to widely held belief.

Through this readmission study, we can conclude that the following kinds of the Medicare diabetic inpatients have a high risk of readmission, those who have cardiovascular disease, kidney disease or those male patients who receive procedures related to heart disease or those who do not get professional medical care after discharge. Therefore, in order to decrease readmission risk, we should take greater care of these kinds of patients. This care might need to include a higher level of reimbursement from Medicare.

Two-way interaction effect analysis is carried out to study the interaction effects between various diabetes complications on Medicare payments, length and frequency of

hospitalization and mortality. Results show that the most interaction effects are significant to costs, while only the interaction between heart disease and other diseases are significant to the frequency of hospitalization. Another discovery is that kidney disease or hypertension has an important influence on the other complications as for length of stay in the hospital. We also find that the interaction between heart disease and kidney disease has a decisive effect on mortality.

We have finished studying outpatients and inpatients, and we will investigate the influence of Medicare, Part D on diabetes medications in chapter VII.

CHAPTER VII

INFLUENCES OF MEDICARE, PART D

Medicare, Part D is the optional prescription drug program. It uses competing private plans to provide beneficiaries access to appropriate drug therapies. In the recently approved healthcare law, the provisions about this plan were also amended. In this chapter, we will estimate the influences of Medicare, Part D since its implementation in 2006 on diabetes medications from usage and cost-effectiveness perspectives as well as on diabetes beneficiaries' health status. Our analyses are based on the MEPS data [19], which cover information about physician visits, inpatients, prescription drugs and demographic information for the year 2005 and the year 2006. The chapter is organized as follows: first, the theories related to survival analysis and cost-effective analysis are introduced, then the impacts of Medicare, part D on the usage of diabetes medications are addressed; finally the influences on the cost-effectiveness of the drugs are presented.

7.1 Basic Theories and Concepts

7.1.1 Survival Analysis

Survival analysis [62] is applied to study the occurrence and timing of events using statistical methods. It is very useful in studying the events in social and natural sciences, such as the onset of disease, births and death. In our research, we will utilize survival analysis to find the first switch of medication.

1. Censoring

Censoring is an important element in survival analysis, which distinguishes this analysis from other statistical methods. Censoring can be counted as an incomplete observations issue and it has three forms: left censoring, right censoring and interval censoring and two types: type I censoring, type II censoring and random censoring as follows:

- Left Censoring: An observation on a variable T is said to be left censored if T is only known when it is smaller than a value t_1.
- Right Censoring: An observation on a variable T is said to be right censored if T is only known when it is greater than a value t_2.
- Interval Censoring: An observation on a variable T is said to be interval censored if T is known when it belongs to the interval $[t_1, t_2]$.
- Type I censoring occurs when the values t_1, t_2 are fixed.
- Type II censoring occurs when observation is terminated after a pre-specified number of events have occurred.
- Random censoring occurs when the time of censoring and the survival time are independent.

In our research, we employ left censoring and type I censoring to study drug usage. We set the end of one year as the censored time; and if some drug user does not switch the drug to another, then the drug is said to be left censored.

2. Descriptions

In this part, we discuss the two standard approaches of describing survival analysis: (1) survival function, (2) hazard function.

(1) Survival Function: It is defined as the probability that an individual survives longer than t,

$$S(t) = P(T > t) = 1 - F(t) \tag{7.1}$$

in which $F(t)$ is a cumulative distribution function of variable t. If the event of interest is a medicine switch, then the survivor function gives the probability that the drug remains to be used beyond time t.

(2) Hazard Function: If a variable under survival analysis is continuous, then the hazard function is preferred and it is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t \le T \le t + \Delta t | T \ge t)}{\Delta t} \tag{7.2}$$

The importance of this function lies in its interpretations. If the occurrence of an event can be repeatable, then the hazard gives an idea about the number of events per interval of time. For a non-repeatable event such as death, its reciprocal tells how much time is left before the event occurs.

3. Estimation of Survival Functions

The estimation of survival functions is a traditional approach including the KM (Kaplan- Meier) method and the life table method. Due to the large data set, we prefer the latter method, in which the survival estimate is obtained by calculating the conditional probabilities of surviving beyond time t, defined by

$$S(t_i) = \prod_{j=1}^{i-1}(1 - q_j) = \prod_{j=1}^{i-1} p_j \tag{7.3}$$

For each time interval i, t_i is the start time, q_i is the conditional probability of failure and p_i is the conditional probability of surviving to t_i or beyond that time.

7.1.2 Cost Effectiveness Analysis

Policy makers want to spend fewer dollars while achieving greater treatment effects. Cost effectiveness analysis, a newly-emerged method in health economics or pharmacoeconomics, can help those people realize their goals. In this section, we will briefly introduce the essential concepts, theories and methods used in this methodology.

1. Basic Concepts

Before discussing the theories, several concepts [63, 64] should be considered.

- Health Intervention: This can refer to a treatment, test, or primary prevention technique, which is used to improve patients' health status or decrease mortality.
- Health Status: If we define a specific state of health, for instance, reducing pains, as a health state, then a health status is the sum of health states. Typically, there are six levels, from pretty healthy to dead shown in Table 7.1.
- QALY (Quality-Adjusted Life Year): It is a measure to estimate quantity and quality of life generated by healthcare interventions. QALY can be calculated as follows:

$$QALY = Weight\ of\ Health\ Status * Life\ Expectancy \tag{7.4}$$

In the above equation, weight of health status is defined by researchers. We define these weights in Table 7.1. Life expectancy is how many years left for a person before s/he dies and it can be looked up in a life table [65].

- ICER (Incremental Cost Effectiveness Ratio): It is the most frequently-used and most important element in health economics, which distinguishes it from other kinds of economics. It is typically defined as a ratio of difference in costs to differences in effects of interventions, shown in (7.5).

$$ICER = (cost_{new} - cost_{old}) / (effect_{new} - effect_{old}) \quad (7.5)$$

ICER indicates the additional costs required to generate one unit of effect. The effect is usually measured by QALY, and the cost is measured in currencies. ICER is used to compare two treatments or two drugs on the same patients in clinical trials, but we apply it to compare two cases for the two years for the same drug treatment with the same patients.

Table 7.1 Weights of Health Status

Level of Health Status	**Weight**
Pretty Healthy	1
Healthy	0.8
Relatively Healthy	0.6
Ill	0.4
Severely Ill	0.2
Dead	0

2. Common Methodologies

In cost effectiveness analysis, two kinds of models [66] are often utilized. One is the decision tree model, different from what we discussed in the previous chapters. The tree model must first be split on two interventions, and then each intervention node is split into several sub-trees about costs and outcomes with corresponding probabilities. The other is a Markov model, assuming that there are finite numbers of defined health states, and at any time, each patient should be assigned to one health state. At the end of each state, the patients can be shifted from one state to another state with a certain probability. Another approach is one-way sensitivity analysis [67] , in which one variable is chosen to change values and the other variables are kept constant each time; the ICER are calculated to see whether the parameter is sensitive or not.

3. Our Approach

In part 2, we introduced several common methods; however, they are not suitable for our studies. Instead, we apply medical resources usage analysis.

In the following section 7.2, we will investigate the impacts on diabetes medications using survival analysis; in section 7.3, we will examine the effects on the cost effectiveness of drugs and patient health.

7.2 Impacts of Medicare, part D on the Usage of Diabetes Medications

In this section, apart from survival analysis, we also perform summary statistics and kernel density estimation to understand the impacts better.

7.2.1 Summary Statistics

We use Summary Statistics to get the average Medicare payment and the average total payment. For comparison, the average Medicare payment in 2005 is expressed in 2006 dollars with 2005 data inflated based on the CPI-U (Consumer Price Index for all Urban Consumers) for prescription drugs [68]. That is to say, once we get the mean values for the year 2005, we multiply them by the index 1.043. To demonstrate this conversion, we put an asterix on the right upper corner of '2005' in Table 7.2.

Table 7.2 Average Overall Payment and Medicare Payment in 2005 & 2006

Year	Variable	Mean	N
2005*	SUM OF PAYMENTS	501.66	1759
	MEDICARE (IMPUTED)	20.58	1759
2006	SUM OF PAYMENTS	558.11	1994
	MEDICARE (IMPUTED)	129.02	1994

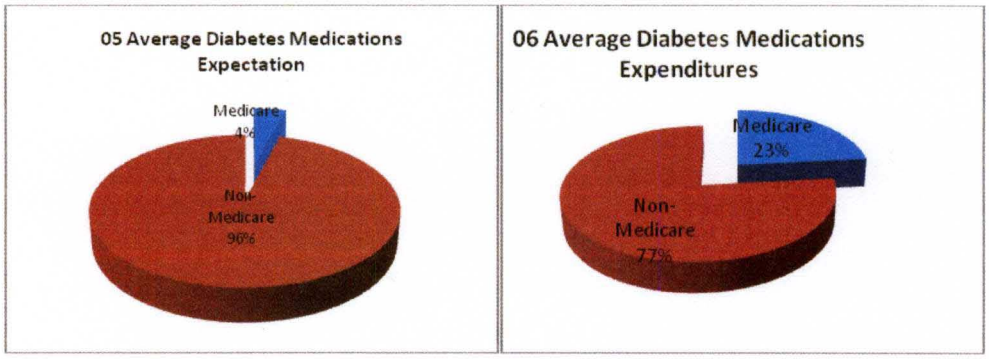


Figure 7.1. Pie Charts of Payments in 2005 &2006

Table 7.2 and Figure 7.1 show that the average of total payments for the prescription increases approximately 12 % from the year 2005 to the year 2006, while the average Medicare payment in 2006 is 6 times as much as that in 2005. The ratio of the average Medicare payment to the total expenditures also increases from 4 % to 23 %. Results demonstrate that the plan, Part D, indeed increases Medicare drug expenditures.

7.2.2 Kernel Density Estimation among Different Clusters of Drugs

Next, we want to see how Medicare payments are distributed among different groups of drugs. We compare Medicare payments in two cases, the general case and the Medicare case in which the beneficiaries join Medicare. We need to preprocess the data sets. First, we convert the variable, NRXNAME, into observations by the transpose procedure, the trim function, the translate function and the concatenation operator. The SAS code [26] is shown below.

```
PROC SORT DATA=SASUSER.SMED06 OUT=SASUSER.SORTMED06;
BY DUPERSID NRXNAME; RUN; OPTIONS OBS=MAX;
DATA SASUSER.SORTMR06; SET SASUSER.SORTMED06;
NRXNAME= TRANSLATE(LEFT(TRIM(NRXNAME)),'_',' '); RUN;
PROC TRANSPOSE DATA=SASUSER.SORTMR06 OUT=SASUSER.TRANMR06
PREFIX=MED_; VAR NRXNAME; BY DUPERSID; RUN;
```

```
DATA SASUSER.CONMR06 (KEEP= DUPERSID SSNRXNAME);
LENGTH SSNRXNAME $ 32767; SET SASUSER.TRANMR06;
ARRAY CONCAT MED_:; SSNRXNAME =LEFT(TRIM(MED_1));
DO I=2 TO DIM(CONCAT); SSNRXNAME=LEFT(TRIM(SSNRXNAME))
||''|| LEFT (TRIM(CONCAT[I])); END; RUN;
```

(1) General Case:

After we get the clusters shown in Figures 7.2 & 7.3 using SAS Text Miner, each cluster is explained in Tables 7.3 & 7.4; we can do kernel density estimation on Medicare payments by clusters.

Clusters

#	Descriptive Terms	Freq	Percentage	RMS Std.
1	glyburide, rosiglitazone, precose, metformin, glyburide-metformin	271	0.154328018...	0.0711406...
2	glyburide-metformin, + supply, starlix, precose, prandin	493	0.280751708...	0.1198462...
3	insulin, pioglitazone, + supply, starlix, prandin	237	0.134965831...	0.0374267...
4	metformin, glimepiride, tolazamide, prandin, rosiglitazone	494	0.281321184...	0.0838507...
5	glipizide, rosiglitazone, precose, starlix, metformin	261	0.148633257...	0.0693840...

Figure7.2. Clusters of Drugs in 2005

Clusters

#	Descriptive Terms	Freq	Percentage	RMS Std.
1	glyburide-metformin, glimepiride, starlix, insulin, prandin	214	0.108961303...	0.2865909...
2	+ supply, glimepiride	358	0.182281059...	0.0148972...
3	glyburide, rosiglitazone, precose, glyburide-metformin, metformin	272	0.138492871...	0.0554535...
4	insulin, + supply, starlix, precose, glyburide-metformin	228	0.116089613...	0.0484771...
5	tolazamide, metformin, glimepiride, prandin, precose	611	0.311099796...	0.0897675...
6	glipizide, rosiglitazone, pioglitazone, precose, metformin	281	0.143075356...	0.0510506...

Figure 7.3. Clusters of Drugs in 2006

Table7.3 Explanation for Clusters in 2005

Cluster #	Label
1	Metformin, Glyburide and their combination
2	Supplies
3	Insulin, Supplies
4	Metformin, Glimepiride
5	Metformin, Glipizide

Table 7.4 Explanation for Clusters in 2006

Cluster#	Label
1	Insulin, Glyburide-metformin
2	Supplies
3	Glyburide, metformin and their combination
4	Insulin, Supplies
5	Metformin, Glimepiride
6	Metformin, Glipizide

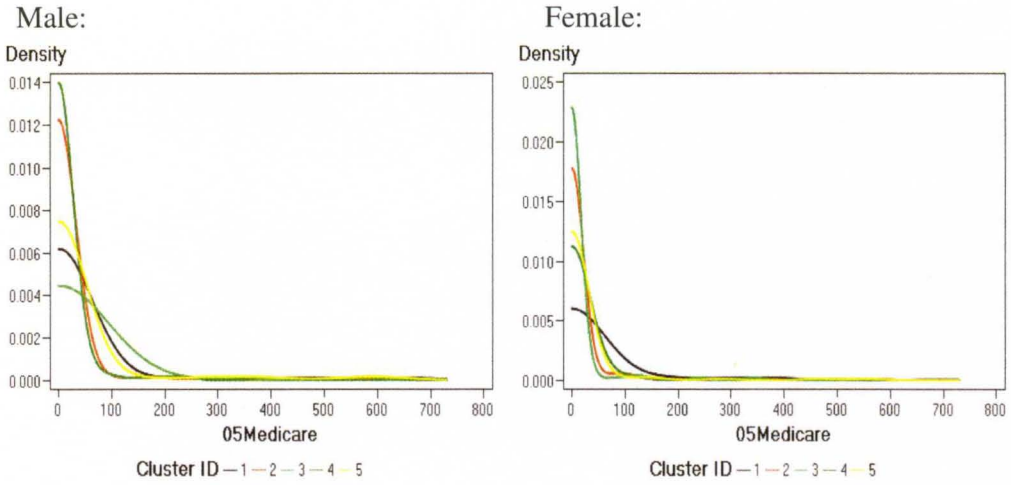


Figure 7.4. Kernel Density Estimation for Medicare in 2005

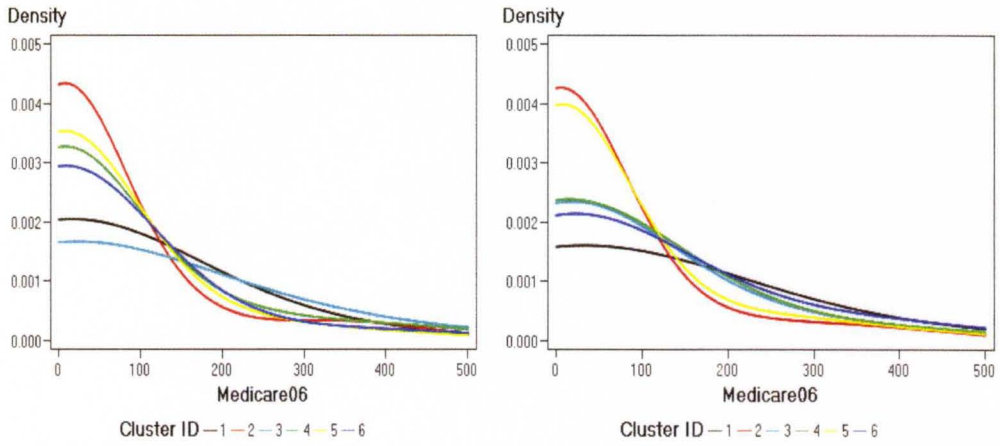


Figure 7.5. Kernel Density Estimation for Medicare in 2006

Figure 7.4 demonstrates that most Medicare payments for diabetes medication are fewer than 200 dollars. For males, most costs of the drugs are lower than 50 dollars. The only exception is cluster 1, which indicates that Medicare pays more for metformin and glyburide, and their combination. For female patients, the densities of clusters #2 and

#4 are higher than those of the other clusters under 50 dollars; after that, the density of cluster 3 is the highest. Hence, female patients spend more on insulin and supplies. In 2006, the ordering for males' expenditures is cluster# 2 > #5> #6>#1>#3 under 120 dollars; after the threshold point, the densities for clusters #1 and #3 become higher than the others. Women spend much more on clusters #2 and # 5 of the drugs under 140 dollars, indicating that metformin and supplies cost females more than the others do. Hence, most Medicare expenditures are on supplies, metformin and glyburide.

(2) Medicare Case

Figures 7.6 and 7. 7 show that in 2006, with the Part D introduction into Medicare, the expenditures on drugs are greatly increased. The costs of the diabetes supplies, metformin and insulin remain higher than the other costs.

Male: Female:

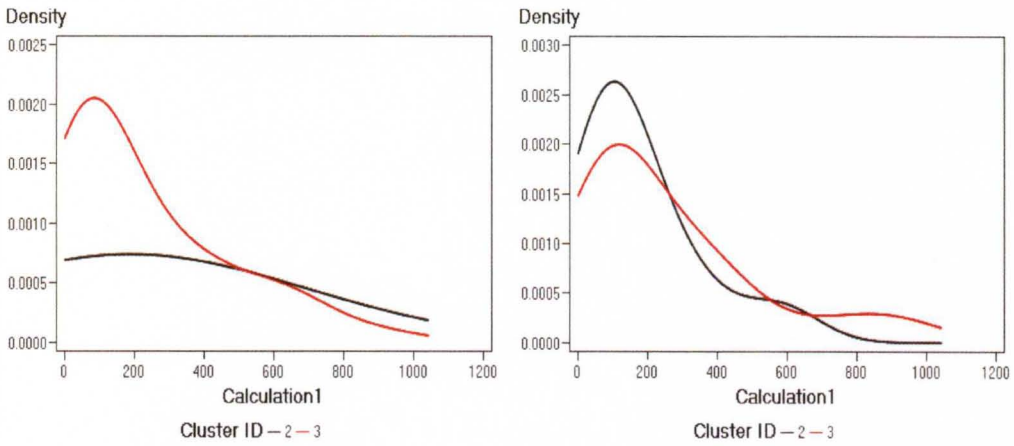


Figure 7.6. Kernel Density Estimation of 2005 Medicare

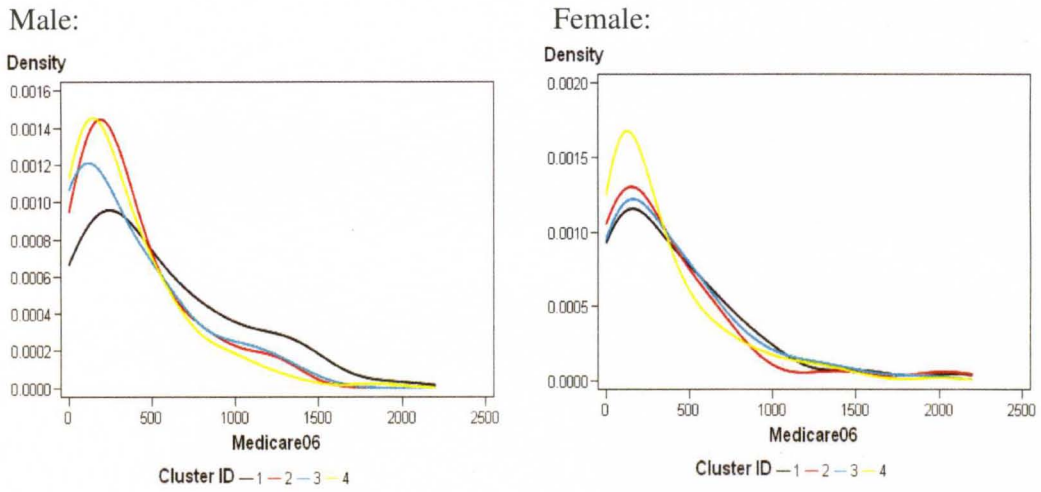


Figure 7.7. Kernel Density Estimation of 2006 Medicare

7.2.3 Association Analysis

In this section, we use the association node in EM 5.2 to get the link graphs for relationship analysis of diabetes oral medications. In each graph, each square stands for a drug; those drugs which are connected to many different drugs form centers and are important in diabetes treatment. The link graphs are displayed in Figures 7.8-7.11.

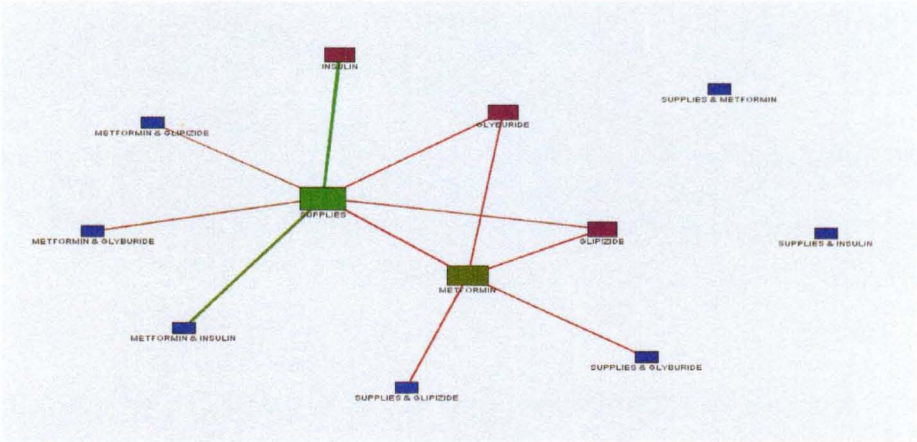


Figure 7.8. Link Graphs for the Drugs in 2005 (General Case)

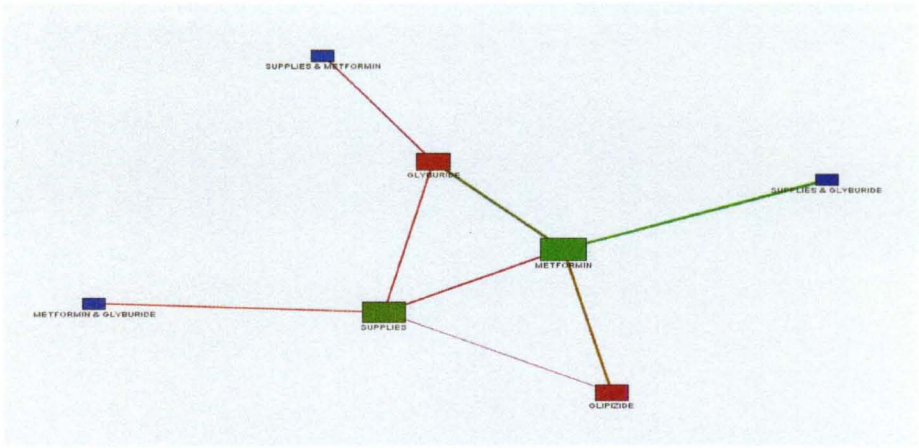


Figure 7.9. Link Graphs for the Drugs in 2005 (Medicare Case)

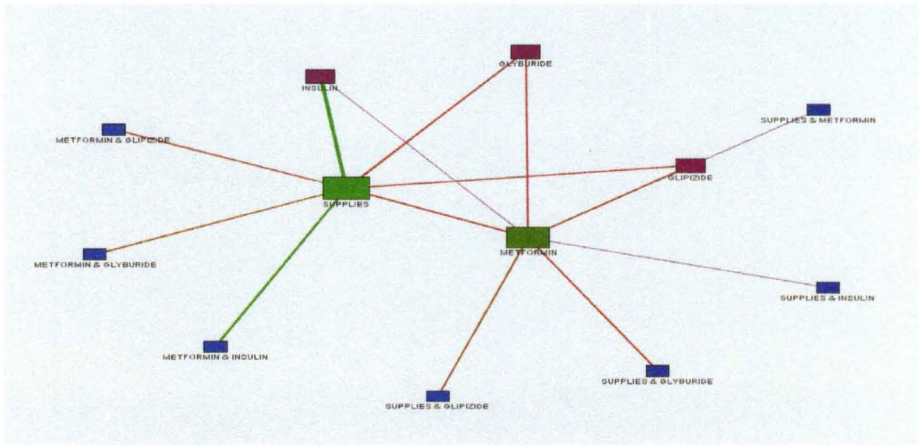


Figure 7.10. Link Graphs for the Drugs in 2006 (General Case)

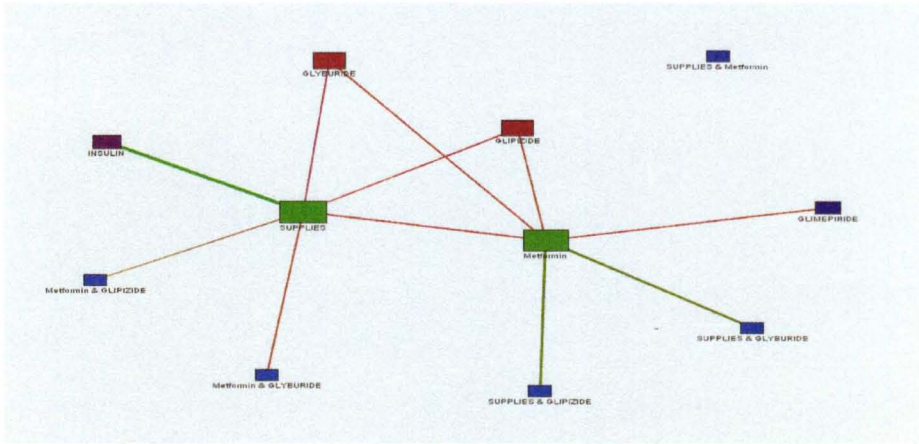


Figure7.11. Link Graphs for the Drugs in 2006 (Medicare Case)

In this analysis, we also discuss the general case and the Medicare case. In 2005 in general case (shown in Figure 7.8), diabetes supplies and metformin were two centers of the graphs. Insulin has a strong relationship to supplies, although it is not related to the other factors. Glyburide, and glipizide are also fundamental. When we study the beneficiaries who get their drugs paid by Medicare (shown in Figure 7.9), metformin becomes more prominent and it is strongly related to glyburide and glipizide. In 2006, the general case (displayed in Figure 7.10) remains almost the same except that the supplies are connected to the combination of insulin and metformin. In the Medicare case (displayed in Figure 7.11), there are fewer connections between different drugs. Figure 7.11 indicates that if insulin is prescribed, then the supplies are very likely to be prescribed, too. Also, once the combinations of the supplies with glyburide are used, then metformin will probably be utilized.

7.2.4 Survival Analysis

Finally, we perform survival analysis by the life test procedure. For a better comparison, we also need physician visit information in 2005. For the year 2006, we process the missing time information using the following SAS code and then sort out the records for the year 2006 to get the data set shown in Figure 7.12.

```
PROC SQL;
CREATE TABLE SASUSER.SMRS06 AS SELECT SRSMED06.DUPERSID,
SRSMED06.RNRXNAME,(( CASE WHEN -1 = SRSMED06.RXBEGDD THEN 1
WHEN -8 = SRSMED06.RXBEGDD THEN 1 WHEN -9 =
SRSMED06.RXBEGDD THEN 1 ELSE SRSMED06.RXBEGDD END)) AS
RXDD,
((CASE WHEN -1 = SRSMED06.RXBEGMM THEN. WHEN -8 =
SRSMED06.RXBEGMM THEN. WHEN -9 = SRSMED06.RXBEGMM THEN.
ELSE  SRSMED06.RXBEGMM END)) AS RXMM,
```

```
((CASE WHEN -1 = SRSMED06.RXBEGYRX THEN 2006 WHEN -14 =
SRSMED06.RXBEGYRX THEN 2006 WHEN -7 = SRSMED06.RXBEGYRX
THEN 2006 WHEN -8 = SRSMED06.RXBEGYRX THEN 2006 WHEN -9 =
SRSMED06.RXBEGYRX THEN 2006 ELSE SRSMED06.RXBEGYRX END)) AS
RXYY
FROM SASUSER.SRSMED06 AS SRSMED06 WHERE CALCULATED RXYY =
2006; QUIT;
```

Next, we suppress the data by removing the repeated information and the SAS code is as follows:

```
PROC SORT DATA=SASUSER.SMRS06 OUT=SASUSER.UNISMRS06
NODUPKEY; BY DUPERSID SSNRXNAME; RUN;
```

	DUPERSID	SSNRXNAME	RXDD	RXMM	RXYY
1	30078019	METFORMIN	1	.	2006
2	30121012	GLIPIZIDE	1	.	2006
3	30149010	METFORMIN	1	.	2006
4	30177026	GLIPIZIDE	27	12	2006
5	30177026	METFORMIN	27	12	2006
6	30180024	GLYBURIDE	1	.	2006
7	30192012	METFORMIN	20	1	2006
8	30206025	GLIMEPIRIDE	1	.	2006
9	30217015	GLYBURIDE	1	2	2006
10	30300013	METFORMIN	30	8	2006

Figure 7.12. Diabetes Medication in 2006

Next, we convert the date into a SAS date by using the MDY function, transpose the data by NRXNAME and DATE and finally merge the two new data sets to get the data displayed in Figure 7.13.

	DUPERSID	VAR	MED_1	MED_2	MED_3	DATE_1	DATE_2	DATE_3
1	30078019	DATE	METFORMIN			.	.	.
2	30121012	DATE	GLIPIZIDE			.	.	.
3	30149010	DATE	METFORMIN			.	.	.
4	30177026	DATE	GLIPIZIDE	METFORMIN		17162	17162	.
5	30180024	DATE	GLYBURIDE			.	.	.
6	30192012	DATE	METFORMIN			16821	.	.
7	30206025	DATE	GLIMEPIRIDE			.	.	.
8	30217015	DATE	GLYBURIDE			16833	.	.
9	30300013	DATE	METFORMIN			17043	.	.
10	30363015	DATE	GLIPIZIDE			.	.	.

Figure 7.13. Analysis Data in 2006

Next, we search for the first switching of the drugs and define the variable, STATUS. During the analysis, we make some assumptions:

- When we use an array statement, we assume the missing date to be the end of the year 2006 and also we convert it into a SAS date.
- If the drug is continued during the survival time, then it is censored and the value of STATUS is 0; otherwise, the value of STATUS is 1.
- If CHMED is equal to the drug, it means the drug is switched to another drug; in other words, it is not censored.
- Due to a lack of information, we set the start date equal to the beginning of the year 2006 and the end date to the end of the year 2006 if such information is unknown.
- We also define the value of STATUS as 0 when the survival time DAYS is equal to 364.
- We suppose the frequency of prescription for the year 2005 is at most 12.

The SAS code [26] is shown below.

```
DATA SASUSER.T06; SET SASUSER.MERGEDATA06;
 ARRAY MEDS(3) MED_1 - MED_3; ARRAY DATES(3) DATE_1 -
DATE_3;
 DO J=1 TO 3; IF DATES(J)=. THEN DATE='31DEC2006'D; END;
   DO I=1 TO 3;
   IF I=1 THEN TEMP=MEDS(I);
   IF MEDS(I) NE TEMP THEN DO;
  MED_NUM=I; DATE_NUM=DATES(I);  CHMED=MEDS(I);
  STATUS=1; I=3;
  END;
END;
/*Define 0-1 indicators and status*/
IF CHMED=' ' THEN STATUS=1;
```

```
IF CHMED='GLYBURIDE' THEN GLYBURIDE=0 AND STATUS=1;ELSE
GLYBURIDE=1; …
/*Define the variables days*/
IF DATE_1^=. THEN SDATE=DATE_1; ELSE SDATE='01JAN2006'D;
IF DATE_2^=. THEN EDATE=DATE_2; ELSE EDATE='31DEC2006'D;
FORMAT SDATE EDATE DATE9; DAYS=DATDIF
(SDATE,EDATE,'ACT/ACT');
IF DAYS=364 THEN STATUS=0; RUN;
```

Finally, we sort the new data by CHMED to get the data shown in Figure 7.14.

	STATUS	GLYBURIDE	METFORMIN	STARLIX	PRECOSE	INSULIN	SDATE	EDATE	DAYS
1	0	1	1	1	1	1	01JAN2006	31DEC2006	364
2	0	1	1	1	1	1	01JAN2006	31DEC2006	364
3	0	1	1	1	1	1	01JAN2006	31DEC2006	364
4	1	1	0	1	1	1	27DEC2006	27DEC2006	0
5	0	1	1	1	1	1	01JAN2006	31DEC2006	364
6	1	1	1	1	1	1	20JAN2006	31DEC2006	345
7	0	1	1	1	1	1	01JAN2006	31DEC2006	364
8	1	1	1	1	1	1	01FEB2006	31DEC2006	333
9	1	1	1	1	1	1	30AUG2006	31DEC2006	123
10	0	1	1	1	1	1	01JAN2006	31DEC2006	364

Figure 7.14. Survival Data for 2006

In 2005, in order to get an accurate conclusion, we filter out the beneficiaries in the office-based visits file rather than in the prescription drug file. We first sort out the enrollees whose Medicare payments are greater than 0 according to the ICD 9 diagnosis codes. Then we get the results shown in Figure 7.15; and we use the same method to get another data set about diabetes patients in the outpatient visit file.

	DUPERSID	OBICD1X	OBICD2X	OBICD3X	OBICD4X
1	30078019	250	-1	-1	-1
2	30121012	401	250	185	530
3	30180024	250	-1	-1	-1
4	30192012	250	-1	-1	-1
5	30201026	250	-1	-1	-1
6	30206025	250	401	-1	-1
7	30363015	401	250	272	716
8	30392041	250	-1	-1	-1
9	30450010	590	429	250	-1
10	30494013	401	250	716	530

Figure 7.15. Diabetes Patients in Office-based Visit

Finally, we use the SQL horizontal join to get all the diabetes beneficiaries and we use these patient IDs to find all Medicare drug plan enrollees. For analysis, we use the life table method, setting the interval at 10 days and stratifying the data by CHMED. The SAS code and some results are shown below.

```
PROC LIFETEST DATA=SASUSER.ST06 OUTSURV=SASUSER.GP06
ALPHA=0.05 METHOD=LIFE WIDTH=10;     STRATA CHMED;
TIME DAYS*STATUS(0); RUN;
```

Table 7.5 Summary of Censored/Uncensored Values for 2005

Summary of the Number of Censored and Uncensored Values					
Stratus	**CHMED**	**Total**	**Failed**	**Censored**	**Percent**
1	GLYBURIDE	4	4	0	0.00
2	GLYBURIDE_METF	4	0	4	100.00
3	INSULIN	8	3	5	62.50
4	METFORMIN	90	10	80	88.89
5	PRECOSE	3	1	2	66.67
6	ROSIGLITAZONE	1	0	1	100.00
7	TOLAZAMIDE	2	0	2	100.00
Total		112	18	94	83.93

Table 7.6 Summary of Censored/Uncensored Values for 2006

Summary of the Number of Censored and Uncensored Values					
Stratu	**CHMED**	**Total**	**Failed**	**Censored**	**Percent**
1	GLYBURIDE	2	0	2	100.00
2	INSULIN	7	0	7	100.00
3	METFORMIN	100	17	83	83.00
4	PIOGLITAZONE	2	2	0	0.00
5	PRANDIN	1	1	0	0.00
6	PRECOSE	5	2	3	60.00
7	STARLIX	3	1	2	66.67
Total		120	23	97	80.83

Results in Tables 7.5 and 7.6 show that the medications are divided into 7 groups in each year by CHMED. In 2005, since the number of prescriptions of rosiglitazone is one, we do not include it. The censored percentages of glyburide- metformin and tolazamide are 100 %, which means that it is hard for the patients to change such medicines once they begin taking them. In 2006, we also discard prandin due to one-time use. The censored rates of glyburide and insulin are 100 %, and the rate of metformin use is 83 %; all of these outcomes demonstrate that the three drugs can seldom be replaced by other medicines. In summary, the metformin and insulin uses are stable in both years. Glyburide itself is unstable in 2005, but stable in 2006. Moreover, the average censored rate in 2005 is a little higher than that in 2006, indicating that the usage of prescribed drugs is more stable in 2005.

Next, we estimate the differences of survival cases among various drugs by survival functions. The survival distribution function (SDF) in 2005 (displayed in Figure 7.16) demonstrates that none of the drug, tolazamide, is switched to the other medicines throughout the whole year. The survival rate of metformin decreases little by little from

100 % to 89 % at the end of the year. During the three periods, 30^{th} – 40^{th}, 120^{th} – 130^{th} and 280^{th} – 290^{th} days, the prescriptions of insulin largely decrease, but in the other time periods, they remain unchanged. The sharp decrease of precose use appears between the 190^{th} day and the 200^{th} day, but before and after that period, the usage is stable. A large number of beneficiaries switch their drugs from glyburide to the other medicines during the following periods, the 40^{th} day – 50^{th} day, the 90^{th} – 100^{th} day and 110^{th} – 120^{th} day, which means that the survival rate of the drug decreases to 20 percent at the end of the year. Therefore, the glyburide usage is very unstable in 2005. The SDF in 2006 (shown in Figure 7.17) shows that insulin and metformin survive longer than the other drugs since the survival rates are higher than that of any other drug throughout the year. None of prescriptions of insulin are changed to another medicine until the end of the year. Only less than 14 % of the prescriptions of metformin are switched to other drugs. Between the 90^{th} day and the 110^{th} day, large quantities of prescriptions of pioglitazone are changed to other drugs; however, after that, no more changes happen. The survival rate of precose goes down to 80 % around the 220^{th} day, to 60 percent around the 280^{th} day and then remains unchanged until the end of the year. The survival rate of starlix sharply decreases on the 320^{th} day and then stabilizes. In general, metformin and insulin uses are more stable than those of the other medicines.

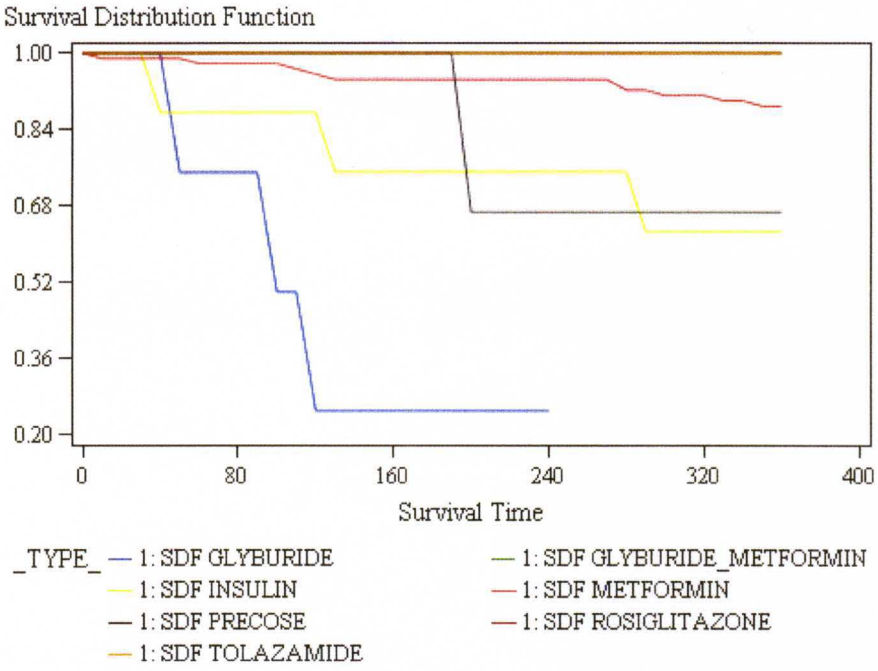


Figure 7.16. Survival Distribution Function for the Year 2005

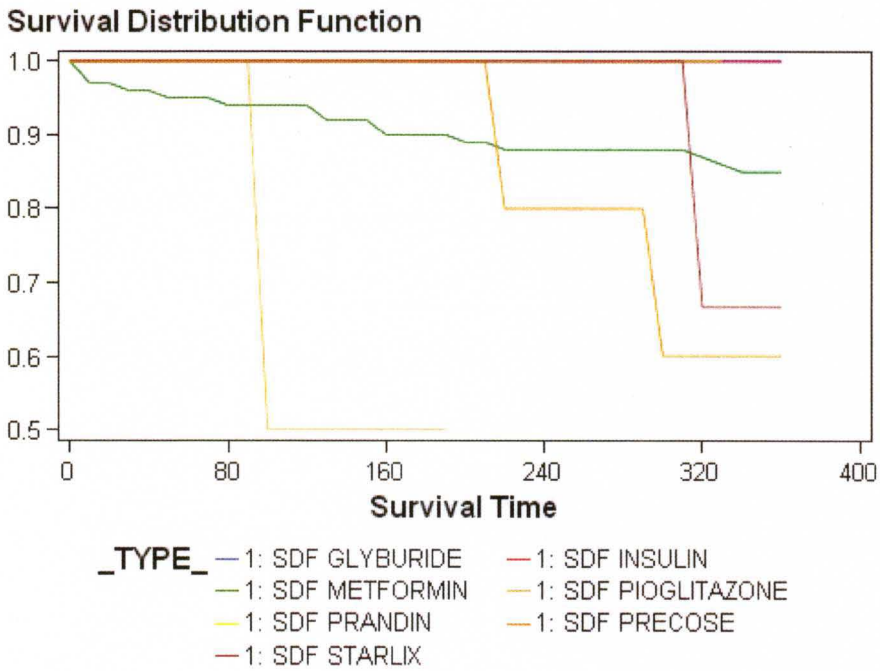


Figure 7.17. Survival Distribution Function for the Year 2006

7.3 Effects of Drug Plan on the Cost Effectiveness of Medications and Health Outcomes

The purpose of this section is to estimate the cost effectiveness of diabetes medications and the health outcomes in Medicare in 2005 and 2006 to examine the impact of Medicare, Part D using the data sets from MEPS [19].

7.3.1 Cost Effectiveness Analysis

First, we need to discover our research subjects who are Medicare drug plan enrollees in 2006 and also joined Medicare in 2005. To keep consistency, we do not consider the Medicare beneficiaries of age 65 in 2006; we also do not consider the patients who switch their drugs from the year 2005 to the year 2006. We use SAS SQL conditional selection to sort out the patients with diabetes according to the ICD9 condition code. Then we use an SQL inner join to combine the full year consolidated data file and prescription drug file by the ID variable, DUPERSID. After we find out the beneficiaries who joined in Medicare, Part D, we use the DUPERSID to match the patients in 2005. Then we get a dataset that we need. After importing a life table, we can combine it with the newly-generated data and calculate ICER1 displayed in Figure 7.18 with the code shown below.

```
/*Combine the life table and 2006 Medicare part D
beneficiary table */
PROC SQL;
CREATE TABLE SASUSER.LE06 AS
SELECT *
FROM SASUSER.LIFETABLE1 AS LT,
SASUSER.BCHWLQ06 AS BC
WHERE LT.AGE=BC.AGE06X; QUIT;
/*Calculate the 2006 QALY for different genders*/
```

```
DATA SASUSER.QALY06; SET SASUSER.LE06;
IF SEX=1 THEN QALY06=MALE*LQ06;
IF SEX=2 THEN QALY06=FEMALE*LQ06; RUN;
PROC SORT DATA=SASUSER.QALY06; BY DUPERSID; RUN;
/*To calculate the ICER */
PROC SQL; CREATE TABLE SASUSER.CICER AS
SELECT DUPERSID, SRXNAME, QALY05, QALY06, TOTMCR05,TOTMCR06,
((TOTMCR06 - TOTMCR05) / (QALY06-QALY05)) AS ICER1
FROM SASUSER.ICER ;  QUIT;
```

	DUPERSID	SRXNAME	QALY05	QALY06	TOTMCR05	TOTMCR06	ICER1
1	30121012	GLIPIZIDE	4.184	3.956	766	76057	-330223.68
2	30136026	GLIPIZIDE	6.8	9.768	0	11084	3734.50135
3	30180024	GLYBURIDE	4.42	4.184	6974	1947	21300.8475
4	30363015	GLIPIZIDE	3.956	3.736	710	4578	-17581.818
5	30386013	GLIPIZIDE	8.934	11.376	0	1782	729.72973
6	30392041	GLYBURIDE	2.71	5.164	2120	6140	1638.14181
7	30437028	GLYBURIDE	1.336	4.008	224	2434	827.095808
8	30489011	STARLIX	11.024	7.86	0	3952	-1249.0518
9	30507010	GLYBURIDE_M...	5.788	5.512	149	783	-2297.1014
10	30516018	GLYBURIDE	7.888	7.576	179	2737	-8198.7179

Figure 7.18. ICER Table

Table 7.7 ICER by Different Diabetes Drugs

SRXNAME	Mean	N
GLIMEPIRIDE	-1268.78	4
GLIPIZIDE	-12573.77	49
GLYBURIDE	-4728.05	45
GLYBURIDE_METFORMIN	1934.47	10
INSULIN	-14203.62	16
METFORMIN	896.1215818	44
STARLIX	-590.0887749	3

Once we get the table, we use the proc means procedure in base SAS to get the average ICER value for each drug shown in Table 7.7. Here, a negative ICER means that

there are savings for the year 2006 over the year 2005. For example, the comparison between the year 2006 and the year 2005 for insulin treatment shows a cost saving of $14,203.62 in 2006. For a positive ICER, the bigger the ICER, the less efficient the new method. Therefore, Table7.7 demonstrates that insulin becomes the most cost-effective in 2006, while Glyburide-metformin is the most inefficient treatment, and metformin is a close second.

7.3.2 Medical Resources Utilization

Next, we evaluate utilizations of healthcare resources by comparing the frequencies of office-based visits, outpatient visits and times of prescription drugs filled as well as the length of stay in the hospital or home health providers separately. We firstly find the data containing the times of office-based visits in these two years, then we use the times of visits in 2005 as the denominator; the difference of the times in these two years is used as numerator to calculate the increasing or decreasing rate. Finally, we get the average rates for each drug. In the same way, we also get the increasing and decreasing rates in the other cases.

Table 7.8 shows that compared to the year 2005, the Medicare diabetes patients receive more drug treatments in 2006 since the drug refill rates increase by an amount varying from 17% to 66%. At the same time, the average LOS (length of stay) in the hospital of the insulin or glipizide users is largely decreased by 80% or 61%, which means that adequate insulin or glipizide usage saves considerable hospitalization resources. However, the average of the prescription frequency and LOS of metformin users increases by 73% and 200 % respectively from the year 2005 to the year 2006. It is also true for glyburide users. In other words, the drug plan makes these two drug

treatments more inefficient. The relationship between the LOS in the hospital and the home health provider for most drug users is negative; the longer the stay using home health providers, the shorter the stay in the hospital. Considering the costs of hospitalization are higher than those of home health providers, the patients should sufficiently utilize the home health agency services.

Table 7.8 Ratios in Utilizations of Healthcare Resources

SRXNAME	OBTRATIO	OPTRATIO	RXTRATIO	LOSRATIO	HHDRATO
GLIMEPIRIDE	0.03	0	0.66	.	.
GLIPIZIDE	0.52	0.19	0.22	-0.61	0.50
GLYBURIDE	1.30	0.82	0.33	0.65	-0.01
GLYBURIDE _METFORMIN	-0.12	0.80	0.20	.	-1.0
INSULIN	0.01	-0.29	0.17	-0.8	-0.07
METFORMIN	0.54	0.29	0.73	2.0	-0.40
STARLIX	0.61	.	0.45	.	.

7.3.3 Health Status

Finally, we use the decision tree model to investigate which factors have vital effects on the beneficiaries' health status. We input all the variables, frequencies of physician visits, drug prescriptions, A1C tests, ER (Emergency Room), LOS in the hospital or home healthcare agency, gender, age and family size. We set the health status as a predicted target. Figure7.19 demonstrates that in 2005, the frequency of A1C tests and the physician visits have vital effects on a patients' health. Figure 7.20 indicates that in 2006, the frequency of drugs filled becomes a key factor to the patient's health status. However, there is something in common between these two years. LOS in the hospital and family size are not important factors to health conditions.

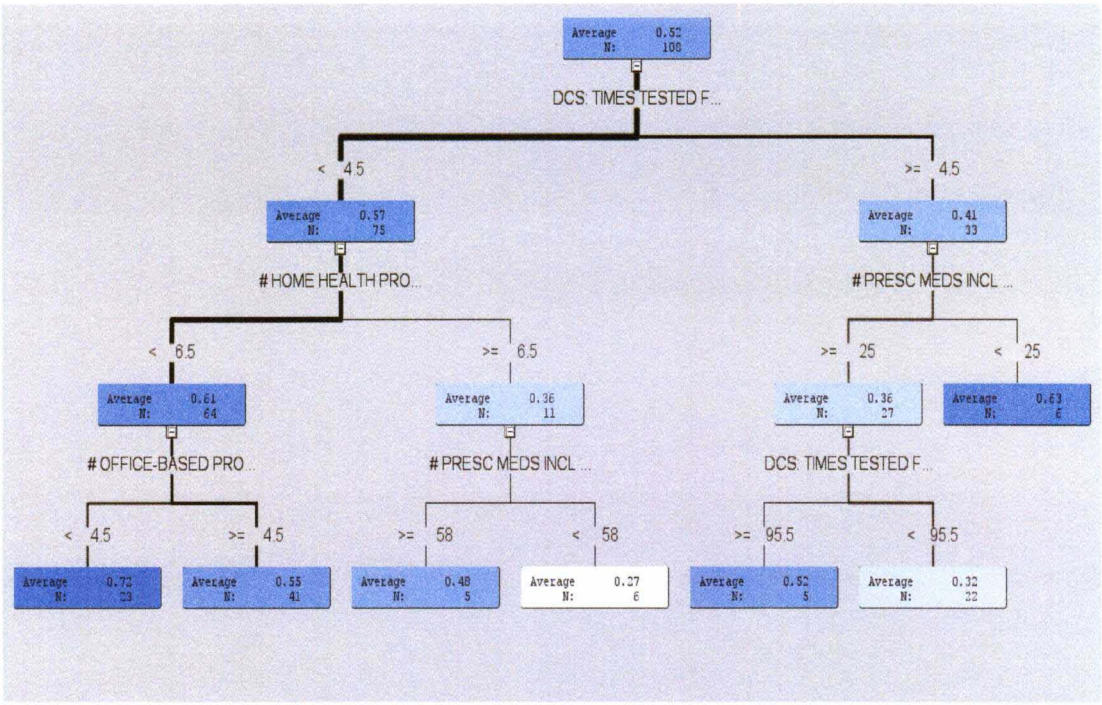


Figure 7.19. Decision Tree for 2005 Health Status

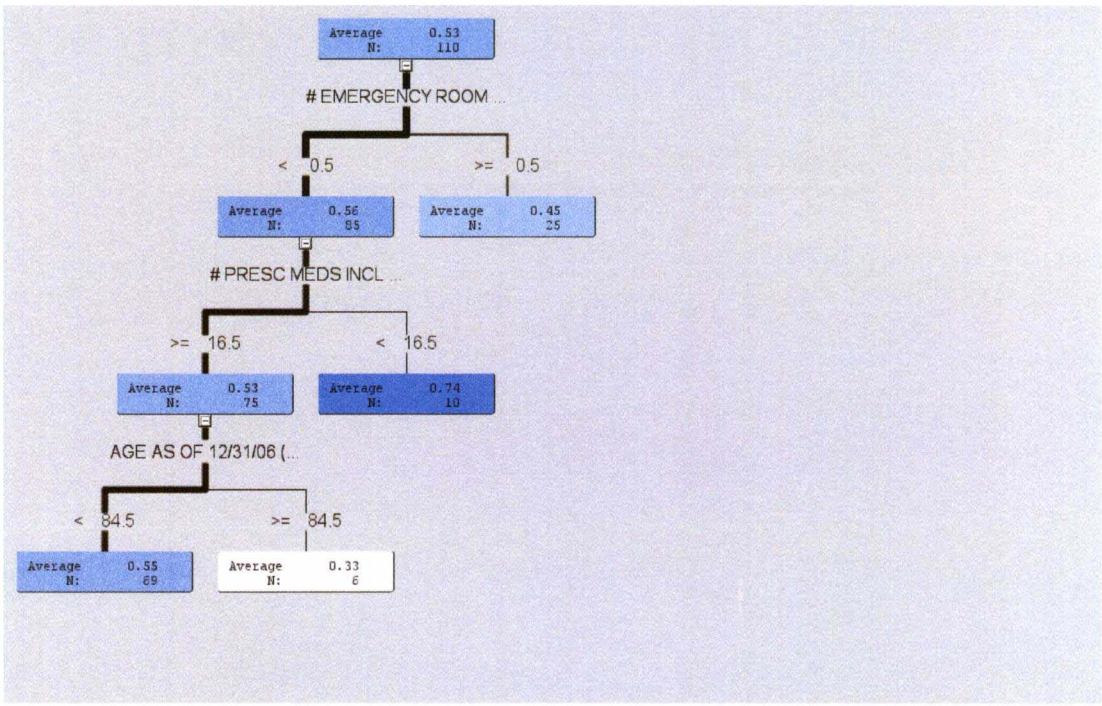


Figure 7.20. Decision Tree for 2006 Health Status

7.4 Conclusion

Based on the analyses in Section 7.2, we can draw the conclusion that Medicare, part D indeed greatly increases the expenditures of Medicare on diabetes medications. The prescription drug plan itself reduces the choices of the medicines for diabetes for each year. We also discover that generally, the usage of insulin and metformin is always more stable than that of other drugs. However, glyburide usage is very unstable in 2005 but stable in 2006. In addition, more drugs are switched into other medicines in 2006, which indicates that the use of drugs is less stable than that in 2005. It is also discovered that in 2005, female patients spend more on insulin and supplies, while the males spend more on metformin. In 2006, the female beneficiaries pay more for metformin.

Cost-effectiveness analysis suggests that Medicare, Part D makes the insulin treatment the most efficient, while the combination of glyburide – metformin is the least effective. Our results also demonstrate that under this drug plan, the Medicare beneficiaries can receive more sufficient drug treatments than ever before. In the meanwhile, enough usage of some drugs such as insulin can decrease the usage of hospital resources. In contrast, the metformin users stay in the hospital for a longer time in 2006. Another discovery is that using the drugs properly can improve the patient health status.

CHAPTER VIII

CONCLUSION

In this dissertation, we introduced pattern recognition analysis and supervised/ unsupervised machine learning algorithms as well various kinds of linear statistical models into the study of diabetes patients in the Medicare population and we proposed several methods to decrease Medicare expenditures while improving healthcare quality.

First, we utilized the linear statistical models to complete a cost analysis of diabetes outpatients in Medicare. We conclude that most predictors that we find are key factors influencing the costs and we suggest that outpatients with diabetes should often monitor their blood glucose level and the patients with such complications as anemia and heart diseases need more medical care. By these means, the costs can be decreased to some extent.

Second, we applied supervised learning approaches such as the decision tree model and market basket analysis to an outcomes study of inpatients. We arrive at the conclusions that the patients who have procedures related to heart diseases often need other procedures; neither age nor the end-stage renal disease is the key factor to mortality, which is contrary to widely held belief.

Third, we employed pattern recognition analysis and the decision tree model in readmission risk analysis. Patients with cardiovascular or kidney co-morbidities have a higher risk of readmission than those with other diseases. Another discovery is that the

patients without professional medical services are more likely to be readmitted to the hospitals after discharge.

Fourth, we performed two- way interaction effects analysis using the generalized linear model with a gamma distribution, the logistic regression model and the Poisson regression model. We find that as for inpatients expenditures, most two –way interaction effects between diabetes complications are significant; when the response variable is frequency of hospitalization, only the effects between kidney disease and the other diseases are significant to the Poisson regression model. We also conclude that the patients who have both heart disease and kidney disease have a much higher risk of dying since the interaction effect between these two diseases are significant to mortality in the logistic regression model.

Fifth, we used survival analysis to analyze diabetes drug usage. Results demonstrate that insulin and metformin are more stable than other drugs in terms of usage and the uses of drugs in 2005 are more stable than those in 2006 with the influence of Medicare, Part D.

Finally, we applied cost-effectiveness analysis to diabetes medications. The results indicate that with the introduction of the Medicare drug plan, insulin becomes the most cost-effective treatment and the combination of glyburide and metformin is the most inefficient. They also demonstrate that metformin users highly increase their length of hospitalization and the frequency of prescriptions from the year 2005 to the year 2006.

In summary, in this dissertation, several new algorithms and methods are introduced or improved and some suggestions are proposed to decrease Medicare costs and improve health outcomes.

REFERENCES

1. Li, S., et al., *Economic effect of following HbA1c testing practice guidelines in the elderly Medicare population: an instrumental variable analysis.* Am J Med Qual, 2010. **25**(3): p. 202-10.

2. Bhattacharyya, S.K. and B.A. Else, *Medical costs of managed care in patients with type 2 diabetes mellitus.* Clinical Therapeutics, 1999. **21**(12): p. 2131-42.

3. Herrin, J., et al., *Cost and effects of performance feedback and nurse case management for medicare beneficiaries with diabetes: a randomized controlled trial.* DISEASE MANAGEMENT, 2007. **10**(6): p. 328-36.

4. Kuo, S., et al., *Trends in Care Practices and Outcomes Among Medicare Beneficiaries with Diabetes.* American Journal of Preventive Medicine, 2005. **29**(5): p. 396-403.

5. McBean, A.M., K. Jung, and B.A. Virnig, *Improved Care and Outcomes Among Elderly Medicare Managed Care Beneficiaries With Diabetes.* THE AMERICAN JOURNAL OF MANAGED CARE, 2005. **11**(4): p. 213-222.

6. Karaca, Z., et al. *The Impact of Medicare Part D on Beneficiaries with Type 2 Diabetes /Drug Utilization and Out-of-Pocket Costs.* 2008 October,2010; Available from: http://www.avalerehealth.net/research/docs/The_Impact_of_Medicare_Part_D_Diabetes_Takeda.pdf.

7. Schmittdiel, J.A., et al., *Patient-provider communication regarding drug costsin Medicare Part D beneficiaries with diabetes: a TRIAD Study.* BMC Health Services Research, 2010. **10**.

8. Centers for Disease Control and Prevention, *National Diabetes Fact Sheet, 2007,* 2007.

9. American Diabetes Association. *Diabetes Basics.* [cited 2010 October 22]; Available from: http://www.diabetes.org/diabetes-basics/.

10. American Diabetes Association. *Diabetes Statistics.* 2010 [cited 2010 October 26]; Available from: http://www.diabetes.org/diabetes-basics/diabetes-statistics/.

11. National Institute of Diabetes and Digestive and Kidney Diseases. *National Diabetes Statistics, 2007.* 2008; Available from: http://diabetes.niddk.nih.gov/dm/pubs/statistics/DM_Statistics.pdf.

12. United States Renal Data System. *USRDS 2007 Annual Data Report.* 2007; Available from: http://www.usrds.org/atlas_2007.htm.

13. American Diabetes Association, *Economic Costs of Diabetes in the U.S.in 2007.* Diabetes Care, 2008. **31**(3).

14. Nathan, D.M., et al., *Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes.* N Engl J Med, 2005. **353**(25): p. 2643-53.

15. Silverberg, A.B. and K.P. Ligaray, *Oral Diabetic Medications and the Geriatric Patient.* Clinic in Geriatric Medicine, 2008. **24**(3): p. 541-549.

16. Medicare Payment Advisory Commision, *A Status Report on Part D for 2009 in Report to the Congress: Medicare Payment Policy*, 2009: Washington, DC.

17. Bihari, M. *Understanding the Medicare Part D Donut Hole :Learn About the Medicare Part D Coverage Gap.* 2010 [cited 2010 August,2010]; Available from: http://healthinsurance.about.com/od/medicare/a/understanding_part_d.htm.

18. Centers for Medicare and Medicaid Services *Chronic Condition Data Warehouse Data [Data File]* 2004.

19. Agency for Healthcare Research and Quality. *Medical Expenditures Panel Survey data [Data File].* 2005 & 2006; Available from: http://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp.

20. Centers for Medicare and Medicaid Services. *Healthcare Common Procedure Coding System (HCPCS) LevelL II Coding Procedures.* 2010; Available from: https://www.cms.gov/MedHCPCSGenInfo/Downloads/LevelIICodingProcedures.pdf.

21. Torrey, T. *What Are CPT Codes? Do CPT Codes Affect Your Healthcare?* 2011 [cited 2011 Janunary 15, 2011]; Available from: http://patients.about.com/od/costsconsumerism/a/cptcodes.htm.

22. Centers for Disease Control and Prevention. *International Classification of Diseases,Ninth Revision, Clinical Modification (ICD-9-CM).* 2009 [cited 2009 October 1st]; Available from: http://www.cdc.gov/nchs/icd/icd9cm.htm.

23. ICD9.Chrisendres.com. *DISEASES OF OTHER ENDOCRINE GLANDS (249-259).* 2009; Available from: http://icd9cm.chrisendres.com/index.php?action=child&recordid=1894.

24. Nisbet, R., J. Elder, and G. Miner, *Handbook of Statistical Analysis & Data Mining Applications* 2009. Burlington: Elsevier, Inc.

25. Refaat, M., *Data Preparation for Data Mining Using SAS*, ed. J. Gary2007, San Francisco: Morgan Kaufmann Publishers.

26. Cerrito, P., *Data Mining Healthcare and Clinical Databases*2010: Data Services Online.

27. BAKER, R.C. *Chapter 5: Data Collection and Sampling [PPT]*. Available from: wweb.uta.edu/insyopma//baker/STATISTICS/Keller7/...7/Chapter05.ppt.

28. Cerrito, P. *The Problem of Regression Assumptions and the Use of Predictive Modeling*. in *SAS Global Forum 2009* 2009. Washington, D.C.

29. Cerrito, P., *Introduction to data mining using SAS Enterprise Miner*2006, Cary, NC: SAS publishing.

30. Muller, K.E. and P.W. Stewart, *Linear Model Theory: Univariate, Multivariate and Mixed Models*. Wiley Series in Probability and Statistics2006, Hoboken, New Jersey: John Wiley & Sons, Inc.

31. Wikipedia. *Durbin–Watson statistic*. 2010 December 25; Available from: http://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic.

32. NIST/SEMATECH, *Anderson-Darling and Shapiro-Wilk tests*, in *Handbook of Statistical Methods*2010.

33. Wikipedia. *Kolmogorov–Smirnov test*. 2010 December 9; Available from: http://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test#Kolmogorov.E2.80.93Smirnov_statistic.

34. Wikipedia. *Autocorrelation*. 2010 [cited 2011 January 26]; Available from: http://en.wikipedia.org/wiki/Autocorrelation.

35. Yaffee, R.A. *Regression Analysis with SAS[Powerpoint slides]*. Available from: www.nyu.edu/its/socsci/Docs/SASREG.ppt.

36. Habing, B. *More on Outlier Diagnostics Supplement to Section 8.9 [Lecture Note]*. 2004 July, 2004; Available from: http://www.stat.sc.edu/curricula/courses/516/516s8p9sup.pdf.

37. Li, B. *Statistics 512: Applied Linear Models:Topic 1[Lecture Note]*. 2008; Available from: http://www.stat.purdue.edu/~boli/stat512/lectures/topic1.pdf.

38. Brinkley, J. *SAS Logistic Regression[Lecture Note].* 2009 Feburary, 2009; Available from: http://www.ecu.edu/cs-dhs/bios/upload/Logistic.pdf.

39. King, G. and L. Zeng, *Logistic Regression in Rare Events Data.* Political Analysis, 2001. **9**: p. 137-163.

40. SAS institute Inc., *SAS/STAT(R) 9.22 User's Guide*2010, Cary, NC, USA: SAS publishing.

41. Li, B. *Statistics 512: Applied Linear Models:Topic 6 [Lecture Note].* 2008 [cited 2010 October 23rd]; Available from: http://www.stat.purdue.edu/~boli/stat512/lectures/topic6.pdf.

42. Li, B. *Statistics 512: Applied Linear Models:Topic 7[Lecture Note].* 2008 [cited 2010 November 5th]; Available from: http://www.stat.purdue.edu/~boli/stat512/lectures/topic7.pdf.

43. Johnston, G. *SAS Software to Fit the Generalized Linear Model* in *SUGI 18.* 1993. New York, USA.

44. Liang, K.-Y. and S. Zeger, *Longitudinal data analysis using generalized linear models.* Biometrika, 1986. **73**(1): p. 13-22.

45. McCullagh, P. and J.A. Nelder, *Generalized Linear Model* 1989: Chapman and Hall.

46. Berk, R. and J. MacDonald, *Overdispersion and Poisson Regression.* Journal of Quantitative Criminology, 2008. **24**(3): p. 269-284.

47. Littell, R.C., W.W. Stroup, and R.J. Freund, *SAS for Linear Models.* 4 ed 2002, Cary, N.C. USA: SAS publishing.

48. Schabenberger, O. *Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models.* in *SUGI 30.* 2005. Philadelphia, Pennsylvania

49. Hill, T. and P. Lewicki, *Statistics: Methods and Applications A comprehensive references for science, industry and data mining*2006, Tulsa, ok: Statsoft.Inc.

50. Anonymous. *DATA MINING.* [cited 2010 November 7]; Available from: http://dataminingarticles.com/.

51. Yuille, A.L. *Detection and Estimation Theory [Lecture Note].* [cited 2010 November 10]; Available from: http://www.stat.ucla.edu/~yuille/courses/Stat153/EMtutorial.pdf.

52. SAS Institute Inc., *SAS Enterprise Miner6.2 help document* 2010.

53. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning:Data Mining, Inference, and Prediction.* 2nd ed 2008, New York, USA: Springer Science+ Business Media, LLC.

54. Anonymous. *CHAID and Exhaustive CHAID Algorithms* [cited 2010 October 22]; Available from: support.spss.com/productsext/spss/.../algorithms/14.0/TREE-CHAID.pdf.

55. Zhang, P.G., *Neural Networks For Data Mining*, in *Data Mining and Knowledge Discovery Handbook* O. Maimon and L. Rokach, Editors. 2005, Springer Science+Business Media: NY, USA.

56. Giudici, P., *Data Mining Model Comparison*, in *Data Mining and Knowledge Discovery Handbook* O. Maimon and L. Rokach, Editors. 2005, Springer Science+Business Media: NY, USA.

57. Scott, D.W., *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley Series in Probability and Statistics1992, New York, USA: John Wiley & Sons,Inc.

58. SAS Institute Inc., *Base SAS(R) 9.2 Procedures Guide: Statistical Procedures.* 3rd ed2010, Cary, NC: SAS Institute Inc.

59. Cerrito, P., *Student Papers in Introductory Statistics for Mathematics Majors*: Data Services Online.

60. Atlantic Information Services, Inc., *CMS Targets Readmission through Payment, Audits; 'Coaching'Model Reduces Rates.* Report on Medicare Compliance, 2008. **17**.

61. Stevens, J.J. *Interaction Effects in ANOVA[Lecture Note].* [cited 2010 December 6]; Available from: http://pages.uoregon.edu/stevensj/interaction.pdf.

62. Allison, P.D., *Survival Analysis Using SAS: A Practical Guide*1995, NC, USA: SAS Publishing.

63. Muennig, P., *Cost-Effectiveness Analysis in Health: A Practical Approach.* 2nd ed2008, CA, USA: Jossey-Bass.

64. Phillips, C. *What is a QALY? [Lecture Note].* [cited 2010 Novemver 16]; Available from: http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/QALY.pdf.

65. The U.S. Social Security Administration. *Period Life Table.* 2006 May 6th, 2010]; Available from: http://www.socialsecurity.gov/OACT/STATS/table4c6.html.

66. Fox-Rushby, J. and D. Fidan, *The Structure of Economics Evaluation* in *Economic Evaluation*, J. Fox-Rushby and J. Cairns, Editors. 2005, Open University Press: NY, USA.

67. Walker, D. and A. Miners, *Basic Sensitivity Analysis* in *Economic Evaluation*, J. Fox-Rushby and J. Cairns, Editors. 2005, Open University Press: NY, USA.

68. Bureau of Labor Statistics, *Annual Average Indexes 2006*.

CURRICULUM VITAE

XIAO WANG

Department of Mathematics

University of Louisville

Louisville, KY, 40292

EDUCATION:

Ph.D.	Applied Mathematics, University of Louisville,	2006 – 2011
M.A.	Applied Mathematics, University of Louisville,	2006 – 2008
M.A.	Economics, Shanghai University of Finance & Economics,	2003 – 2006
B.A.	Finance, Qingdao University,	1998 – 2002

SKILLS:

SAS modules (SAS/BASE, SAS/ STAT, Enterprise Miner, Enterprise Guide, Text Miner)

MS office Suite (Word, Excel, Access and Power Point)

SQL

AWARDS:

Mid-west SAS User Group conference scholarship	Oct. 2010
SAS Student Ambassador	Apr. 2010
Mid-west SAS User Group conference scholarship	Oct. 2009
SAS Student Ambassador	Mar. 2009

PUBLICATIONS:

BOOK

1. Cerrito, PB, Wang, X. *Problems and Issues with Comparative Effectiveness Analysis.* Accepted by IGI Publishing, Hershey, PA. Tentative completion date: March, 2012.

BOOK CHAPTER

1. Wang, X. (2010). Analyzing the Relationship between Diagnosis and the Cost of Diabetic Patients. In Cerrito, P.(Ed.), *Cases on Health Outcomes and Clinical Data Mining: Studies and Frameworks.* : IGI Global.

CONFERENCE PAPERS (Published in Conference Proceedings)

1. Wang, X. (2010, October). *To Investigate the Impact of Medicare, Part D on the Cost- Effectiveness of Diabetes Medications and Health Outcomes with SAS.* Paper presented at MWSUG2010, Milwaukee, WI.
2. Wang, X. (2010, April). *Investigating the Impact of Medicare, Part D on the Diabetes Medications Using Enterprise Miner and Survival Analysis.* Paper presented at SAS Global Forum 2010, Seattle, WA.
3. Wang, X. (2009, October). *Outcome Research for Diabetic Inpatients with SAS Enterprise Miner 5.2.* Paper presented at MWSUG2009, Cleveland, OH.
4. Wang, X. (2009, March). *Using SAS Enterprise Guide 4.1 to Reduce the Cost of Diabetic Outpatients in Medicare.* Paper presented at SAS Global Forum 2009, Washington, DC.

PRESENTATIONS

1. Wang, X. (2011, May). *Two-way Interaction Effect Analysis of Diabetes Complications on Health Costs and Health Outcomes in Medicare Inpatients.* Poster session to present at ISPOR 16th Annual International Meeting, Baltimore, MD.
2. Wang, X. (2010, October). *To Score the Risk of Readmission of Diabetes Patients in Medicare with SAS.* Poster session presented at M2010 Data Mining Conference, Las Vegas, NV.
3. Wang, X. (2010, May). *Cost-effectiveness Analysis of the Diabetes Medications in Medicare with SAS.* Poster session presented at ISPOR 15th Annual International Meeting, Atlanta, GA.
4. Wang, X. (2009, October). *Investigating the Impact of Medicare, part D Using SAS and Enterprise Miner.* Poster session presented at M 2009 Data Mining Conference, Las Vegas, NV.

5. Wang, X. (2009, May). *Use of SAS to Analyze the Relationship between Diagnosis and the Cost of Diabetic Outpatients*. Poster session presented at ISPOR 14th Annual International Meeting, Orlando, FL.
6. Wang, X. (2008, October). *Using SAS Enterprise Guide 4.1 to Reduce the Cost of Diabetic Outpatients in Medicare*. Poster session presented at M2008 Data Mining Conference, Las Vegas, NV.