Chapter 1
Preprocessing the Data

GENERAL INTRODUCTION

In this book, we provide tools that are needed to investigate administrative and clinical databases that are routinely collected in the support of patient treatment. Often, these databases are large and require non-traditional methodology to investigate. In addition, as they are collected for purposes other than research, there is considerable preprocessing that is required in order to use the data for the purpose of analysis in order to find important results that can improve the quality of patient care.

Therefore, we will show by example just how to preprocess the data, and how non-traditional statistical methods can be used to investigate the data and to extract meaning from the databases.

We will show details and programming code necessary to complete the preprocessing, and we will discuss the type of preprocessing necessary to use each statistical method and data mining technique.

We will also demonstrate techniques such as time series analysis and survival data mining that can be used to extract meaningful information concerning patient treatments. Time series analysis can be used to investigate general trends in terms of prescribing or treatments that are prescribed to patients. Survival data mining can investigate the progression of chronic diseases in relationship to patient outcomes. This would be necessary to investigate the relationship of Type II medications to the progression of the diabetes, considering the time to renal or heart failure, or the shift from Type II medications to insulin and eventually to death.

DOI: 10.4018/978-1-61520-905-7.ch001
What we want to demonstrate in this book is how the complexity of the patient record as depicted in billing or clinical databases can be used to investigate the entire patient condition, and the sequence of patient healthcare needs. We will also demonstrate how the databases can be used to investigate general trends in the data. We want to show how meaningful results can be found when examining the patient record in its entire complexity. We want to use these techniques to drill down into individual details and to examine long-term trends rather than to focus on groups and short-term outcomes. We want to investigate patients as they are actually treated in terms of disease progression.

One of the biggest problems in using these databases is that there is often a one-to-many relationship in the data. That means that there are multiple records per patient. For example, a pharmacy database defines its observational unit as a prescription. A patient with multiple prescriptions will have many different observations. Similarly, an insurer’s observational unit is a claim, with multiple claims for any one patient. It is necessary to modify the database to construct a one-to-one relationship so that the patient is the observational unit. This can be done in multiple ways. Since different databases record this information in different ways, we will be particularly careful in discussing the preprocessing required to make these transitions.

In particular, we will examine data from a wound care center, from a cohort of approximately 30,000 patients, from the Centers for Medicare and Medicaid, and from a listing of inpatient discharges. Each of these databases provide different types of information. We will look at trends in the treatment of individuals, and also trends in the general reliance upon medical resources. In particular, we will examine the impact of Medicare, part D on the use and cost of medications for chronic illnesses. In the initial examination of the data, we find a 10-fold increase in the use of Medicare for prescription costs in 2006, the first year of Medicare, part D, compared to the year, 2005. There are shifts from Medicaid, self-payment, and private insurance payments to Medicare, part D. However, these shifts can vary from medication to medication. We will also examine trends in medication prescriptions generally.

Throughout this text, we will emphasize the preprocessing of data. Before any statistical analysis can take place, the data need to be extracted. This is particularly true if we are extracting information about primary versus secondary conditions. It is very important that when studying a subsample, that the subsample is exactly the one sought. The coding and techniques used in this book should be adapted to your data and explorations.

Another common preprocessing task is to extract a subsample of patients based upon primary and secondary patient conditions. Frequently, propensity scores are used to define matched samples. This particular preprocessing will be examined in great detail throughout this text.

**BACKGROUND**

**Statistical Methods with Large Healthcare Databases**

Generally, the focus of previous texts relate to the use of computers as an aid in clinical decision making. (Berner 2006) One text discusses the use of outcomes data to determine best practices. (Epstein and Blumenfield 2002) Many studies now rely on publicly available, national databases, but use the traditional statistical methodology. Given the number of research publications that use these national databases, including some databases that are available commercially at very high cost, there should be some information available on how to examine the information within the databases. However, there are no texts that focus on the preparation of the data for analysis. Research publications tend to spend very little time discuss-