12-2011

Acute myocardial infarction patient data to assess healthcare utilization and treatments.

Pedro Ramos
University of Louisville

ACUTE MYOCARDIAL INFARCTION PATIENT DATA TO ASSESS HEALTHCARE UTILIZATION AND TREATMENTS

By

Pedro Ramos

B.S. Delta State University, 1999

M.S. Middle Tennessee State University, 2005

A Dissertation

Submitted to the Faculty of the

College of Arts and Sciences of the University of Louisville

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

Department of Mathematics

University of Louisville

Louisville, Kentucky

December 2011

MINING ACUTE MYOCARDIAL INFARCTION PATIENT DATA TO ASSESS HEALTHCARE UTILIZATION AND TREATMENTS

By

Pedro Ramos

B.S. Delta State University, 1999

M.S. Middle Tennessee State University, 2005

A dissertation approved on

November 18, 2011

By the following Dissertation Committee

Dr. Patricia Cerrito (Director)

Dr. Ryan Gill (Co-Director)

Dr. Kiseop Lee

Dr. Jiaxu Li

Dr. Ibrahim Imam

Dr. Thomas Riedel

ABSTRACT

MINING ACUTE MYOCARDIAL INFARCTION PATIENT DATA TO ASSESS HEALTHCARE UTILIZATION AND TREATMENTS

Pedro Ramos

November 18, 2011

The goal of this study is to use a data mining framework to assess the three main treatments for acute myocardial infarction: thrombolytic therapy, percutaneous coronary intervention (percutaneous angioplasty), and coronary artery bypass surgery. The need for a data mining framework in this study arises because of the use of real world data rather than highly clean and homogenous data found in most clinical trials and epidemiological studies. The assessment is based on determining a profile of patients undergoing an episode of acute myocardial infarction, determine resource utilization by treatment, and creating a model that predicts each treatment resource utilization and cost.

Text Mining is used to find a subset of input attributes that characterize subjects who undergo the different treatments for acute myocardial infarction as well as distinct resource utilization profiles. Classical statistical methods are used to evaluate the results of text clustering. The features selected by supervised learning are used to build predictive models for resource utilization

and are compared with those features selected by traditional statistical methods for a predictive model with the same outcome. Sequence analysis is used to determine the sequence of treatment of acute myocardial infarction. The resulting sequence is used to construct a probability tree that defines the basis for cost effectiveness analysis that compares acute myocardial infarction treatments. To determine effectiveness, survival analysis methodology is implemented to assess the occurrence of death during the hospitalization, the likelihood of a repeated episode of acute myocardial infarction, and the length of time between reoccurrence of an episode of acute myocardial infarction or the occurrence of death.

The complexity of this study was mainly based on the data source used: administrative data from insurance claims. Such data source was not originally designed for the study of health outcomes or health resource utilization. However, by transforming record tables from many-to-many relations to one-to-one relations, they became useful in tracking the evolution of disease and disease outcomes. Also, by transforming tables from a wide-format to a long-format, the records became analyzable by many data mining algorithms. Moreover, this study contributed to field of applied mathematics and public health by implementing a sequence analysis on consecutive procedures to determine the sequence of events that describe the evolution of a hospitalization for acute myocardial infarction. This same data transformation and algorithm can be used in the study of rare diseases whose evolution is not well understood.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE
PAGE

CHAPTER 1

Motivation

This study has the objective of investigating the various treatments for emergent myocardial infarction based on claims databases. During the treatment of emergent myocardial infarction, different procedures are required to stabilize the patient, and these procedures have various outcomes associated with them. Part of this investigation examines the sequence of procedures involved in the treatment of myocardial infarction and their corresponding outcomes. In addition, lots of variability exists in the cost of treatment of emergent myocardial infarction; this dissertation aims to build cost predictive models and resource utilization models of acceptable accuracy. Since various treatments have been developed and perfected in the past decade, the effectiveness of these treatments will be assessed. To comprehensively study these treatments, an economic evaluation thereof will be conducted in the form of a cost-effectiveness analysis.

Claims databases and patient registries can be considered retrospective observational studies with content about a patient's health conditions, particular disease, and receipt of a particular treatment. They can be used for understanding natural history, assessing quality of care, provider performance,

and assessing cost-effectiveness. Claims databases involve a diverse group of patients in contrast to clinical trials that seek uniformity in the participating subjects. Therefore, these databases better reflect real-world management practices and outcomes than randomized control trials. Indeed, their information can build on clinical trial data to update decision models and cost-effectiveness models to provide a more realistic picture of a method/intervention/product's value.

Whereas clear guidelines to conduct, analyze, and report on clinical trials and cohort studies exist, similar guidelines to design and conduct studies based on registries and claims databases have not been developed. Without the structure of a clinical trial or cohort protocol, a daunting number of challenges to analyze and assess *real-world* data remains. For one, clinical trials descriptive analyses are based on standard assessment tools and fixed assessment times, and rely on randomization to allow group comparisons free of bias due to patient selection and other factors. Cohort studies, which capture time related changes, predetermine a fixed set of confounders to be measured in order to adjust the causal-effect analysis and control for bias. However, these methodologies fail to capture the complexity of the evolving real world and confounders that are not considered in the analysis. Dynamically complex problems often involve long delays between cause and effect as well as multiple goals and interests that may conflict, in various ways, with each other. Although such methods are successful in some areas of the health enterprise in the United States — and the world — they are confined to solving a problem without fully considering *spilling* effects.

In other words, this approach addresses problems in a piece-meal fashion. A short excerpt of Utopia by Sir Thomas More better describes this situation concisely [1]:

> *"By applying one remedy to one sore, you will provoke another; and that which removes the one ill symptom produces others, whereas the strengthening one part weakens the rest."*

In this dissertation, I propose a data mining framework, in the analysis of claims databases that encompasses knowledge discovery, inference, and prediction to develop decision models to guide private and public health policy decision makers. To empirically examine their feasibility, claims databases will be examined with respect to outcomes, resource utilization, cost of illness, and disparities in myocardial infarction patients.

First, classical methods such as literature reviews and possible meta-analysis techniques should be considered in the exploratory phase in order to account for known confounders and avoid *reinventing the wheel.* In addition, exploratory data mining methods should be implemented to determine unknown confounders and possibly new outcomes. If possible, linkage among disparate databases should be implemented. Secondly, if temporal information is found in the database, this should be explored and analyzed to account for risk factors, diseases, and health resources that are in a continuous state of interaction and continuous change. Similarly, spatial features, if available, should be data mined to understand the space-dependency of the various confounders. This should be followed by the construction of influential diagrams that incorporate time and space. Data modeling and simulation should be used to incorporate robustness

in inference and to implement a *what-if* analysis that is able to reproduce historical patterns, but also able to generate insights. To explore possible futures, predictive modeling is paramount. Diagnostics, sensibility analysis, and performance measurements must be conducted.

Myocardial Infarction

The heart is the most active muscle in the body and needs a constant supply of blood, which the cardiac conducting system controls by transporting electrical impulses through the cardiac muscle to trigger a series of contractions [2]. The impulse of each heartbeat starts in the sinoatrial node, and it flows quickly through the atria and causes them to contract — atrial systole. Electricity does not pass directly between the atria and ventricles; instead, it is channeled into the atrioventricular node where it is delayed slightly. This ensures that the atrial contraction has finished before the ventricles start to contract. Although the heart chambers are always full of blood, the blood cannot reach all of the cells of its walls, so the heart has its own blood vessels, the coronary circulation. The coronary arteries that supply the heart are forced to close under the pressure of the contracting muscle. They therefore can only fill when the heart is relaxed during diastole.

A heart attack or myocardial infarction (MI) occurs when blood supply to the heart is disrupted by an occlusion in one or more of the coronary arteries. This deprivation of blood to the heart muscle causes damage or possibly death to the heart's tissues known as myocardium [3].It is been well established that the longer the heart is deprived of blood, more of the heart muscle is damaged and

killed. The axiom cardiologists have is that time saved is heart saved and thus establishing the need to open the coronary arteries occluded as quickly as possible [4].

Some of the symptoms of a heart attack are chest pain or angina, shortness of breath, profuse sweating, a burning sensation in the esophagus, and radiating pain to the arms and legs, especially on the left side of the body. Individuals who have previously been diagnosed with angina pectoris are aware of these symptoms and know that they must manage the pain with nitroglycerin and aspirin [4-6]. For this group of individuals, it is not recommended that they seek medical attention unless the pain does not subside within five minutes. If this is the case, then the individuals should visit an emergency department to alleviate pain and to receive treatment.

STEMI

One of the least predictable and most severe heart attacks is classified as ST, elevated myocardial infarction (STEMI). It is caused by sudden clots, known as thrombotic occlusions, in the coronary arteries that had not experienced any narrowing previously [7]. They are indicated by the electrocardiogram (ECG), which is performed upon admission to the emergency department, when the ECG displays an abnormal elevation in the "ST segment" of the electrical heart wave. STEMI is considered the most severe type of heart attack because it is caused by a complete occlusion of one of the coronary arteries. The less blood that flows into the heart and the longer the diminished flow lasts, the greater the damage to the heart muscle (myocardium) and the less likely the patient will

recuperate. For individual who have not had a history of angina pectoris, it is essential that they seek medical attention as the maxim cardiologists follow is that time saved is heart saved.

Diagnosis of a MI

Once a patient is at the emergency department, emergency personnel will follow a well established and studied algorithm to determine if the individual is experiencing an MI. Nonetheless, emergency departments routinely begin treatment for an MI, believing a false-positive (treating an MI while the patient truly has, for example, heart burn) is a less severe mistake to make than a false-negative (treat for heart burn while the patient is experiencing an MI) [8]. The patient (and/or their relatives) will be asked several questions concerning the patient's medical history to determine the risk of a heart attack. In addition, a series of diagnostic tests will be conducted immediately upon arrival to the hospital: electrocardiogram, blood test, and echocardiogram. The electrocardiogram (ECG) and echocardiogram (ECHO) are usually performed in tandem. During an electrocardiogram, electrodes are positioned on the chest and limbs in such a way that electrical currents in all areas of the heart can be monitored. The recording displays the voltage between pairs of electrodes. In a typical ECG, each heartbeat produces three distinctive waves, P, QRS, and T that show a regular beat. Electrical activity in the sinoatrial node instigates atrial systole. The P-wave represents the spreading of electrical impulses from the sinoatrial node, through the atria, to the atrioventricular node. Then, electrical activity, represented by the QRS-wave, continues from the atrioventricular node

through the ventricles to produce ventricular contraction. Finally, the electrical impulse recedes as the heart resets itself, represented by T-wave, and both the atria and ventricles relax completely.

The ECG can also identify the site of any damage that disturbs the flow of electricity, as the waves will form an unusual pattern. While an electrocardiogram assesses the electrical activity of the heart, an echocardiogram uses ultrasound to produce images of the heart structures. The benefits of these tests will be discussed below, but the results of these tests allow cardiologists to determine the exact type of MI a patient is experiencing (e.g., STEMI). The most definitive test to establish if an MI is occurring is the confirmatory blood tests. The human body only produces changes in the levels of the enzymes, troponin and creatine kinase, if the heart muscle has recently been damaged. Therefore, if these enzymes show in the blood test, it is a very definitive indication that an MI has occurred [8]. There are many reasons for the presence of creatine kinase, including kidney failure. Troponin type TnI and type TnT are produced ONLY in the heart. They are normally present in the blood. When a heart attack occurs, their levels spike within three hours. They tend to remained elevated for several days thereafter:

- Type 1: spontaneous and related to ischemia due to a coronary event.
- Type 2: secondary to ischemia due to an imbalance between oxygen demand and supply.

- Type 3: sudden cardiac death with symptoms of ischemia and new ST elevation or LBBD, verified thrombus by angiography, or autopsy.
- Type 4a: associated with previously performed PCI.
- Type 4b: associated with verified stent thrombosis.
- Type 5: associated with CABG.

However, for practical purposes, and in clinical settings, an MI is classified based on results from an electrocardiographic diagnostic test that assesses the heart's various electrical waves to determine the precise site of the infarction[6, 9]:

- Inferior (or diaphragmatic) wall: II, II and aVF
- Septal: V1 and V2
- Anteroseptal: V1, V2, Vf3 and sometimes V4
- Anterior: V3, V4 and sometimes V2
- Apical: V3, V4 or both
- Lateral: I, aVL, V5 and V6
- Extensive anterior: I, aVL and V1 through V6

According to the ECG reading, the severity of the MI can also be determined as either STEMI, the most severe, or NSTEMI [4]. STEMI will display an elevation in the ST segment of the T-wave, which represents a total occlusion of a coronary artery.

Treatment of STEMI

Currently, there are three options to treat patients experiencing a STEMI heart attack: thrombolysis, percutaneous coronary intervention (PCI), also known

as balloon angioplasty, and coronary artery bypass graft surgery (CABG) if three or more occlusions have occurred [5-7].

Thrombolysis consists of injecting the patient with clot-diluting drugs in order to open the blocked artery. However, thrombolysis has been considered a less effective and less efficacious reperfusion technique since it takes an extended amount of time to begin to work [5, 10]. By the time thrombolysis dilutes the clot, too much of the myocardium is dead or damaged severely. Thrombolysis was the first strategy developed to combat occlusions. However, more recently, PCI has been repeatedly shown in numerous differing populations to be more effective and efficacious when compared with thrombolysis, in preserving more of a patient's myocardium. PCI has shown to have superior clinical outcomes such as lower mortality rates, lower rates of recurrence of thrombotic occlusion, lower rates of re-infarction, and shorter recovery times to a productive life [7, 11-14]. The evidence on cost-effectiveness of thrombolysis versus angioplasty has not been conclusive. A study by Hartwell et al. favored percutaneous coronary intervention over thrombolysis. However, a study conducted by Reeder et al in 1994 concluded that these alternatives are equally cost-effective.

Development of Thrombolytic Therapy

Thrombolytic or fibrinolytic therapy was originally studied and developed in the 1950's with promising results by Dr. Sol Sherry who demonstrated that streptokinase could dissolve blood clots that typically occluded coronary arteries. However, more recent reperfusion theory demonstrates that thrombosis is a

secondary event in ischemic events. Long after the advance that a coronary occlusion was an etiology of acute myocardial infarction, the Gruppo Italiano per lo Studio della Streptochinasi nell'Infarctomiocardio (GISSI) study (in 1986) established through a multicenter randomized clinical trial that in-hospital mortality was reduced by more than 3% following streptokinase administration [5]. Yet, streptokinase and other first generation thrombolytic agents came with the limitation of being immunogenic and a third of the patients receiving therapy were unresponsive to such therapy.

DNA research advances led to fibrin specific plasminogen activators, which resulted in much lower rates of mortality and risk reduction when compared to streptokinase to treat AMI leading to extended hope for thrombolysis. Similarly, aspirin demonstrated similar survival benefits to those of streptokinase alone in several clinical trials. The use of these drugs have become part of the American College of Cardiology/American Heart Association guidelines for the treatment of MI as evidence based quality of care measures; however, no study has been conducted in the cost-effectiveness thereof [10].

Percutaneous Coronary Intervention

PCI, or angioplasty, was developed in 1977 when the Swiss radiologist, Andreas Gruentzig, performed the first percutaneous transluminal coronary angioplasty on a thirty-eight year old patient with a left coronary artery lesion [15]. Compared to today's equipment, the catheters were large and could easily injure

a blood vessel. In addition, no guidewires were used and balloon catheters would discharge suddenly at low-pressure levels, which rendered this type of intervention available to only ten percent of patients needing revascularization.

Fortunately, during the early 1980s, the catheters were manufactured to a smaller diameter and balloons were designed to inflate at higher pressures, which made balloon angioplasty available to nearly 50% of patients in need of revascularization. However, one of the drawbacks of balloon angioplasty is that it fractures plaque that is causing the current occlusion. The fissured atheroma may cause the formation of a new thrombus later. Another limitation is the weakening of the vessel that may cause the vessel to recoil, resulting in restenosis. As a result, new devices aimed to combat these limitations were developed and manufactured. The transluminal extraction catheter and excimer laser enabled the development of atherectomy, although they were not associated with a reduction in the incidence of restenosis, rendering these instruments as niche tools. It was not until the advent of coronary stents in 1986 that PCI was shown to be safe and more effective by significantly reducing restenosis. Improvements to stents such as pharmacologic coating and various levels of flexibility and strength followed in the 1990s and the first decade of the twenty-first century.

PCI (emergency angioplasty) is performed in a series of steps with slight variations for an individual case. Once the patient presents at the emergency facility, the patient is treated with a fibrinolytic agent unless contraindicated by the patient's clinical condition such as ventricular arrhythmias or cardiogenic

shock[4, 5, 10, 15, 16]. Subsequently and immediately, an angiogram is performed. During an angiogram, a catheter is inserted into one of the femoral arteries (located in the groin/thigh) and guided to the coronary arteries. The femoral arteries are the largest arteries in the body. Injuries to the femoral artery used can result very rapidly in severe blood loss and possible death and thus creating the need for PCI's to be performed by experienced surgeons, well-staffed and experienced catheter labs, and at well-equipped facilities. A contrasting substance is then injected into the vessel to make the area surrounding the heart clear in the X-ray images. It is this tool that allows the emergency personnel to determine what vessels are blocked. After the affected area has been determined, a balloon catheter is inserted and guided to the blocked vessel. Once in place, the balloon is inflated to open out the walls of the blood vessel and crush the clot. Also, it is recommended to place a tubular mesh, know as a stent, in the affected segment of the blood vessel to prevent the collapse of the vessel's walls. Finally, the catheter is removed and the entry-puncture sealed or the catheter may be left in place up to twelve hours, depending on the length of time needed to prevent the patient's blood from thinning with further complication of severe hemorrhage or death. After successful angioplasty, most patients are discharged within 24 hours of the procedure.

One key aspect necessary for PCI to be successful is the door-to-balloon time (DTB)[16-21].This is the length of time between the patient arriving at the emergency facility and the moment the balloon is inflated in the affected segment

of a blood vessel. After many clinical studies, it has been determined that the DTB should be less than 90 minutes. The reason for this timeframe is that it provided, in multiple clinical trials, lower in-hospital death rates, reduced 30-day mortality rates, shorter average lengths of stay in the hospital, lower rates of re-infarction, and lower rates of re-occlusion[17, 19, 21-23].

Effectiveness of PCI

Several clinical trials have been conducted to investigate the effectiveness of PCI compared to adjunctive therapy in NSTEMI patients. A key idea in these results was the stratification of the patients according to the Thrombolysis Myocardial Infarction (TIMI) score, which classifies patients as high, medium, or low risk. High-risk patients, with TIMI scores of five or greater, are those with three or more of the following characteristics: prolonged chest pain, hypotension, 65-years of age or older, ST-segment changes, and elevated biomarkers levels [2].

Retrospective studies based on the Global Registry of Acute Coronary Events (GRACE) and the Euro Heart Survey of Acute Coronary Syndromes found that the proportion of patients with an admission diagnosis of unstable angina or MI quickly progress to NSTEMI (24% and 21% respectively) or to STEMI (6% and 69% respectively) [16, 24]. These two studies also found that hospitals with a catheterization lab treat 53% (in the US) and 25.4% (in Europe) of patients with a diagnosis of NSTEMI with PCI. Researchers from the clinical trial, TIMI IIIB [25], found three significant independent predictors of late positive

troponin, a confirmatory biomarker for MI, (1) levels for patients presenting with ACS and early negative troponin levels: TIMI 3 (OR=3.52, 95% CI = 2.38-5.23, p<0.001), (2) ST-deviations (OR=2.91, 95% CI =1.92-4.40, p<.001), and (3) no prior use of beta blockers (OR=1.74, 95% CI=1.15-2.63, p=.008).

For patients with a high-risk score, most studies have found that an early aggressive reperfusion therapy, PCI or CABG, resulted in better clinical outcomes when compared to a conservative therapy of pharmacologic agents (Table 1). In FRISC II, a clinical trial comparing early invasive with non-early invasive strategies at one year follow-up found that the composite outcome of death or MI occurred in 10.4% of those in the early invasive arm and 14.1% of those in the non-early invasive arm with a p-value of 0.015 [26]. A new meta-analysis by Dr. Keith Fox found that an aggressive approach leads to better long-term outcomes when compared to a more selective conservative strategy (Table 2) [27]. The decision between PCI and CABG is primarily based on anatomical and physiological characteristic of the injury. For example: left main coronary artery disease, compromised left ventricular function, diabetes, and/or three or more vessel injuries are treated with CABG, while two or fewer vessel injuries are reperfused with PCI [2].

Table 1. 30-day and 6-month mortality rates stratified by aggressive and conservative therapy.

Study	30 - days			6 - Months		
	Aggressive Therapy	Conservative Therapy	OR/RR	Aggressive Therapy	Conservative Therapy	OR/RR
Neumann et al. [28]	5.9%	11.6%	1.96 (1.01,3.82) p=0.025	NA	NA	NA
Cannon et al.[29]	7.4%	10.5%	.67 (.5, .91) p=0.009	15.9%	19.4%	.78 (.62, .97) p=0.025
Morrow et al. [30]	7.4%	16.2%		15.3%	25%	.54 (.4, .73) p<0.001

Table 2. Cox regression of outcomes stratified by selective and routine invasive procedures.[28-30]

Outcomes	Selective Invasive	Routine Invasive	Hazard Ratio	p-value
MI	12.9%	10.0%	0.77	.001
CV death	8.1%	6.8%	0.83	.068
CV death/MI	17.9%	14.7%	0.81	.002
All-cause mortality	11.7%	10.6%	0.90	.190
All-cause mortality/MI	20.9%	18.1%	0.85	.008

Similarly, a myriad of clinical studies and retrospective studies based on registries have been conducted to investigate the difference in clinical outcomes between primary PCI and thrombolytic therapy. A vast majority of the studies have concluded that PCI resulted in better long-lasting outcomes when compared with fibrinolytic therapy (Table 3). In a small trial by Zijlstra, tests for differences in unconventional outcomes that relate to an individual's long-term functioning were performed showing those managed with PCI had less unstable

angina (5.7% vs. 19.4% p=0.02), more left ventricular ejection (51 vs. 45 p=0.004) and more patency in the related artery (91 vs. 68 p=0.001) [14]. Likewise, the GUSTO IIb analysis concluded that PCI was superior to thrombolytic therapy. In 30-day follow-ups, a composite outcome of death, non-fatal re-infarction, and disabling stroke was measured in patients randomized to PCI and thrombolysis showing PCI's superiority in this composite outcome (9.1% vs. 13.7%, p=0.013). The DANAMI-2 clinical trial also showed that PCI was superior to fibrinolytic therapy when it was stopped in the second interim study with a p-value less than 0.009 in favor of PCI [31].

Furthermore, the advent of stents and drug-eluting stents has resulted in superior outcomes in the latest trials and PCI-registry analyses [2]. In a meta-analysis comparing randomized clinical trials of primary PCI versus Thrombolysis, researchers found that the odds ratio of mortality at six weeks was 0.56 (0.33, 0.94), demonstrating that PCI was favorable [32].

A key aspect of these studies is the time to reperfusion subsequent to hospital admission. The optimal time to reperfuse via PCI technique has been determined to be 90 minutes or less [10]. For instance, a cohort study based on the American College of Cardiology's National Data Registry indicated that longer door-to-balloon times were associated with a higher risk of mortality in hospitals [33]. The risk of in-hospital mortality increased from 3.0% with a 30-minute door-to-balloon time to 8.4% with a door-to-balloon time of 180 minutes in non-linear fashion and with a p-value less than 0.001.

Table 3. Long-term Outcomes: PCI vs. Fibrinolytic Therapy.

Authors	Outcomes	In-hospital			6 – Months / 1 – Year		
		PCI	Thrombolysis	p-value	PCI	Thrombolysis	p-value
Grines et al.[20]	Death	2.6%	6.5%	0.06	NA	NA	NA
	Reinfarction	5.1%	12.0%	0.02	8.5%	16.8%	0.02
Zijlstra et al.[14]	Death	0.0%	6.0%	0.13	NA	NA	NA
	Reinfarction	0.0%	12.5%	0.003	NA	NA	NA
Ribichini et al.[13]	Death	1.8%	5.5%	0.6	3.6%	7.3%	0.7
	Reinfarction	1.8%	9.1%	0.2	5.5%	45.5%	0.0001
Le May et al.[34]	Death	4.8%	3.3%	1.0	4.8%	3.3%	1.0
	Reinfarction	11.3%	42.6%	0.001	14.5%	49.2%	0.001
Schoming et al.[35]	Death	NA	NA	NA	8.5%	23.2%	
	Reinfarction	NA	NA	NA	10%	34.9%	

Issue with Primary PCI in the hospital without onsite CABG Capabilities

An inherent risk to performing primary PCI is accessing or puncturing the femoral artery, which is a large artery, and may lead to a patient "bleeding out" very fast and leading to the patient's death if not corrected quickly. Also, issues may arise during the procedure such as an abrupt closure of the previously opened artery. Patients with elevated risks for MI and other co-morbidities may need to be transferred to an operating room that is capable and equipped to perform open-heart surgeries. Therefore, it is recommended that facilities performing PCI's should have appropriate onsite back-up surgical capabilities[10].

The decision to allow hospitals without CABG capabilities to perform primary PCI rests on each state's regulatory body. While the American College

of Cardiology/American Heart Association (ACC/AHA) guidelines give primary angioplasty without surgical backup a class 2b indication ("probably reasonable"), 30 states currently allow hospitals without surgical open heart surgery (SOS) capabilities to perform primary (emergency) PCI. A myriad of studies, mostly retrospective, have shown that no difference in clinical outcomes exists between hospitals with CABG capabilities and hospitals without surgical backup for primary PCI[22, 36-43]. In the study "The Cost-effectiveness of the Kentucky Pilot Project of Allowing Primary PCI at Hospitals without Onsite CABG Capabilities" by Ramos in 2010, it is investigated how cost-effective it is to allow select facilities lacking emergency CABG capabilities to perform emergency PCI given that these facilities meet recommendations concerning screening criteria, surgeons' experience, and facility's volume [44]. By performing a meta-analysis, robust event rates estimates were derived, and the National Inpatient Sample 2005 was used to obtain respective therapeutic alternatives cost estimates. Total charges from this dataset were used as a surrogate to costs since the latter are considered proprietary information to insurance companies; and neither transportation nor patient's personal caregiver expenditures were estimated due to the lack of a source thereof. This study was based on cost per death adverted instead of quality of live measures. However, quality of life data from patients such as preference on proximity of the facility providing PCI, reduction on door-to-balloon time, and access to primary care provider could have increased the cost-effectiveness of allowing facilities without onsite CABG capabilities to perform emergent PCI. This study concluded that the alternative to allow

Regional Hospitals as well to perform primary PCI dominated the other alternative of *Only Allowing Hospitals with Onsite CABG* to perform PCI. This means that allowing regional hospitals to perform primary PCI both incur fewer costs while also averting more deaths. The incremental cost-effectiveness ratio of allowing regional hospitals to perform PCI was -$41K per death averted, when compared to the option of *Only Hospitals with Onsite CABG* (Table 4). Hence, allowing regional hospitals to perform PCI will save $41K per death averted. This evidence establishes this alternative as a cost-effective way in which to provide primary PCI in the State of Kentucky and supports the allowing of regional hospitals (that meet the recommendations outlined in Myers et al) to perform primary PCI.

Table 4. Cost Effectiveness Analysis

Strategy	Data		Incremental Comparison		
	Cost / 1000	Lives Saved	Cost	Deaths Adverted	C/E Ratio
Only Hospitals with onsite CABG	$54,687.38 K	954			
All PCI equipped Hospitals	$54,401.25 K	961	-$286.19 K	7	-$41,164.25

Many of the retrospective studies have found that many cases of MI originate in rural areas where most hospitals lack open-heart surgery capabilities [22, 40, 42]. Between 1998 and 2002, Alamance Regional Medical Center offered primary PCI under the guidance of Duke University Medical Center as part of a pilot program [40]. The inclusion criteria were those of low to moderate-risk patients according to the literature. A total of 561 interventions were performed with a success rate of 98%, adverse events included one death due to

acute renal failure after successful PCI, while the other 2% of the patients were transferred to Duke Medical Center where 0.7% had to undergo bypass surgery (successfully). Magic Valley Regional Medical Center (MVRMC) in Twin Falls, ID, implemented a program of providing PCI for acute coronary syndromes in a rural setting without surgical backup starting in 2003 [37]. It compared its outcomes with those of the ACC registry of 2004, which are considered the standard of care. The distributions of door-to-balloon times were statistically significant with that of MVRMC lower than that of the ACC registry. The in-hospital death rates were 1.96% for MVRMC and 1.16% for the ACC registry ($p=0.1446$) indicating a not significant difference. Only one patient had to be transferred for open-heart surgery after the perforation of the left anterior artery. The transport time to Boise was 100 minutes, and the surgery did not have any complications.

In 2000, the New York State Department of Health, which has a Certificate of Need system for limiting the number of hospitals in which CABG surgery and PCI can be performed, began to allow a limited number of hospitals to perform emergency (primary) PCI for patients with STEMI. By 2006, a total of 11 hospitals were certified to perform PCI without surgical backup (henceforth called P-PCI centers). This study compares patient outcomes for patients with STEMI in those hospitals with patient outcomes in full service (FS) cardiac hospitals (hospitals that perform CABG surgery as well as P-PCI and elective PCI) (Table 5) [38]. The PAMI-No SOS study showed that primary PCI in high-risk STEMI patients in hospitals without on-site cardiac surgery is safe, effective and faster

than primary PCI after transfer to a surgical facility[43]. Locally, the state of Kentucky conducted a pilot study to investigate the soundness of allowing select facilities in the state to perform primary PCI despite being devoid of onsite emergency backup capabilities [36]. It concluded that clinical outcomes were not different between the local hospitals and the national standard and that study populations were similar to those of national PCI registries.

Table 5. Short-term Outcomes: PCI-Centers vs. Full-Service Centers. [38]

	P-PCI Center	FS - Center	p-value
In-hospital/30-day mortality	2.31%	1.91%	0.40
Same/Next day CABG	0.23%	0.69%	0.046
Emergency CABG	0.06%	0.35%	0.06

Cost Benefit of PCI vs. Thrombolytic Therapy

As early as 1994, the Mayo Clinic conducted a study to compare costs of immediate PCI and thrombolytic therapy for acute coronary syndromes [45]. The Mayo Clinic's hypothesis was that thrombolysis followed by adjunctive medical treatment was more cost-effective than angioplasty. However, the researchers arrived at the conclusion that there was no difference in cost-effectiveness within a twelve-month period.

In 1993, the Health Technology Assessment Programme in the UK embarked in a long-term cost-effectiveness analysis of primary angioplasty taking the perspective of the UK's National Health System (NHS) with the subsequent results published in 2005 [46]. Their results consistently and

convincingly demonstrated a clinical advantage of immediate PCI over thrombolysis. No evidence was found suggesting that services should be concentrated to large hospitals in metropolitan areas despite evidence that larger volumes of the procedure resulted in lower levels of mortality. The economic evaluators compared PCI with thrombolytic therapy for people with AMI; they found that PCI was more cost-effective than thrombolysis even when taking into account variations in the cost of drugs and health status of the patients. In 2009, Khot et al. conducted a small prospective analysis of the costs related to reducing the door-to-balloon time for STEMI patients at St. Francis Hospital and Health Center (Beech Grove and Indianapolis, IN)[47]. They concluded that the payers obtain all of the financial benefits of reducing DTB in STEMI patients, both during initial hospitalization and after one-year follow-up.

Coronary Artery Bypass Graft Surgery

During angiography, the number of obstructed vessels is determined. If the left main coronary artery is seriously narrowed or obstructed, then coronary artery bypass graft surgery is recommended. CABG is also the treatment of choice when severe coronary artery disease is revealed during angiography. However, coronary artery bypass surgery remains controversial as a treatment for ongoing acute myocardial infarction. CABG is now considered when ongoing ischemia persists in patients who have been previously treated with angioplasty or fibrinolytic therapy. CABG is a procedure that creates new routes around the

blocked or severely narrowed arteries by transplanting a healthy blood vessel from another part of the patient's body.

The development of coronary artery bypass graft surgery may be traced to the Vineberg procedure, which was first implemented in April 1950. The Vineberg procedure consisted of implanting an internal mammary artery into a left ventricular myocardium. However, this procedure was not received with enthusiasm within the medical community, as its physiologic benefits could not be fully documented; coronary angiography had not been developed yet. The first true graft surgery recorded seems to have taken place in 1958, and it seemed that it was in response to a procedural accident. William Longmire was performing a coronary endarterectomy when a right coronary artery disintegrated. He then decided to graft an internal mammary artery to restore blood flow. It was not until 1964 that the first aortocoronary saphenous vein graft was successfully recognized; Doctors Debakey and Garret had then an experience similar to that of doctor Longmire. They performed a left anterior descending coronary endarterectomy, which had to be salvaged by an aortocoronary saphenous vein graft. Their patient survived the surgery and had an unobstructed aortocoronary saphenous vein graft when restudied eight years later. With the development of angiography by Sones and Shirey in 1958, the need for proof of increased perfusion began to be met. By 1964, Mason Sones had accumulated an extensive number of cineangiograms that demonstrated the safety and efficacy of a saphenous vein graft for single-vessel coronary artery disease. A dramatic increase in the application of these techniques ensued, and

coronary artery bypass operations became the most frequent surgical procedure in the United States within a decade.

The cardiac surgeon will make an incision, usually about 10-inches long, in the middle of the patient's chest. Then, the surgeon proceeds to separate the breastbone to have access to the heart and aorta. At this point, the patient is to be connected to a heart-lung bypass machine. The surgeon stops the patient's heart and cross clamps the aorta. To minimize damage, the heart is cooled with a solution of iced water and salt, and a preservative solution is injected into the coronary arteries. Plastic tubes are connected to the right atrium to channel venous blood out of the body to the heart-lung machine, which adds oxygen to the blood and circulates it through the patient's body.

The doctor proceeds to remove a vein or artery from another part of the body, usually the saphenous vein from the leg, the radial artery from the forearm, or a portion of the internal mammary artery from the chest. The surgeon creates a new blood route by sewing one end of the recently grafted vessel to the coronary artery beyond the lesions and the other end to the aorta. At this point, the surgical team will restore blood flow to the heart, which usually starts beating again on its own or is electrically shocked. The patient is disconnected from the heart-lung machine, and tubes are inserted into the patient's chest to drain fluid. The surgeon proceeds to reconnect the breastbone with wire, which remains within, and sew the incision closed. Coronary artery bypass graft surgery usually takes between four and six hours; afterwards, the patient is transferred to an intensive care unit for at least two days.

Goals and Objectives

The goal of this study is to use a data mining framework to assess the three main treatments for acute myocardial infarction and their variants: thrombolytic therapy, percutaneous coronary intervention (percutaneous angioplasty), and coronary artery bypass surgery. The need for a data mining framework in this study arises because of the use of real world data rather than highly clean and homogenous data found in most clinical trials and epidemiological studies. The assessment is based on determining a profile of patients undergoing an episode of acute myocardial infarction, determine resource utilization by each treatment, and creating a model that predicts each treatment resource utilization and cost for a distinct subject.

Text Mining and unsupervised clustering is used to find a subset of input attributes that characterize subjects who undergo the different AMI treatments as well as distinct resource utilization profiles. Classical statistical methods are used to evaluate the results of the clustering techniques. The features selected by unsupervised learning will be used to build predictive models for resource utilization and compared with those features selected by traditional statistical methods for a predictive model with the same outcome. Sequence, Association, and Link analysis is used to determine the sequence of treatment of acute myocardial infarction. The resulting sequence is used to construct a probability tree that defines the basis for cost effectiveness analysis that compares AMI treatments and their variants. To determine effectiveness, survival analysis methodology is used to assess the occurrence of death during the

hospitalization, the likelihood of a repeated episode of acute myocardial infarction, and the length of time between reoccurrence of an episode of acute myocardial infarction or the occurrence of death.

CHAPTER 2

Data Mining

Data mining represents the combined efforts of several well-established fields: classical statistical analysis, machine learning, artificial intelligence, and the development of large databases. Because the convergence of these domains makes up the whole of data mining methodology, it means that inductive and deductive methods have active roles in knowledge discovery, which is better defined by Fayyad [48] as "*the non-trivial process of identifying valid, novel, potential useful, and ultimately understandable patterns of data.*" In turn, data mining seeks to leverage these patterns into models that lead to actions and decisions to increase benefit in some form. However, data mining goes further because it involves an iterative closed-loop system for the evaluation of models and their ability to be adapted to other data sets.

The first step in data mining focuses on the discernment of faint patterns by analysis algorithms that can evaluate nonlinear relationships between predictor variables themselves and their targets. Typically, machine learning enables these tasks of exploratory data analysis and they include rule discovery, clustering, descriptive generalizations, feature selection, and classification. Exploratory data analysis is one of the major activities within a data mining

project. It includes interactive and visual techniques that allow one to view a data set in terms of summary statistical parameters and graphical display to get a feel for any patterns or trends that are in the data set. After interactively exploring a data set, descriptive modeling follows. This undertaking involves forming data set views at a higher level. For instance, one may wish to determine overall probability distributions of the data, also known as density estimations. One may also want to create dependency models describing the relationship between variables.

Another activity encompassing descriptive modeling is that of data partitioning. It consists in either cluster analysis or data segmentation. Cluster analysis tries to find natural groups with many clusters or a user's specified number of clusters. For segmentation, the goal is to find homogeneous groups related to the variable to be modeled. For instance, in business a highly sought segment is that of customers who are big-spenders.

A third activity of data mining is predictive modeling. This activity consists of building a model where the value of one variable can be predicted from the value of other variables. When the variable whose value is to be predicted consists of categories, such as Yes/No, multiple choice, 'from 0 to 5', the model is referred to as classification. On the other hand, when the predicted variable consists of a continuous range of values such as values between 0 and 1, age of a person, blood pressure, number of visits to a primary doctor, the model is referred to as regression. Data mining also involves discovering patterns and rules as well as the activity known as *retrieval by content*. The former attempts

to find combinations of items that occur frequently together in transaction databases such as finding genetic patterns in DNA microarray assays. The latter, retrieval by content, begins with a known patterns of interest and follows the goal to find similar patterns in the new data set. It is mainly used in pattern recognition within text material or image data.

The method followed in data mining process for analytics is a blend of mathematical and scientific methods [49]. Both the mathematical and scientific methods include five basic steps. Mathematical methodology involves the five basic steps of understanding, analysis, synthesis, review, and extension, while the scientific method includes the five basic steps of characterization from experience and observation, hypothesis generation, deduction, test, and experimentation. The basic data mining process flow follows the mathematical method, but some steps from the scientific method are included, for instance, characterization, test, and experimentation. This process has been characterized in numerous but similar formats. The format followed in this dissertation is known as SEMMA: Sampling, Explore, Modify, Model, and Asses [49]. The step of sampling involves the creation of one or more data tables. These samples should be large so that they contain significant information; yet they should be small enough so that they can be processed in timely fashion. In the exploration step, one searches for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas. Modifying the data includes not only selecting existing variables but also transforming existing variables and creating new ones upon which the model selection process

focuses. Data modeling uses analytical tools to search for combinations of the data that dependably predicts a desired outcome. Finally, while assessing the data, one evaluates the usefulness and reliability of the finding from the data mining process.

Rule Discovery: Association and Sequence Rules

Rule discovery, also known as association rules, has as a goal the detection of relationships or associations between specific values of categorical data variables in large databases. Association rules fall in the realm of unsupervised learning that seeks to describe the joint distribution of *N* variables. In other words, if one has a set of *N* observations (x_1, x_2, ...,x_N) of a random vector X with joint density Pr(X), one seeks to infer the properties of this probability density without the help of a *supervisor* or *teacher* providing correct answers or a degree-of-error for each observation [50]. Because of the high dimensionality of large databases, most nonparametric methods fail to estimate this density distribution appropriately. In the case of association rules, the problem is somewhat simplified by changing the goal to finding the joint values of the variables $Z = \{Z_1, Z_2, \ldots, Z_N\}$ appear most frequently within the database. Variables are abstracted to binary values $Z_j \in \{0, 1\}$, where $Z_{ij} = 1$ if the j^{th} variable is present or $Z_{ij} = 0$ otherwise in observation *i*. This method is more clearly described in the context of sales transaction analysis where its origins lie. Each variable is assigned one of two values; $Z_{ij} = 1$ if the j^{th} item is purchased as part of a transaction, whereas $Z_{ij} = 0$ if it was not purchased. Those variables

that frequently have joint values of one represent items that are frequently purchased together. This information can be very useful in the segmentation of customers based on buying patterns. In the case of health outcome analytics, this information can be used in the segmentation of a population based on diseases and treatments.

In more general terms, the goal of association rule analysis is to find a collection of prototype X-values v_1, v_2, ...,v_L for the feature vector X, such that the probability density of $Pr(v_l)$ evaluated at each of those values is relatively large. The algorithm seeks regions of the X-space with high probability content relative to their size or support. Let S_j represent the set of all possible values of the j^{th} variable, its support, and let s_j be a subset of the values in S_j. The goal can be stated as finding the subsets of variable values s_1, s_2, ...,s_p such that the probability of each of the variables simultaneously assuming a value within its respective subset $\Pr\left[\bigcap_{j=1}^{p}\left(X_j \in s_j\right)\right]$ is relatively large.

The intersection of subsets $\bigcap_{j=1}^{p}\left(X_j \in s_j\right)$ is called a conjunctive rule.

However, general approaches are not feasible for very large databases where $p \approx 10^4$ and $N \approx 10^8$. A further simplification is to consider the integers $J=\{1, 2, \ldots, p\}$ and corresponding values v_{oj}, j in J such that $\Pr\left[\bigcap_{j\in J}\left(X_j = v_{0_j}\right)\right]$ is large. Binary-valued dummy variables are used in seeking a solution. The support for S_j is finite for each variable X_j. A new set of variables Z_1, ...,Z_k is created, one such variable for each of the values v_{ij} attainable by each of the original variables X_1,

$X_2, \ldots, X_p$. The number of dummy variables K is $K = \sum_{j=1}^{p} |S_j|$, where $|S_j|$ represents the number of distinct values attainable by X_j. Each dummy variable Z_k is assigned a value of one if the variable with which it is associated takes on the corresponding value to which Z_k is assigned, and Z_k is assigned to zero otherwise. This means that $\Pr\left[\bigcap_{j \in J}(X_j = v_{0j})\right]$ is transformed into finding the integers $K \subset \{1,2,\ldots,K\}$ such that $\Pr\left[\bigcap_{k \in K}(Z_k = 1)\right] = \Pr\left[\prod_{k \in K} Z_k = 1\right]$ is large. Here, the set K is called an *item set.* The number of variables Z_k in the item set is called its *size*. This means that the estimated value of the last expression is a fraction of observations in the database for which the conjunction is true:

$\hat{\Pr}\left[\prod_{k \in K}(Z_k = 1)\right] = \frac{1}{N}\sum_{i=1}^{N}\prod_{k \in K} z_{ik}$, where z_{ik} is the value of Z_k for this i^{th} case. This is also called the support or prevalence $T(K)$ of the item set K. Note that an observation *i* for which $\prod_{k \in K} z_{ik} = 1$ is said to *contain* the item set K. Thus, the algorithm returns high support item set K's in the form of *association rules.* The items Z_k, k in K, are partitioned into two disjoint subsets, $A \cup B = K$, and written A → B where A is called the *antecedent* and B the *consequent.*

Each association rule has several properties based on the prevalence of the antecedent and consequent item sets in the database [49]. These properties are numbers whose magnitude describes the *strength* of the association rule. The support of the rule A → B is the fraction of observations in the union of the

antecedent and consequent, which is simply the support of the transaction or the item set K from which it is derived. This is an estimate of the probability of simultaneously observing both item sets $A \cap B$ in a randomly selected rule, $\Pr(A \cap B)$, which is the number of transactions containing both A and B divided by the number of antecedents. The confidence of A → B is its support divided by the support of the antecedent: $\frac{\Pr(A \cap B)}{\Pr(A)} = \Pr(B \mid A)$. The third number or measure of strength of an association rule is the lift, which is equal to the ratio of the confidence to the expected confidence or support of B: $\frac{\Pr(B \mid A)}{\Pr(B)}$.

Furthermore, sequence analysis is concerned with the order in which a group of items was purchased and focuses on discovering temporal structural relationships. Note that the general sequence analysis, which deals with a multitude of transactions, can be adapted to frequent episode discovery. The latter addresses the problem of finding frequent episodes in a temporal data sequence. Here, an episode is considered a sequence of events appearing in a specific order within a specific time window; for instance, the occurrence of myocardial infarction followed by cardiac arrest.

In health analytics, this can be abstracted to the order of disease diagnoses or the order of procedures to treat an ailment. Useful sequence rules are not always obvious, and sequence analysis helps one extract these rules regardless of how hidden they may be in the database. The visualization of these rules, link analysis, aids one to construct a *map* of the connections between items, and the strength of such connections is represented by the

thickness of each line connecting any two *items* linked together in an association or sequence rule.

Clustering

The purpose of clustering methods is to detect similar subgroups among a large collection of cases and to assign those observations to the clusters in such a manner that cases within a group should be more similar to each other than to cases in other clusters [49]. In addition, cluster analysis helps one form descriptive statistics to ascertain whether or not data consist of a set of distinct subgroups with each cluster representing items with substantially different properties. At the core of clustering is the idea of grouping objects based on a definition of similarity within the algorithm. This definition is based, in turn, on the choice of distance or affinity measure between two objects [50].

Frequently, one has measurements x_{ij} for i = 1, 2, ..., N, on variables j = 1, 2, ..., p. First, pairwise dissimilarities between observations are denoted by $d_j(x_{ij}, x_{i'j})$ between the values of the j^{th} variable. Then one defines

$D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij}, x_{i'j})$ as the dissimilarity between objects i and i'. Some common choices to measure lack of affinity are:

1. Square Distance: $d(x_i, x_{i'}) = (x_i - x_{i'})^2$
2. Absolute Error: $d(x_i, x_{i'}) = |x_i - x_{i'}|$
3. Correlation: $d(x_i, x_{i'}) = \dfrac{\sum_j \left(x_{ij} - \bar{x}_i\right)\left(x_{ij} - \bar{x}_{i'}\right)}{\sqrt{\sum_j \left(x_{ij} - \bar{x}_i\right)^2 \sum_j \left(x_{ij} - \bar{x}_{i'}\right)^2}}, where \quad \bar{x}_i = \dfrac{\sum_j x_{ij}}{p}$

4. Categorical Variables: assume the variable has M distinct values; then,

$$d(x_i, x_{i'}) = 1 \text{ if } X_i = X_{i'}$$

$$d(x_i, x_{i'}) = 0 \text{ otherwise.}$$

5. Ordinal Variables: usually represented by contiguous integers as an ordered set. If there are *N* original values, then their realizations are replaced by $\frac{j - 1/2}{N}$, $j \in \{1, ..., N\}$ and treated as quantitative variables on this scale.

One sees that the clustering algorithm partitions the observations into groups so that the pairwise dissimilarities between those assigned to the same cluster tend to be smaller than those assigned to different clusters. However, the algorithms fall in three distinct types. Combinatorial algorithms work directly on the data with no direct reference to an underlying probability model. Mixture modeling supposes the data define an independent and identically distributed sample from some population described by a probability density function, which is parameterized and fitted by maximum likelihood or Bayesian approaches. The third type is the mode seeker algorithm, which is not parameterized. This method attempts to estimate distinct modes of the probability density function; observations nearest to each respective mode are then clustered together.

Text Mining and Clustering

Text mining is the process of discovering new, previously unknown, potentially useful information from a variety of unstructured data sources, including business documents, customer comments, web pages, and XML files.

Language, although complex, can be studied by considering its simpler properties based on counting and text pattern matching. The latter identifies letters or characters in a document or string; however, it may focus on rejecting words that contain a string of letters of interest that are not related. The process of counting the number of matches to a text pattern occurs repeatedly in text mining, such that one can compare two different documents by counting how many times different words occur in each document. In addition, the language can be checked against large language collections, called corpora. After the computer algorithm has found the example of usage, analysts can attribute meaning to these discovered patterns. This blend of human and computer discovery is a back and forth process that is repeated as many times as necessary similar to the design of experiment process in industrial process control and quality control. Following a given design, one may proceed through several iterations, harvesting preliminary insight from previous iterations of the experiment, then adjusting input parameters and running the experiment again until the results show an F-value significant at the 95% level of confidence.

Mining documents' text helps one understand them without reading every single word thereof. This allows the discovery of the underlying themes and concepts within collections of large documents [49]. Within text mining, there are two basic applications: predictive text mining and descriptive text mining. The latter focuses on uncovering the themes and concepts within a collection of texts. One of the main outcomes of descriptive text mining is that of clustering documents into meaningful groups and reporting the concepts that discovered

within the clusters. On the other hand, predictive text mining encompasses the classification of documents into categories and using the implicit information in the text for decision-making. Predictive text mining involves examining historical data to predict future outcomes.

In health sciences, a large part of the documentation of patient care exists in the form of charts and notes, which must be extracted manually for billing and insurance purposes. In this case scenario, text mining may offer a lower cost process to transfer that information. Another interesting use of text mining is to gather data from the Internet in the form of medical journal articles meeting predetermined criteria with the purpose of conducting a meta-analysis on the subject of interest.

Text Clustering defines a small number of groups of documents so that document within the same cluster are related and documents in different groups are not closely related [49]. This is usually achieved by looking at terms within each document. Documents within a cluster are represented by a list of terms, and those terms appear in most of the document within the group with a large frequency.

In medical databases, patients' conditions and treatments are often recorded with standard numeric codes, called ICD-9 codes. These codes consist of five digits and can be considered nominal data represented in text format. ICD-9 codes have stemming properties resembling words and text. Some of its digits identify a general category, while other digits identify a more specialized category. For instance, the first three digits of 410 represent acute myocardial

infarction (AMI); 4101 represents an acute myocardial infarction of the anterior wall while 41001 represents an acute myocardial infarction of the anterior lateral wall. The fifth digit takes, in this case, takes on the values 1 and two representing the first episode of care and subsequent episodes of care respectively. This means that by combining the codes for a hospital stay, a text document for this stay is built and records all the illnesses or diagnoses the patient experienced during the inpatient visit. These documents can be analyzed with text mining to find similarities between codes in patient conditions, taking full advantage of the stemming properties contained within ICD-9 codes.

Kernel Density Estimation

Often, analysts assume that the underlying distribution of covariates is bell-shaped, which is often not true when the population is not homogeneous. In these cases, nonparametric estimation is usually more reliable [49]. Kernel density estimation is an unsupervised, nonparametric learning procedure that estimates the probability density function of a variable or feature. Let $X_1, \ldots, X_n$ denote the observed data, a sample from unknown f. The kernel density estimator is defined to be $\hat{f}(x)_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\mathrm{K}\left(\frac{x - X_i}{h}\right)$ where h is a positive number, called the bandwidth; and K is a kernel or smoother function with the following properties:

1. $\mathrm{K}(x) \geq 0$
2. $\int \mathrm{K}(x)dx = 1$
3. $\int x \cdot \mathrm{K}(x)dx = 0$

4. $\sigma_K^2 \equiv \int x^2 \cdot K(x)dx > 0$

The kernel density estimator puts a smoothed-out lump of mass of size 1/n over each data point X_i. The bandwidth h controls the amount of smoothing. When h is close to zero, $\hat{f}_n$ consists of a set of spikes, one at each data point. The height of the spikes tends to infinity as h tends to zero. When h tends to infinity, $\hat{f}_n$ tends to a uniform density. The kernel K can be any density function; however, the Gaussian kernel is usually preferred despite many studies indicating the limited impact of the value of K.

Generalized Linear Models

The goal of generalized linear models is to describe explanatory variables effects on a response variable. These are regression models encompassing non-normal response distributions and modeling functions of the mean [51]. Within this framework, effects are evaluated, relevant interactions are included, and smoothed estimates of responses are provided. Generalized Linear Models consist of three components: random component, systematic component, and a link function. Hence, a generalized linear model is a linear model for a transformed mean of a response variable that has a distribution in the natural exponential family.

The random component consists of the response variable denoted by *Y* and its probability distribution [52]. The latter must belong to the natural exponential family of probability density functions with the following form:

$$f(y_i; \theta_i, \varphi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \phi)\right\} \quad (1)$$

The components of this family of probability distributions can be described as

- θ_i is the natural parameter, which characterizes central tendency.
- φ is a parameter of dispersion.
- $i = \{1, 2, 3, \dots, N\}$.

When φ is known, then (1) simplifies to

$$f(y_i;\ \theta_i) = d(\theta_i) \cdot h(y_i)\ \cdot \exp\{y_i Q(\theta_i)\} \quad (2)$$

In this case, one can identify the components of equation (2) with the components of equation (1) as follows:

- $d(\theta_i) = exp\left\{-\frac{b(\theta_i)}{a(\varphi)}\right\}$
- $h(y_i) = exp\{c(y_i;\ \varphi)\}$
- $Q(\theta_i) = \frac{\theta_i}{a(\varphi)}$
- θ_i may vary for $i = \{1, 2, 3, \dots, N\}$ depending on the value of explanatory variables.
- $Q(\theta_i)$ is a function of the natural parameter.
- Some examples of distributions belonging to the natural exponential family are then normal distributions, the binomial distribution, and the Poisson distribution.

The systematic component relates a vector $(n_1, n_2, n_3, \dots, n_N)$ to the explanatory variables through a linear model [51]. Let X_{ij} be the value of the predictor for $j = \{1, 2, \dots, p\}$ and for the subject $i = \{1, 2, 3, \dots, N\}$; this means that the systematic component is expressed as the linear combination of explanatory variables:

$$n_i = \sum_j \beta_j x_{ij}\,, i = 1, 2, \ldots, N$$

In other words, the systematic component is defined as a linear predictor.

The link function is simply a function *g* of the mean $\mu = E(Y_i)$, which connects the expected value to the systematic component μ_i to n_i in the following fashion:

$$n_i = g(\mu_i) \quad (3)$$

The function *g* must be monotonic and differentiable on its support. This means that the link function expresses a transformation of the expected value of the outcome as a linear combination of predictors. When the link function transforms the mean to the natural parameter, the link function is called the canonical link:

$$g(\mu_i) = Q(\theta_i) = \sum_j \beta_j x_{ij}\ ;\ i = 1, 2, \ldots, N \textit{ is the canonical case} \quad (4)$$

In the case of a continuous response, GLM's model the response under the assumption of a continuous distribution with constant variance across input levels, which is in the family of natural exponential distributions, including dispersion parameters. Then the normal distribution of the outcome has the following probability density function:

$$f(y;\, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} \quad (5)$$

To see that the normal probability density function belongs to the natural exponential family, one can identify the components of (1) with those in (5) as follows:

- $\theta = \mu$

- $\varphi = \sigma$
- $a(\sigma) = \sigma^2$
- $b(\mu) = \frac{\mu^2}{2}$
- $c(y\,;\,\sigma) = -\frac{y^2}{2\sigma^2} + log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)$

The mean is its natural parameter. In turn, this means that an ordinary regression model for $E(Y)$ is a GLM using the identity link:

$$g(\mu_i) = \mu_i = \sum_j \beta_j x_{ij}\,;\; i = 1, 2, \ldots, N \; is \; the \; linear \; case \quad (6)$$

Even in the case of a categorical explanatory variable with a normally distributed random component, the regression model is a GLM more commonly known as Analysis of Variance (ANOVA). Furthermore, if the explanatory variable is mixed, continuous and categorical, the statistical model falls under the family of GLM's and is known as Analysis of Covariance (ANCOVA).

Binary data representing success and failure outcomes by 1 and 0 can be modeled by generalized linear models as well. The distribution, which is Bernoulli, describes probabilities of success and failure as $P(Y = 1) = \pi$ and $P(Y = 0) = 1 - \pi$ respectively and the expected value of Y as $E(Y) = \pi$ [52]. The variance for a binary response is given by $var(Y) = \pi \cdot (1 - \pi)$. It is well known that the probability mass function is described by:

$$f(y;\pi) = \pi^y(1-\pi)^{1-y} = (1-\pi)[\pi/(1-\pi)]^y$$
$$= (1-\pi)exp\big(y\, log(\pi/(1-\pi))\big) \quad (7)$$

where *y = 0* or *y = 1*. This function belongs to the natural exponential family (1). The components of this mass function can be identified with the components of equation (2) as follows:

- $\theta_i = \pi \; for \; i = \{1, 2, 3, \ldots, N\}$
- $d(\theta_i) = d(\pi)$
  $= 1 - \pi$
- $h(y_i) = 1$
- $Q(\pi) = log\left(\frac{\pi}{1-\pi}\right)$

In this case, the natural parameter $Q(\pi) = log\left[\frac{\pi}{(1-\pi)}\right]$ is called the log odds of response 1 or the *logit* of π [51]. The *logit* function is the canonical link and is used for constructing the GLM model as follows:

$$log\left(\frac{\pi}{1-\pi}\right) = \sum_j \beta_j x_{ij} \; (8)$$

This model is the called the logistic regression model. It reflects a nonlinear relationship between π and the predictors x_{ij}, which is monotonic. This means that π increases continuously or π decreases continuously as the predictors x_{ij} increase. While π must be a number between 0 and 1, the logit can be any real number. This makes the logit model free of structural problems that are present in a linear probability model.

There are cases when the outcome of interest represents possible counts such as the number of visits to a primary physician. The Poisson distribution is the simplest probability distribution, belonging to the natural exponential family, which can represent possible counts of an event. A key property of the

outcomes is that they can take any nonnegative integer value. To build this model, one lets *Y* represent a count, and its mean is denoted by $\mu = E(Y)$ and variance $var(Y) = \mu$ [51]. The Poisson probability mass function for the response variable *Y* is

$$f(y;\mu) = \frac{e^{-\mu} \cdot \mu^{y}}{y!}$$
$$= \exp(-\mu) \cdot \left(\frac{1}{y!}\right) \exp(y \log(\mu)) \qquad (9)$$
$$where\ y = 0,1,2,....$$

It is easy to see that (9) has the form of the natural exponential family of probability distributions by equating its components as

- $\theta_i = \mu$
- $d(\mu) = exp(-\mu)$
- $h(y) = \frac{1}{y!}$
- $Q(\mu) = log(\mu)$

Clearly, the natural parameter is $log(\mu)$. This means that the canonical link function *g* is the *log link*. When the GLM uses this link, the model is called a Poisson loglinear model because the logarithm of the mean is represented as a linear combination of inputs.

$$log(\mu_i) = \sum_j \beta_j x_{ij} \ , where\ i \in \{1, 2, \dots, N\} \quad (10)$$

The *log mean* can take any real value like the linear predictor $\sum_j \beta_j x_{ij}$. For this model, the mean satisfies the exponential relationship

$$\mu = exp\left(\sum_j \beta_j x_{ij}\right) = e^{\beta_1 x_{i1}} \cdot e^{\beta_2 x_{i2}} \cdot \ldots \cdot e^{\beta_p x_{ip}}$$

From this relationship, one can see that a one-unit increase in a predictor has a multiplicative effect of e^{β_j}: the mean at $x_{ij} + 1$ equals the mean at x_{ij} multiplied by e^{β_j}.

Generalized Linear Model parameter estimates are based on the paradigm of maximum likelihood estimation, which chooses values of the model parameters in a way that the distribution produced gives the observed data the greatest probability [53]. Given that GLM's are restricted to the exponential family of distributions for the random component *Y*, a single algorithm based on the least squares can be implemented to the entire family of models regardless of link function. The attractiveness of maximum-likelihood estimators is based on their asymptotic properties [51]. *Consistency* of the estimators means that the estimators converge in probability to the value being estimated. *Normality at infinity* means that as the sample size increases, the distribution of the ML estimator tends to the normal distribution centered at the true value of the parameter and variance of the reciprocal of the Fisher information function. Similarly, these parameters are *asymptotically efficient* because they achieve the Cramer-Rao lower bound as the sample size increases to infinity.

The ML estimators for the exponential family or the random component are found by using the log-likelihood function $L = \sum_i L_i$ where $L_i = log\big(f(y_i;\, \theta_i, \varphi)\big)$ denotes one observation's contribution to the log-likelihood:

$$L_i = \frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi) \qquad (11)$$

To find the estimates, the log-likelihood is maximized by finding the first and second derivatives of (11):

$$\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\varphi)} \qquad (12)$$

$$\frac{\partial^2 L_i}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a(\varphi)} \qquad (13)$$

where $b'(\theta_i)$ and $b''(\theta_i)$ represent the first two derivatives of the function $b(\cdot)$ evaluated at θ_i, then one applies the general likelihood results to (12) and (13)

$$E\left(\frac{\partial L}{\partial \theta}\right) = 0 \qquad (14)$$

$$-E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = E\left(\frac{\partial L}{\partial \theta}\right)^2 \qquad (15)$$

These hold under regularity conditions, which are satisfied by the exponential family of distributions. If one applies the above results to one observation, one obtains from the first formula

$$E\left(\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)}\right) = 0$$

$$\mu_i = E(Y_i) = b'(\theta_i) \qquad (16)$$

and from the second equation

$$\frac{b''(\theta_i)}{a(\varphi)} = E\left(\frac{Y_i - b'(\theta_i)}{a(\varphi)}\right)^2 = var\left(\frac{Y_i}{[a(\varphi)]}\right)$$

$$var(Y_i) = b''(\theta_i)a(\varphi) \qquad (17)$$

Clearly, one can notice that the function $b(\cdot)$ determines all the moments of Y_i. Now, when N independent observations are considered, the log-likelihood is given by

$$L(\boldsymbol{\beta}) = \sum_i L_i = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of parameters to be estimated, which reflects the dependence of $\boldsymbol{\theta}$ on the model parameters $\boldsymbol{\beta}$. These results give one the following likelihood equations:

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, 2, \ldots, p. \qquad (18)$$

Explicitly $\boldsymbol{\beta}$ is not present in the above equations; however, it is there implicitly through $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$. Since a GLM's random component is restricted to the natural exponential family of distributions, the likelihood equations depend on the distribution of Y_i only through μ_i and $var(Y_i)$. The latter depends on the mean through a particular functional form $var(Y_i) = v(\mu_i)$ for some function v, such as $v(\mu_i) = \mu_i$ for the Poisson distribution $v(\mu_i) = \mu_i(1 - \mu_i)$ for the Bernoulli distribution, and $v(\mu_i) = \sigma^2$ for the normal distribution.

The maximum likelihood estimates $\widehat{\boldsymbol{\beta}}$ of GLM parameters are found by solving the nonlinear equations (18). To solve them, a general-purpose iterative method called the Newton-Raphson Method is used two ways to determine the maximum of a likelihood function [51]. This method begins with an initial guess for the solution. A second guess is obtained by approximating the function to be maximized in a neighborhood of the initial guess by a second-degree polynomial

and finding the location of that polynomial's maximum value. Then, it approximates the function in a neighborhood of the second guess by another second-degree polynomial, and the third guess is the location of its maximum. Thus, the method generates a sequence of guesses converging to the location of the maximum when the function is suitable and the initial guess is good. Mathematically, the Newton-Raphson method determines the value of $\widehat{\boldsymbol{\beta}}$ in the following manner. Let $\boldsymbol{\mu}' = \left(\partial L(\boldsymbol{\beta})/\partial\beta_1, \partial L(\boldsymbol{\beta})/\partial\beta_2, \ldots, \partial L(\boldsymbol{\beta})/\partial\beta_p\right)$and **H** denote the matrix having entries $h_{ab} = \partial^2 L(\boldsymbol{\beta})/\partial\beta_a\partial\beta_b$ called the hessian matrix. Let $\mu^{(t)}$ and $\boldsymbol{H}^{(t)}$ be μ and $\boldsymbol{H}$ evaluated at $\boldsymbol{\beta}^t$, which is the guess t for $\widehat{\boldsymbol{\beta}}$. Step t in the iterative process approximates $L(\boldsymbol{\beta})$ near $\boldsymbol{\beta}^t$ by terms up to second order in its Taylor series expansion:

$$L(\boldsymbol{\beta}) \approx L(\boldsymbol{\beta}^t) + \boldsymbol{\mu}^{(t)'}(\boldsymbol{\beta} - \boldsymbol{\beta}^t) + \left(\frac{1}{2}\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^t)'\boldsymbol{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^t)$$

Then, one solves $\partial L(\boldsymbol{\beta})/\partial\beta \approx \mu^{(t)} + \boldsymbol{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^t) = 0$ for $\boldsymbol{\beta}$, which yields the next guess. This can be expressed as $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\boldsymbol{H}^{(t)}\right)^{-1}\boldsymbol{\mu}^{(t)}$, assuming that $\boldsymbol{H}^{(t)}$ is nonsingular. Iterations proceed until changes in between successive cycles are sufficiently small.

Model selection within the GLM framework is based on the deviance statistic and evaluation of residuals [52, 53]. Given a specific GLM model and observations $\boldsymbol{y} = (y_1, y_2, y_3, \ldots, y_N)$, the log-likelihood function denoted by $L(\boldsymbol{\mu}, \boldsymbol{y})$ is expressed in terms of the means $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$. The maximum of the log-likelihood for the model can be denoted by $L(\widehat{\boldsymbol{\mu}}; \boldsymbol{y})$. If one considers all the possible models, one realizes that $L(\boldsymbol{y}; \boldsymbol{y})$ is the maximum that the function can

achieve. This occurs for the most general model, having a separate parameter for each observation and the perfect fit $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$, which is called the saturated model. This model is not useful since it does not provide data reduction [51]. However, it serves as baseline for comparison with other model fits. Clearly, a saturated model explains all variation by the systematic component of the model. For a particular unsaturated model, the corresponding ML estimates can be denoted by $\tilde{\theta}_i$ and $\hat{\mu}_i$. For a particular unsaturated model, one can denote its maximized log-likelihood by $L(\hat{\boldsymbol{\mu}};\ \boldsymbol{y})$ and for the saturated model by $L(\boldsymbol{y};\ \boldsymbol{y})$. The likelihood-ratio statistic, which tests the null hypothesis that the model holds against the alternative that a more general model holds is given by

$$-2log\left(\frac{maximum\ likelihood\ for\ model}{maximum\ likelihood\ for\ saturated\ model}\right) = -2[L(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) - L(\boldsymbol{y}, \boldsymbol{y})].$$

This is called the scaled deviance or simply the deviance. The greater the scaled deviance, the poorer the fit is [51]. For most GLM's, the scaled deviance has an approximate chi-squared distribution. This statistic describes the lack of fit. One can explicitly express this statistic in terms of the natural exponential family of distributions (11) by

$$-2[L(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) - L(\boldsymbol{y}, \boldsymbol{y})] = 2\sum_i \frac{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)}{a(\varphi)} - 2\sum_i \frac{y_i\hat{\theta}_i - b(\hat{\theta}_i)}{a(\varphi)} = D(\boldsymbol{y};\ \hat{\boldsymbol{\mu}}). \quad (19)$$

If an overall goodness-of-fit test provides evidence that a GLM model fits poorly, one can examine residuals, which highlight where the fit is poor. The main type of residual evaluated is that proposed by Pearson, namely the Pearson residual:

$$e_i = \frac{y_i - \widehat{\mu_i}}{(var(y_i))^{1/2}}.$$

When the model holds, Pearson residuals are less variable than standard normal because these residual compare y_i to the fitted means rather than the true mean. For instance, the denominator estimated $(var(y_i))^{1/2} = (var(y_i - \mu_i))^{1/2}$ rather than $(var(y_i - \hat{\mu}_i))^{1/2}$. On the other hand, standardized residual divide the ordinary residual by their asymptotic standard errors.

An important activity of data mining is predictive modeling, more specifically classification. It helps organizations identify different types of customers, physicians predict which patients are at high risk of cardiovascular disease, and businesses determine whether sales personnel are committing fraud [49]. Classification mainly aims to include or exclude observations in a small set of specific categories. Many methods for data classification exist; some of the most commonly used are neural networks, decision trees, and regression [50]. Given the myriad of classification methods, a very important aspect of this activity is to compare results to determine the best means of classification for a problem. One way to do this is by comparing the rates of correct classification and choosing the technique with the highest rate. However, when data are used to define the model, accuracy tends to be inflated. For instance, one can define a predictive model that is 100% accurate on a training set but 0% accurate on a validation set. This means that validation is paramount. Another difference between classical statistical inference and predictive modeling within a data-mining framework is the use of a holdout sample. Its purpose is to reserve a

portion of the data to be used to test the accuracy of the model after the model has been defined. Thus, data mining focuses less on model assumptions and more on the model's ability to actually predict outcomes. In other words, assumptions are not as important as outcomes.

Within health plans databases, the process of predictive modeling has gained a great deal of interest. Many constituencies operate under the principle of high risk and risk assessment; these include actuary, underwriting, medical management, marketing, and sales. In order to enhance these services, data mining leverages inductive and deductive methods for description and forecasting purposes, which align with organic systems such as the health care industry very well. With classical statistical models, this process involves rules based grouping and indirect correlations to complete the risk assessment and predictions. In each of these processes, the approach is usually linear and the units measured generally are in non-weighted amounts resulting in minimal predictive value. In turn, data mining algorithms provide prominent advantages to these classical modeling techniques: multiple weights for complex variables, multiple endpoints consideration with complex weighting, and identification of continuous variable discrimination points. In addition, data mining measures model accuracy in results rather than validating model assumptions. These differences create improved results measured by relative characteristic curves (ROC) AND R^2 values. In addition, data mining techniques such as neural networks, decision trees, fuzzy logic, rule induction, and principal components,

allow the modeling of variables in nonlinear relationships, which are more prevalent in complex clinical-financial relationships in the real world.

CHAPTER 3

MarketScan® Commercial Claims and Encounters Database Description

It contains person-specific clinical utilization and expenditures throughout outpatient, inpatient, and prescription drug services from approximately 45 large employers, health plans, and public organizations [54]. The database links paid claims and encounter data to patient information across types of providers and service sites over time. It includes private sector health data from approximately 100 payers annually. These data represent the healthcare experience of insured employees, early retirees, and COBRA continues. The types of plans covered include fee-for-service plans, fully capitated plans, and partially capitated plans. The structure of the annually collected data consists of four databases: Medical/Surgical database, Outpatient Pharmaceutical Claims database, and Populations/Enrollment database. The Medical/Surgical database consists of three tables: Inpatient Admissions Table, Inpatient Services Table, and Outpatient Services Table.

The data contained in the medical/surgical tables are of the upmost interest in this study. The Inpatient Admission Table summarizes a hospital admission, which is a group of medical/surgical services that include a room and board claim. Each record is constructed after all of the encounters or claims associated with an admission are fully identified. Included within each record is a

summary of facility and professional payment information. Combined, the tables for the years 2000 and 2001 contain 494,106 records and seventy-seven variables, see appendix A for a detail descriptions of all the variables. Of main interest are the fifteen diagnosis variables and fifteen procedures variable, which are applied chronologically based on service dates. The Inpatient Service Tables contain 8,719,966 records and sixty-four variables, see appendix B for a detail descriptions of all the variables. This table details each individual facility and professional encounter and service that comprise the inpatient admission record. The inpatient admission table and inpatient service table are linked by the identifier CASEID within the same year.

The records in the Outpatient Service Table correspond to encounters and claims for services provided in a doctor's office, hospital outpatient facility, emergency room or other outpatient facility. The Outpatient Service Tables have 117,781,365 records and fifty-four variables which include five diagnosis, one procedure, physician specialty, and place of service. However, a small percentage of these observations truly represent an inpatient admission, which is reflected by checking the values in the variable 'place of service.' The reason for this issue is that some claims do not show an explicit 'room and board' charge and hence are excluded from the inpatient admission table. A detailed description of these variables can be found in appendix C.

The Outpatient Pharmaceutical Claims tables contain 751,148,837 records and fifty-two variables, which are described in detail in appendix D. The records in these tables are linked to the medical/surgical tables by the variable

ENROLID, enrollee ID. Each record represents a card program or mail order prescription claim; however, prescription drug plans with capitation arrangements are not represented in this database. To determine whether or not individuals are enrolled in plans with drug information each year, the RX variable within the Surgical/Medical Tables can be used. This variable will have a value of one if the individual is enrolled in health care plan, which encompasses a prescription plan while a value of zero will indicate otherwise.

Table 6. ICD-9 Codes for Sample Selection.

ICD-9	Description
410.XX	Acute myocardial infarction. The following fifth-digit subclassification is for use with category 410: 0 episode of care unspecified, 1 initial episode of care, 2 subsequent episode of care.
410.0X	Acute myocardial infarction. ST elevation myocardial infarction (STEMI) of anterolateral wall.
410.1X	Acute myocardial infarction. Infarction: anterior wall (with contiguous portion of intraventricular septum), anteroapical (with contiguous portion of intraventricular septum) ,anteroseptal (with contiguous portion of intraventricular septum), ST elevation myocardial infarction (STEMI) of other anterior wall.
410.2X	Acute myocardial infarction. ST elevation myocardial infarction (STEMI) of inferolateral wall.
410.3.X	Acute myocardial infarction. ST elevation myocardial infarction (STEMI) of inferoposterior wall.
410.4X	Acute myocardial infarction. Infarction: diaphragmatic wall (with contiguous portion of intraventricular septum), inferior wall (with contiguous portion of intraventricular septum), ST elevation myocardial infarction (STEMI) of other inferior wall.
410.5X	Acute myocardial infarction. Infarction: apical-lateral, basal-lateral, high lateral posterolateral, ST elevation myocardial infarction (STEMI) of other lateral wall.
410.6X	Acute myocardial infarction. Infarction: posterobasal, strictly posterior, ST elevation myocardial infarction (STEMI) of true posterior wall.

Data Preprocessing

The Inpatient Admission table contained 7,622 observations corresponding to episodes of acute myocardial infarction (AMI). These

observations were extracted based on ICD9 codes in table 6. However, after removing observations with missing values for principal procedure, 6,746 records are left. For these observations, average age, average length of stay, and average total pay were 55.14 years, 4.43 days, and $19,464 respectively (Table 7). Age of the subjects seems to follow a bell shape distribution given that its mean and media, 56 years, are close together and the box-plot is fairly symmetric (Figure 1). However, length of stay seems to be severely skewed with a large standard deviation, 6.23 days, with respect to its mean. Similarly, the distribution of total pay has a large spread and a large difference between its mean and media. Peculiarly, the minimum values for both age and total pay are zero. These observations must be further investigated before including them in the statistical analysis or machine learning algorithms. It must also be noted that length of stay had a maximum value of 368 days, which must be investigated before including this observation in further analysis. In addition, most observations correspond to male subjects, 7,837 (73.55%). Geographically, most cases of AMI occurred in the southeast, 37.54%, and closely followed by the northeast, 34.07% (Figure 2). Moreover, most observations come from the insured employee, 69.82%, rather than a relative covered within the same health care plan (Figure 3). When stratified by occupation, most observations are within industries traditionally labeled as blue collar: manufacturing, 53.24%, and transportation, 18.07% (Figure 4). After conducting a frequency count on the variable RX, drug plan indicator, it is determined all these subjects have a drug

plan as part of their health care plan. This will help later in the analysis of their prescription history.

Table 7. Summary Statistics from Inpatient Admissions Sample

Variable	Mean	Std. Dev	Min.	Max.	Median
Age	55.14	7.39	0	94	56
Length of Stay	4.43	6.23	1	368	3
Total Pay	$19,464.52	$25,445.68	0	$1,046,147.00	$12,757.04

Figure 1. Distribution of Age within AMI Sample

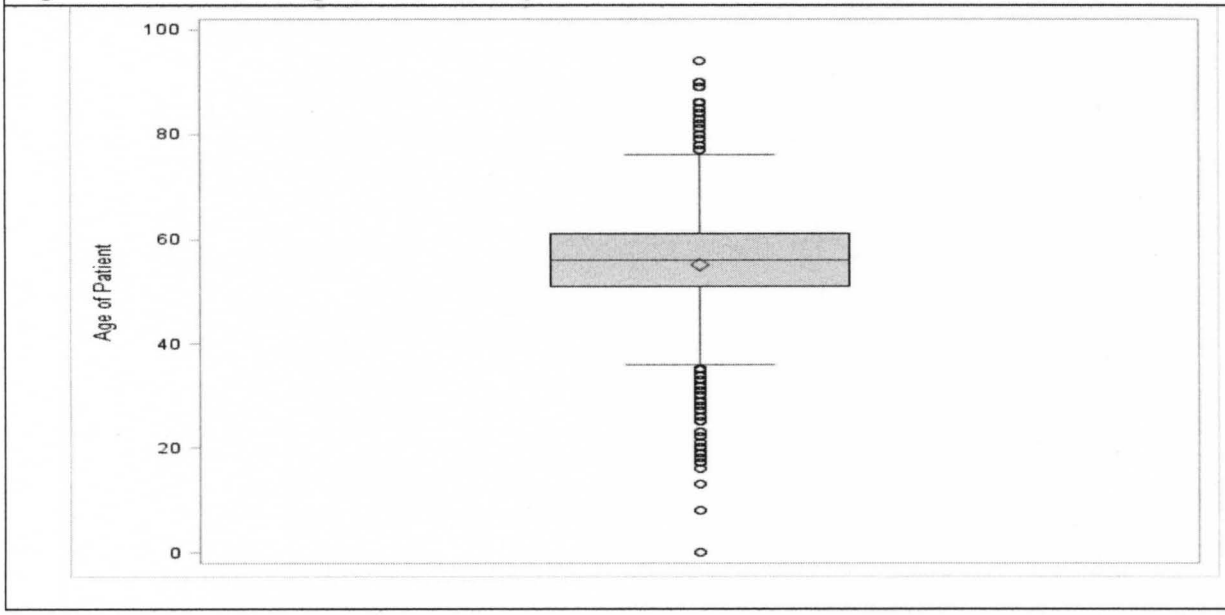


Figure 2. Distribution of AMI cases by Region.

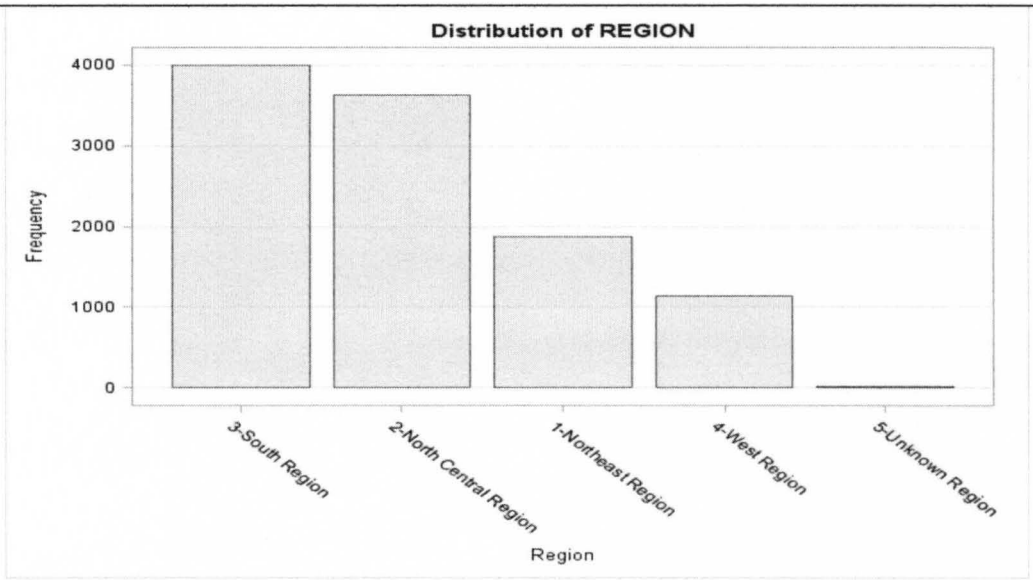


Figure 3. Distribution of AMI cases by relationship to insured employee.

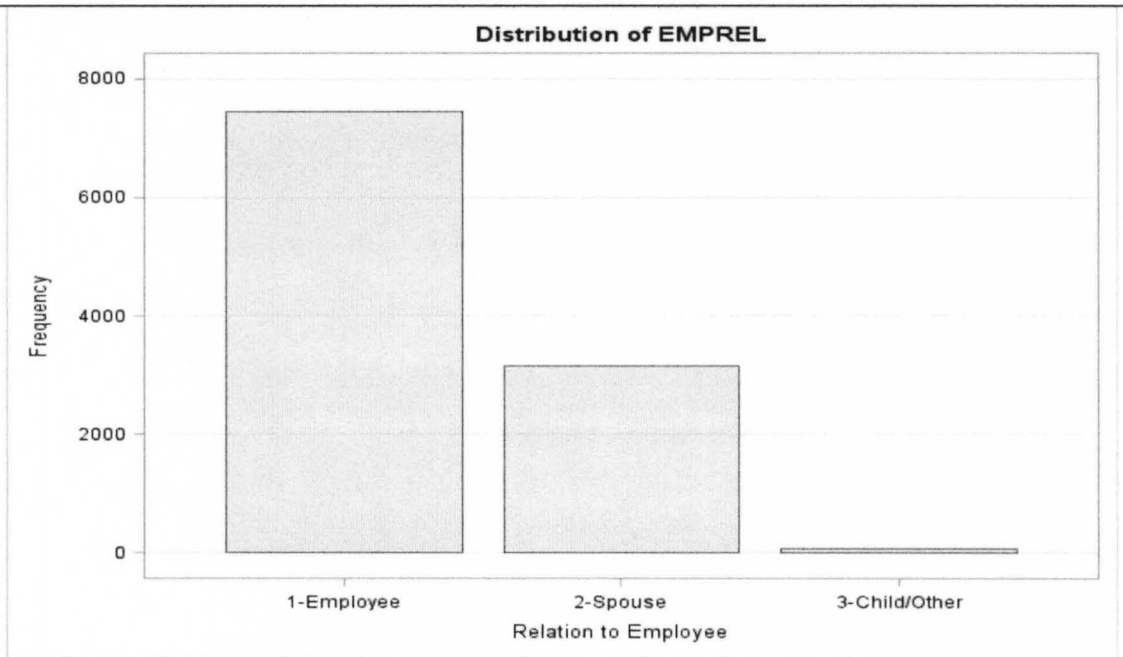
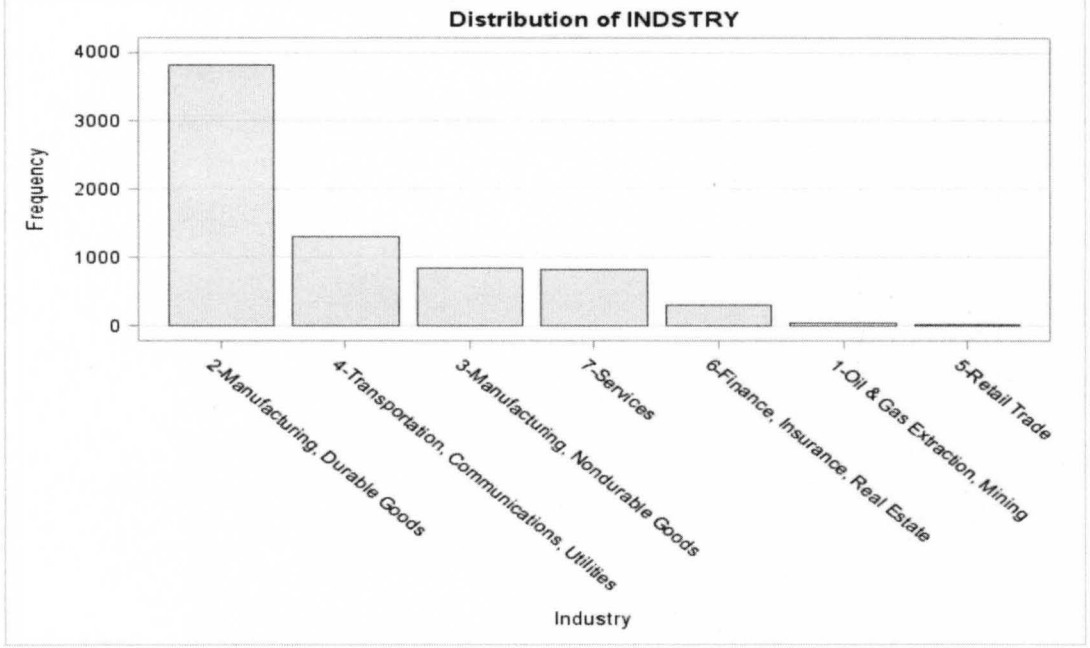


Figure 4. Distribution of AMI cases by employee's industry.



A master table containing unique id's from the inpatient admission table was created. This table was used to filter the outpatient service table for observations corresponding to those subjects who had a record of acute myocardial infarction in the inpatient service table. The resulting table contained 703,793 observations with corresponding records in the inpatient service table. At first glance, one can notice that the ratio of this number to the number of subjects in the inpatient service table is 81, which implies that the average number of outpatient visits within these two years is 81 (Table 8, Figure 5).

Figure 5. Outpatient Observations counts corresponding to AMI patients in Inpatient Admission.

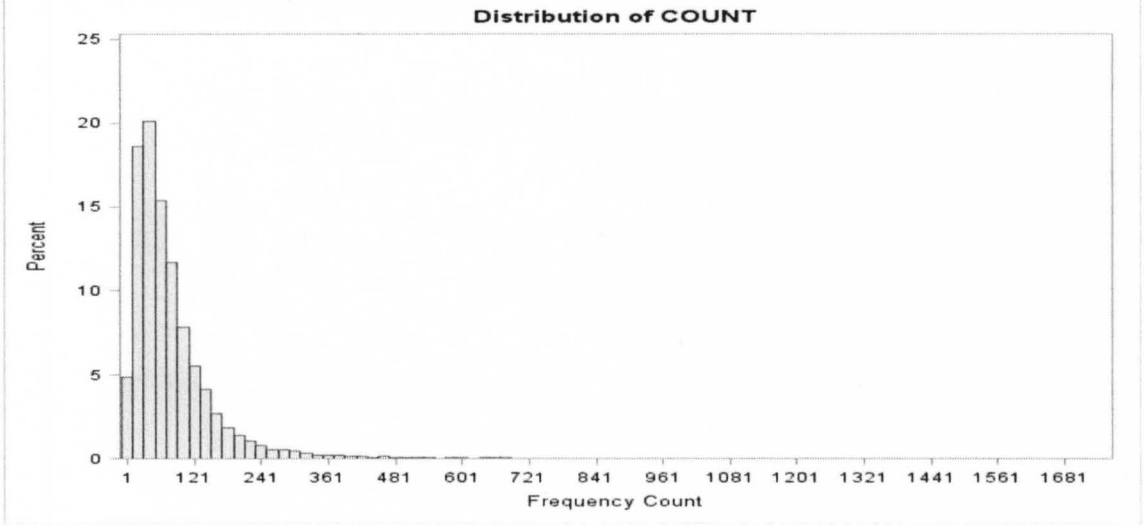


Table 8. Number of Outpatient Records from AMI patients in the Inpatient Admission table.

Mean	Std. Dev	Mode	Lower Quartile	Median	Upper Quartile
81.41	87.28	29	32	58	100

This count has a severely skewed distribution and rather inconsistent with other studies about health care resource utilization. After sorting this table by id, service date, and provider id, many consecutive records share a single user id, provider id, and service date. This indicates that many of these records correspond to a single outpatient visit. This is consistent with the fact that the outpatient service table contains only one column corresponding to a medical/surgical procedure for payment purposes (Table 9).

Table 9. Sample from Outpatient Services after filtering for Inpatient Services observations

Row number	ENROLID	PROVID	SVCDATE	DX1	DX2	DX3	DX4	DX5	PROC1
1	32001	.	02/23/2000	2859	2859				
2	32001	118551932	04/19/2000	41401					99213
3	32001	946809934	07/20/2000	72981					93970
4	32001	262495705	08/08/2000	4919					99283
5	32001	541765934	08/28/2000	4919	4919				
6	32001	675372933	09/01/2000	41400					99213
7	32001	890068434	11/18/2000	4293					71010
8	32001	890068434	11/18/2000	4019					99231
9	32001	890068434	11/19/2000	41401					71010
10	32001	890068434	11/19/2000	41090					93010
11	32001	890068434	11/19/2000	41400					99232
12	32001	890068434	11/19/2000	41400					99232
13	32001	890068434	11/19/2000	41400					99232
14	32001	890068434	11/19/2000	41090					99233
15	32001	890068434	11/20/2000	41400					93010

To be able to analyze the records within the MarketScan databases, code descriptions were necessary for the understanding of diagnoses and procedures. The Center for Medicaid and Medicare Services (CMS) makes available through its website ICD9 code tables in limited tab text files formats. These data consist of three hierarchical text files: diagnosis codes, procedure codes, and diagnosis supplement codes. For this study, both diagnosis and procedure codes are necessary. The tables consist of a code followed by its description, which can extend over several lines. The diagnosis codes vary between three and five digits with a decimal point after the third digit when their length is greater than three digits. To build a one-to-one correspondence with the codes available in

the MarketScan databases, the decimal point is removed and text is left-aligned. The procedure codes consist of three or four digits with a decimal point between the second and third digit. Their processing is similar to that of diagnosis code. In addition, procedure coding is more common with Current Procedural Terminology codes (CPT). These codes are five digits long and come in a tab delimited text file available through the American Medical Association website.

To analyze prescription pattern within the MarketScan database, a description for the National Drug Codes (NDC) was necessary. This was downloaded via the Federal Drug Administration website. The data for these codes consists of three tables: manufacturer information, name and ingredient information, and packaging information. NDC codes from the FDA tables are ten digits long and come in three formats: 4-4-2, 5-3-2, and 5-4-1. The first segment identifies the manufacture, the second segment along with the first segment identifies a drug, and the third segment identifies a packaging/dose. Within the MarketScan database, however, NDC codes have been formatted to an eleven digit 5-4-2 format for computer processing. To construct a one-to-one correspondence to the FDA lookup table, its NDC codes must be reformatted by padding zeros within the different code segments and removing the hyphen within the code (Table 10).

Table 10. Conversion scheme for NDC codes utilization.

FDA Format	Format Display	MarketScan Format (Display)
4-4-2	XXXX-XXXX-XX	0XXXX-XXXX-XX
5-3-2	XXXXX-XXX-XX	XXXXX-0XXX-XX
5-4-1	XXXXX-XXXX-X	XXXXX-XXXX-0X

Data Derivation

Some of the outcomes of interest are not contained explicitly within the databases used for this study. In these cases, other features are explored with the aim of deriving the outcomes of interest. For this study some outcomes and variables of interest that must be deduced from other variables are re-infarction, days to re-infarction, number of prescriptions before and after AMI, number of outpatient visits before and after AMI, number of diagnoses during hospitalization, prescription history as text document, diagnosis history as text document, and cost for the main procedure to treat AMI.

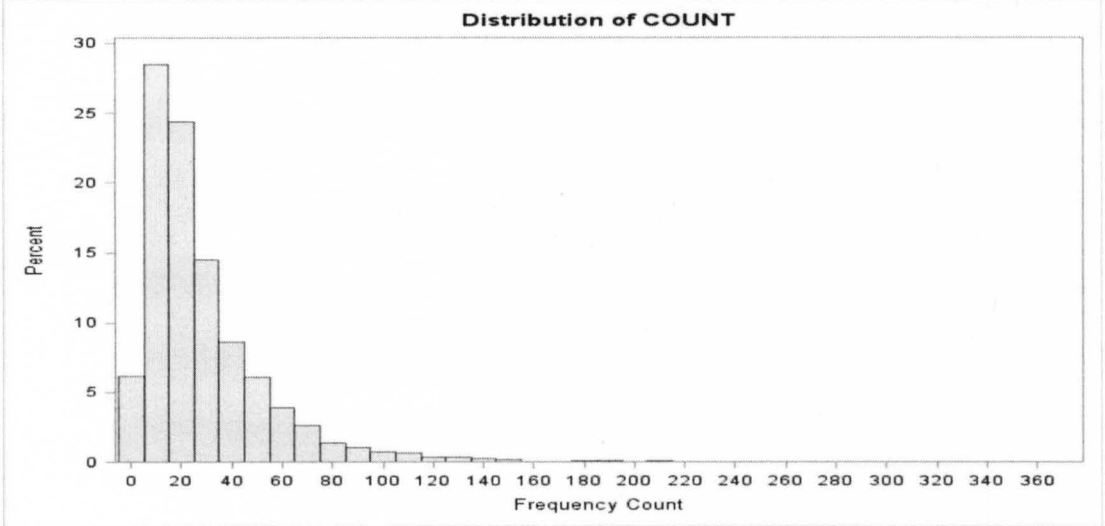
Reinfarction was derived by comparing the first admission per subject to all its following inpatient admission and checking the presence of an ICD9 code corresponding to acute myocardial infarction. When a following inpatient visit corresponded to AMI, a binary variable was created. The value of one corresponds to reinfarction while a value of zero corresponds to not-reinfarction. Days to reinfarction was computed by obtaining the difference between the first admission and the following admission corresponding to reinfarction. For those observations without reinfarction, the days to reinfarction were censored at the end of follow-up: December 31, 2001. To obtain a true count of outpatient visits, the outpatient service table had to be transformed from a wide format to a long format (Appendix I). After conducting this transformation, a much more credible count is obtained (Table 11, Figure 6). The resulting count still has a skewed distribution; however, it is not as severe and has a much smaller standard deviation, which is consistent with other studies regarding health care utilization.

A similar algorithm was utilized to obtain prescription counts before the first episode of AMI and after it.

Table 11. Number of Outpatient Visits from AMI patients in the Inpatient Admission Table.

Mean	Std. Dev	Mode	Lower	Median	Upper
28	26.33	9	11	20	36

Figure 6. Distribution of Counts of Outpatient Visits corresponding to AMI patients.



Visual and Statistical Data Exploration

The first step in data mining is to explore the data in order to gain information about the possible relationships among the feature thereof [55]. Univariate and bivariate visualizations are informative. These techniques allow one to have a feeling for the properties of the features and types of association among variables, which in turn help define a predictive model. First, the univariate distributions of numerical features are investigated with box-plots and summary statistics (Figures 7 - 11, Table 12). It is clear that these distributions do not follow a bell-shape curve and are severely skewed with the exception of

patient's age. The summary statistics reveal that values of zero occur for payment. Further investigation clarify that these values correspond to subject in a capitated plan and must be removed when modeling total pay later in the data mining process. An extreme value of $479K is present in the total payment variable as well as a value of 90 years in the variable age occurs.

Figure 7. Age of Patient.

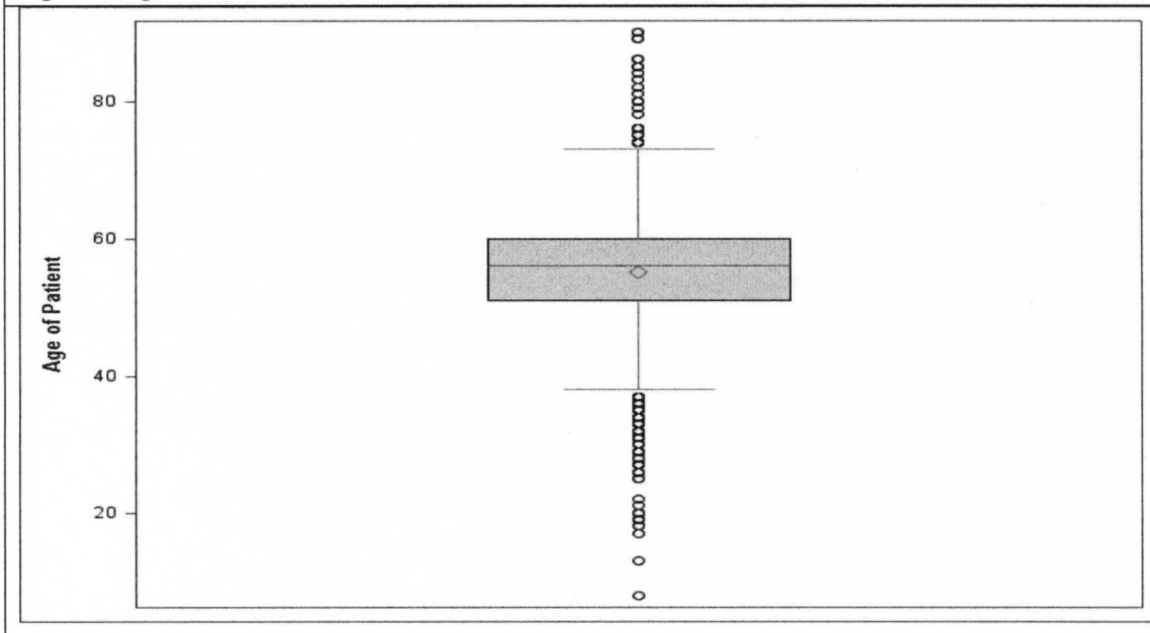


Figure 8. Length of Stay.

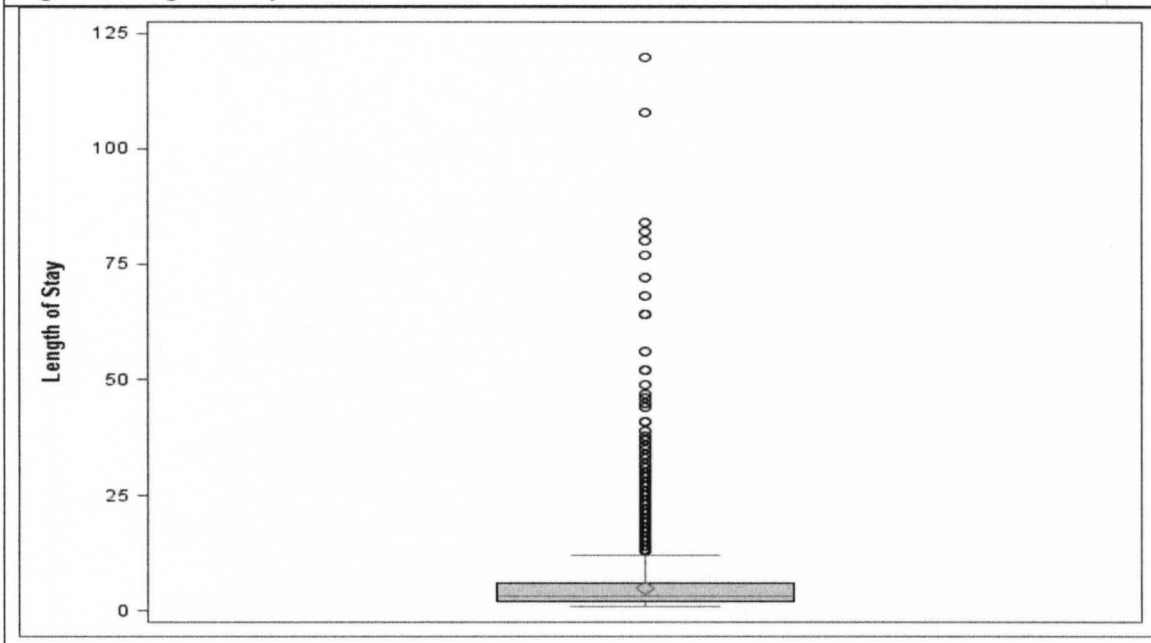


Figure 9. Number of Outpatient visits after first AMI episode.

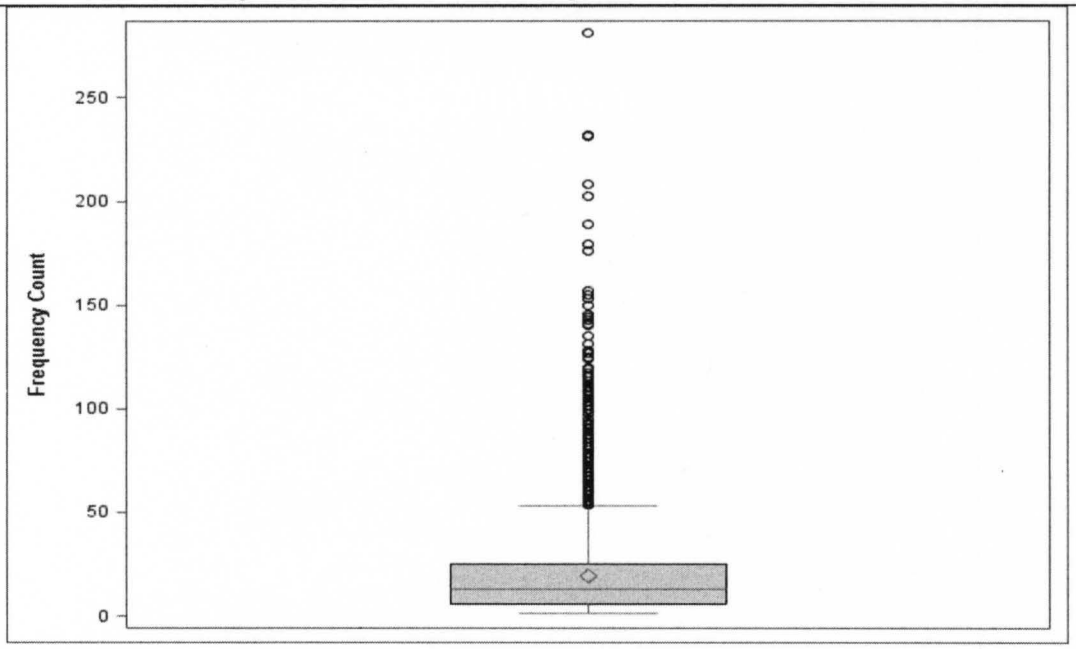


Figure 10. Number of Prescriptions before first AMI episode.

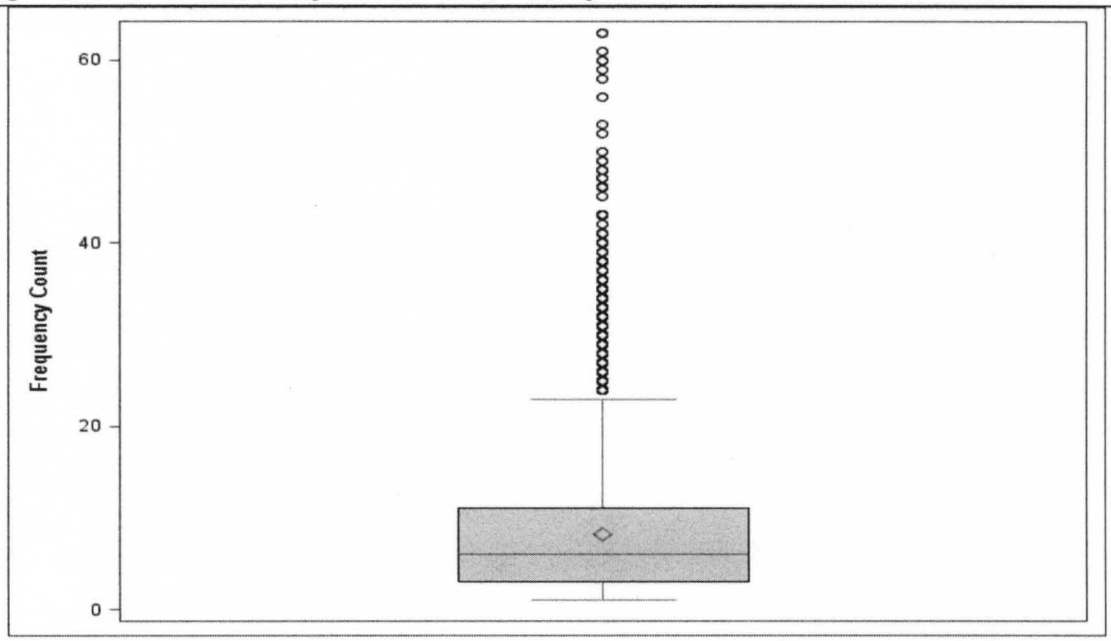


Figure 11. Number of Prescriptions after first AMI episode.

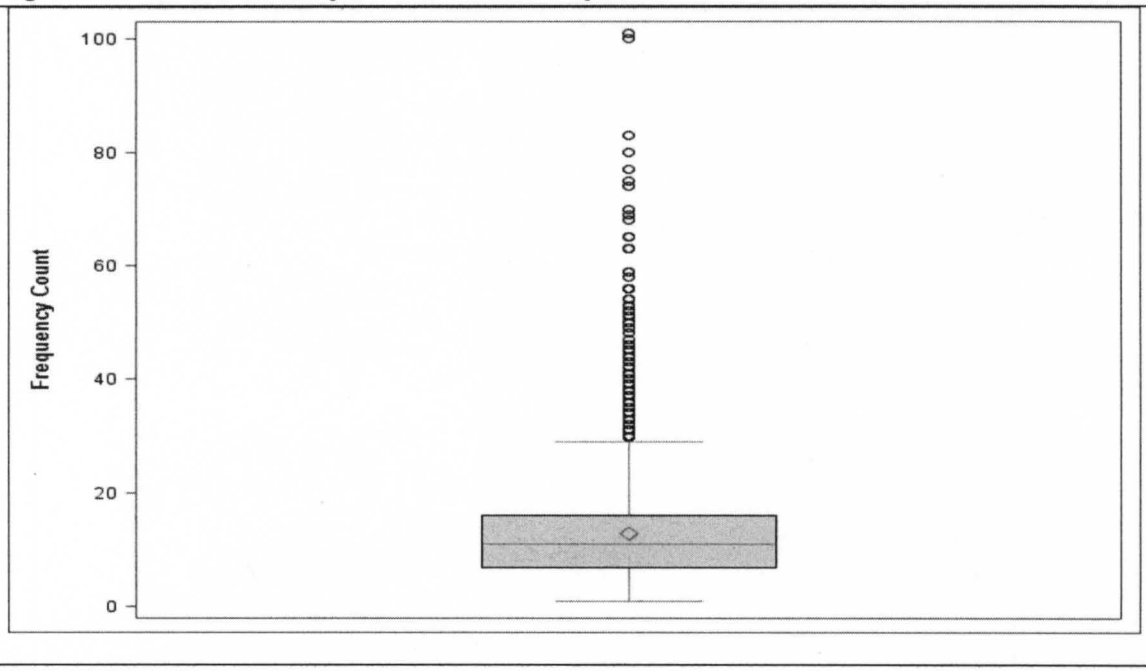


Table 12. Summary Statistics for Continuous Variables (Overall Sample).

Variable	Mean	Std. Dev	Median	Min	Max
Age	55.03	7.30	56.00	8.00	90.00
Length of Stay	4.80	5.33	3.00	1.00	120.00
Pay	23,064.57	25,653.06	16,127.50	0	478,580.32
No. of Visits	19.10	19.92	13.00	1.00	281.00
No. of Prescriptions (Before)	8.23	7.77	6.00	1.00	63.00
No. of Prescriptions (After)	12.78	8.82	11.00	1.00	101.00

Secondly, visualizing combination of two features at time is informative and simple to understand. Pair of numeric variables are visually analyzed with scatter plots to investigate the type of relationship or association between each pair [55]. The scatter plot matrix below reveals that linear relationships are nonexistent among these variables with the exception of a mild linear association between the number of outpatient visits after the first case of AMI and the number of prescription after the first case of AMI. In particular, length of stay does not have a monotonic association with age of patient.

Because this study focuses on the three main treatments of acute myocardial infarction, subgroup analysis can reveal further information about the distributions and association of the features within the dataset. The box-plots for the numerical distributions per treatment do not become more bell-shaped but remain skewed with smaller spread, though. The exception is the age of the patient, which becomes slightly more normalized within each treatment (Figures 12-17). It is also interesting that the least invasive intervention, thrombolytic therapy, has a larger spread for length of stay than the other two treatments do.

To explore categorical features, bar graphs are used. The bar graph exploring region and treatment shows a pattern of higher preference for PCI in

the northeastern and southeastern region. A chi-square test for treatments and state of hospital reveals a statically significant ($p < .001$) association between geographical area and treatment as well. Similarly, a chi-square test is significant between reinfarction and treatment ($p<.001$). In contrast, discharge status of alive or dead is not statistically significantly associated with treatment ($p = .501$).

Figure 12. Scatter Matrix of Continuous Variable.

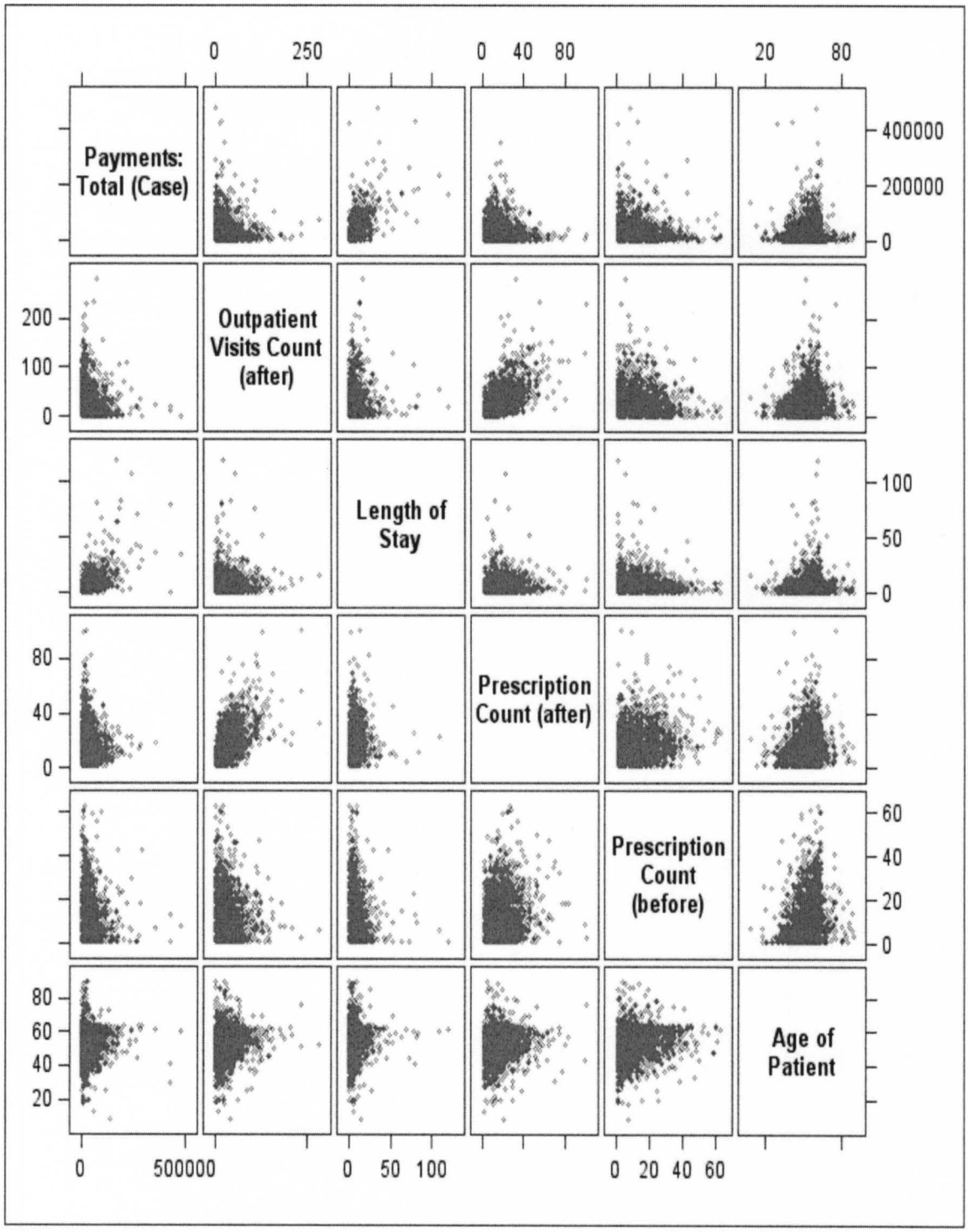


Figure 13. Age of Patient per Treatment.

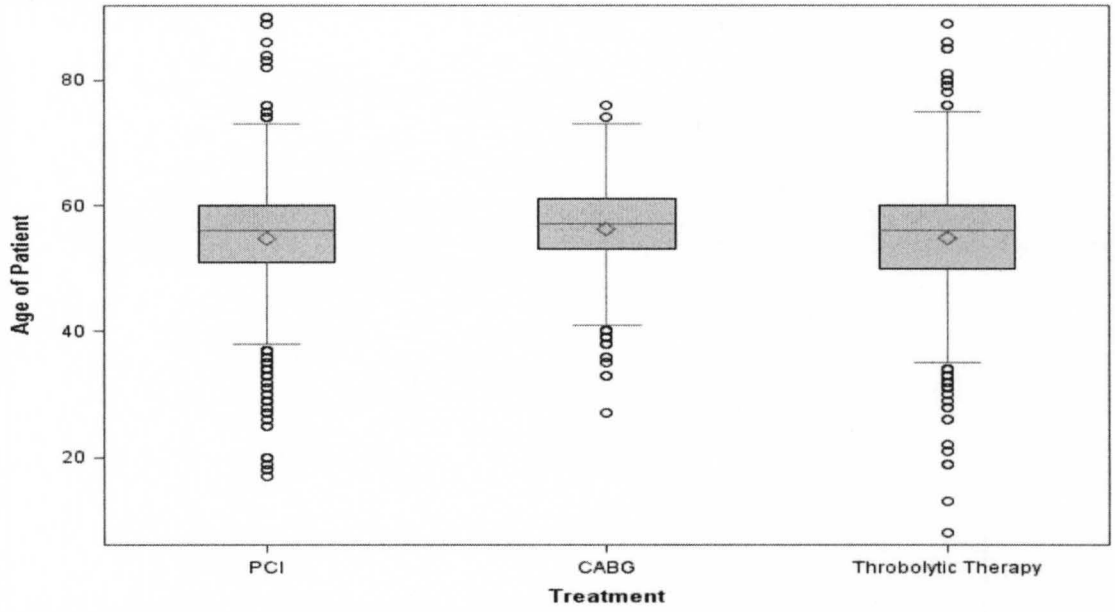


Figure 14. Length of Stay per Treatment.

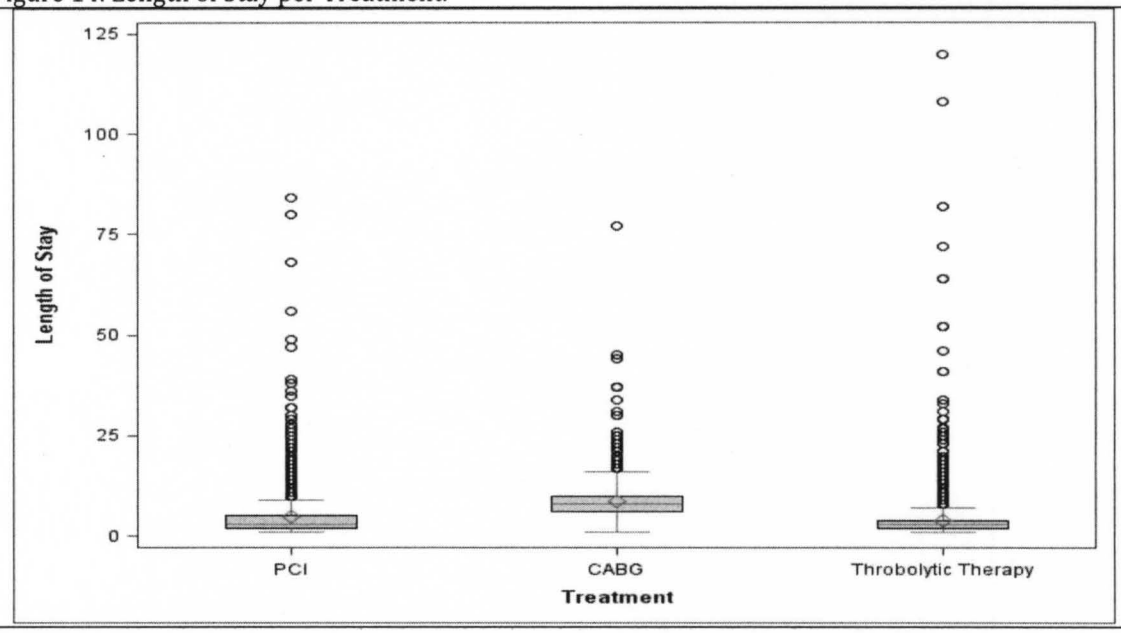


Figure 15. Total Pay per Treatment.

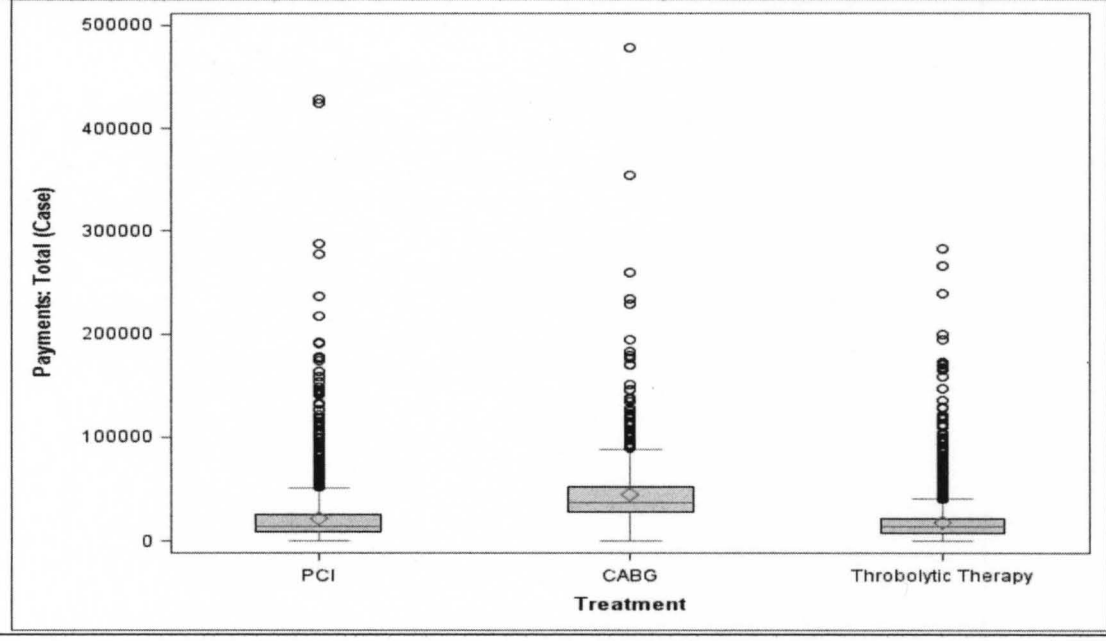


Figure 16. Number of Outpatient Visits before first episode of AMI.

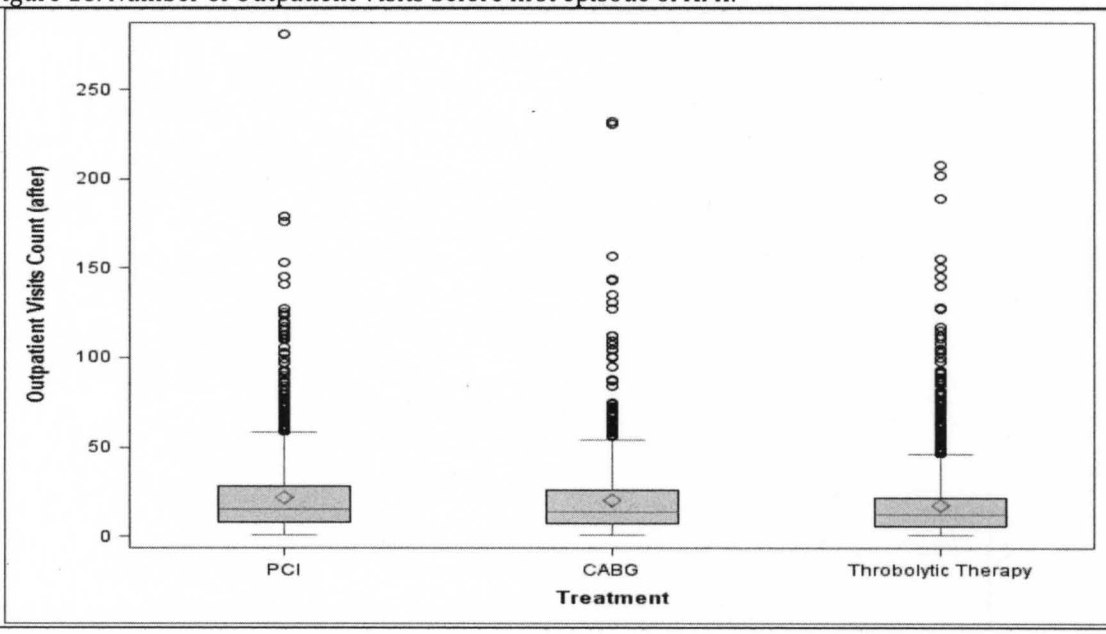


Figure 17. Number of Outpatient Visits after first episode of AMI.

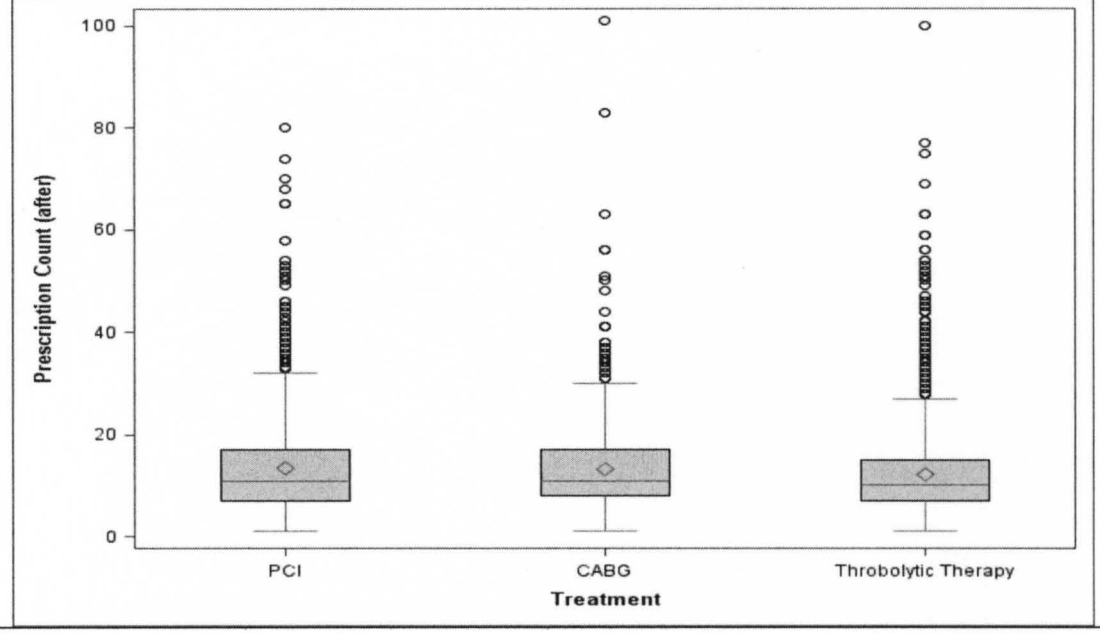


Figure 18. Number of Outpatient Visits per Month after first episode of AMI.

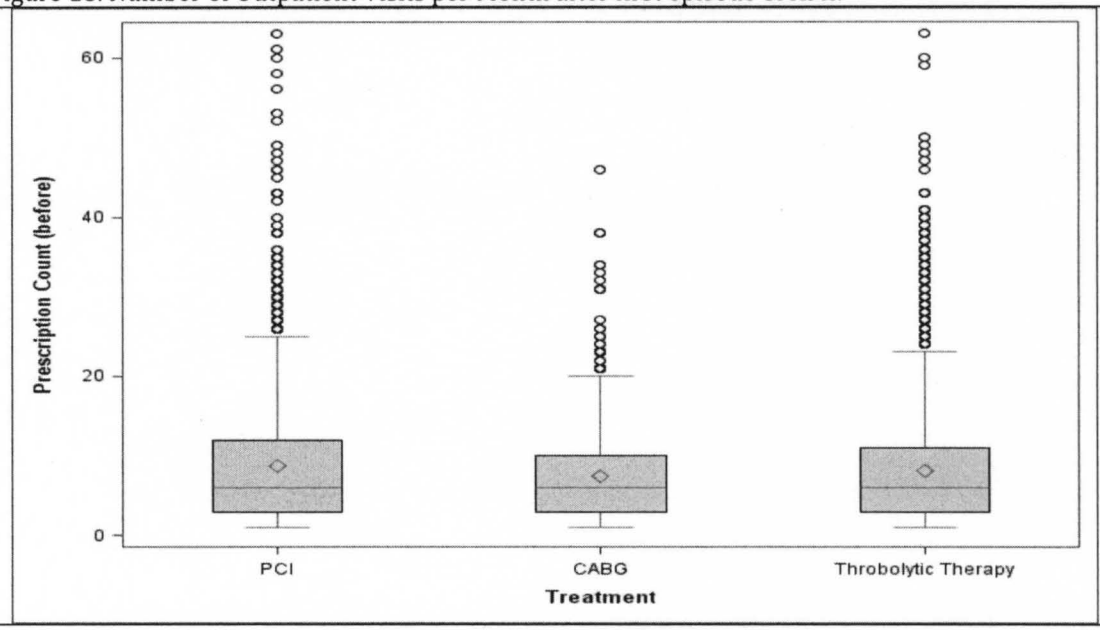


Figure 19.Distribution of Treatment by Region.

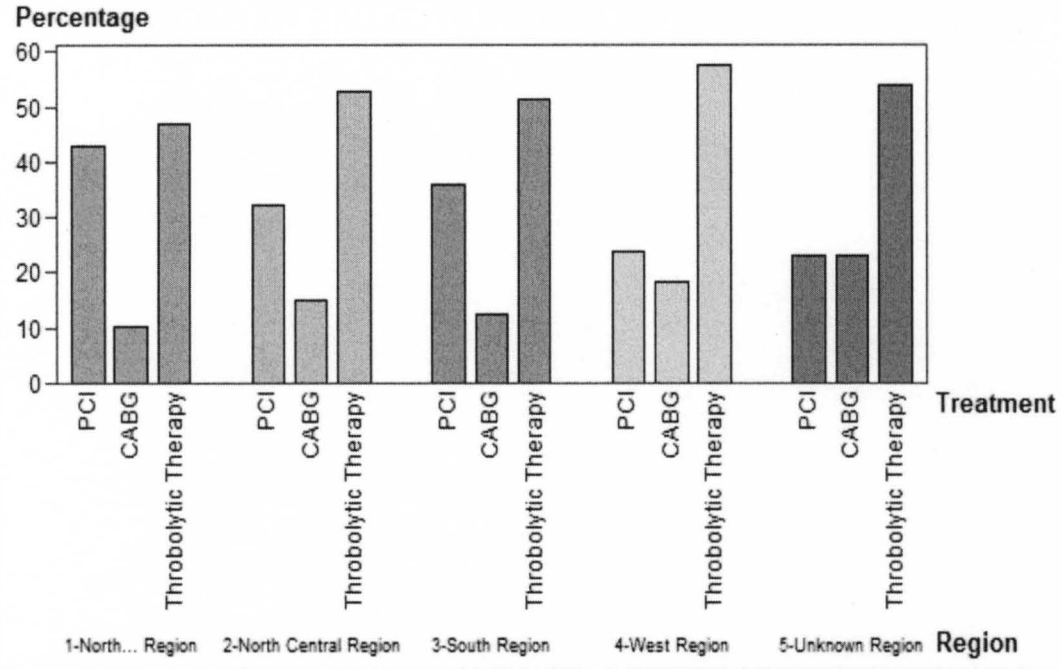


Figure 20.Distribution of Treatment by Gender.

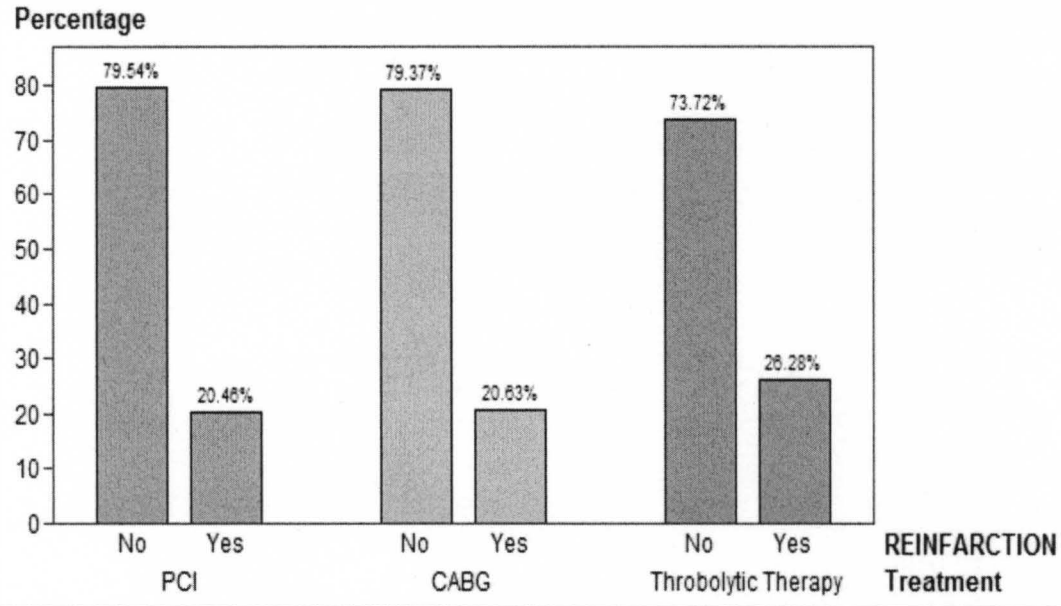
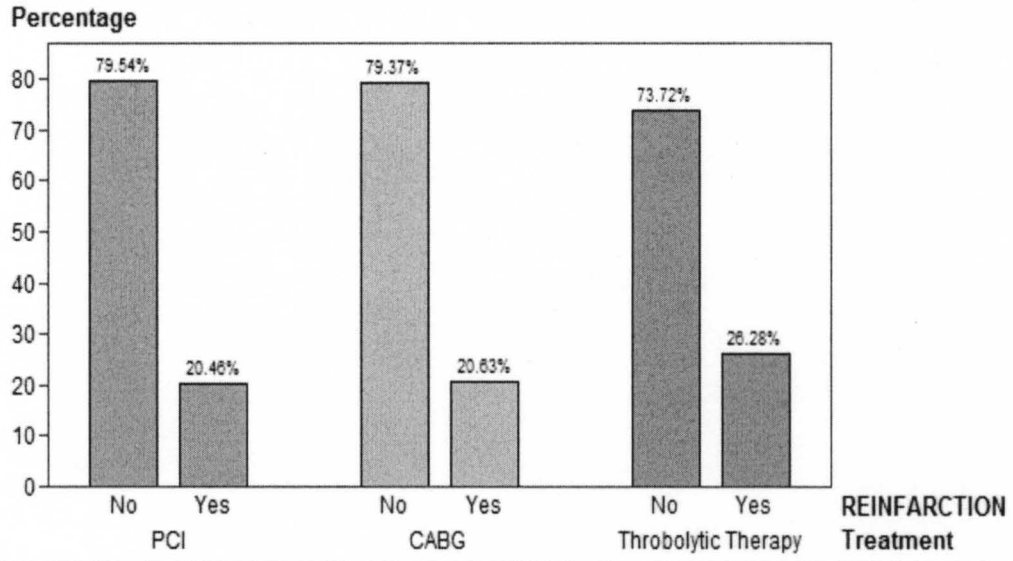


Figure 21.Distribution of Reinfarction by Treatment.



Data Cleansing

The observations corresponding to the maximum values for total pay by treatment have been removed from the sample due to the presence of numerous invalid diagnosis and procedure codes. Similarly, those observations with values of $0 and $1 for total pay have been removed. In this case, these records correspond to capitated plans unsuitable for analysis of financial variables. The three observations with largest length of stay values for thrombolytic therapy have also been removed due to a large number of missing values in demographics and invalid diagnosis codes. Since geographical location seems to play a role in the values of financial variables, observations with unknown values for state and region have been removed for building predictive models.

CHAPTER 4

Text Mining

Using text as input for classification has become very important in the past decade. As more professionals capture observations in the form of open-question surveys, computer systems and word processor have made it easy to capture the written word digitally. This phenomenon has predominantly been observed in the health care industry as more doctors document their observation in smaller handheld computers and new nearly-universal health record systems are implemented [55].

Traditionally, in health science the presence of one or more co-morbidities within a record is examined by checking for the presence of one or more of the corresponding diagnosis codes. Then, this presence within the patient's record is study individually to determine its association with a disease of interest after adjusting for previously found risk factors [55]. This methodology is cumbersome to implement to study a patient's whole health record history as many more diagnoses are added to it year after year. With text mining, all diagnoses can be compiled into a document and taken into account for classification and profiling purposes.

Text Transformation for Data Mining

One of the main issues in text mining is to represent text documents in a format suitable to data mining algorithms. First, features that represent a document must be extracted thereof. Obviously, there are thousands of features that could be extracted to represent a document such as each word, its frequency, occurrence of group of words, or occurrence of various character combinations. However, the best known representation of a document for data mining purposes is that of a vector space model [56]. In this format, each document is represented by *N significant* features X_i and the collection of documents is represented by a matrix in which each row represents a document.

The first step to transform a collection of documents is to construct a dictionary of all the terms in the training collection of documents along with its document frequency. An important part of parsing the documents is the use of stemming and stop-words to achieve feature reduction. Stemming is the use of morphological variants of the words within a document. For instance, cardiac and cardiology have the same linguistic root *cardi*; hence, they are put together and count as one item. Similarly, stop-words is a list of words that are considered useless or of low-information content such as conjunctions, adverbs, and pronouns. Whenever one of these words is found in the stop-words table, it is removed from the document's feature representation.

However, after using stemming and stop-words tables, many features representing a document still remain. The next step in feature reduction is to determine a numerical representation of these features by determining a weight

[56]. Mainly, there are two popular methodologies: binary and logarithmic. After converting the dictionary into a document-by-term table in which each cell represents the frequency of the term in the corresponding document, a weight function is applied. The binary approach assigns a value of 1 if the term appears in a document and a value of zero otherwise, which removes the effect of terms that appear more often in a document. The logarithmic approach takes the logarithm of the frequency of the term in a document, which decreases the effect of terms occurring repeatedly in a document. The last step in the feature extraction phase is to determine a feature's importance by term-weighting methods [56]. Two effective weighting methods are *entropy (1)* and *mutual information (2)*.

$$w_i = 1 + \frac{\sum_j \left(g_{ij}/h_i\right) log_2 \left(g_{ij}/h_i\right)}{log_2(m)} \quad (1)$$

where h_i is the frequency of term *i* in the collection of documents, m is the number of documents in the collection, and g_{ij} is the frequency of the term *i* in document *j*. Entropy gives greater weight to terms that rarely occur in the collection of documents. Notice that in this formula the logarithm is taken to be zero when the frequency is zero. The weights are

$$w_i = max_{C_k}\left[log\left(\frac{P(t_i \cdot C_k)}{P(t_i) \cdot P(C_k)}\right)\right] \quad (2)$$

where $P(t_i)$ is the proportion of documents that contain term t_i, $P(C_k)$ is the proportion of documents that belong to category C_k, and $P(t_i, C_k)$ is the proportion of documents that contain term t_i and belong to category C_k. Here the logarithm is taken to be 0 if $P(t_i{*}C_k)=0$ or $P(C_k)=0$. Mutual information weighting is used

when a target categorical variable is of interest from the start. These weights are proportional to the similarity of the distribution of documents that contain the term to the distribution of documents that are contained in the respective category.

In this study, the interest is to profile AMI victims according to their diagnosis record by ICD9 codes. Therefore, after constructing the documents vector representation, they are submitted to an unsupervised clustering algorithm. Four collections of documents are explored: the diagnosis history before the first AMI event recorded in the database, the collection of diagnosis recorded during hospitalization, the prescription history before the first AMI event in recorded in the database, and the prescription history after the first AMI event recorded in the database.

Results

The clustering algorithm used for the collection of diagnosis before the first AMI record in the database was expectation-maximization (EM), which assumes that a model based on mixtures approximates the distribution of the data by fitting n cluster density functions [55]. Three clusters are found with 45%, 32%, and 23% of the observations respectively (Table 13). From their descriptive terms based on ICD-9 codes they are labeled upper respiratory infections, metabolic disorders and ischemic heart disease, and anemia and non-inflammatory joint pain. One way to evaluate the homogeneity of the clusters is by looking at their standardized root mean squared, which measures the distance among the instances in the cluster (Figure 22).

Table 13. Diagnosis Cluster prior first AMI.

Cluster	Percentage	Descriptive Terms	Label	Severity Ranking
1	45%	4660 78079 53081 41090 7295 7862 4659 4619 78651 78609	Upper respiratory infections	B
2	32%	25000 25001 25002 2720 2724 4011 4111 4139 41400 41401	Metabolic disorders and ischemic heart disease	A
3	23%	2859 4019 71941 7231 78650 0 4280 7295 78609 7862	Anemia and non-inflammatory joint pain	C

The most homogenous cluster is that corresponding to those patients who have been diagnosed with metabolic disorders such as diabetes and hyperlipidemia prior their hospitalization for AMI, followed by those who were diagnosed with anemia or joint pain, those who experience upper respiratory infections respectively. Also, there were more cases of upper respiratory infections than metabolic disorders and anemia accompanied with joint pain. To examine differences within clusters, kernel density estimation for continuous variable and bar graphs for categorical variables are used on outcomes of interest. Cluster three had the highest percentage of females among the cluster which is congruent with the fact that anemia is more prevalent in women.

Figure 22. Cluster Frequency by Root Mean Squared.

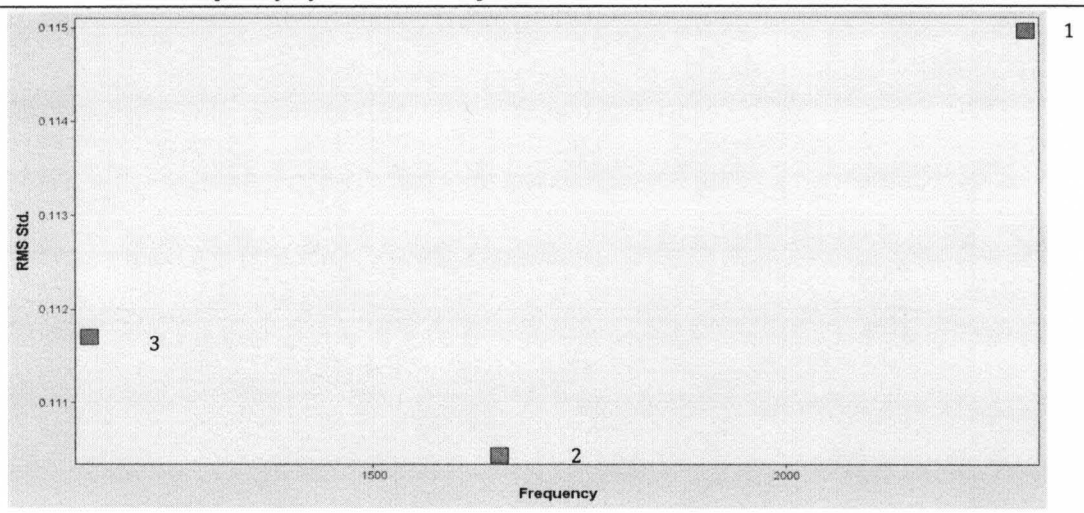


Figure 23. Gender by Diagnosis cluster.

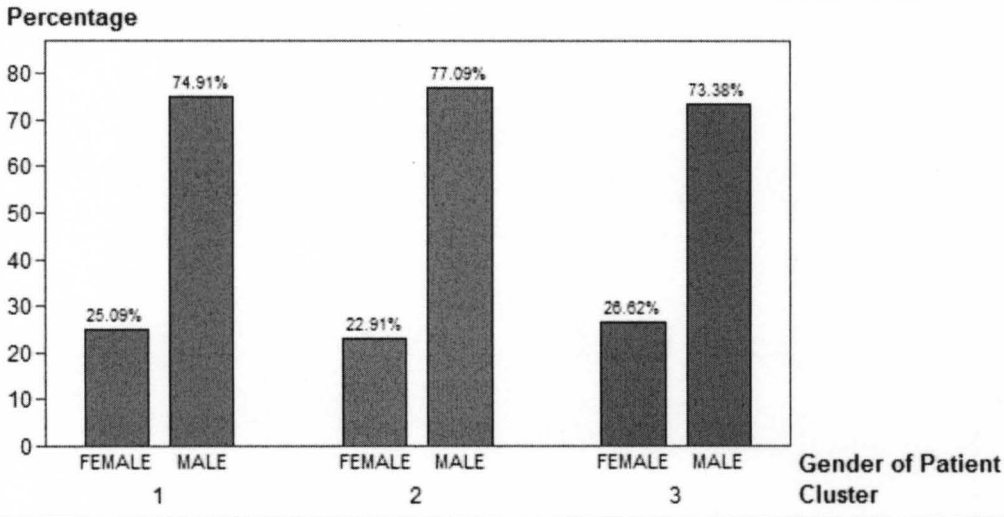


Figure 24 reveals that individuals in cluster 3 tend to be slightly younger than those of cluster 1 and 2. Figure 25, on the other hand, shows that cluster 2 has a higher probability of longer stays at the hospital during the first AMI event. Similarly, individuals in cluster 2 and 3 have a higher probability of payment $17,000 and $27,000 than those in cluster 1 (Figure 26). This supports the severity ranking from A to C where A is the highest severity and C the least.

Figure 24. Age distribution by cluster.

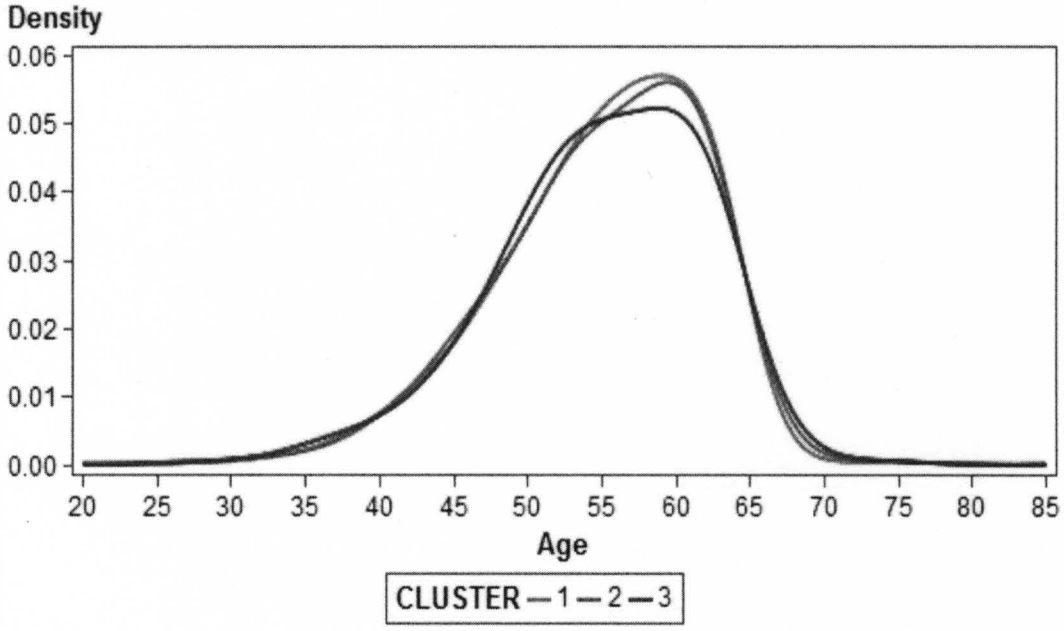
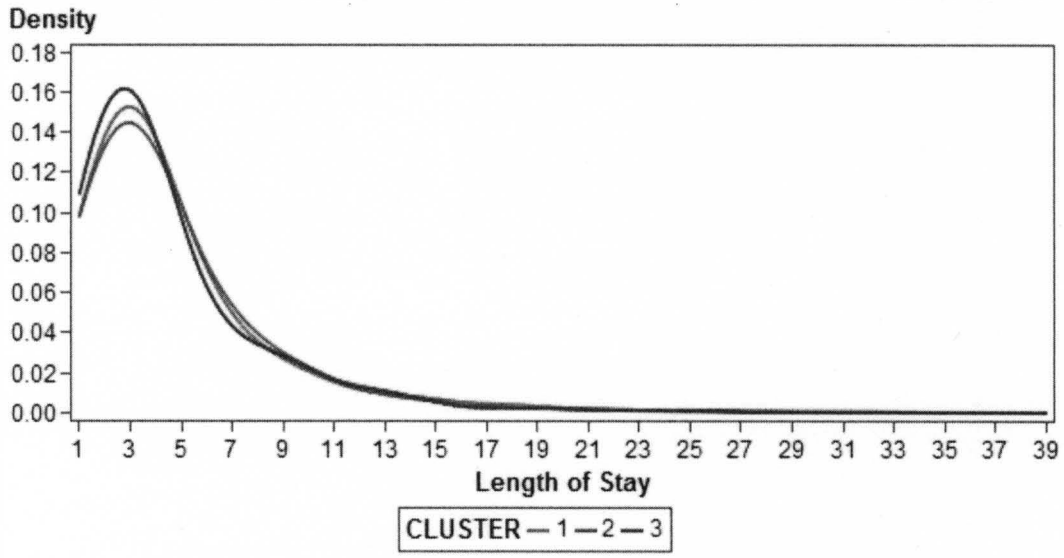


Figure 25. Length of Stay distribution by cluster.



The severity ranking is also validated when discharge status at the first episode status is examined (Figure 27). The proportion of patients who died during hospitalization is much lower for those in the lowest severity rank C compared to severity rank A and B, both of which almost identical proportions.

Figure 26. Total Pay for first AMI hospitalization distribution by cluster.

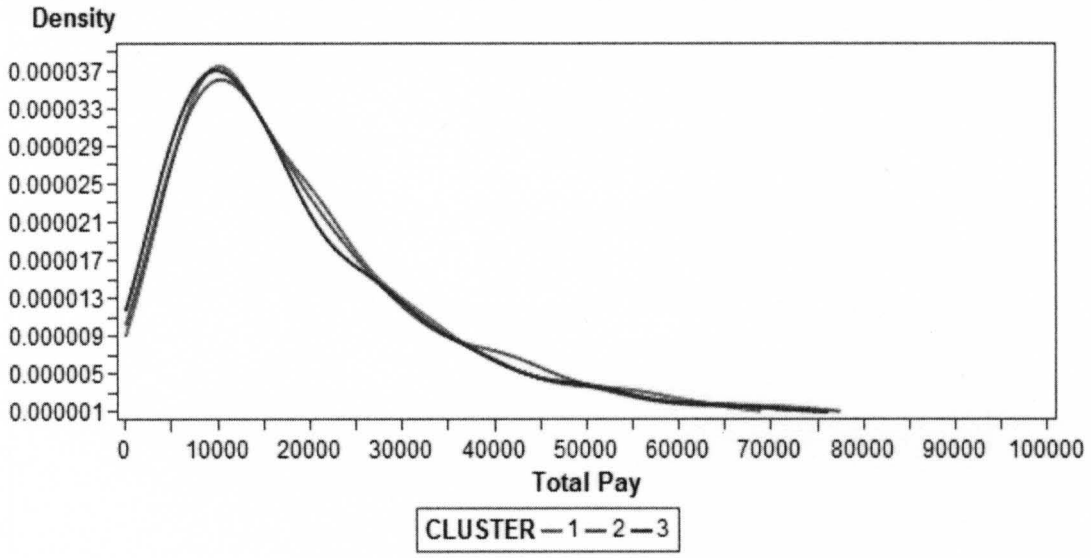
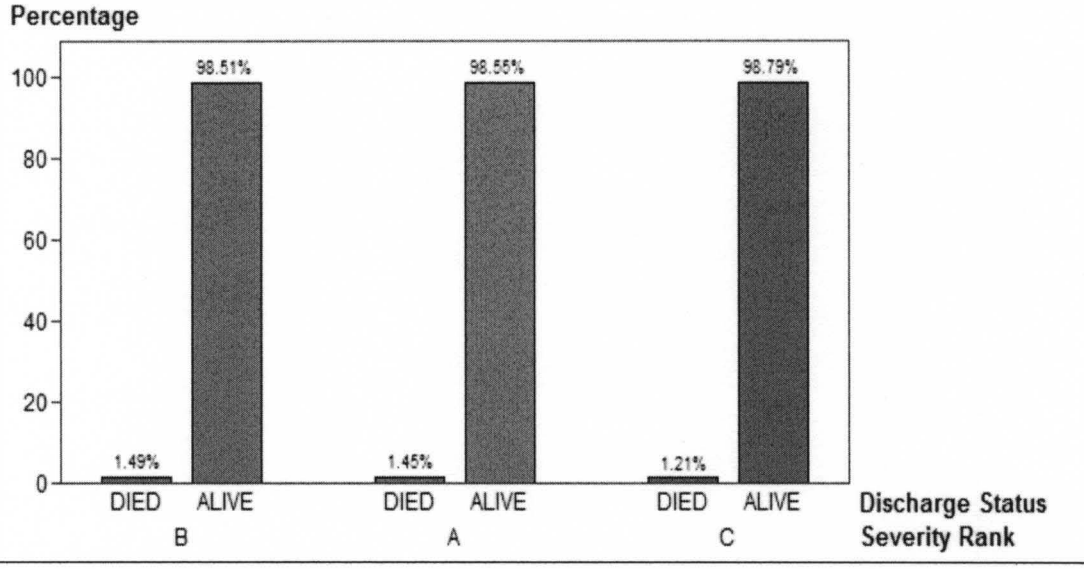


Figure 27. Discharge Status by Severity Rank.



To validate this unsupervised method, the prescription history prior the first episode of AMI was mined and clustered. It resulted in five clusters that are in line with the clusters resulting from mining the patients' diagnosis history. Upper respiratory infections were a common theme between the two along with ischemic heart disease and disease related to the metabolic system such as high blood pressure and gout.

Table 14. Prescription clusters prior first AMI

Cluster	Percentage	Diagnosis related to medications
1	31%	Chest pain, high blood pressure, skin inflammation, nausea
2	21%	HBP, HF, upper respiratory infection
3	27%	Anxiety, anemia, upper respiratory infection
4	17%	Organ transplant, osteoporosis, gout, menopause, HBP
5	4%	Anxiety, gout, skin inflammation, upper respiratory infection, indigestion

The cross tabulation of the clusters shows that diagnosis cluster one and prescription clusters two and three share many patients who were diagnosed with upper respiratory infections and were prescribed antibiotics to treat these infections (Table 15). Similarly, prescription clusters five and diagnosis cluster two have many subjects in common. Both prescription cluster three and diagnosis cluster three have a high count of subjects in their intersection.

Table 15. Cross-tabulation of Prescription clusters and diagnosis clusters.

		Diagnosis Cluster		
		1	2	3
Prescription CluPrsters	1	259	187	126
	2	181	133	74
	3	219	166	116
	4	124	117	80
	5	41	29	7

CHAPTER 5

Modeling Counts

A main goal of this study is to analyze resource utilization by patients having experienced and episode of acute myocardial infarction. In particular, number of outpatient visits following hospitalization, number of prescriptions following hospitalization, and total cost involving outpatient visits following hospitalization are of interest. The first two outcomes involve count data; these measures with highly skewed distributions are not appropriately analyzed with traditional methods such as ordinary least squares. Rather, count data must be analyzed using distributions that reflect the underlying nature of data such as Poisson regression [57]. By using count data models, significant relationships are revealed. These relationships are usually missed by ordinary least squares.

Poisson Regression for Count Data

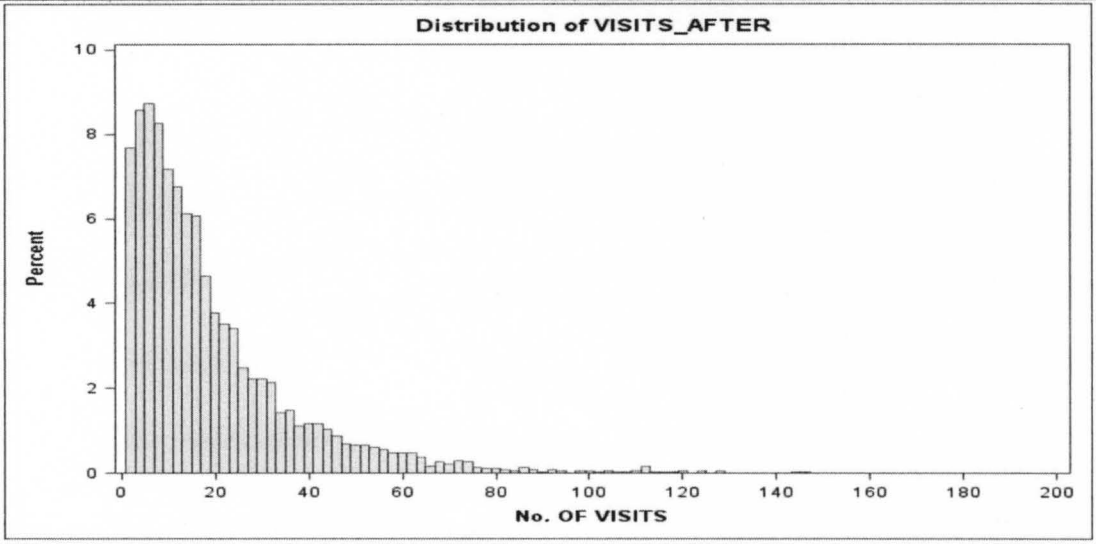
The Poisson distribution is used to model the probability of a number of events occurring in a given time frame or space-dimension [58]. This model shares many similarities with traditional statistical models such as ordinary least squares. However, ordinary least squares transform the dependent variable, outcome of interest, to normalize the residuals; instead, the log transformation used in the Poisson distribution assures that the predicted values of the dependent variable will be zero or positive as its underlying nature of count data.

Poisson regression models the parameter μ which represents both the mean and variance of the distribution. Based on the generalized linear model, the log function is used as the link function, so the log of μ is modeled as a linear combination of the parameters.

$$log(\mu) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}$$

In this study, subjects were *observed* for a period of two years during which they experienced an episode of acute myocardial infarction at different points in time. Hence, the outcomes of number of outpatient visits after hospitalization and number of prescriptions after hospitalization are not appropriate to model directly. Rather, the number of prescriptions per month and number of outpatient visits per month following hospitalization are much appropriate to model in order to account for the differences from person to person in follow up time (Figure 28, 29).

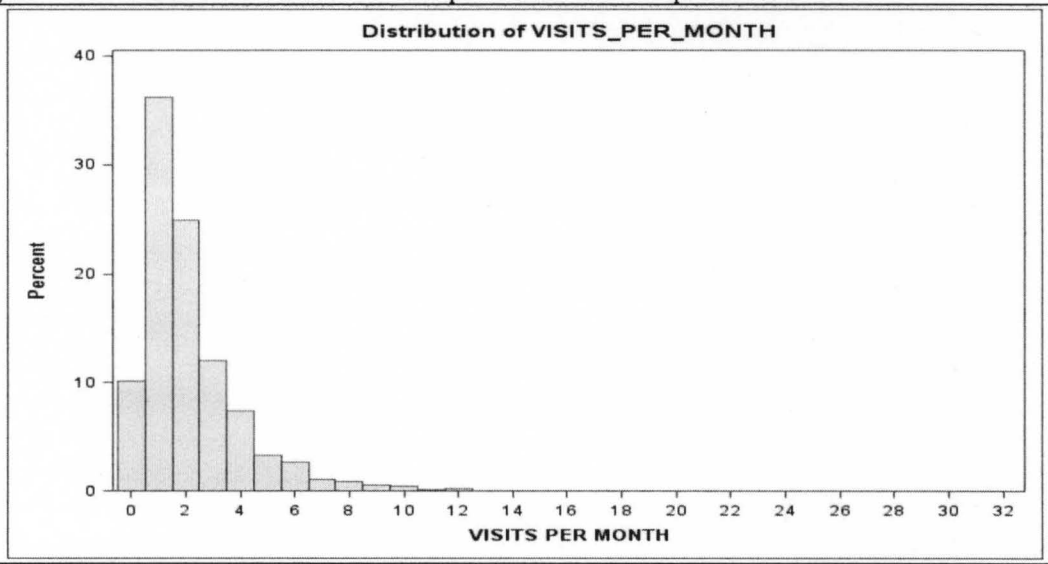
Figure 28. Distribution of number of visits after hospitalization.



Results

Another reason for modeling the number per month is the issue of overdispersion where the variance is significantly greater than the mean, which is violation of the model assumptions. From Figure 28 and 29, the distribution of number of visits has a mean of 19 and variance of 300 while the distribution of number of visits per month has a mean of 3 and variance of 3.5, which is more in line with the Poisson model.

Figure 29. Distribution of number of visits per month after hospitalization.



This same issue arises when modeling the number of prescriptions after hospitalization for acute myocardial infarction. The raw number of prescription is not suitable to analyze because of the differences in follow up time among subjects. In addition, number of prescriptions after hospitalization suffers from over dispersion (Figure 30). Therefore, number of prescriptions per month, which takes into account the differences in follow up time, will be analyzed. The

empirical distribution of number of prescriptions per month has a mean of 2.01 and variance of 3.05, Figure 31, while the distribution of the number of prescriptions after hospitalization have a mean of 12.77 and variance 77.69 in violation of Poisson model.

Figure 30. Distribution of prescription counts after hospitalization.

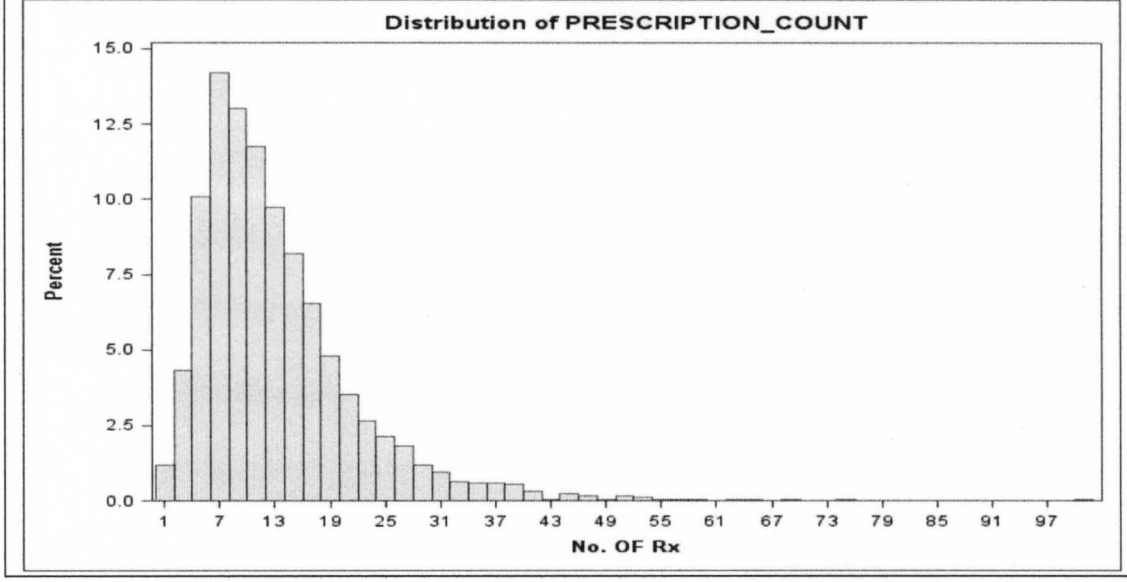
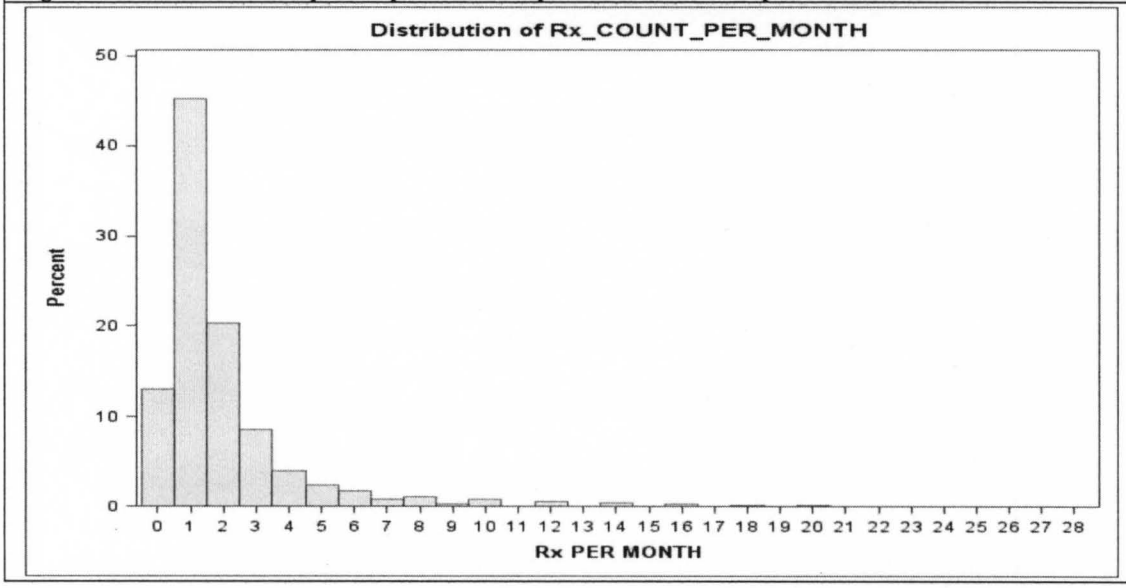


Figure 31. Distribution of prescription counts per month after hospitalization.



After fitting the number of outpatient visits per month using Poisson regression and maximum likelihood estimation, the mode estimates are found in Table 24 while the model goodness of fit is found in Table 16. In contrast to most studies based on clinical data, age is not found to be significant. This implies that in practice the age of a patient does not play a role in predicting the number of outpatient visits as a measure of health care resource utilization. The statistics for goodness of fit are deviance and Pearson's chi-square with values of 0.85 and 1.02 after dividing by degrees of freedoms respectively. The closer to one the deviance is, the better the fit of the model. This means the model is an excellent fit for explaining the relationship between the predictors and the outcome.

Table 16. Analysis Of Maximum Likelihood Parameter Estimates for Visits per Month count.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	P-value
Intercept		1	3.1526	0.1069	869.99	<.0001
AGE		1	-0.0007	0.0005	2.04	0.1529
DAYS		1	0.0155	0.0005	1006.10	<.0001
TRT	CABG	1	0.1365	0.0108	160.00	<.0001
TRT	PCI	1	0.1789	0.0098	329.89	<.0001
TRT	Thrombolytic Therapy	0	0.0000	0.0000	.	.
REGION	Northeast Region	1	0.2180	0.1031	4.47	0.0346
REGION	North Central Region	1	-0.0867	0.1028	0.71	0.3992
REGION	South Region	1	-0.0743	0.1029	0.52	0.4703
REGION	West Region	1	0.0916	0.1035	0.78	0.3762
REGION	Unknown Region	0	0.0000	0.0000	.	.
SEX	Male	1	-0.1710	0.0080	456.80	<.0001
SEX	Female	0	0.0000	0.0000	.	.
INDSTRY	Oil & Gas Extraction, Mining	1	-0.5358	0.0741	52.34	<.0001
INDSTRY	Manufacturing, Durable Goods	1	-0.2200	0.0109	407.46	<.0001
INDSTRY	Manufacturing, Nondurable Goods	1	-0.1329	0.0146	82.45	<.0001
INDSTRY	Transportation, Communications	1	-0.1520	0.0128	141.15	<.0001
INDSTRY	Retail Trade	1	-0.5415	0.0667	65.99	<.0001
INDSTRY	Finance, Insurance, Real Estate	1	-0.2289	0.0199	132.44	<.0001

Table 17. Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	4384	3719.2319	0.8484
Pearson Chi-Square	4384	4388.0000	1.0209

Table 18. LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
AGE	1	4384	0.17	0.6772	0.17	0.6772
DAYS	1	4384	67.25	<.0001	67.25	<.0001
TRT	3	4384	11.99	<.0001	35.97	<.0001
REGION	4	4384	20.14	<.0001	80.58	<.0001
SEX	1	4384	37.85	<.0001	37.85	<.0001
INDSTRY	6	4384	7.05	<.0001	42.31	<.0001

The resulting Poisson model for the number of prescriptions after hospitalization is summarized in tables 28 and table 29. The deviance and Pearson's chi-square have values of 0.9203 and 1.002 after dividing by degrees of freedom respectively. Again, this model fits the data very well. The estimates table reveals that the explanatory variables influence the number of visits per month after hospitalization for acute myocardial infarction.

Table 19.Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	4215	3879.130	0.9203
Pearson Chi-Square	4215	4225.000	1.002

Table 20. LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
AGE	1	4215	6.24	0.0125	6.24	0.0320
TRT	3	4215	5.08	0.0016	15.25	0.2568
REGION	4	4215	3.82	0.0042	15.28	0.0039
SEX	1	4215	205.24	<.0001	205.24	<.0001
INDSTRY	6	4215	5.22	<.0001	31.35	<.0001

Table 21. Analysis Of Maximum Likelihood Parameter Estimates.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	P-value
Intercept		1	2.8548	0.2423	138.81	<.0001
AGE		1	0.0032	0.0013	6.19	0.0128
DAYS		1	0.0138	0.0168	82.56	<.0001
TRT	CABG	1	0.0771	0.0256	9.07	0.0026
TRT	PCI	0	-0.0288	0.0246	1.36	0.2428
TRT	Thrombolytic Therapy	1	0.0000	0.000	.	.
REGION	Northeast Region	1	-0.2543	0.2312	1.21	0.2714
REGION	North Central Region	1	-0.2289	0.2299	0.99	0.3196
REGION	South Region	1	-0.1700	0.2303	0.55	0.4604
REGION	West Region	1	-0.2755	0.2324	1.41	0.2358
REGION	Unknown Region	0	0.0000	0.0000	.	.
SEX	Male	1	-0.2872	0.0196	213.59	<.0001
SEX	Female	0	0.0000	0.0000	.	.
INDSTRY	Oil & Gas Extraction, Mining	1	-0.0377	0.1513	0.06	0.8032
INDSTRY	Manufacturing, Durable Goods	1	0.0125	0.0276	0.20	0.6510
INDSTRY	Manufacturing, Nondurable Goods	1	-0.0171	0.0375	0.21	0.6481
INDSTRY	Transportation, Communications	1	-0.0810	0.0332	5.94	0.0148
INDSTRY	Retail Trade	1	-0.1220	0.1597	0.58	0.4446
INDSTRY	Finance, Insurance, Real Estate	1	-0.1976	0.0526	14.09	0.0002
INDSTRY	Services	0	0.0000	0.0000	.	.

On the other hand, modeling the number of outpatient visits per month with ordinary least squares fails to find many significance relationships among the predicting features and the outcome of interest. Table 31 displays the results

of fitting the number of outpatient visits per month after hospitalization. Clearly, all the covariates are found non-significant for explaining the number of outpatient visits. This highlights the utility of modeling count data with large numbers of single digit counts and overdispersion with appropriate distribution that match the underlying nature of the data rather than forcing a model for convenience.

Table 22. Analysis Of Ordinary Least Square Parameter Estimates.

Effect		Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		2.7451	0.9276	4214	2.96	0.0031
AGE		0.006467	0.005165	4214	1.25	0.2106
DAYS		0.01718	0.008152	4214	2.11	0.0351
SEX	1-Male	-0.6717	0.08420	4214	-7.98	<.0001
SEX	2-Female	0	.	.	.	.
REGION	1-Northeast Region	-0.7898	0.8815	4214	-0.90	0.3703
REGION	2-North Central Region	-0.5850	0.8766	4214	-0.67	0.5046
REGION	3-South Region	-0.5184	0.8777	4214	-0.59	0.5548
REGION	4-West Region	-0.8129	0.8859	4214	-0.92	0.3589
REGION	Missing/Unknown	0	.	.	.	.
INDSTRY	1-Oil & Gas Extraction, Mining	0.9304	0.5415	4214	1.72	0.0858
INDSTRY	2-Manufacturing, Durable Goods	0.4143	0.1133	4214	3.66	0.0003
INDSTRY	3-Manufacturing, Nondurable Goods	-0.00083	0.1551	4214	-0.01	0.9957
INDSTRY	4-Transportation, Communications, Utilities	-0.09138	0.1365	4214	-0.67	0.5033
INDSTRY	5-Retail Trade	0.1437	0.5566	4214	0.26	0.7963
INDSTRY	6-Finance, Insurance, Real Estate	0.06654	0.1967	4214	0.34	0.7351
INDSTRY	7-Services	0	.	.	.	.
TRT	Missing/Unknown	-0.7512	0.09609	4214	-7.82	0.05
TRT	CABG	-0.1515	0.1099	4214	-1.38	0.05
TRT	PCI	-0.5940	0.09820	4214	-6.05	0.05
TRT	Thrombolytic Therapy	0	.	.	.	.

CHAPTER 6

Predictive Modeling

Predictive modeling is concerned with problems of classification and estimation of future outcomes [55]. The learning style is considered supervised learning in that instances or records with known output attributes, also known as dependent variables in classical statistics, and known input attributes, independent variables, are submitted to the learning algorithm and a correction is made when the predicted outcome does not match the observed outcome. In this study, health outcomes and health care utilization measures are of interest. Thus far, number of outpatient visits per month and number of prescriptions per month after hospitalization have been modeled with generalized linear models for count data. Now, the health outcomes of reinfarction after hospitalization for acute myocardial infarction and death during hospitalization for acute myocardial infarction are undertaken. Also, monthly cost for outpatient visits after hospitalization for AMI is modeled.

The outcome of reinfarction was not directly observed in the database, but rather was derived during the data exploration phase of the project. The variable reinfarction was encoded in binary format with a value of 1 indicating reinfarction after first hospitalization and a value of zero if no reinfarction was found during the follow-up period. On the other hand, death during hospitalization was

explicitly part of the instance for each inpatient admission record within the attribute *discharge status*. The MarketScan database contains a table without patient service records that corresponds to each procedure performed during the outpatient visits. Therefore, the total outpatient costs per month was not directly recorded but had to be derived as well. The outpatient service table was filtered to contain records that corresponded to inpatient records and that took place after the date of discharge for the hospitalization due to AMI.

Logistic Regression

Reinfarction after hospitalization and death during hospitalization are binary outcomes which lend themselves to be modeled with a generalized linear model, specifically logistic regression. Logistic regression models the natural logarithm of the odds of an event [55]. In other words, it models the natural logarithm of the ratio of the expected number of times an event will occur to the expected number of times an event will not occur as linear combination of the parameters and input attributes (Equation 1).

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \quad (1)$$

where p_i is the probability that $y_i = 1$. When this equation is solved for p_i, one obtains

$$p_i = \frac{1}{1+\exp\left(-\alpha - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik}\right)} \quad (2)$$

which is the logistic function with values between zero and one for all real numbers. This makes the logistic regression ideal to model probability values which are between zero and one respectively. Advantages to this model are that

the parameter estimates are simple to interpret in terms of odds ratios. Furthermore, it has desirable sampling properties that were discussed in chapter two.

Decision Trees

Another data mining predictive model is implemented by decision trees. In particular for this study, binary decision trees are implemented. A key characteristic of decision trees is that they only use those attributes that best differentiate the *concepts* to be learned [55]. The idea behind it is that of processing the data as a sequence of simple questions, and their answers determine what the next question is, if any. Each *question* is represented by a node that is connected to a following node, *child link*, in the shape of an *inverted tree*. Intrinsically, the construction of decision trees has cross-validation embedded in the algorithm. This means that at the starting point a subset of instances is selected for training and another subset is selected for testing. Consequently, the algorithm assesses the goodness of fit more accurately, possibly at each stage of the building process. At a high level, most decision trees are constructed with an algorithm as follows:

1. Let K be a set of training observation.
2. Select a variable that best differentiates the observations within K.
3. Make a tree node represented by the selected variable:
    a. Make child nodes representing unique values or interval for the selected variable.

b. Use the child nodes to further subdivide the observation into smaller classes.

4. For each class in step three:
   a. If the observations satisfy predetermined criteria, specify the classification for new observations following this decision path.
   b. Otherwise, there must exist at least a variable the further subdivides the path of the tree. In this case, return to step 2.

The splitting rule at each node is a key element that distinguishes one decision tree algorithm from another. Commonly implemented rules are *entropy*, explained in chapter four, Chi-square test for classification, and F-test for estimates of continuous outcome variables. Another feature that defines a decision tree algorithm is its *stopping rule*. Some common stopping rules are minimum number of cases, a certain fraction of number of instances, achieving a maximum number of levels of splitting, or reaching a maximum number of nodes. Furthermore, conditions under which an algorithm cannot further split a node are: cases are all duplicates of one another, only one observation is left, and complete agreement of all target values.

Decision trees have many advantages as predictive models. They tend to be nonparametric or semi-parametric, and data does not need to specifically follow a probability distribution. Attributes for the model are not preselected but rather selected automatically by the algorithm. Data transformation is not necessary to feed the data to the algorithm, and mathematical monotonic transformations do not affect the outcome. However, despite not being affected

by outliers in the input space, decision trees algorithms are affected locally by outliers in the output space.

Neural Networks

Neural networks are another data mining method for prediction that has gained wider acceptance in the field of medical science and the business field of insurance. They were conceived on the early understanding of the human brain. Its building block is termed a *neurode*, a mathematical representation of the human brain's neurons [59]. These are connected in a network grid, usually fewer than one hundred, to imitate the functioning of the human brain, which in contrast contains billions thereof. The current understanding of neuron cells is that they receive electrical impulses from cells around them and accumulate these charges until a threshold is surpassed. Then, the neuron cell *fires* an electrical impulse to neighboring cells. These actions are controlled by biochemical processes changing over time with aging of the autonomic nervous system, which is in charge of the means humans learn. In contrast, artificial neurons connected in networks imitate these processes by numerical aggregation and mathematical functions, and knowledge is represented as set of layers of interconnected neurodes.

Network architecture, or shape, and learning algorithm characterize most modern neural networks [59]. In this study only the neural network commonly known as the *perceptron* is implemented for supervised learning. At least two layers of neurodes are required, one for inputs and one for outputs. However,

the strength of neural networks lies in their ability to handle nonlinear relationships which are achieve by adding layers of neurodes, known as hidden layers, between the inputs and output(s) layers.

The perceptron is a fully connected feed-forward neural network. *Fully connected* indicates that neurodes in one layered are connected to all neurodes in the next layer. *Feed-forward* means that input instances flow in one direction from the input layer towards the output layer and never backwards. In addition, neurodes in a common layer are never connected to each other. The connections between pairs of neurodes represent weights which model the relationships between the input neurodes and output node(s).

A particular issue with neural networks is input and output representation. Input nodes must be numeric and normalized or standardized between values of zero and one or between values of negative one and positive one respectively. Basically, there are two ways to represent categorical data. One method is to divide the interval from zero to one into equal size subintervals corresponding to the various categories of an attribute. However, this methodology embeds an ordinal structure between categories which may not exist, such as color and gender. Another method uses binary representation for categories. In this case, the number of neurodes required to represent a feature is equal to m - 1, where m is the number of categories within an attribute. The transformation of continuous attribute is usually handled by *normalization*, described below:

$$NewInput = \frac{InputValue - MinimumValue}{MaximumValue - MinimumValue} \quad (3)$$

where

- *NewInput* is the computed value to feed the network.
- *InputValue* is the original value to be converted.
- *MinimumValue* is the smallest possible value for the variable.
- *MaximumValue* is the largest possible value for the variable.

Similarly, output format has the same issues. However, the solutions are the same. For categorical variables, binary representation tends to work best, and continuous data is converted back to scale by reversing the process in equation (3):

$$OutputValue = NeurodeOutput \times (MaximumValue - MinimumValue) + MimumValue$$

During the feed-forward phase, the neural network accepts input values. The neurodes of the input layer pass these values to the following hidden layer unchanged. In turn, the neurodes in the hidden layer takes in the inputs and weights them with the corresponding weights between each pair of neurodes and combines them additively. This sum becomes the input to the neurode's evaluation function, usually the sigmoid function (4), which outputs a value close to one when *sufficiently excited* and a value close to zero otherwise.

$$f(x) = \frac{1}{1 + \exp(-x)} \quad (4)$$

This activity continues through each layer of neurodes until an output for each output node is produced; then the back-propagation phase takes place. During this phase, the output error is determined; that is the difference between the estimated value and the observed value is computed. If an error is observed, it is assumed that every network link associated with the output is at fault to

some degree. This triggers a modification in weight values beginning at the output layer and moving backwards through the hidden layers. The error at *target* nodes is computed as follows:

$$Error(t) = (t - o_t) \times f'(x_t)$$

where

t = target output

o_t = computed output

$f'(x_t)$ = first-ordered derivative of the sigmoid function

In turn, the error in a hidden layer neurode *k* is computed as follows:

$$Error(k) = \left(\sum Error(j) W_{kj}\right) f'(x_k)$$

where

Error(j) = computed output error at node j

W_{kj} = the weight between nodes k and j

$f'(x_k)$ = the first-order derivative of the sigmoid function

x_k = the input to the sigmoid function at node k

Finally, the appropriate weight correction at each node during the back-propagation is computed as follows:

$$W_{kj}(new) = W_{kj} + r \times Error(j) \times o_k$$

Where

W_{jk}(new) = the new weigh between nodes j and k

W_{jk} = the current weight between nodes j and k

r = learning rate

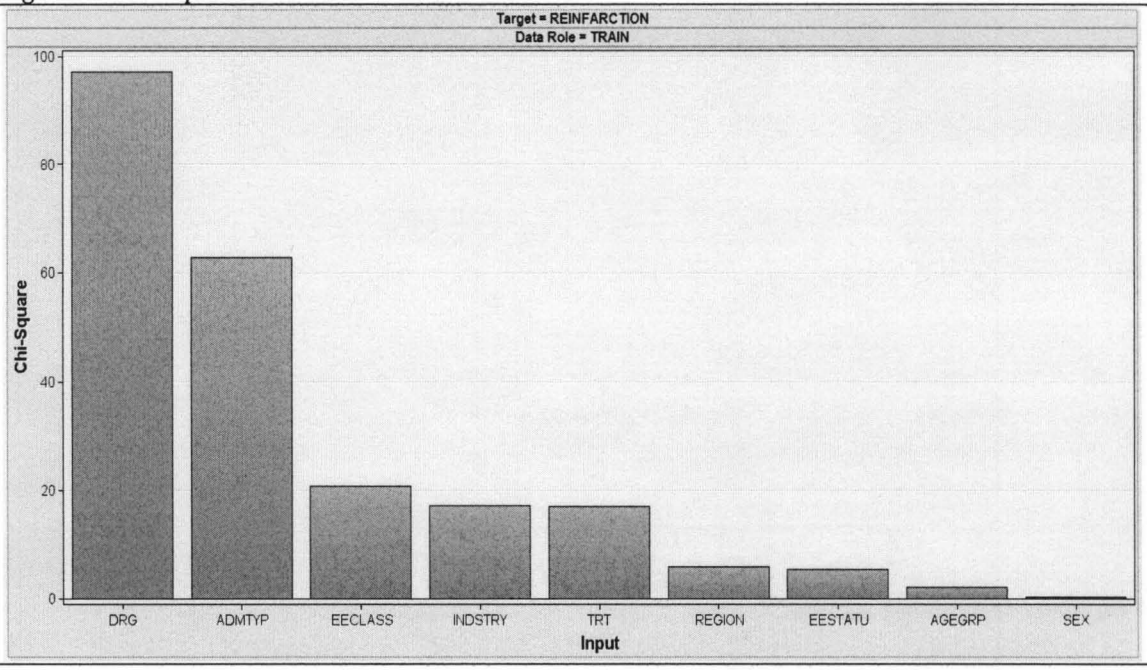
Error(j) = computed error at node j

o_k = output of node k

Feature Selection and Sampling

In this section of the study, the goal is to produce effective predictive models for reinfarction and monthly outpatient costs. A key element is building predictive models is feature selection. One crude method is to start by using the chi-square statistic; another is by fitting a decision tree and using the resulting variables worth, which is equivalent to the sum of square errors [55]. Figure 31 displays the results for the chi-square. Diagnosis related group, which is a code system developed by Centers for Medicare & Medicaid Services for reimbursement purposes, is highly associated with reinfarction. It is closely followed by admission type. Much less associated, by this measure, are employee's classification (i.e. full-time, union, etc..), industry, and treatment. However, industry will not be used as some observations correspond to employee's dependents and a different approach should be followed. It is worth noticing is that age and gender are not very associated with the outcome, in contrast to most clinical studies. Three predictive models are developed for the outcome of reinfarction: logistic regression, decision tree, and multilayer perceptron.

Figure 32. Chi-square in relation to reinfarction.



Another issue in predicting reinfarction is its rare occurrence. Most predictive models are very sensitive to rare occurrences [55]. To circumvent this weakness, careful sampling must take place. In this case less than eight percent of the instances have a positive indicator for reinfarction. Therefore, all these instances must be included in the sample for fitting and validation (n=366). Another sample of 366 instances with a negative indicator for reinfarction must be taken. Since AMI treatment is of interest, this random sample will be stratified by treatment so that it remains representative. Finally, these two samples are merged and submitted to the algorithms (n=732).

Results

The decision tree implemented is a binary tree. Its splitting rules are based on entropy for categorical variables and minimum variance for continuous

data with a threshold of 0.2. It resulted on two splits. The first split was based on diagnosis related group, and the second on length of stay during hospitalization for AMI. The resulting *English Rules* that characterized the resulting tree are in Table 23. Its misclassification rate for the outcome of reinfarction was 0.35 in the training set and 0.39 in the validation set. This means that the tree generalizes well.

Figure 33. Decision Tree for Reinfarction.

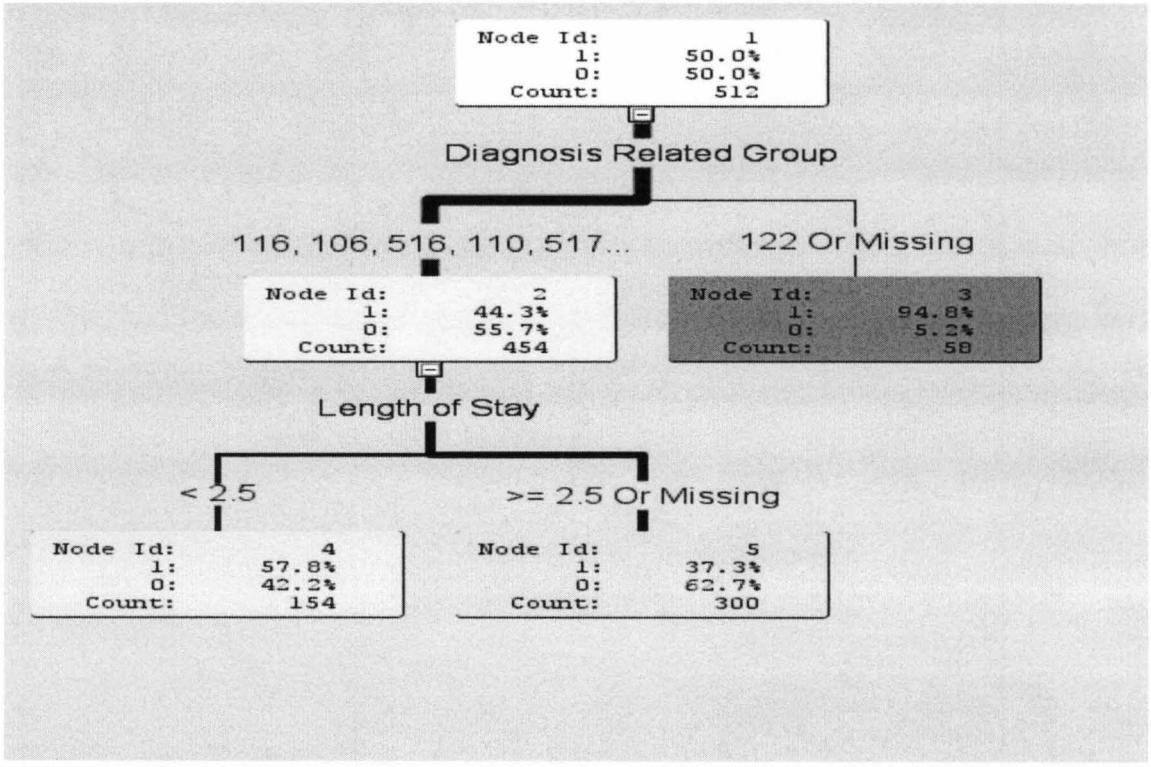


Table 23. Decision Tree – English Rules.

Node 3
if Diagnosis Related Group IS ONE OF: 122 or MISSING then Tree Node Identifier = 3 Number of Observations = 58 Predicted: REINFARCTION=1 = 0.95 Predicted: REINFARCTION=0 = 0.05
Node 4
if Length of Stay < 2.5 AND Diagnosis Related Group IS ONE OF: 116, 106, 516, 110, 517, 112, 107, 109, 104 then Tree Node Identifier = 4 Number of Observations = 154 Predicted: REINFARCTION=1 = 0.58 Predicted: REINFARCTION=0 = 0.42
Node 5
if Length of Stay >= 2.5 or MISSING AND Diagnosis Related Group IS ONE OF: 116, 106, 516, 110, 517, 112, 107, 109, 104 then Tree Node Identifier = 5 Number of Observations = 300 Predicted: REINFARCTION=1 = 0.37 Predicted: REINFARCTION=0 = 0.63

The logistic regression model was first fit with stepwise forward selection, but it did not find any significant variables at the 0.05 significance level. Backward selection had the same results. Thus, the model was fitted the variables of importance from the decision tree: diagnosis related group and length of stay. In addition, AMI treatment and gender were included. However, none of these variables were found to have a significant effect on the outcome of interest: reinfarction. This is congruent to the findings of stepwise-forward selection. Table 24 displays the results from fitting this model. This model had a misclassification rate of 0.37 in the training set and 0.45 in the validation set. This indicates that the logistic regression model does not generalize well.

Table 24. Logistic Regression Analysis of Effects

Effect	DF	Chi-Square	P-value
Length of Stay	1	1.642	0.2001
DRG	17	21.98	0.1854
Sex	1	1.493	0.2837
Treatment	2	0.5395	0.7636

The multilayer perceptron architecture consisted of a layer of inputs for diagnosis related codes, length of stay, treatment, number of visits per month before and after hospitalization, age, and number of prescriptions before and after hospitalization. In addition, a hidden layer with three neurodes was implemented between the input layer and the output layer. The activation function for hidden layer nodes was the logistic function. Since the outcome was dichotomous, the binomial activation function was used on output nodes. The resulting neural network had a misclassification rate of 0.221 in the training set and 0.385 in the validation set. This may indicate that over fitting may have occurred; however, the neural network generalizes well when compared to the other two models (Table 25).

Table 25. Misclassification Rates for Predictive Models of Reinfarction.

Model	Misclassification Rate (Training Set)	Misclassification Rate (Validation Set)
Decision Tree	0.352	0.390
Logistic Regression	0.367	0.455
Neural Network	0.221	0.385

During data exploration, it was found that outpatient costs per month after hospitalizations were not bell-shaped distributed but rather skewed. For modeling purposes, a data transformation is applied. The natural log is applied to the outpatient costs per month and clearly a bell-shape distribution is achieved

(Figure 34). After determining correlations, it is found that total pay during hospitalization, length of stay during hospitalization, and age are positively correlated to outpatient costs (Figure 35).

Figure 34. Distribution for logarithmic transformation for outpatient costs.

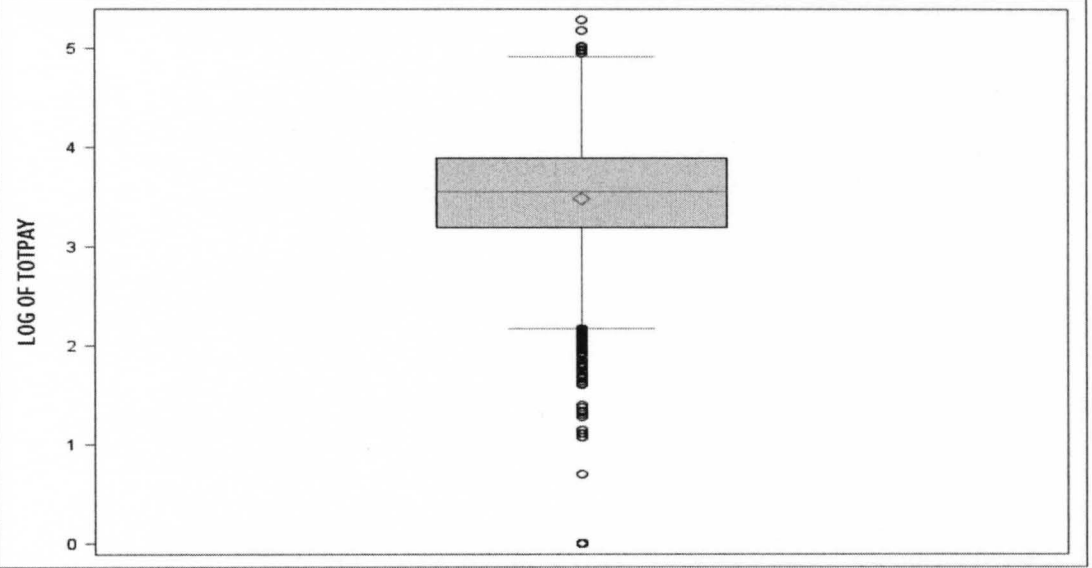
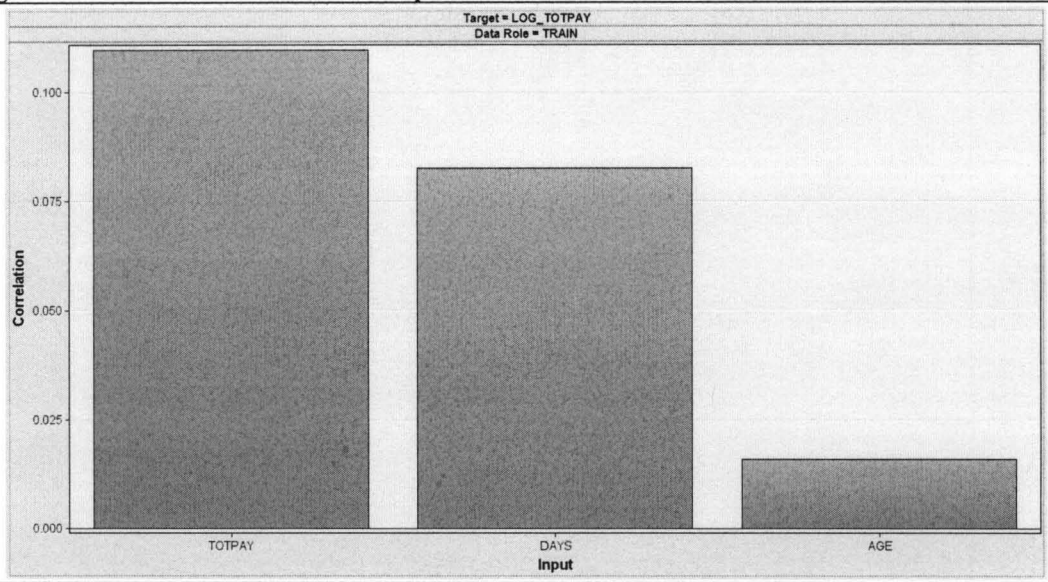


Figure 35. Pearson correlations with inpatient costs.



The same strategy for feature selection is implemented for modeling outpatient costs per month after hospitalization. That is, a tree is implemented, and the features selected by it are used in implementing a generalized linear model and a neural net. The features selected by the decision tree in order of importance are total payment during hospitalization for AMI, diagnosis related code, monthly number of prescriptions prior to hospitalization, length of stay during hospitalization, and gender. A stepwise-forward selection is also implemented resulting in the selection of four common features with the decision tree: diagnosis related group, monthly number of visits prior to hospitalization, total pay during hospitalization, and gender (Table 26). In addition, geographical region was found significant while the decision tree did not find this feature important. The multilayer perceptron was implemented with a single hidden layer consisting of four neurodes. The activation function for hidden neurons was the logistic function while the activation function for the output layer was linear function. The criterion for model selection is the root mean square error (Table 27). Clearly, the multilayer perceptron was superior to the generalized linear model with a mean square error of 1.29. It is important to observe that both model generalized well in the validation set, though.

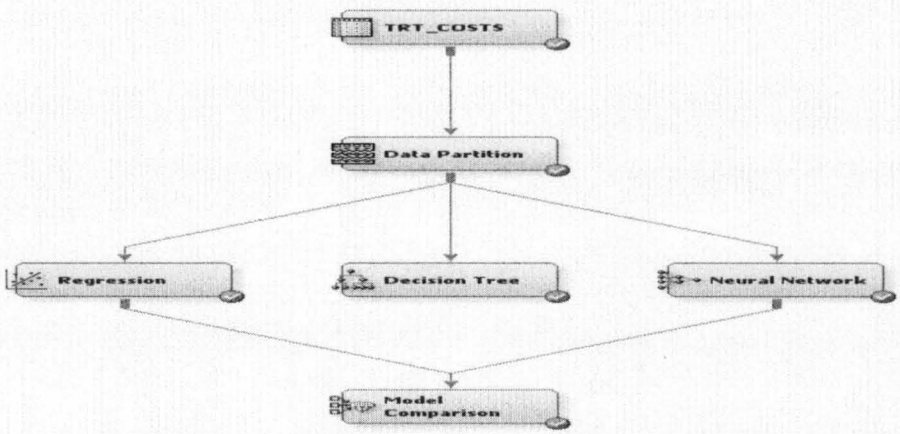
Table 26. Analysis of Effects for Linear Regression

Effect	DF	Sum of Squares	F Value	P-value
DRG	29	306.43	5.93	< .0001
Region	4	50.71	7.21	< .0001
Sex	1	10.78	6.06	.0139
Total Pay	1	97.29	54.68	< .0001
Prior Visits per Month	1	131.98	74.17	< .0001

Table 27. Model Selection Criterion for Predictive Models of Outpatient Costs.

Model	Mean Square Error (Training Set)	Mean Square Error (Validation Set)
Linear Regression	1.34	1.33
Neural Network	1.29	1.28

Figure 36. Enterprise Miner Predictive Modeling Workflow.



CHAPTER 7

Treatment Effectiveness

Thus far, the data set features have been explored for predictive purposes in cases of acute myocardial infarction. In this section, the aim is to explore the effectiveness for the various treatments for acute myocardial infarction when due to coronary occlusion. A statistically powerful method to evaluate effectiveness for treatment is survival analysis, which involves studying whether or not an event of interest occurs and when the event of interest occurs. In this case, the event of interest is reinfarction after hospital discharge for treatment of myocardial infarction. This is possible with this data set because the date of discharge and date of admission are readily available in the inpatient admission table. To identify the occurrence of an event of interest, the diagnosis within each record are checked against the criteria of interest, in this case an ICD9 code corresponding to acute myocardial infarction.

There is a characteristic of this data that makes it appropriate for an application of survival analysis in studying the effectiveness of a treatment for myocardial infarction. All subjects are followed for a period of two years only. This means that many of them did not have a reinfarction during this time but could have had one after this time. In turn, this means that such observations are censored at the end of the observational period. Logistic regression is not

appropriate because it does not take into account how the *chances* of experiencing a reinfarction changes over time while ordinary least square regression cannot take into account censored observations without losing statistical power.

Survival Analysis: Cox Proportional Hazards Models

The specific survival analysis model implemented in this research is that of Cox proportional hazards. It allows for the inclusion of subjects starting observation at different times such as this scenario as not all individual experienced an episode of infarction the same day. This method models the hazard function, which can be thought of as the number of events per interval of time.

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \le T < t + \Delta t \mid T \ge t\}}{\Delta t} \quad (1)$$

This function is modeled in the following form:

$$h_i(t) = \lambda_0(t)\exp\{\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}\} \quad (2)$$

The function in (2) says that for an individual i at time t, the hazard is the product of a baseline function $\lambda_0(t)$, a nonnegative unspecified function, and the natural exponential function of a linear combination of k variables. In this case, $\lambda_0(t)$ is regarded as the hazard function for an individual with all variables taking values of zero. The coefficients β_1 through β_k are estimated using the method of partial likelihood. This consists of writing the product of the likelihoods for all the events that are observed rather than writing the product of all the likelihoods for all individuals in the sample:

$$PL = \prod_{i=1}^{n} \left[\frac{e^{\beta x_i}}{\sum_{j=1}^{n} Y_{ij} e^{\beta x_j}} \right]^{\delta_i} \quad (3)$$

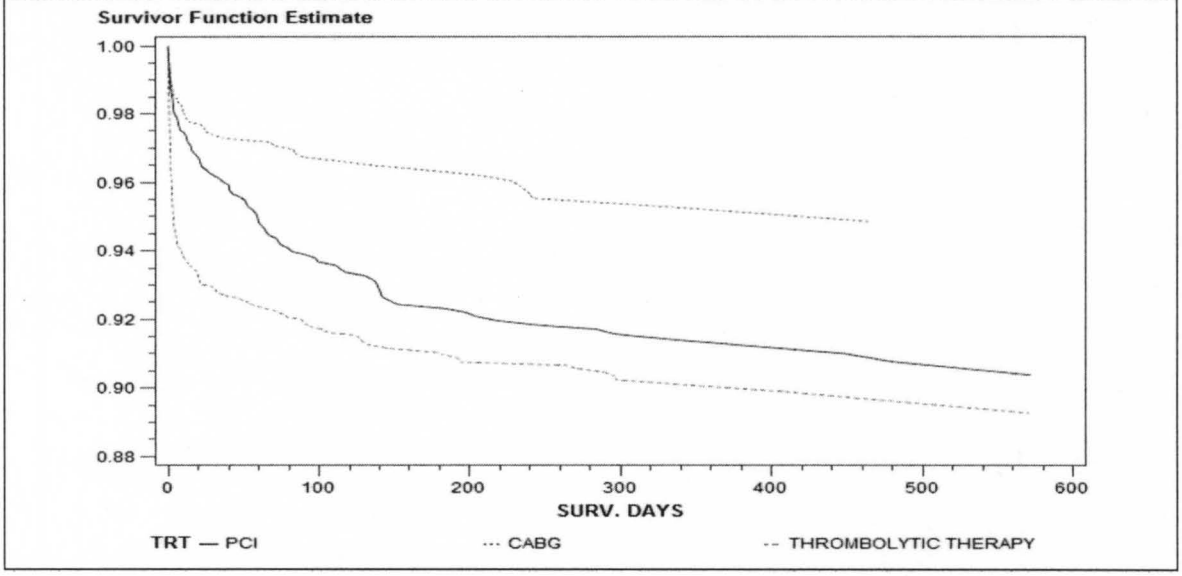
where i represents the individual observation, i = {1, ..., n}, t_i is the time of event or censoring time for observation i, δ_i is 1 if t_i is uncensored or if t_i is censored, x_i is a vector of k covariate values, and β is a vector of coefficients. The coefficients are finally estimated by the Maximum likelihood method explained in chapter 2.

The covariates of interest analyzed for effectiveness of treatments are treatment itself, gender, age, and length of stay during hospitalization. Table 27 contains the results after fitting the Cox proportional hazards model to the MarketScan data set. Both treatment and length of stay during hospitalization have a significant effect on time to reinfarction while the other covariates were found to be statistically insignificant. One can see that the hazard of reinfarction for patients treated with PCI is only 82% that of the patients treated with thrombolytic therapy, and the hazard of reinfarction for patients treated with CABG is 58% of that of patients treated with thrombolytic therapy. Similarly, it is evident in Figure 37 that the three groups have distinct survival curves with CABG treated patients having a longer time to reinfarction compared to PCI and thrombolytic therapy treated patients.

Table 27. Analysis of Cox Proportional Hazard Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
TRT	CABG	1	-0.54089	0.19425	7.7533	0.0054	0.582
TRT	PCI	1	-0.19432	0.12326	2.4854	0.1149	0.823
AGE		1	0.00609	0.00778	0.6121	0.4340	1.006
SEX	1	1	-0.05063	0.12667	0.1597	0.6894	0.951
REGION	1	1	-0.36232	1.00969	0.1288	0.7197	0.696
REGION	2	1	-0.65314	1.00512	0.4223	0.5158	0.520
REGION	3	1	-0.38735	1.00488	0.1486	0.6999	0.679
REGION	4	1	-0.41520	1.01396	0.1677	0.6822	0.660
Length of Stay		1	-0.07179	0.02315	9.6175	0.0019	0.931

Figure 37. Survival Curves by Treatment



Cost-effectiveness Analysis

During the last decade many interventional cardiologists and public health administrators have promoted the idea of more aggressive interventions in treating acute myocardial infarction with the aim of reducing mortality and rates of

reinfarction. One of the objectives of this study is to determine the cost-effectiveness of percutaneous coronary intervention and coronary artery bypass graft surgery as more aggressive alternatives to thrombolytic therapy. Cost-effectiveness analysis is better defined in *Cost-Effectiveness in Health and Medicine* [60] as

> *"Cost-effectiveness analysis is a method designed to assess the comparative impacts of expenditures on different health interventions... based on the premise that 'for any given level of resources available, society wishes to maximize the total aggregate health benefits conferred.' The central measure used in cost-effectiveness analysis is the cost-effectiveness ratio. Implicit in the cost-effectiveness ratio is a comparison between alternatives. The cost-effectiveness ratio for comparing the two alternatives is the difference in their costs divided by the difference in their effectiveness, or C/E."*

In this part of the study, effectiveness is measured in number of deaths averted within hospitalization since patient's id in the database cannot be matched to death certificates therefore making assertion of death out of a hospitalization record impossible. Examination of benefits and costs of the three distinct interventions for acute myocardial infarction is based on the model constructed using the data mining method of sequence rules analysis explained in chapter two.

To apply this data mining technique to the MarketScan data set, the inpatient admission records were converted from a wide format to a long format (Appendix J). As described previously, each record in this table contains up to fifteen medical and surgical procedures related to the hospital admission, which is a wide format. To obtain the long format, the data set was transposed. Each record in the resulting data set consisted of a patient's id and a procedure in order of execution, which is a long format. The *a priori* algorithm requires a chain size and minimum confidence level and support level. The chain size consists of the number of items, procedures in this case, that can be included in a sequence; the confidence level represents the empirical conditional probability of one procedure being performed given that another procedure has been performed while the support level denotes the percent of all transactions, inpatient admissions, that contain a given sequence.

As explained in chapter one, many procedures are performed during a diagnosis of acute myocardial infarction, such as electrocardiograms and other lab tests. These are standard procedures, which are expected to have a high support and confidence level within the data set. However, the structure of the model for cost-effectiveness seeks to describe the performance of procedures that deviate from the intended procedure used to treat a patient experiencing an acute myocardial infarction. For instance, sometimes the administration of thrombolytic agents is unsuccessful in treating a thrombus and emergency percutaneous coronary angioplasty must be performed to save the patient's life. Therefore, both the support and confidence level were set to 1% while the

maximum chain size was set to three. In addition, all CPT and ICD9 codes for PCI, CABG, and Thrombolytic therapy were compressed into one code for each procedure respectively in order to obtain more meaningful association rules.

Table 28. Derived Sequences

Sequence	Confidence
Thrombolytic Therapy → Rescue PCI	2.01%
PCI → Rescue CABG	1.33%
Thrombolytic Therapy, Rescue PCI → Alive at discharge	95.71%
PCI, Rescue CABG → Alive at discharge	94.18%

Given the resulting sequences and their respective confidence levels, a model in the form of a decision probability tree is constructed (Figure 38). Cost-effectiveness framework dictates that the decision tree must be binary at each decision node. This means that the probabilities at each decision node are complementary. That is, they add up to one. Following this criteria, subsets from the inpatient admission table were created to match the events described by the derived sequences. From these subsets, event probabilities were estimated from the observed frequencies and event costs were estimated by respective events' sample means (Table 29, 30).

Figure 38. Cost-Effectiveness Model

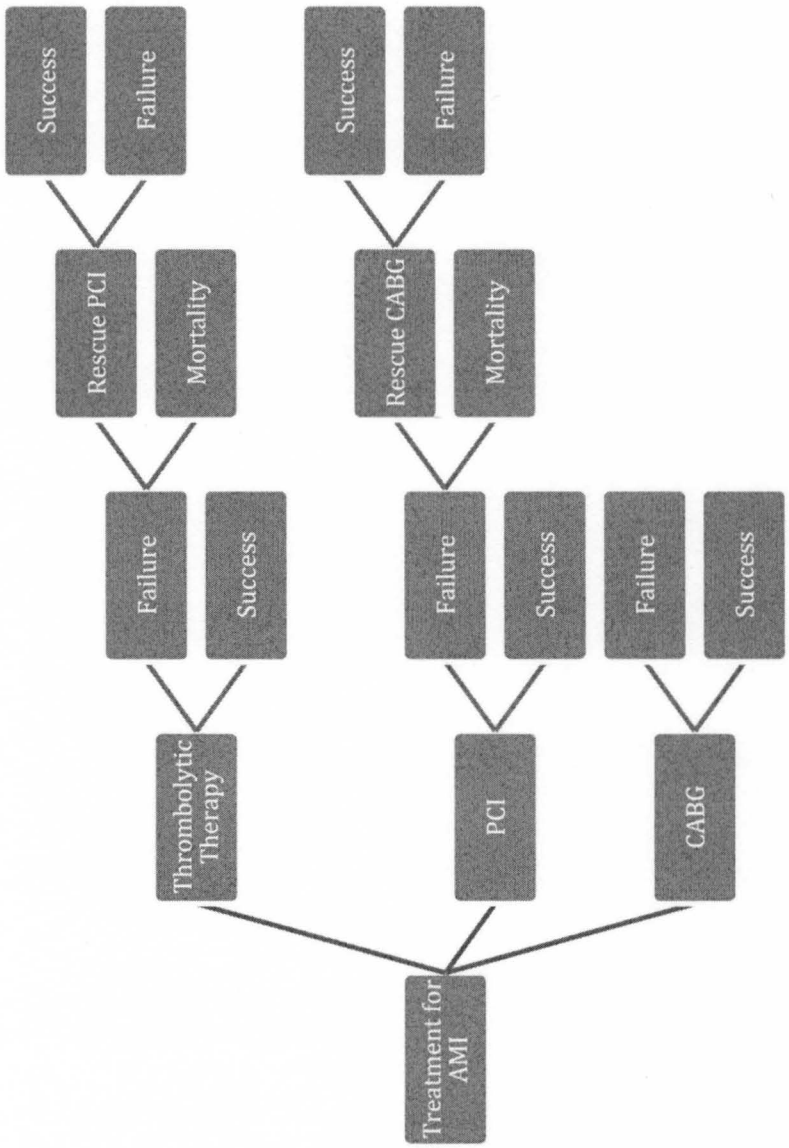


Table 29. Cost-Effectiveness Model Probabilities

Event	Probability
Success given Thrombolytic Therapy	0.9899
Failure given Thrombolytic Therapy	0.0101
Rescue PCI given Failed Thrombolytic Therapy	0.0201
Mortality given Failed Thrombolytic Therapy	0.9799
Failure given Rescue PCI	0.0429
Success given Rescue PCI	0.9571
Success given PCI	0.9956
Failure given PCI	0.0044
Rescue CABG given Failed PCI	0.0133
Mortality given Failed PCI	0.9807
Failure given Rescue CABG	0.0582
Success given Rescue CABG	0.9418
Success given CABG	0.9818
Failure given CABG	0.0182

Figure 39. Sequence Analysis Diagram – SAS Enterprise Miner

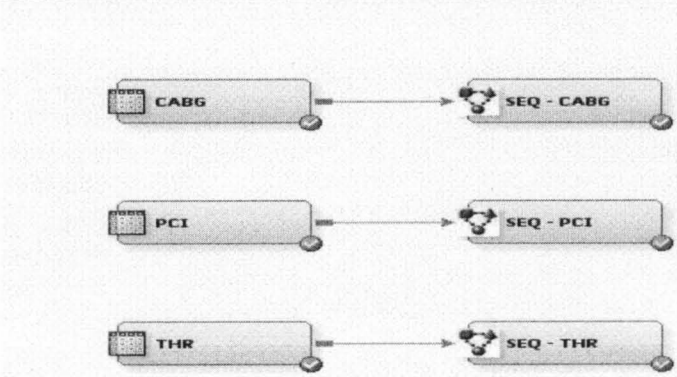


Table 30. Cost Estimates for Cost-effectiveness model

Treatment(s)	Cost Estimate Mean (Std.Dev)
Thrombolytic Therapy	$18,242 (13,127)
Thrombolytic Therapy + Rescue PCI	$19,172 (12,903)
PCI	$18,917 (13,044)
PCI + Rescue CABG	$47,929 (36,894)
CABG	$44,751 (32,606)

To evaluate the cost-effectiveness of these three alternative treatments for acute myocardial infarction the decision model (Figure 38) had to be rolled back. That is, the joint probability of each outcome (e.g., Rescue PCI given

Thrombolytic Therapy) was multiplied by the utility of the outcome (probability of alive) and summed over all outcomes; these values were then stratified by alternatives. The incremental cost-effectiveness then was evaluated by dividing the incremental costs by the incremental effectiveness. The following expression computes the effectiveness thrombolytic therapy:

$$Effectiveness = P(rescue\ PCI \mid Throbolytic\ therapy) \cdot P(alive \mid rescue\ PCI) + P(no\ rescue\ PCI \mid thrombolytic\ therapy) \cdot P(alive \mid no\ rescue\ PCI)$$

This represents the long-term proportion of saved lives while the following expression computes the cost-effectiveness ratio between therapy A and therapy B:

$$C/E = \frac{Expected\ Cost\ for\ therapy\ A - Expected\ Cost\ for\ therapy\ B}{Effectiveness\ for\ therapy\ A - Effectiveness\ for\ therapy\ B}$$

Thus, to analyze the results of the model, a table with the various expected costs and effectiveness, in descending order, for each treatment is constructed (Table 31).

Table 31. Cost-effectiveness for Treatment for AMI

Treatment	Cost	Effectiveness	Incremental Cost	Incremental Effectiveness	C/E Ratio
PCI	$18,918	0.995655			
Thrombolytic Therapy	$18,242	0.990094	$6,76	.0055607	$121,569
CABG	$44,759	0.98180	-$25,841	.013855	-$41,081

From Table 31, it is evident that it costs $121,569 per life saved were PCI implemented over thrombolytic therapy in the long term. On the other hand, it would save $41,081 per life saved were PCI implemented over CABG in the long term. Clearly, percutaneous coronary intervention is more cost-effective than

coronary artery bypass graft surgery. In contrast, it is not clear whether percutaneous coronary intervention is more cost-effective than thrombolytic therapy. This would be possible once a benchmark is set. However, the United States of America does not have such benchmark, and such decision rests in the hands of consumers and health insurance providers.

CHAPTER 8

Conclusion

This study implemented a data mining framework to determine and asses the main treatments for acute myocardial infarction. With the help of literature review, it was determined that thrombolytic therapy, percutaneous coronary intervention (angioplasty), and coronary artery bypass graft surgery were the three main treatments for victims of sudden heart attacks. Traditionally, health outcomes and utilization have been studied under the framework of classical statistics, which makes great assumptions about the distributions of the variables under study. Under ideal circumstances, such as clinical trials, most of these assumptions are satisfied. However, in day-to-day practice most of these assumptions fail to hold true. Exploratory data mining revealed that the distributions of most continuous outcomes did not follow a Gaussian distribution but rather severely skewed distributions, which do not conform to the assumptions of classical statistical models. Consequently, the need for a different analytical framework such as data mining was required.

The assessment was based on determining a profile of patients undergoing an episode of acute myocardial infarction, determine resource

utilization by each treatment, and creating a model that predicts each treatment resource utilization and cost for a distinct subject.

Text Mining and unsupervised clustering defined feature-based profiles that characterized subjects who underwent different treatments for acute myocardial infarction. Likewise, with the implementation of text mining along with clustering analysis prescription-based profiles were created which were associated with corresponding diagnosis clusters. This was demonstrated with the use of kernel densities, a nonparametric statistical method, by estimating the empirical distributions of outcomes by clusters of diagnosis and severity.

Decision trees, a data mining method, were used not only for their predictive modeling value but also established the most logical thresholds for variables; and they defined a hierarchy of features selected as inputs for other predictive models. It was evident that traditional features selection such as stepwise forward selection did not perform as well when used in non-homogenous data sets. In addition, it was demonstrated the importance of carefully considering the nature of the data being study in the use of appropriate statistical models for count data. The use of a Poisson model, a model for distribution of counts, was shown to be empirically superior to ordinary least squares in determining covariates that influence the distribution of the specific outcome.

The role of neural networks in building superior predictive models was confirmed as more appropriate in establishing higher predictive power when real world data is used. However, neural networks do have some drawbacks that

were clearly exposed. The mathematical complexity renders the model somewhat difficult to interpret for most stakeholders. In contrast to traditional regression models that define a relationships between outcome and an input based on the significance of corresponding coefficients and often linear associations which fail to hold in the medical world.

Finally, contribution to the field of applied mathematics and public health was achieved by carefully transforming wide-format records to long-format and implementing sequence analysis on them. This resulted in the determination of the sequence of treatment of acute myocardial infarction. This, in turn, was used to construct a probability tree that defined the basis for a cost-effectiveness analysis that compared treatments for acute myocardial infarction. The semi-parametric model known as Cox Proportional Hazards was implemented to evaluate the effectiveness of these treatments within a survival analysis methodology. It found that thrombolytic therapy had superior outcomes when compared to angioplasty and open-heart surgery. These finding are similar to those in studies contemporaneous to the data set. However, it must be noted that literature review of more recent studies have shown that as percutaneous coronary intervention has been refined, it has also become more effective in treating episodes of sudden heart attacks.

REFERENCES

1. More T: **Utopia**, Revised Edition edn. Cambrige, UK: Cambrige University Press; 1941.
2. Topol E: **Textbook of Interventional Cardiology**, 5 edn. La Jolla: Saunders; 2007.
3. Association AH: **Heart Attack**. 2010.
4. Tcheng J: **Primary Angioplasty in Acute Myocardial Infarction**, 1 edn. Totowa: Humana Press; 2002.
5. King S, Yeung A: **Interventional Cardiology**, 1 edn: McGraw-Hill Professional; 2006.
6. Manson J, Ridker P, Gaziano M, Hennekens C: **Prevention of myocardial infarction**, 1 edn. New York: Oxford University Press; 1996.
7. DeGeare VSMDFa, Dangas GMDPFb, Stone GWMDFb, Grines CLMDFc: **Interventional procedures in acute myocardial infarction. [Miscellaneous Article]**. *American Heart Journal January* 2001, **141**(1):15-25.
8. Thygesen K, Alpert JS, White HD, Jaffe AS, Apple FS, Galvani M, Katus HA, Newby LK, Ravkilde J, Chaitman B *et al*: **Universal definition of myocardial infarction**. *Circulation* 2007, **116**(22):2634-2653. Epub 2007 Oct 2619.
9. **Myocardial Ischemai, Injury and Infarction** [http://www.americanheart.org/presenter.jhtml?identifier=251]
10. Antman EM, Anbe DT, Armstrong PW, Bates ER, Green LA, Hand M, Hochman JS, Krumholz HM, Kushner FG, Lamas GA *et al*: **ACC/AHA Guidelines for the Management of Patients With ST-Elevation Myocardial Infarction--Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the 1999 Guidelines for the Management of Patients With Acute Myocardial Infarction)**. *Journal of the American College of Cardiology* 2004, **44**(3):671-719.
11. Nielsen PH, Maeng M, Busk M, Mortensen LS, Kristensen SD, Nielsen TT, Andersen HR: **Primary Angioplasty Versus Fibrinolysis in Acute Myocardial Infarction. Long-Term Follow-Up in the Danish Acute Myocardial Infarction 2 Trial**. *Circulation* 2010, **22**:22.
12. Ribeiro EE, Silva LA, Carneiro R, D'Oliveira LG, Gasquez A, Amino JG, Tavares JR, Petrizzo A, Torossian S, Duprat Filho R *et al*: **Randomized**

trial of direct coronary angioplasty versus intravenous streptokinase in acute myocardial infarction. *J Am Coll Cardiol* 1993, **22**(2):376-380.
13. Ribichini F, Steffenino G, Dellavalle A, Ferrero V, Vado A, Feola M, Uslenghi E: **Comparison of thrombolytic therapy and primary coronary angioplasty with liberal stenting for inferior myocardial infarction with precordial ST-segment depression: immediate and long-term results of a randomized study**. *J Am Coll Cardiol* 1998, **32**(6):1687-1694.
14. Zijlstra F, de Boer MJ, Hoorntje JC, Reiffers S, Reiber JH, Suryapranata H: **A comparison of immediate coronary angioplasty with intravenous streptokinase in acute myocardial infarction**. *N Engl J Med* 1993, **328**(10):680-684.
15. Grech ED: **ABC of interventional cardiology: percutaneous coronary intervention. I: history and development**. *Bmj* 2003, **326**(7398):1080-1082.
16. Fox KA, Goodman SG, Klein W, Brieger D, Steg PG, Dabbous O, Avezum A: **Management of acute coronary syndromes. Variations in practice and outcome; findings from the Global Registry of Acute Coronary Events (GRACE)**. *Eur Heart J* 2002, **23**(15):1177-1189.
17. Blankenship JC, Skelding KA, Scott TD, Buckley J, Zimmerman DK, Temple A, Sartorius J, Jimenez E, Berger PB: **ST-elevation myocardial infarction patients can be enrolled in randomized trials before emergent coronary intervention without sacrificing door-to-balloon time**. *Am Heart J* 2009, **158**(3):400-407.
18. Bohmer E, Hoffmann P, Abdelnoor M, Arnesen H, Halvorsen S: **Efficacy and safety of immediate angioplasty versus ischemia-guided management after thrombolysis in acute myocardial infarction in areas with very long transfer distances results of the NORDISTEMI (NORwegian study on DIstrict treatment of ST-elevation myocardial infarction)**. *J Am Coll Cardiol* 2010, **55**(2):102-110. Epub 2009 Sep 2010.
19. Gibbons RJ, Holmes DR, Reeder GS, Bailey KR, Hopfenspirger MR, Gersh BJ: **Immediate angioplasty compared with the administration of a thrombolytic agent followed by conservative treatment for myocardial infarction. The Mayo Coronary Care Unit and Catheterization Laboratory Groups**. *N Engl J Med* 1993, **328**(10):685-691.
20. Grines CL, Browne KF, Marco J, Rothbaum D, Stone GW, O'Keefe J, Overlie P, Donohue B, Chelliah N, Timmis GC *et al*: **A comparison of immediate angioplasty with thrombolytic therapy for acute myocardial infarction. The Primary Angioplasty in Myocardial Infarction Study Group**. *N Engl J Med* 1993, **328**(10):673-679.
21. Lotfi M, Mackie K, Dzavik V, Seidelin PH: **Impact of delays to cardiac surgery after failed angioplasty and stenting**. *Journal of the American College of Cardiology* 2004, **43**(3):337-342.

22. Jamal SM, Shrive FM, Ghali WA, Knudtson ML, Eisenberg MJ: **In-hospital outcomes after percutaneous coronary intervention in Canada: 1992/93 to 2000/01**. *Can J Cardiol* 2003, **19**(7):782-789.
23. Rathore SS, Curtis JP, Nallamothu BK, Wang Y, Foody JM, Kosiborod M, Masoudi FA, Havranek EP, Krumholz HM: **Association of Door-to-Balloon Time and Mortality in Patients >=65 Years With ST-Elevation Myocardial Infarction Undergoing Primary Percutaneous Coronary Intervention**. *The American Journal of Cardiology* 2009, **104**(9):1198-1203.
24. Hasdai D, Behar S, Wallentin L, Danchin N, Gitt AK, Boersma E, Fioretti PM, Simoons ML, Battler A: **A prospective survey of the characteristics, treatments and outcomes of patients with acute coronary syndromes in Europe and the Mediterranean basin; the Euro Heart Survey of Acute Coronary Syndromes (Euro Heart Survey ACS)**. *Eur Heart J* 2002, **23**(15):1190-1201.
25. Januzzi JL, Jr., Newby LK, Murphy SA, Pieper K, Antman EM, Morrow DA, Sabatine MS, Ohman EM, Cannon CP, Braunwald E: **Predicting a late positive serum troponin in initially troponin-negative patients with non-ST-elevation acute coronary syndrome: clinical predictors and validated risk score results from the TIMI IIIB and GUSTO IIA studies**. *Am Heart J* 2006, **151**(2):360-366.
26. Wallentin L, Lagerqvist B, Husted S, Kontny F, Stahle E, Swahn E: **Outcome at 1 year after an invasive compared with a non-invasive strategy in unstable coronary-artery disease: the FRISC II invasive randomised trial. FRISC II Investigators. Fast Revascularisation during Instability in Coronary artery disease**. *Lancet* 2000, **356**(9223):9-16.
27. **New meta-analysis: Routine invasive strategy betters selective care for non-ST elevation ACS** [http://www.theheart.org/article/1055735.do]
28. Neumann FJ, Kastrati A, Pogatsa-Murray G, Mehilli J, Bollwein H, Bestehorn HP, Schmitt C, Seyfarth M, Dirschinger J, Schomig A: **Evaluation of prolonged antithrombotic pretreatment ("cooling-off" strategy) before intervention in patients with unstable coronary syndromes: a randomized controlled trial**. *Jama* 2003, **290**(12):1593-1599.
29. Cannon CP, Weintraub WS, Demopoulos LA, Vicari R, Frey MJ, Lakkis N, Neumann FJ, Robertson DH, DeLucca PT, DiBattiste PM *et al*: **Comparison of early invasive and conservative strategies in patients with unstable coronary syndromes treated with the glycoprotein IIb/IIIa inhibitor tirofiban**. *N Engl J Med* 2001, **344**(25):1879-1887.
30. Morrow DA, Cannon CP, Rifai N, Frey MJ, Vicari R, Lakkis N, Robertson DH, Hille DA, DeLucca PT, DiBattiste PM *et al*: **Ability of Minor Elevations of Troponins I and T to Predict Benefit From an Early Invasive Strategy in Patients With Unstable Angina and Non-ST Elevation Myocardial Infarction: Results From a Randomized Trial**. *JAMA* 2001, **286**(19):2405-2412.

31. Andersen HR, Nielsen TT, Vesterlund T, Grande P, Abildgaard U, Thayssen P, Pedersen F, Mortensen LS: **Danish multicenter randomized study on fibrinolytic therapy versus acute coronary angioplasty in acute myocardial infarction: rationale and design of the DANish trial in Acute Myocardial Infarction-2 (DANAMI-2)**. *Am Heart J* 2003, **146**(2):234-241.
32. Michels KB, Yusuf S: **Does PTCA in acute myocardial infarction affect mortality and reinfarction rates? A quantitative overview (meta-analysis) of the randomized clinical trials**. *Circulation* 1995, **91**(2):476-485.
33. Rathore SS, Curtis JP, Chen J, Wang Y, Nallamothu BK, Epstein AJ, Krumholz HM, for the National Cardiovascular Data Registry: **Association of door-to-balloon time and mortality in patients admitted to hospital with ST elevation myocardial infarction: national cohort study**. *BMJ* 2009, **338**(may19_1):b1807-.
34. Le May MR, Labinaz M, Davies RF, Marquis JF, Laramee LA, O'Brien ER, Williams WL, Beanlands RS, Nichol G, Higginson LA: **Stenting versus thrombolysis in acute myocardial infarction trial (STAT)**. *J Am Coll Cardiol* 2001, **37**(4):985-991.
35. Schomig A, Kastrati A, Dirschinger J, Mehilli J, Schricke U, Pache J, Martinoff S, Neumann FJ, Schwaiger M: **Coronary stenting plus platelet glycoprotein IIb/IIIa blockade compared with tissue plasminogen activator in acute myocardial infarction. Stent versus Thrombolysis for Occluded Coronary Arteries in Patients with Acute Myocardial Infarction Study Investigators**. *N Engl J Med* 2000, **343**(6):385-391.
36. Myers J, Brock G, Appana S, Gray L: **Kentucky pilot project for primary PCI without onsite CABG**. *J Ky Med Assoc* 2009, **107**(11):451-458.
37. Brown DC, Mogelson S, Harris R, Kemp D, Massey M: **Percutaneous coronary interventions in a rural hospital without surgical backup: report of one year of experience**. *Clin Cardiol* 2006, **29**(8):337-340.
38. Hannan EL, Zhong Y, Racz M, Jacobs AK, Walford G, Cozzens K, Holmes DR, Jones RH, Hibberd M, Doran D *et al*: **Outcomes for Patients With ST-Elevation Myocardial Infarction in Hospitals With and Without Onsite Coronary Artery Bypass Graft Surgery: The New York State Experience**. *Circ Cardiovasc Interv* 2009, **2**(6):519-527.
39. Ong SH, Lim VY, Chang BC, Lingamanaicker J, Tan CH, Goh YS, Tan KS: **Three-year experience of primary percutaneous coronary intervention for acute ST-segment elevation myocardial infarction in a hospital without on-site cardiac surgery**. *Ann Acad Med Singapore* 2009, **38**(12):1085-1089.
40. Paraschos A, Callwood D, Wightman MB, Tcheng JE, Phillips HR, Stiles GL, Daniel JM, Sketch MH, Jr.: **Outcomes following elective percutaneous coronary intervention without on-site surgical backup in a community hospital**. *Am J Cardiol* 2005, **95**(9):1091-1093.

41. Singh KP, Harrington RA: **Primary percutaneous coronary intervention in acute myocardial infarction**. *Med Clin North Am* 2007, **91**(4):639-655; x-xi.
42. Thompson CR, Humphries KH, Gao M, Galbraith PD, Norris C, Carere RG, Knudtson ML, Ghali WA: **Revascularization use and survival outcomes after cardiac catheterization in British Columbia and Alberta**. *Can J Cardiol* 2004, **20**(14):1417-1423.
43. Wharton TP, Jr., Grines LL, Turco MA, Johnston JD, Souther J, Lew DC, Shaikh AZ, Bilnoski W, Singhi SK, Atay AE *et al*: **Primary angioplasty in acute myocardial infarction at hospitals with no surgery on-site (the PAMI-No SOS study) versus transfer to surgical centers for primary angioplasty**. *J Am Coll Cardiol* 2004, **43**(11):1943-1950.
44. Ramos P: **The Cost-effectiveness of the Kentucky Pilot Project of Allowing Primary PCI at Hospitals without Onsite CABG Capabilities**. In. Louisville: University of Louisville; 2010.
45. Reeder GS, Bailey KR, Gersh BJ, Holmes DR, Jr., Christianson J, Gibbons RJ: **Cost comparison of immediate angioplasty versus thrombolysis followed by conservative therapy for acute myocardial infarction: a randomized prospective trial. Mayo Coronary Care Unit and Catheterization Laboratory Groups**. *Mayo Clin Proc* 1994, **69**(1):5-12.
46. Hartwell D, Colquitt J, Loveman E, Clegg AJ, Brodin H, Waugh N, Royle P, Davidson P, Vale L, MacKenzie L: **Clinical effectiveness and cost-effectiveness of immediate angioplasty for acute myocardial infarction: systematic review and economic evaluation**. *Health Technol Assess* 2005, **9**(17):1-99, iii-iv.
47. Khot UN, Johnson-Wood ML, Geddes JB, Ramsey C, Khot MB, Taillon H, Todd R, Shaikh SR, Berg WJ: **Financial impact of reducing door-to-balloon time in ST-elevation myocardial infarction: a single hospital experience**. *BMC Cardiovasc Disord* 2009, **9**:32.
48. Fayyad UM, Peatetsky-Shapiro G, Smyth P, Uthurusamy R: **Advances in Knowledge Discovery and Data Mining**, 1 edn. Cambridge: MIT Press; 1996.
49. **Introduction to Data Mining Using SAS Enterprise Miner** [https://ezproxy.siast.sk.ca:443/login?url=http://proquest.safaribooksonline.com/9781590478295]
50. Hastie TJ, Tibshirani RJ, Friedman JH: **The elements of statistical learning**. New York: Springer; 2001.
51. Gill J: **Generalized linear models :a unified approach**. Thousand Oaks, Calif. :: Sage Publications, Inc.; 2001.
52. Jorgensen B: **The Theory of Dispersion Models (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)**: Chapman and Hall/CRC; 1997.
53. Hogg RV, Craig AT: **Introduction to mathematical statistics**. Englewood Cliffs (N.J.): Prentice-Hall international; 1995.
54. **MarketScan® Database**. In. Ann Arbor.

55. **Introduction to data mining using SAS Enterprise Miner** [http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=180233]
56. **Fundamentals of predictive text mining** [http://site.ebrary.com/id/10396806]
57. Cameron AC, Trivedi PK: **Regression analysis of count data**. Cambridge, UK; New York, NY, USA: Cambridge University Press; 1998.
58. Greene WH: **Accounting for excess zeros and sample selection in Poisson and negative binomial regression models**. New York: New York University, Leonard N. Stern School of Business; 1994.
59. De Wilde P: **Neural network models : theory and projects**. London; New York: Springer; 1997.
60. Gold MR: **Cost-effectiveness in health and medicine**. New York: Oxford University Press; 1996.

APPENDIX A
MARKETSCAN DATA: INPATIENT ADMISSION TABLE VARIABLES DESCRIPTION

Inpatient Admissions Table - Variables			
Name	**Type**	**Label**	**Length**
ADMDATE	Numeric	Date of Admission	5
ADMTYP	Character	Admission Type	1
AGE	Numeric	Age of Patient	4
AGEGRP	Character	Age Group	1
CASEID	Numeric	Case and Services Link	7
DATATYP	Numeric	Data Type	4
DAYS	Numeric	Length of Stay	5
DOBYR	Numeric	Patient Birth Year	4
DRG	Numeric	Diagnosis Related Group	4
DSTATUS	Character	Discharge Status	2
DX1	Character	Diagnosis 1	5
DX10	Character	Diagnosis 10	5
DX11	Character	Diagnosis 11	5
DX12	Character	Diagnosis 12	5
DX13	Character	Diagnosis 13	5
DX14	Character	Diagnosis 14	5
DX15	Character	Diagnosis 15	5
DX2	Character	Diagnosis 2	5
DX3	Character	Diagnosis 3	5
DX4	Character	Diagnosis 4	5
DX5	Character	Diagnosis 5	5
DX6	Character	Diagnosis 6	5
DX7	Character	Diagnosis 7	5
DX8	Character	Diagnosis 8	5
DX9	Character	Diagnosis 9	5
EECLASS	Character	Employee Classification	1
EESTATU	Character	Employment Status	1
EGEOLOC	Character	Geographic Location Employee	2
EIDFLAG	Character	Enrollee ID Derivation Flag	1
EMPCTY	Numeric	County Employee	5
EMPREL	Character	Relation to Employee	1
EMPZIP	Numeric	Zipcode Employee 3 Digit	4
ENRFLAG	Character	Enrollment Flag	1
ENROLID	Numeric	Enrollee ID	7
EPISODE	Numeric		3
HOSPCTY	Numeric	County Hospital	5
HOSPPAY	Numeric	Payments: Hospital	7
HOSPZIP	Numeric	Zipcode Hospital 3 Digit	4

INDSTRY	Character	Industry	1
LASTADM	Numeric	Days from Prior Discharge	5
MDC	Character	Major Diagnostic Category	2
MSA	Numeric	Metropolitan Statistical Area	5
NEXTADM	Numeric	Days to Next Admission	5
PATFLAG	Character	Patient Indistinct Flag	1
PATID	Numeric	Patient ID	7
PDX	Character	Diagnosis Principal	5
PHYFLAG	Character	Physician Specialty Coding Flag	1
PHYSID	Numeric	Physician ID	7
PHYSPAY	Numeric	Payments: Physician	7
PLANKEY	Numeric	Benefit Plan Link	5
PLANTYP	Numeric	Plan Indicator	4
PPROC	Character	Procedure Principal	5
PROC1	Character	Procedure Code 1	5
PROC10	Character	Procedure 10	5
PROC11	Character	Procedure 11	5
PROC12	Character	Procedure 12	5
PROC13	Character	Procedure 13	5
PROC14	Character	Procedure 14	5
PROC15	Character	Procedure 15	5
PROC2	Character	Procedure 2	5
PROC3	Character	Procedure 3	5
PROC4	Character	Procedure 4	5
PROC5	Character	Procedure 5	5
PROC6	Character	Procedure 6	5
PROC7	Character	Procedure 7	5
PROC8	Character	Procedure 8	5
PROC9	Character	Procedure 9	5
REGION	Character	Region	1
RX	Character	Cohort Drug Indicator	1
SEQNUM	Numeric	Sequence Number	7
SEX	Character	Gender of Patient	1
STATE	Character	State of Hospital	2
TOTNET	Numeric	Payments: Net (Case)	7
TOTPAY	Numeric	Payments: Total (Case)	7
TRIMLOS	Numeric	Trim Flag Length of Stay	5
TRIMPDM	Numeric	Trim Flag Per Diem	5
UNIHOSP	Numeric	Hospital ID Number (MDST)	7
VERSION	Character	Version	2

WGTKEY	Numeric	MarketScan National Weight Link	4
YEAR	Numeric	Date Year Incurred	4

APPENDIX B
MARKETSCAN DATA: INPATIENT SERVICES TABLE VARIABLES DESCRIPTION

Inpatient Services Table - Variables			
Name	**Type**	**Label**	**Length**
ADMDATE	Numeric	Date of Admission	5
ADMTYP	Character	Admission Type	1
AGE	Numeric	Age of Patient	4
AGEGRP	Character	Age Group	1
CASEID	Numeric	Case and Services Link	7
COB	Numeric	COB and Other Savings	7
COPAY	Numeric	Copayment	7
DATATYP	Numeric	Data Type	4
DEDUCT	Numeric	Deductible	7
DOBYR	Numeric	Patient Birth Year	4
DRG	Numeric	Diagnosis Related Group	4
DSTATUS	Character	Discharge Status	2
DX1	Character	Diagnosis 1	5
DX2	Character	Diagnosis 2	5
DX3	Character	Diagnosis 3	5
DX4	Character	Diagnosis 4	5
DX5	Character	Diagnosis 5	5
EECLASS	Character	Employee Classification	1
EESTATU	Character	Employment Status	1
EGEOLOC	Character	Geographic Location Employee	2
EIDFLAG	Character	Enrollee ID Derivation Flag	1
EMPCTY	Numeric	County Employee	5
EMPREL	Character	Relation to Employee	1
EMPZIP	Numeric	Zipcode Employee 3 Digit	4
ENRFLAG	Character	Enrollment Flag	1
ENROLID	Numeric	Enrollee ID	7
HOSPCTY	Numeric	County Hospital	5
HOSPZIP	Numeric	Zipcode Hospital 3 Digit	4
INDSTRY	Character	Industry	1
MDC	Character	Major Diagnostic Category	2
MSA	Numeric	Metropolitan Statistical Area	5
NETPAY	Numeric	Payments Net	7
PATFLAG	Character	Patient Indistinct Flag	1
PATID	Numeric	Patient ID	7
PAY	Numeric	Payment	7
PDDATE	Numeric	Date Claim Paid	5
PDX	Character	Diagnosis Principal	5
PHYFLAG	Character	Physician Specialty Coding Flag	1

PLANKEY	Numeric	Benefit Plan Link	5
PLANTYP	Numeric	Plan Indicator	4
PPROC	Character	Procedure Principal	5
PROC1	Character	Procedure Code 1	5
PROCMOD	Character	Procedure Code Modifier	2
PROCTYP	Character	Procedure Code Type	1
PROVCTY	Numeric	County Provider	5
PROVID	Numeric	Provider ID	7
PROVZIP	Numeric	Zipcode Provider 3 Digit	4
QTY	Numeric	Quantity of Services	5
RECFLAG	Character	Record Flag	1
REGION	Character	Region	1
REVCODE	Character	Revenue Code	3
RX	Character	Cohort Drug Indicator	1
SEQNUM	Numeric	Sequence Number	7
SEX	Character	Gender of Patient	1
STDPLAC	Numeric	Place of Service	4
STDPROV	Numeric	Provider Type	4
STDSVC	Numeric	Service Type	4
SVCDATE	Numeric	Date Service Incurred	5
TOTPAY	Numeric	Payments: Total (Case)	7
TSVCDAT	Numeric	Date Service Ending	5
UNIHOSP	Numeric	Hospital ID Number (MDST)	7
VERSION	Character	Version	2
WGTKEY	Numeric	MarketScan National Weight Link	4
YEAR	Numeric	Date Year Incurred	4

APPENDIX C
MARKETSCAN DATA: OUTPATIENT SERVICES TABLE VARIABLES DESCRIPTION

Outpatient Services Table - Variables			
Name	**Type**	**Label**	**Length**
AGE	Numeric	Age of Patient	4
AGEGRP	Character	Age Group	1
COB	Numeric	COB and Other Savings	5
COPAY	Numeric	Copayment	5
DATATYP	Numeric	Data Type	4
DEDUCT	Numeric	Deductible	5
DOBYR	Numeric	Patient Birth Year	4
DX1	Character	Diagnosis 1	5
DX2	Character	Diagnosis 2	5
DX3	Character	Diagnosis 3	5
DX4	Character	Diagnosis 4	5
DX5	Character	Diagnosis 5	5
EECLASS	Character	Employee Classification	1
EESTATU	Character	Employment Status	1
EGEOLOC	Character	Geographic Location Employee	2
EIDFLAG	Character	Enrollee ID Derivation Flag	1
EMPCTY	Numeric	County Employee	5
EMPREL	Character	Relation to Employee	1
EMPZIP	Numeric	Zipcode Employee 3 Digit	4
ENRFLAG	Character	Enrollment Flag	1
ENROLID	Numeric	Enrollee ID	7
INDSTRY	Character	Industry	1
MDC	Character	Major Diagnostic Category	2
MSA	Numeric	Metropolitan Statistical Area	5
NETPAY	Numeric	Payments Net	5
PATFLAG	Character	Patient Indistinct Flag	1
PATID	Numeric	Patient ID	7
PAY	Numeric	Payment	5
PDDATE	Numeric	Date Claim Paid	5
PHYFLAG	Character	Physician Specialty Coding Flag	1
PLANKEY	Numeric	Benefit Plan Link	5
PLANTYP	Numeric	Plan Indicator	4
PROC1	Character	Procedure Code 1	5
PROCGRP	Numeric	Procedure Code Group	4
PROCMOD	Character	Procedure Code Modifier	2
PROCTYP	Character	Procedure Code Type	1
PROVCTY	Numeric	County Provider	5
PROVID	Numeric	Provider ID	7

	PROVZIP	Numeric	Zipcode Provider 3 Digit	4
	QTY	Numeric	Quantity of Services	5
	RECFLAG	Character	Record Flag	1
	REGION	Character	Region	1
	REVCODE	Character	Revenue Code	3
	RX	Character	Cohort Drug Indicator	1
	SEQNUM	Numeric	Sequence Number	7
	SEX	Character	Gender of Patient	1
	STDPLAC	Numeric	Place of Service	4
	STDPROV	Numeric	Provider Type	4
	STDSVC	Numeric	Service Type	4
	SVCDATE	Numeric	Date Service Incurred	5
	TG	Character	Treatment Group	2
	VERSION	Character	Version	2
	WGTKEY	Numeric	MarketScan National Weight Link	4
	YEAR	Numeric	Date Year Incurred	4

APPENDIX D
MARKETSCAN DATA: OUTPATIENT PHARMACEUTICAL CLAIMS VARIABLES DESCRIPTION

Outpatient Pharmaceutical Claims Table - Variables			
Name	**Type**	**Label**	**Length**
AGE	Numeric	Age of Patient	4
AGEGRP	Character	Age Group	1
AWP	Numeric	Average Wholesale Price	7
COB	Numeric	COB and Other Savings	7
COPAY	Numeric	Copayment	7
DATATYP	Numeric	Data Type	4
DAWIND	Character	Dispense as Written Indicator	2
DAYSUPP	Numeric	Days Supply	5
DEACLAS	Character	DEA Classification	1
DEDUCT	Numeric	Deductible	7
DISPFEE	Numeric	Dispensing Fee	7
DOBYR	Numeric	Patient Birth Year	4
EECLASS	Character	Employee Classification	1
EESTATU	Character	Employment Status	1
EGEOLOC	Character	Geographic Location Employee	2
EIDFLAG	Character	Enrollee ID Derivation Flag	1
EMPCTY	Numeric	County Employee	5
EMPREL	Character	Relation to Employee	1
EMPZIP	Numeric	Zipcode Employee 3 Digit	4
ENRFLAG	Character	Enrollment Flag	1
ENROLID	Numeric	Enrollee ID	7
GENERID	Numeric	Generic Product ID	7
GENIND	Character	Generic Indicator	1
INDSTRY	Character	Industry	1
INGCOST	Numeric	Ingredient Cost	7
MAINTIN	Character	Maintenance Indicator	1
METQTY	Numeric	Metric Quantity	5
MSA	Numeric	Metropolitan Statistical Area	5
NDCNUM	Character	National Drug Code	11
NETPAY	Numeric	Payments Net	7
PATFLAG	Character	Patient Indistinct Flag	1
PATID	Numeric	Patient ID	7
PAY	Numeric	Payment	7
PDDATE	Numeric	Date Claim Paid	5
PHARMID	Numeric	Pharmacy ID	7
PHCLASS	Numeric	Pharmacy Class Code	4
PHRMCTY	Numeric	County Provider	5
PHRMZIP	Numeric	Zipcode Provider 3 Digit	4

PHYFLAG	Character	Physician Specialty Coding Flag	1
PLANKEY	Numeric	Benefit Plan Link	5
PLANTYP	Numeric	Plan Indicator	4
QTY	Numeric	Quantity of Services	5
REFILL	Numeric	Refill Number	4
REGION	Character	Region	1
SALETAX	Numeric	Sales Tax	7
SEQNUM	Numeric	Sequence Number	7
SEX	Character	Gender of Patient	1
SVCDATE	Numeric	Date Service Incurred	5
THERCLS	Numeric	Therapeutic Class	4
THERGRP	Character	Therapeutic Group	2
VERSION	Character	Version	2
YEAR	Numeric	Date Year Incurred	4

APPENDIX E
MARKETSCAN DATA: INPATIENT ADMISSION SAMPLE

Row number	ENROLLID	ADMDATE	AADMTYP	AGE	DX1	DX10	DX11	DX12	DX13	DDXX11	DX2	DX3	DX4	DX5	DX6	DX7	DX8	DX9	EECSLTAASTSU	EEGEOLAC	EEGIDLFAG
1	36173 2401	02/10/ 2000	2	58	411 81						410 00	441 01	414 01	578 1	715 90	786 51			914 9		1
2	36182 5601	05/16/ 2000	2	60	410 71						411 1	436	414 01						916 2		1
3	36182 5601	05/26/ 2000	1	60	410 71						427 1	436	414 01	250 90	411 1				916 2		1
4	36186 6001	12/27/ 2000	2	54	411 1						410 92	786 59	401 9						996 2		1
5	36192 7501	04/15/ 2000	2	53	410 71						411 1	414 01	414 9						916 2		1
6	36192 7501	04/16/ 2000	1	53	410 41						410 90	786 50	414 00	414 01					916 2		1
7	16334 7001	03/24/ 2000	1	60	414 01						410 30	413 9	414 00	786 52	411 1	V45 81			916 2		1
8	16302 9701	03/07/ 2000	2	54	410 01						410 11	428 0	786 52	427 89					916 2		1
9	16302 9701	03/10/ 2000	2	54	410 01						410 00	410 11	428 0						916 2		1
10	16302 9701	03/12/ 2000	1	54	410 01						410 00	410 11	511 8	414 01	428 0	411 1			916 2		1
11	16302 9701	03/21/ 2000	1	54	414 01						410 11	428 0	511 8	786 09	786 50	995 2	411 1	42 94	916 2		1
12	36097 9701	02/15/ 2000	2	64	410 42						410 41	410 90	427 5	348 1					936 4		1
13	36101 6601	06/02/ 2000	1	42	410 41	998 11	402 10	410 90			410 11	786 50	414 00	414 01	423 9	V58 49	518 89	51 19	915 2		1
14	36185 4501	02/09/ 2000	2	62	410 71						411 1	414 00	428 0	414 01	V45 81				935 2		1
15	36196 5004	11/11/ 2000	1	62	410 41	786 50	V45 81	V58 49	250 90		410 71	486	414 01	793 1	414 00	414 9	231 8	51 19	915 2		2
16	36225 2201	04/24/ 2000	1	52	410 71	276 6	411 1	V72 82	410 90		410 11	414 01	786 50	518 3	428 0	429 2	518 89	51 80	995 2		1
17	36225 2201	07/17/ 2000	2	52	410 41						410 71	410 90	518 89	414 9	414 02	V45 81			915 2		1

Row number	ENROLID	ADMDATE	AAGDE	DX1	DX10	DX11	DX12	DX13	DDXX11	DX2	DX3	DX4	DX5	DX6	DX7	DX8	DX9	EEEGEOSLTOAACSTSU	EIDLFLAG
18	36233 9702	02/19/ 2000	2 4 1	410 41						401 9								9 1 5 2	1
19	36233 9702	02/24/ 2000	1 4 1	410 41						414 01								9 1 5 2	1
20	36243 4702	11/16/ 2000	2 6 0	410 41						786 50	401 9							9 1 6 2	1

Row number	ENROLID	AGEGRP	CASEID	DATATYP	DAYS	DOBYR	DRG	DSTATUS	EMPCTY	EMPREL	EMPZIP	ENRFLAG	EPISODE	HOSPCTY	HOSPPAY	HOSPZIP	INDST	LASTADM	MDC	MSA	NEXTADM	PAFLAG	PDX
1	361732401	5	1560	1	2	1942	124	01	48499	1	754	1	0	48113	3219	753	6	.05	0000	.0		0	4111
2	361825601	5	1586	1	2	1940	121	01	06041	1	949	1	1	06075	17006	941	6	17 05	7360	8		0	41071
3	361825601	5	1587	1	1	1940	112	01	06041	1	949	1	1	06075	28296	941	6	8 05	7360	122		0	41071
4	361866001	4	1595	1	1	1946	124	01	06065	1	922	1	2	06037	776	911	6	.05	6780	.0			4111
5	361927501	4	1624	1	1	1947	122	05	06041	1	949	1	1	15009	4867	967	6	.05	7360	.0		0	41071
6	361927501	4	1625	1	2	1947	116	01	06041	1	949	1	1	15003	15661	968	6	.05	7360	.0			41041
7	163347001	5	1644	1	1	1940	112	01	06081	1	944	1	0	06075	2882	941	6	69 05	7360	9		0	41401
8	163029701	4	1742	1	2	1946	121	05	06001	1	945	1	1	06075	35233	941	6	.05	5775	.0		0	41001
9	163029701	4	1743	1	1	1946	121	05	06001	1	945	1	1	06075	5112	941	6	.05	5775	.0		0	41001
10	163029701	4	1744	1	9	1946	107	05	06001	1	945	1	1	06075	29160	941	6	.05	5775	.0		0	41001
11	163029701	4	1745	1	1	1946	109	01	06001	1	945	1	1	06075	74506	941	6	.05	5775	.0		0	41401
12	360979701	5	1802	1	7	1936	123	20	41049	1	978	1	2	41059	8640	978	6	.05	0000	.0		0	41042
13	361016601	3	1808	1	1	1958	107	01	04013	2	852	1	1	04013	26465	850	6	.05	6200	.0		0	41041
14	361854501	5	2099	1	3	1938	121	01	04023	1	856	1	1	04019	4025	857	6	.05	0000	.0		0	41071
15	361965004	5	2143	1	1	1938	107	01	04013	3	852	1	1	04013	58095	850	6	13 05	6200	.0			41041
16	362252201	4	2209	1	10	1948	107	01	04013	1	852	1	1	04013	23650	850	6	.05	6200	74		0	41071
17	362252201	4	2210	1	6	1948	122	01	04013	1	852	1	1	04013	13990	850	6	74 05	6200	79		0	41041
1	362339703		222	1	4	195	12	0	04012		852	1	1	04014	4260	856	6	.0	620	.0			4104

Row number	ENROLID	AGEGRP	CASEID	DATATYP	DAYS	DOBYR	DRG	DSTATUS	EMPCTY	EMPREL	EMPZIP	ENRFLAG	EPISOED	HOSPCTY	HOSPPAY	HOSPZIP	INDSTRY	LASTADM	MDC	MSA	NEXTADM	PAFLAG	PDX
8	02	5		9	2	5	3		2			3		2		5	0		1				
19	362339702	3	2226	1	1	1959	112	01	040113	2	852	1	1	040113	1250	852	6	.05	6200			.01	41041
20	362434702	5	2238	1	1	1940	122	05	061111	2	913	1	1	060307	1117	900	6	.05	8735			.01	41041

Row number	ENROLLID	PHYSID	PHYSPAY	PPLAANNKTEYYP	PPROC	PROC1	PROC10	PROC11	PROC12	PROC13	PROC14	PROC15	PROC2	PROC3	PROC4	PROC5	PROC6	PROC7	PROC8	PROC9
1	361732401	752254247	1005	.5	93510	93510	80049	80061	82310	83735	84100	84443	3722	82550	82553	84484	85610	85730	93010	99222
2	361825601	942157429	1565	.5	3722	3722	93018	99238					93510	93543	93545	93555	93556	99223	99233	93016
3	361825601	942157429	369	.5	3605	3605							92960	99222	99238					
4	361866001	330452991	570	.5	93545	93545							3722	93010	99223	93510	93539	93543	93555	
5	361927501	953044080	510	.5	99254	99254							99232							
6	361927501	990331208	5466	.5	3601	3601	99222	99238					93010	99252	92980	93510	93543	93545	93555	93556
7	163347001	943180015	2448	.5	92982	92982	93555	93556	99253	93042	99239	99356	3601	71020	75896	93010	93510	93539	93543	93545
8	163029701	568961081	796	.5	3722	3722							71010	93010	93307	93320	93325	99291	99292	
9	163029701	568961081	531	.5	3722	3722							99291	99292						
10	163029701	568961081	3183	.5	33533	33533	93510	93543	93545	93555	93556	99255	3722	99291	93307	93320	93325	99233	99232	36489
11	163029701	943268944	3347	.5	3612	3612	99292	99233	32020	71030	99254	99232	33518	33533	71010	93010	93307	93320	93325	99291
12	360979701	930425580	826	.5	9671	9671							99223	99291	99292	71010	99232	99239	99302	

Row number	ENROLID	PHYSID	PHYSPAY	PPLLAANNKTEYYP	PPROC	PROC1	PROC10	PROC11	PROC12	PROC13	PROC14	PROC15	PROC2	PROC3	PROC4	PROC5	PROC6	PROC7	PROC8	PROC9
13	36101 6601	86090 2679	6221	.5	3613	3613	93543	93545	93555	93556	99231	71020	71010	99254	93307	93320	93325	99233	99232	93510
14	36185 4501	86088 8560	1119	.5	3722	3722							99223	99233	93510	93543	93545	93555	93556	99239
15	36196 5004	86090 2679	4504	.5	3615	3615	93555	93556	71020	99231	00562	33519	71010	99223	99291	99233	99232	93510	93543	93545
16	36225 2201	86090 2679	3754	.5	3612	3612	93545	93555	93556	00562	33518	33533	99222	99232	99254	71020	93307	93320	93510	93543
17	36225 2201	86051 6994	996	.5	3722	3722	93543	93545	93555	93556	99232	71010	99222	99233	93307	93320	93325	93510	93539	93540
18	36233 9702	.	0	.5																
19	36233 9702	86039 4149	17	.5	3601	3601							93010							
20	36243 4702	95274 1568	190	.5	9925	9925							71010							

Row number	ENROLID	REGION	RRX	SEQNUM	SEX	STATE	TOTNET	TOTPAY	TRIMLOS	TRIMPDM	UNIHOSP	VERSION	YEAR
1	361732401	3	1	1560	2	49	4545	4745	0	0	696719315	10	2000
2	361825601	4	1	1586	1	62	18371	18571	0	1	451727315	10	2000
3	361825601	4	1	1587	1	62	28565	28665	0	1	451727315	10	2000
4	361866001	4	1	1595	2	62	890	1346	0	-2	.	10	2000
5	361927501	4	1	1624	1	63	5277	5377	-1	0	906735317	10	2000
6	361927501	4	1	1625	1	63	20927	21127	0	0	.	10	2000
7	163347001	4	1	1644	1	62	5396	5496	0	1	166702315	10	2000
8	163029701	4	1	1742	1	62	36304	36604	0	2	21703315	10	2000
9	163029701	4	1	1743	1	62	5243	5643	-1	0	21703315	10	2000
10	163029701	4	1	1744	1	62	38640	38640	0	0	21703315	10	2000
11	163029701	4	1	1745	1	62	79307	79807	1	0	.	10	2000
12	360979701	4	1	1802	2	64	8987	9487	.	.	941706317	10	2000
13	361016601	4	1	1808	1	52	36784	36784	1	-1	.	10	2000
14	361854501	4	1	2099	1	52	4769	5144	0	0	446734316	10	2000
15	361965004	4	1	2143	1	52	67370	67937	1	0	.	10	2000
16	362252201	4	1	2209	1	52	32550	32550	0	-2	.	10	2000
17	362252201	4	1	2210	1	52	15505	15505	1	-1	.	10	2000
18	362339702	4	1	2225	1	52	3860	4260	0	-1	246718316	10	2000
19	362339702	4	1	2226	1	52	1167	1267	0	-1	246966816	10	2000
20	362434702	4	1	2238	1	62	1333	1333	-1	-1	166944315	10	2000

APPENDIX F
MARKETSCAN DATA: OUTPATIENT SERVICE SAMPLE

Row number	ENROLID	PROVID	SVCDATE	DX1	DX2	DX3	DX4	DX5	PROC1
1	32001	.	02/23/2000	2859	2859				
2	32001	118551932	04/19/2000	41401					99213
3	32001	946809934	07/20/2000	72981					93970
4	32001	262495705	08/08/2000	4919					99283
5	32001	541765934	08/28/2000	4919	4919				
6	32001	675372933	09/01/2000	41400					99213
7	32001	890068434	11/18/2000	4293					71010
8	32001	890068434	11/18/2000	4019					99231
9	32001	890068434	11/19/2000	41401					71010
10	32001	890068434	11/19/2000	41090					93010
11	32001	890068434	11/19/2000	41400					99232
12	32001	890068434	11/19/2000	41400					99232
13	32001	890068434	11/19/2000	41400					99232
14	32001	890068434	11/19/2000	41090					99233
15	32001	890068434	11/20/2000	41400					93010
16	32001	890068434	11/21/2000	41400					93010
17	32001	890068434	11/21/2000	4271					93510
18	32001	890068434	11/21/2000	4271					93539
19	32001	890068434	11/21/2000	4271					93540
20	32001	890068434	11/21/2000	4271					93545
21	32001	890068434	11/21/2000	4271					93556
22	32001	890068434	11/22/2000	41401					92980
23	32001	890068434	11/22/2000	41400					93010
24	32001	890068434	11/22/2000	41011					93307
25	32001	890068434	11/22/2000	41011					93320
26	32001	890068434	11/22/2000	41011					93325
27	32001	890068434	11/23/2000	41400					93010
28	32001	890068434	11/27/2000	41401					71020
29	32001	890068434	11/27/2000	4271					93620
30	32001	890068434	11/30/2000	4271					33249
31	32001	890068434	11/30/2000	4271					93641
32	42601	569732410	08/09/2000	8830					99203
33	42601	569732410	08/09/2000	8830					J1885

APPENDIX G
SAS PROGRAM TO EXTRACT INPATIENT ADMISSION SAMPLE

```
LIBNAME MYDATA 'H:\Data Mining\Fall 2011\Data';
LIBNAME MS 'H:\MSdata';

%MACRO NAMES(NAME= , NUMBER= );
	%DO N=1 %TO &NUMBER;
		&NAME&N
	%END;
%MEND NAMES;

/* EXTRACT AMI OBSERVATION FROM INPATIENT ADMISSION TABLES */

DATA IN1;
	SET MS.INPATIENTADMISSION0;
	LENGTH EPISODE 3.;
	EPISODE = 3;
	IF MISSING(ENROLID) = 0 THEN
	DO;
		IF SUBSTRN(DX1,1,3) = '410' THEN
			IF LENGTH(DX1) = 5 THEN EPISODE =
SUBSTRN(DX1,5,1);
			ELSE EPISODE = 0;
		ELSE IF SUBSTRN(DX2,1,3) = '410' THEN
			IF LENGTH(DX2) = 5 THEN EPISODE =
SUBSTRN(DX2,5,1);
			ELSE EPISODE = 0;
	END;
	IF EPISODE NE 3;
RUN;

DATA IN2;
	SET MS.INPATIENTADMISSION1;
	LENGTH EPISODE 3.;
	EPISODE = 3;
	IF MISSING(ENROLID) = 0 THEN
	DO;
		IF SUBSTRN(DX1,1,3) = '410' THEN
			IF LENGTH(DX1) = 5 THEN EPISODE =
SUBSTRN(DX1,5,1);
			ELSE EPISODE = 0;
		ELSE IF SUBSTRN(DX2,1,3) = '410' THEN
			IF LENGTH(DX2) = 5 THEN EPISODE =
SUBSTRN(DX2,5,1);
			ELSE EPISODE = 0;
	END;
	IF EPISODE NE 3;
RUN;
```

```
/* COLLECT SMALLER SUBSETS (ABOVE) INTO ONE LARGE SET OF
OUTPATIENTSERVICE OBSERVATIONS */

DATA MYDATA.INALL;
      SET %NAMES(NAME=IN,NUMBER=2);
RUN;

/* CREATE MASTER TABLE OF UNIQUE ENROLID'S */

DATA INP_MT;
      SET MYDATA.INALL (KEEP= ENROLID);
RUN;

PROC SORT DATA=INP_MT OUT=MYDATA.MASTER_TABLE NODUPKEY;
      BY ENROLID;
RUN;
```

APPENDIX H
SAS PROGRAM TO COMPUTE INPATIENT SAMPLE SUMMARY STATISTICS

```
PROC MEANS DATA= MYDATA.INALL;
        FW=12
        PRINTALLTYPES
        CHARTYPE
        QMETHOD=OS
        VARDEF=DF
                MEAN
                STD
                MIN
                MAX NONOBS
                MEDIAN      ;
        VAR AGE DAYS TOTPAY;

RUN;
ODS GRAPHICS ON;
TITLE;
TITLE1 "Summary Statistics";
TITLE2 "Box and Whisker Plots";
PROC SGPLOT DATA= MYDATA.INALL;
        VBOX AGE;
RUN;QUIT;
PROC SGPLOT DATA= MYDATA.INALL;
        VBOX DAYS;
RUN;QUIT;
PROC SGPLOT DATA= MYDATA.INALL;
        VBOX TOTPAY;
RUN;QUIT;
ODS GRAPHICS OFF;
```

APPENDIX I
SAS PROGRAM TO EXTRACT OUTPATIENT RECORDS

```
LIBNAME MYDATA 'H:\Data Mining\Fall 2011\Data';
LIBNAME MS 'H:\MSdata';

/* FOR TABLE CORRESPONDING TO 2001 */

%MACRO NAMES(NAME= , NUMBER= );
	%DO N=1 %TO &NUMBER;
		&NAME&N (DROP = WGTKEY)
	%END;
%MEND NAMES;

/* FOR TABLE CORRESPONDING TO 2000 */

%MACRO NAMES1(NAME= , NUMBER= );
	%DO N=1 %TO &NUMBER;
	&NAME&N (DROP=PAYIND PCPID PCPSPEC PHYCLAS PMGID
REFIND REFTYP)
	%END;
%MEND NAMES1;

/* GET OUTPATIENT SERVICE OBSERVATION CORRESPONDING TO */
/* MASTER TABLE ENROLID VARIABLE */
%MACRO GET_OUTPAT_SVC(NAME1= , NAME2= , NAME3= , NUMBER= );

	%DO N=1 %TO &NUMBER;

		PROC SORT DATA=&NAME1&N;
			BY ENROLID;
		RUN;

		DATA &NAME3&N;
			MERGE &NAME1&N (IN=LEFT) &NAME2 (IN=RIGHT);
			BY ENROLID;
			IF LEFT AND RIGHT;
		RUN;

	%END;

%MEND GET_OUTPAT_SVC;


/* CREATE A SINGLE OUTPATIENT SERVICE TABLE */
%GET_OUTPAT_SVC(NAME1=MS.OUTPATIENTSERVICE0,
NAME2=MYDATA.MASTER_TABLE, NAME3=OUT0, NUMBER=6);
%GET_OUTPAT_SVC(NAME1=MS.OUTPATIENTSERVICE1,NAME2=MYDAT
A.MASTER_TABLE, NAME3=OUT1, NUMBER=8);
```

```
DATA MYDATA.OUTALL;
        SET %NAMES1(NAME=OUT0,NUMBER=6)
%NAMES(NAME=OUT1,NUMBER=8);
RUN;
DATA OUTSAMPLE;
        SET MYDATA.OUTALL (KEEP=ENROLID SVCDATE PROVID DX1-DX5
PROC1);
RUN;
PROC SORT DATA=OUTSAMPLE;
        BY ENROLID SVCDATE PROVID;
RUN;
/* TRANSFORM SET FROM WIDE TO LONG */
DATA OS1;
        SET OUTSAMPLE (KEEP=ENROLID SVCDATE PROVID DX1
RENAME=(DX1=DX));
        IF MISSING(DX)=0;
RUN;
DATA OS2;
        SET OUTSAMPLE (KEEP=ENROLID SVCDATE PROVID DX2
RENAME=(DX2=DX));
        IF MISSING(DX)=0;
RUN;
DATA OS3;
        SET OUTSAMPLE (KEEP=ENROLID SVCDATE PROVID DX3
RENAME=(DX3=DX));
        IF MISSING(DX)=0;
RUN;
DATA OS4;
        SET OUTSAMPLE (KEEP=ENROLID SVCDATE PROVID DX4
RENAME=(DX4=DX));
        IF MISSING(DX)=0;
RUN;
DATA OS5;
        SET OUTSAMPLE (KEEP=ENROLID SVCDATE PROVID DX5
RENAME=(DX5=DX));
        IF MISSING(DX)=0;
RUN;
DATA OS6;
        SET OUTSAMPLE (KEEP=ENROLID SVCDATE PROVID PROC1);
RUN;

DATA OUT1;
        SET OS1 OS2 OS3 OS4 OS5;
RUN;
```

```
PROC SORT DATA=OUT1;
        BY ENROLID SVCDATE PROVID;
RUN;

/* TRANSFORM SET FROM LONG TO WIDE */

PROC TRANSPOSE DATA=OUT1 OUT=TRANS1 (DROP=_NAME_ _LABEL_)
PREFIX=DX;
        VAR DX;
        BY ENROLID SVCDATE PROVID;
RUN;

PROC TRANSPOSE DATA=OS6 OUT=TRANS2 (DROP=_NAME_ _LABEL_)
PREFIX=PROC ;
        VAR PROC1;
        BY ENROLID SVCDATE PROVID;
RUN;

PROC SORT DATA=TRANS1;
        BY ENROLID SVCDATE PROVID;
RUN;
PROC SORT DATA=TRANS2;
        BY ENROLID SVCDATE PROVID;
RUN;

DATA MYDATA.OUT_BY_ID_DATE;
        MERGE TRANS1 (IN=LEFT) TRANS2 (IN=RIGHT);
        BY ENROLID SVCDATE PROVID;
        IF LEFT AND RIGHT;
RUN;

DATA OUT_VISIT_BEFORE OUT_VISIT_AFTER;
        SET MYDATA.OUT_BY_ID_DATE;
        IF SVCDATE < ADMDATE THEN
                OUTPUT OUT_VISIT_BEFORE;
        ELSE
                OUTPUT OUT_VISIT_AFTER;
RUN;

PROC FREQ DATA=OUT_VISIT_BEFORE;
        TABLES ENROLID;
RUN;
PROC FREQ DATA=OUT_VISIT_AFTER;
        TABLES ENROLID;
RUN;
```

APPENDIX J
SAS PROGRAM DATA DERIVATION

```
/* OUTPUT: FIRST OBSERVATION (EVENT) FROM INPATIENT ADMISSION */
/* VAR REINFARCTION: 1 IF THERE IS A SECOND AMI OBSERVATION
CORRESPONDING TO THE SAME ENROLID */
/* 0 IF THERE IS ONLY ONE OBSERVATION FOR THE CORRESPONDING
ENORLID                */

PROC SORT DATA=MYDATA.INALL (KEEP=ENROLID ADMDATE)
OUT=SORT_INALL;
        BY ENROLID ADMDATE;
RUN;
DATA MYDATA.REINFARCTION_INALL;
        SET SORT_INALL;
        BY ENROLID;
        LENGTH REINFARCTION 3;
        IF FIRST.ENROLID = 1 AND LAST.ENROLID = 1 THEN
                DO;
                        REINFARCTION = 0;
                        OUTPUT;
                END;
        ELSE IF FIRST.ENROLID = 1 AND LAST.ENROLID = 0 THEN
                DO;
                        REINFARCTION = 1;
                        OUTPUT;
                END;
RUN;
/* OUTPUT: NUMBER OF DAYS BETWEEN FIRST ADMISSION AND
READMISSION */

PROC SORT DATA=MYDATA.INALL_SURV (KEEP=ENROLID ADMDATE
LASTIN NEXTIN) OUT=SORT_INALLSURV;
        BY ENROLID ADMDATE;
RUN;

DATA MYDATA.INALL_SURV_LOH;
        SET SORT_INALLSURV;
        LENGTH LOH 3;
        BY ENROLID;
        IF FIRST.ENROLID = 1 THEN
                DO;
                        LOH = NEXTIN - ADMDATE;
                        OUTPUT;
                END;
RUN;

%MACRO OBTAINDRUGNAME(INTABLE, NAMETABLE, BYVARIABLE,
OUTTABLE);
```

```
 PROC SORT DATA=&INTABLE;
  BY &BYVARIABLE;
 RUN;

 PROC SORT DATA=&NAMETABLE;
  BY &BYVARIABLE;
 RUN;

DATA RX_INTERMEDIATE;
 MERGE &INTABLE (IN = LEFT) &NAMETABLE (KEEP = NDC
PROPRIETARYNAME SUBSTANCENAME);
 BY &BYVARIABLE;
 IF LEFT=1 AND MISSING(PROPRIETARYNAME) = 0;
RUN;

DATA &OUTTABLE;
        SET RX_INTERMEDIATE;
        LENGTH NAME $ 99;
        IF MISSING(SUBSTANCENAME) = 0 THEN
                NAME = SUBSTANCENAME;
        ELSE
                NAME = PROPRIETARYNAME;
        DROP PROPRIETARYNAME SUBSTANCENAME;
RUN;

%MEND;

PROC SORT DATA=MYDATA.REINFARCTION_INALL (KEEP= ENROLID
ADMDATE) OUT=SORT_REINFARCTION;
        BY ENROLID;
RUN;

PROC SORT DATA=MYDATA.AMI_PAT_RXRECORDS OUT=RXRECORDS;
        BY ENROLID;
RUN;

/*CREATE TABLE OF ALL RX RECORDS FOR AMI PATIENTS WITH THE
DATE OF THE EVENT: FIRST AMI */

DATA MYDATA.RXRECORD_EVENTDATE;
        MERGE RXRECORDS SORT_REINFARCTION;
        BY ENROLID;
RUN;

/* OUTPUT: PRESCRIPTION DOCUMENTS */
```

```
DATA RX_BEFORE RX_AFTER;
	SET MYDATA.RXRECORD_EVENTDATE (KEEP = ENROLID NDCNUM
SVCDATE ADMDATE);
	IF SVCDATE < ADMDATE THEN
		OUTPUT RX_BEFORE;
	ELSE
		OUTPUT RX_AFTER;
RUN;


DATA RX_BEFORE_NDC;
 SET RX_BEFORE (KEEP=ENROLID NDCNUM);
 LENGTH NDC $ 9;
 NDC = SUBSTR(NDCNUM,1,9);
 DROP NDCNUM;
RUN;

DATA RX_AFTER_NDC;
 SET RX_AFTER (KEEP=ENROLID NDCNUM);
 LENGTH NDC $ 9;
 NDC = SUBSTR(NDCNUM,1,9);
 DROP NDCNUM;
RUN;

%OBTAINDRUGNAME(RX_BEFORE_NDC, MYDATA.NDCOLD, NDC,
RX_BEFORE_NAME);
%OBTAINDRUGNAME(RX_AFTER_NDC, MYDATA.NDCOLD, NDC,
RX_AFTER_NAME);

/* TRANSPOSE PRESCRIPTIONS: EACH PATIENT HAS ONE ROW WITH
MANY PRESCRIPTIONS */

PROC SORT DATA=RX_BEFORE_NAME;
	BY ENROLID;
RUN;

PROC SORT DATA=RX_AFTER_NAME;
	BY ENROLID;
RUN;

PROC TRANSPOSE DATA=RX_BEFORE_NAME OUT=TRN_RX_BEFORE
PREFIX=NAME;
	BY ENROLID;
	VAR NAME;
RUN;
```

```
PROC TRANSPOSE DATA=RX_AFTER_NAME OUT=TRN_RX_AFTER
PREFIX=NAME;
        BY ENROLID;
        VAR NAME;
RUN;

DATA MYDATA.PRESCRIPTIONS_BEFORE;
        SET TRN_RX_BEFORE;
        LENGTH NEWNAME $ 99;
        LENGTH PX $ 2000;
        ARRAY P_X[*] NAME::;
        PX = TRANSLATE(LEFT(TRIM(P_X[1])),'_', ' ');
        DO I=2 TO DIM(P_X);
                NEWNAME = TRANSLATE(LEFT(TRIMN(P_X[I])), '_', ' ');
                PX = CATX(' ', PX, NEWNAME);
        END;
        KEEP ENROLID PX;
RUN;

DATA MYDATA.PRESCRIPTIONS_AFTER;
        SET TRN_RX_AFTER;
        LENGTH NEWNAME $ 99;
        LENGTH PX $ 2000;
        ARRAY P_X[*] NAME::;
        PX = TRANSLATE(LEFT(TRIM(P_X[1])),'_', ' ');
        DO I=2 TO DIM(P_X);
                NEWNAME = TRANSLATE(LEFT(TRIMN(P_X[I])), '_', ' ');
                PX = CATX(' ', PX, NEWNAME);
        END;
        KEEP ENROLID PX;
RUN;
/* OUTPUT: PRESCRIPTION COUNTS */

DATA BEFORE_EVENT AFTER_EVENT;
        SET MYDATA.RXRECORD_EVENTDATE (KEEP=ENROLID SVCDATE
ADMDATE NDCNUM);
        LENGTH NDC $ 9;
        NDC = SUBSTR(NDCNUM,1,9);
        IF SVCDATE < ADMDATE THEN
                        OUTPUT BEFORE_EVENT;
        ELSE
                        OUTPUT AFTER_EVENT;
RUN;
```

```
PROC SORT DATA=BEFORE_EVENT (DROP=NDCNUM)
OUT=BEFORE_EVENT_COMP NODUPKEY;
        BY ENROLID NDC;
RUN;

PROC SORT DATA=AFTER_EVENT (DROP=NDCNUM)
OUT=AFTER_EVENT_COMP NODUPKEY;
        BY ENROLID NDC;
RUN;
/* CREATE INPATIENT DIAGNOSIS STRING */

PROC SORT DATA=MYDATA.INALL (KEEP=ENROLID ADMDATE DX1-DX15)
OUT=IN_OBS;
        BY ENROLID ADMDATE;
RUN;

DATA FIRST_IN_ADM;
        SET IN_OBS;
        BY ENROLID;
        IF FIRST.ENROLID = 1 THEN OUTPUT;
RUN;

DATA MYDATA.IN_DX_STRING;
        SET FIRST_IN_ADM;
        LENGTH DXT $ 80;
        ARRAY DX[*] DX1-DX15;
        DXT = TRIM(DX[1]);
        DO I=2 TO 15;
                DXT = TRIM(CATX(' ', DXT, DX[I]));
        END;
        KEEP ENROLID DXT;
RUN;
PROC SORT DATA=MYDATA.REINFARCTION_INALL (KEEP= ENROLID
ADMDATE) OUT=SORT_REINFARCTION;
        BY ENROLID;
RUN;

PROC SORT DATA=MYDATA.OUT_BY_ID_DATE (KEEP=ENROLID
SVCDATE) OUT=SORT_OUT;
        BY ENROLID;
RUN;

/*CREATE TABLE OF ALL OUTPATIENT VISITS WITH THE DATE OF THE
EVENT: FIRST AMI */

DATA OUT_WITH_ADMDATE;
```

```
        MERGE SORT_OUT SORT_REINFARCTION;
        BY ENROLID;
RUN;

DATA OUT_VISIT_BEFORE OUT_VISIT_AFTER;
        SET OUT_WITH_ADMDATE;
        IF SVCDATE < ADMDATE THEN
                OUTPUT OUT_VISIT_BEFORE;
        ELSE
                OUTPUT OUT_VISIT_AFTER;
RUN;
/* WIDE TO LONG FORMAT -- PCI CABG THROMBOLYTIC THERAPY ONLY
*/
DATA TMP1 (KEEP=ID PR TRT);
        SET MYDATA.ALL (KEEP=ENROLID PROC1-PROC15 TRT
WHERE=(TRT NOT = 0));
        ARRAY PRC [15] PROC1-PROC15;
        DO I=1 TO 15;
                IF MISSING(PRC[I]) = 0 THEN
                DO;
                        ID = ENROLID;
                        PR = PRC[I];
                        OUTPUT;
                END;
        END;
RUN;

/* SEPARATE ICD9 RECORDS FROM CPT RECORDS */

DATA ICD CP;
        SET TMP1;
        IF LENGTH(PR) = 5 THEN
                OUTPUT CP;
        ELSE
                OUTPUT ICD;
RUN;

/* FILTER TO VALID CPT CODES */

PROC SORT DATA=MYDATA.CPT;
        BY CODE;
RUN;

PROC SORT DATA=CP;
        BY PR;
RUN;
```

```
DATA TMP2;
        MERGE CP (IN=LEFT) MYDATA.CPT (KEEP=CODE IN=RIGHT
RENAME=(CODE=PR));
        BY PR;
        IF LEFT AND RIGHT;
RUN;

/* GET DISCHARGE STATUS */

DATA TMP3 (KEEP=ID PR TRT);
        SET MYDATA.ALL (KEEP=ENROLID ALIVE TRT
RENAME=(ENROLID=ID) WHERE=(TRT NOT = 0));
        PR = PUT(ALIVE, $5.);
RUN;

/* APPEND ICD9 TO CPT */

DATA TMP4;
        SET TMP2 ICD;
RUN;

/* APPEND DISCHARGE STATUS */

DATA TMP5;
        SET TMP4 TMP3;
RUN;

PROC SORT DATA=TMP5 OUT=CP_ICD;
        BY ID;
RUN;
```

APPENDIX K
SAS PROGRAM – REGRESSION AND SURVIVAL ANALYSIS

```
DATA OUT_VISITS;
	SET MYDATA.ALL (KEEP=ENROLID ADMDATE AGE DAYS ALIVE TRT
VISITS_BEFORE VISITS_AFTER REGION SEX INDSTRY);
	MONTHS = INTCK('MONTH', ADMDATE,'31DEC2001'D, 'C');
	LOGRISK = LOG(MONTHS/24 + .5);
	VISITS_PER_MONTH = VISITS_AFTER/(MONTHS + .5);
	LABEL VISITS_PER_MONTH = 'VISITS PER MONTH'
		VISITS_AFTER = 'No. OF VISITS';
	OBS+1;
	FORMAT TRT TRT.
		SEX $SEX.
		INDSTRY $INDSTRY.
		REGION $REGION.;

	WHERE ALIVE=1;
RUN;

DATA RX_COUNT;
	SET MYDATA.ALL (KEEP=ENROLID ADMDATE AGE DAYS ALIVE TRT
RX_CNT_B RX_CNT_A REGION SEX INDSTRY
RENAME=(RX_CNT_A=PRESCRIPTION_COUNT));
	MONTHS = INTCK('MONTH', ADMDATE, '31DEC2001'D, 'C');
	LOGRISK = LOG(MONTHS/24 + .5);
	Rx_COUNT_PER_MONTH = PRESCRIPTION_COUNT/(MONTHS + .5);
	LABEL Rx_COUNT_PER_MONTH = 'Rx PER MONTH'
		PRESCRIPTION_COUNT = 'No. OF Rx';
	OBS+1;
		FORMAT TRT TRT.
		SEX $SEX.
		INDSTRY $INDSTRY.
		REGION $REGION.;
	WHERE ALIVE = 1;
RUN;



PROC GENMOD DATA=OUT_VISITS PLOTS=(PREDICTED);
	TITLE1 'OUT VISITS PER MONTH';
	CLASS TRT INDSTRY REGION SEX;
	MODEL VISITS_AFTER = AGE DAYS TRT REGION SEX INDSTRY /
DIST=POISSON OFFSET=LOGRISK TYPE3;
RUN;

PROC GENMOD DATA=OUT_VISITS PLOTS=(PREDICTED);
	TITLE1 'OUT VISITS PER MONTH ADJUSTED';
	CLASS TRT INDSTRY REGION SEX;
```

```
        MODEL VISITS_AFTER = AGE DAYS TRT REGION SEX INDSTRY /
DIST=POISSON OFFSET=LOGRISK SCALE=PEARSON TYPE3;
        OUTPUT OUT=SASDATA.VISITS PREDICTED=PRED;
RUN;


PROC GENMOD DATA=RX_COUNT PLOTS=(PREDICTED );
        TITLE1 'RX COUNT PER MONTH';
        CLASS TRT INDSTRY REGION SEX;
        MODEL PRESCRIPTION_COUNT = AGE DAYS TRT REGION SEX
INDSTRY / DIST=POISSON OFFSET=LOGRISK TYPE3;
RUN;

PROC GENMOD DATA=RX_COUNT PLOTS=(PREDICTED);
        TITLE1 'RX COUNT PER MONTH ADJUSTED';
        CLASS TRT INDSTRY REGION SEX;
        MODEL PRESCRIPTION_COUNT = AGE DAYS TRT REGION SEX
INDSTRY / DIST=POISSON OFFSET=LOGRISK SCALE=PEARSON TYPE3;
        OUTPUT OUT=SASDATA.RXCNT PREDICTED=PRED;
RUN;
PROC FORMAT;
        VALUE TRT
                1='PCI'
                2='CABG'
                3='THROMBOLYTIC THERAPY';
RUN;

DATA SURVDATA;
        SET MYDATA.ALL (KEEP=ENROLID LOH REINFARCTION AGE SEX
TRT REGION ALIVE DAYS WHERE=(TRT NOT = 0));
        FORMAT TRT TRT.;
        LABEL LOH = 'SURV. DAYS';
RUN;

PROC PHREG DATA=SURVDATA;
        CLASS REGION SEX TRT;
        MODEL LOH*REINFARCTION(0) = TRT AGE SEX REGION DAYS/
TIES=EFRON;
        TITLE1 'ORIGINAL SET';
RUN;

PROC PHREG DATA=SURVDATA;
        MODEL LOH*REINFARCTION(0) = TRT / TIES=EFRON;
        STRATA TRT;
        BASELINE OUT=A LOGLOGS=LLS SURVIVAL=S;
        TITLE1 'STRATIFIED SET';
```

```
RUN;

PROC GPLOT DATA=A;
        PLOT S*LOH=TRT;
        SYMBOL1 INTERPOL=JOIN COLOR=BLACK LINE=1;
        SYMBOL2 INTERPOL=JOIN COLOR=BLUE LINE=2;
        SYMBOL3 INTERPOL=JOIN COLOR=GREEN LINE=3;
RUN;
```

CURRICULUM VITAE

PEDRO GREGORIO RAMOS

Education

Middle Tennessee State University Department of Mathematical Sciences Murfreesboro, TN	M.S. in Mathematics, August 2005 Concentration: General Mathematics Minor: Computer Science
Delta State University School of Business Cleveland, MS	B.S. in Business, May 1999 Major: Business Administration Minor: Information Systems and Finance
Nashville State Community College Department of Computer Information Systems Nashville, TN	A.A.S. in Computer Information Systems, July 2000 Concentration: Applications Developer
Broward Community College Fort Lauderdale, FL	Associates of Arts, July 1997 Concentration: Business Administration

Academic Professional Experience

Institution	Rank	Dates
University of Louisville	Graduate Teaching Assistant	August 2007 – Present
University of Louisville	Research Assistant	December 2009 – December 2010
University of Louisville	Virtual Math Center Manager	August 2006 – July 2007
Cumberland University	Mathematics Instructor	January 2006 – May 2006
Nashville State Community College	Mathematics Instructor	January 2006 – August 2006

Publications

Chapter 10: Outcomes Research in Gastrointestinal Treatments

Outcomes and Clinical Data Mining: Studies and Frameworks. Cerrito PB, Editor. IGI Publishing. 2010

Presentations and Posters

APHA 2010 "Administrating TDaP during Pregnancy Increases a Newborn's Protection against Pertussis, Diphtheria and Tetanus"	Denver, CO
Annual Pediatric Nursing Conference 2010 "Increasing Burn Prevention Knowledge via Online Education"	Philadelphia, PA
MWSUG 2010 "Gastrointestinal Diseases: Diagnoses, Misdiagnoses, and Comorbidities"	Milwaukee, WI
ISPOR 2010 "Patterns in Prescriptions of Antipsychotics and their Outcomes"	Atlanta, GA
MWSUG 2009 "Irritable Bowel Syndrome and Mood Disorders"	Cleveland, OH
MWSUG 2008 "Relations between IBS and Depression"	Indianapolis, IN
SESUG 2007 "Gasoline Prices in Kentucky and Worldwide Oil Prices"	Hilton Head, SC

Community Service and Extracurricular Activities

Broward Community College	Swim Team, August 1995 – July 1997 Swimming Instructor, Summer 1996 and Summer 1997
Delta State University	Swim Team, August 1997 – May 1999
YMCA of Middle Tennessee	Swimming Instructor, August 2000 to July 2001
Siloam Family Health Center	English-Spanish Medical Interpreter, July 2004 to July 2006