Preface

The purpose of this text is to show how the investigation of healthcare databases can be used to examine physician decisions to develop evidence-based treatment guidelines that optimize patient outcomes. This book is intended for interested healthcare researchers who want to start working with available health outcomes databases, but who do not know how to get started. It is for the researcher who may not have a lot of computing resources available and who will have to perform all needed analyses, but who lacks many of the computing skills required to complete the project. Therefore, we are providing some basic information on processing the data with step-by-step instructions so that you will be able to use your own data and follow the instructions to find meaningful results.

Clinical trials cannot be used to investigate all possible treatment decisions that are required for patients. In particular, clinical trials have inclusion/exclusion criteria that ensure that the treatment population is not representative of the general population. For example, there are studies that look at the use of non-steroidal anti-inflammatory medications (NSAIDs) to treat patients with rheumatoid arthritis. At the same time, there are studies that indicate that NSAIDs should be avoided if the patient has congestive heart failure. However, what about the patient who has both rheumatoid arthritis and congestive heart failure? Are NSAIDs indicated, or should they be avoided? There is an absence of any studies that address this sub-population of patients. Usually, patients with congestive heart failure are excluded from clinical trials for arthritis medications. However, there are many such patients in the treated population, and because of a lack of guidelines, there is considerable variability in the treatment decisions that are made. It is possible, then, to investigate the relationship between treatment decisions and patient outcomes to find which treatment choices are optimal. An example of that is the treatment of osteomyelitis with resistant infection. In some cases, the treatment will be amputation while in others, it will be antibiotics. We can look at the dose, duration, and type of antibiotic used to see if there is recurrence that results in amputation, or if the antibiotic treatment combination prevents the need for amputation.

One of the most important activities when using health outcomes databases is the necessary preprocessing of the data, including extraction of specific patients using inclusion/exclusion criteria. If the preprocessing is done improperly, the results will have little value. Unfortunately, in most of the medical literature, the preprocessing steps are rarely discussed, so that it is impossible for anyone to evaluate the quality of the preprocessing. Frequently, the preprocessing is 90% of the entire analysis. Therefore, it is a crucial aspect that needs to be considered carefully. We will spend considerable time and effort in this book discussing and demonstrating preprocessing requirements.

Another important problem is to extract information about patient conditions and patient procedures from the databases. While most databases identify the primary diagnosis and procedure, co-morbidities can exist in multiple columns of data. These columns have to be combined in some way so that the
extraction can take place. Another problem is that the treatment for chronic diseases is sequential based upon the progression of the disease. Episodes of treatment need to be identified, and time-dependent variables need to be utilized in any analysis; they require the use of time-dependent techniques.

Probably the most important aspect of the preprocessing of large, healthcare databases is how to handle all of the reported patient conditions and co-morbidities. While patients can be matched on the basis of demographic information, it is much more difficult to match based upon all of the co-morbidities. Should patients be matched exactly based upon conditions, or should they be matched on equivalent levels of severity? The problem is that defining a ranking of severity is also difficult. Is a patient on dialysis more or less severe compared to a patient with congestive heart failure? Should acute conditions be considered in defining the level of severity (including severe infections), or should chronic conditions only be used to define severity? We suggest a number of alternative methods that can be used to examine the patient’s co-morbidities. In particular, we provide a method that can use all of the patient co-morbidities instead of relying on a compressed list of codes. This method utilizes the linkages between codes rather than to make false assumptions that co-morbidities are independent of each other when it is well known that patients with certain conditions are more likely to have other co-morbidities.

Because it is not always possible to write a specific hypothesis when you do not yet know enough about the data to define one, this text takes more of an exploratory approach to investigating the data. In this way, it is also possible to generate hypotheses that can then be analyzed using additional data. One of the advantages of using many of the available databases for health outcomes research is that the datasets are large and can be partitioned in many ways so that part of the data can be used to generate hypotheses and another part can be used to validate them.

Because of the large size of many of these databases, it is not possible to use traditional statistical methods. Such methods work with relatively small datasets. The administrative and clinical databases can have thousands or millions of patient records. Any statistical test has four parameters: Type I error, Type II error (or power), sample size, and effect size. Consideration of the effect size is often omitted when performing statistical inference. Specifying three of the parameters fixes the fourth. If the sample size is very large, the effect size is virtually zero, meaning that the hypothesis will be statistically significant because the statistical test finds a minute difference between groups.

In clinical trials, a power analysis computes the sample size after first fixing Type I error, power, and effect size. The effect size is half the width of the confidence limit surrounding the hypothesized population measure. For a simple test of the mean, $H_0: \mu = \mu_0$, the effect size is equal to $\frac{s}{n}$, where $s$ is the sample standard deviation and $n$ is the sample size. Note that as $n$ increases, the value of $\frac{s}{n}$ decreases so that the effect size begins to converge to zero. For this reason, the p-value will be statistically significant, but the effect size will be so small that the result will have no practical importance. This is true when using the general linear model or the generalized linear model with a different link function. The standard error is still computed based upon assumptions involving the Central Limit Theorem and the definition of standard error.

Another problem with these large samples is that there is an assumption that the data entry is uniform. While that might be a reasonable assumption when the data are from just one healthcare provider, uniformity is not a reasonable assumption across providers. Therefore, we must be very careful about drawing conclusions when comparing different providers, especially in matters of quality and reimbursements. Instead, we can look to other techniques that do not require this assumption of uniformity.

We intend to examine the totality of patient care and outcomes, focusing on patients with chronic illnesses who receive multiple treatments, both as inpatients and as outpatients. This book represents
a continuation of our earlier texts, Text Mining Healthcare and Clinical Databases published by SAS Press, Inc and Text Mining Techniques for Healthcare Provider Quality Determination: Methods for Rank Comparisons published by IGI Publishing, Inc. Therefore, we will focus on additional data mining techniques as well as the needed preprocessing required to use data mining techniques for healthcare research.

Once the data have been preprocessed, we apply data mining and text mining tools to extract useful knowledge from the data. These include predictive modeling, market basket analysis, and clustering the patient observations. We will demonstrate how these techniques are important in the use of large, electronic databases. In particular, one of the biggest issues is how to classify and define a patient’s condition when it is identified using a series of diagnosis codes. Is a patient with diabetes more or less severe compared to a patient with asthma? How can we tell or decide which patient is more severe? Because the datasets are large and contain information about patient co-morbidities, we can also drill down into the data to examine the potential for confounding factors. Confounders are often a problem with regard to observational studies. It is the major reason that observational studies provide erroneous conclusions; confounders are ignored and the conclusion is more directly related to the confounders than the variables studied.

Chapter 1 gives a general introduction to the use of data mining techniques to investigate data. It also gives an introduction concerning the necessary and required preprocessing needed to investigate the data. In particular, we will use data from the Medical Expenditure Panel Survey, the National Inpatient Sample, Medpar, and the Centers for Medicare and Medicaid. We will use SAS Enterprise Guide and SAS Enterprise Miner to investigate the data. Enterprise Guide is a point-and-click user interface to the statistical software, SAS. It provides sufficient tools to preprocess data. Enterprise Miner is the SAS component used for data mining. Chapter 2 discusses the problems of missing values, and also of errors in the dataset.

Chapters 3, 4, and 5 examine some of the basic preprocessing requirements for the Medical Expenditure Panel Survey, Medpar data, and the National Inpatient Sample respectively. In particular, the chapters discuss how to extract a subsample of the dataset, and how to compute basic summary information.

In these chapters, we include information concerning the need to audit the data, the need to cross-reference variables in the dataset, and the use of propensity scores. We discuss the extraction of subsamples by diagnosis or procedure, and the use of matched samples for comparison purposes. In particular, in Chapter 5, we will take advantage of the rich information contained in the National Inpatient Sample to investigate the relationship between DRG (Diagnosis Related Groups) codes used for billing and ICD9 codes used to specify patient conditions. We examine the results of DRG Grouper software.

Chapters 6 and 7 discuss some of the preprocessing problems of changing the observational unit in the data. The data that are collected are not often in an optimal form for data mining. Before the data mining tools can be used, the observational unit has to be modified. For example, in a prescription database, the observational unit is often the prescription. We might want to modify that so that the patient is the observational unit. Then, we need to find a way to relate all of the prescriptions for one patient into one observation. This is called a one-to-many relationship. A many-to-many relationship may be in a claims database with a table of inpatient treatments and another table of prescriptions. Some patients will have multiple prescriptions and multiple inpatient episodes. We need to be able to relate all of this information to the patient, so that it can be analyzed by patient. Chapter 7 discusses the many-to-many relationship in the context of investigating data from the Medical Expenditure Panel Survey (MEPS). It demonstrates how datasets from different healthcare providers can be used to extract information concern-
ing the totality of care for the patient. It can be used, for example, to investigate the relationship between patient treatment for a chronic disease and the need for the patient to have emergency and inpatient care.

Another subject discussed briefly in Chapter 8 are some issues that are important when dealing with large datasets. In particular, it is important to determine the influence of outliers, and to see if the outliers define the entire model. If this occurs, these outliers need to be removed to ensure that the model gives accurate results. We will provide a possible solution. We also discuss how to extract patients with specific conditions, and demonstrate why problems occur if extraction is based upon the primary procedure or diagnosis. We show the difference between extracting from the primary diagnosis and extracting from all diagnoses.

Chapter 9 examines the use of time series methods that are helpful in examining general trends in the data. It can be used to examine trends in prescribing medications, and it can be used to investigate how a new drug can permeate the market and for what patient conditions it is used. While medications generally have approved uses, physicians can prescribe them for other conditions. Chapter 9 examines another aspect of data in sequence. Patients with chronic diseases generally have multiple episodes of care that can result in switching or adding medications, emergency or inpatient treatment, or disease progression. This progression exists in the data; it can be extracted to find the relationship between treatment and episodes.

Chapter 10 adds the issue of patient compliance, and demonstrates how compliance can be defined using information in the databases. In particular, we can examine compliance through prescription records. A patient who takes a medication daily should have a total of 365 doses dispensed in a year’s time. A patient who has just two thirds of the 365 doses will not be completely compliant. We can then use cutpoints, or percentages of full compliance, to rank the level of compliance. Care must be taken to ensure that it is a lack of prescriptions and not switching to a new medication in the middle of the year that we use to define compliance.

Another big issue in working with these large, complex databases is the fact that patients have many co-morbidities. The different codes that are used to document a patient condition can have thousands of different possible values. In order to work with these conditions in predictive models or in statistical models, there has to be some way to compress these conditions to reduce the total number to a few. In the past, patient severity scores have been defined and consensus panels were formed to define such a severity index. In Chapter 11, we give an alternate methodology that allows us to use all patient conditions while also not assuming that different providers code these conditions uniformly. Chapter 12 discusses more standard techniques for defining patient severity indices. In this chapter, we demonstrate some of the problems with current severity scores, and demonstrate a novel approach treating the codes as text and using text analysis software.

Chapter 13 investigates the use of data to estimate probabilities of treatment outcomes and adverse events to develop decision trees and treatment guidelines. Typically, probabilities are currently estimated using physician consensus panels and surveys. We can take advantage of the variability of treatment decisions to relate them to outcomes. The large databases can then be used to compute probability estimators. We want to demonstrate how the use of actual data can be used to find optimal pathways that can be used in the place of consensus panels. We can estimate the risk of adverse events. In addition, we will discuss the use of decision trees in comparative effectiveness analysis and how it leads to the rationing of care.

In chapters 14-16, we give examples of analyses to investigate the relationship between treatment and outcome. We investigate wound care, especially the treatment of osteomyelitis. We also examine
the treatment of asthma and COPD in relationship to the medications used to treat the diseases. We are especially interested in whether there are medication treatments that reduce the need for emergency and inpatient care. We want to demonstrate how the totality of observational data can be used to extract meaningful information about patient care. These examples will show how the different data mining techniques can be used to discover new knowledge concerning patient conditions and their treatments. We use the techniques of Market Basket Analysis (also called association rules), text analysis, and survival data mining to investigate the patient conditions.

The final chapter gives a general discussion of the methods and their further potential in investigating health outcomes. It generally summarizes what was developed throughout the textbook. It will also give suggestions for additional study using these databases. In particular, we advocate an exploratory rather than inferential approach when investigating the data.

It is our interest in writing this book to demonstrate how concern for details in the datasets, and the use of data mining techniques can extract important and meaningful knowledge from the data. We anticipate that this knowledge can be used to improve patient care.