Preface

This book gives several case studies developed by faculty and graduates of the University of Louisville's PhD program in Applied and Industrial Mathematics. The program is generally focused on applications in industry, and many of the program members focus on the study of health outcomes research using data mining techniques. Because so much data are now becoming readily available to investigate health outcomes, it is important to examine just how statistical models are used to do this. In particular, many of the datasets are large, and it is no longer possible to restrict attention to regression models and p-values.

Clinical databases tend to be very large. They are so large that the standard measure of a model's effectiveness, the p-value, will become statistically significant with an effect size that is nearly zero. Therefore, other measures need to be used to gauge a model's effectiveness. Currently, the only measure used in medical studies is the p-value. In linear regression or the general linear model, it would not be unusual to have a model that is statistically significant but with an r^2 value of 2% or less, suggesting that most of the variability in the outcome variable remains unaccounted for. It is a sure sign that there are too many patient observations in a model when most of the p-values are equal to '<0.00001'. The simplest solution, of course, is to reduce the size of the sample to one that is meaningful in regression. However, sampling does not utilize all of the potential information that is available in the data set and a reduction in the size of the sample requires a reduction in the number of variables used in the model so as to avoid the problem of over-fitting. Unfortunately, few studies that have been published in the medical literature using large samples take any of these problems into consideration.

An advantage of using data mining techniques is that we can investigate outcomes at the patient level rather than at the group level. Typically in regression, we look to patient type to determine those at high risk. Patients above a certain age represent one type. Patients who smoke represent another type. However, with data mining, we can examine and predict specific outcomes for a patient of a specific age who smokes 10 cigarettes a week, who drinks one glass of wine on weekends, and who is physically in good shape.

The purpose of using data mining is to explore the data so that the information gathered can be used to make decisions. In healthcare, the purpose is to make decisions with regard to patient treatment. Decision making does not necessarily require that a specific hypothesis test is generated and proven (or disproven). Exploration without a preconceived idea as to what will be discovered is also a valid means of data investigation.

A measure of the relationship of treatment decisions to patient outcomes that we can consider stems from the fact that physicians vary in how they treat similar patients. That variability itself can be used to examine the relationship between physician treatment decisions and patient outcomes. Once we determine which outcome is "best" from the patient's viewpoint, we can determine which treatment decisions

are more likely to lead to that decision. This is particularly true for patients with chronic illness where there is a sequence of treatment outcomes followed by multiple patient outcomes. For example, a patient with diabetes can start with medication tablets, and then progress to insulin injections. Such patients can potentially end up with organ failure: failure of the heart, kidney, and so on. We can examine treatments that prolong the time to such organ failure. In this way, data mining can find optimal treatments as a decision making process.

Health outcomes research depends upon datasets that are routinely collected in the course of patient treatment, but are observational in nature and are very large. Traditional statistical methods were developed for randomized trials that are typically small in terms of the number of subjects where the main focus is on just one outcome variable. Only a few independent, input variables were needed because of the property of randomness. With large, observational datasets, there are some very important issues that cannot be disregarded. In particular, there is always the potential of confounding factors that must be considered.

There are many examples in the medical literature of observational studies that did ignore confounding factors. For example, the study of cervical cancer initially focused on the birth control pill, ignoring the reasons that women chose to use the pill. More recently, the association between the HPV infection and cervical cancer has been established. Because of a general perception that bacteria cannot exist in the acid content of the stomach, there was a general perception that peptic ulcers were caused by stress. The treatment offered was psychological, and H.pylori was not even considered as a possibility. More recently, hormone replacement therapy was considered as a way to reduce heart disease in women until a randomized trial debunked the treatment. It became popular because many women with heart disease were initially denied the therapy because of a perception that the therapy could increase heart problems. Observational studies that ignore confounders and rely on the standard regression models can often result in completely wrong conclusions.

Large data sets are required to examine rare occurrences. There needs to be a sufficient number of rare occurrences in the database to be comparable. For example, if a condition occurs 0.1% of the time, there would be approximately one such occurrence for every 1000 patients, 10 occurrences for 10,000 patients, and so on. A minimum of 100,000 patients in the dataset would be required to find 100 occurrences. However, all 100,000 patients cannot be used in a model to predict these occurrences. The model would be nearly 99% accurate, but would predict nearly every patient as a non-occurrence. In the absence of large samples and long-term follow up, surrogate endpoints are still used. For example, instead of looking at the mortality rate or rate of heart attacks to test a statin medication, the surrogate endpoint of cholesterol level is used. Instead of testing a new vaccine to see if there is a reduction in the infection rate, blood levels are measured.

Instead of using traditional statistical techniques, the studies in this book use exploratory data analysis and data mining tools. These tools were designed to find patterns and trends in observational data. In large datasets, data mining can examine enough variables to investigate potential confounders. One of the major potential confounders is the collection of co-morbidities that many patients have. Interactions between medications and conditions needs to be examined within the model, and such interactions are costly in terms of degrees of freedom in traditional regression models. They require large samples for analysis.

Chapter 1 gives a brief introduction to the data mining techniques that are used throughout the cases. The methods are used to drill down and discover important information in the datasets that are investigated in this casebook. These techniques include market basket analysis, predictive modeling, time

series analysis, survival data mining, and text mining. In particular, it discusses an important, but little used technique known as kernel density estimation. This is a means of estimating the entire population distribution. While it is typical to assume that the population has a normal distribution with a bell-shaped density curve, that assumption is not valid if the population is heterogeneous, or is skewed. Using observational data concerning patient treatment, the population is always heterogeneous and skewed. Therefore, the standard assumptions used for defining linear and regression models are not valid. Other techniques must be used instead.

Generally, the cases in this book use datasets that are publicly available for the purpose of research. These include the National Inpatient Sample (NIS) and the Medical Expenditure Panel Survey (MEPS) available via the National Center for Health Statistics. The National Inpatient Sample contains a stratified sample of all inpatient events from 1000 different hospitals scattered over 37 states. It is published every year, two years behind the admission dates. The Medical Expenditure Panel Survey collects information about all contacts with the healthcare profession for a cohort of 30,000 individuals scattered over approximately 11,000 households. A different cohort has been collected each year since 1996. Each individual has data collected for two years. It contains actual cost and payment information; most other publicly available datasets contain information about charges only. Therefore, the MEPS is used to make estimates on healthcare expenditures by the population generally.

In addition, data from Thomson Medstat were used for some of the cases. Thomson Medstat collects all claims data from 100 insurance providers for approximately 40 million individuals. These individuals are followed longitudinally. These data were used in several of the cases as well. The remaining data were from local sources and used to investigate more specific questions of healthcare delivery. Thomson has a program to make its data available for student dissertation research, and we greatly appreciate the support.

The first section of the casebook contains various studies of outcomes research to investigate physician decision making. These case studies include an examination into the treatment of osteomyelitis, cardiovascular by-pass surgery versus angioplasty, the treatment of asthma, and the treatment of both lung cancer and breast cancer. In addition, there is a chapter related to the use of physical therapy as an attempt to avoid surgery for orthopedic problems and a study related to patient compliance with treatment in relationship to diagnosis. In particular, it discusses the importance of data visualization techniques as a means of data discovery and decision making using these large healthcare datasets.

One of the major findings from this section is that amputation is in fact the primary treatment for osteomyelitis for patients with diabetes, as discussed in detail in Chapter 2. Physicians are reluctant to prescribe antibiotics and often use inappropriate antibiotics for too short durations, resulting in recurrence of the infection. Amputation is assumed to eradicate the infection even though the amputations can often become sequential. This study demonstrates very clearly how treatment perception can be used for prescribing in the absence of information from the study of these outcomes datasets.

Chapter 3 examines the results in cardiovascular surgery where the major choice is CABG (cardiovascular bypass graft) or angioplasty. The introduction of the eluting stent in 2002 changed the dynamics of that choice. It appears that the eluting stent yields results that are very comparable to bypass surgery. The data here were examined using survival data mining. It shows the importance of defining an episode of care from claims datasets, and to be able to distinguish between different episodes of treatment.

The next chapter, 4, examines treatment choices for the chronic condition of asthma. This chapter investigates the various medications that are available for treatment, and how they are prescribed by

physicians. It also examines the treatment of patients in the hospital for patients who have asthma. This study used both the NIS and MEPS to investigate both medication and inpatient treatment of asthma.

The next two chapters look at two different types of cancer, breast cancer and lung cancer. The purpose is to examine different treatment choices. In the first case, the purpose is to investigate the choice between a lumpectomy and mastectomy, and the patient conditions that might be related to these choices. In the second, we are also looking at treatment choices and the various regimens of chemotherapy.

A similar question motivates Chapter 7, which looks at the tendency to require physical therapy with the intent of preventing the need for surgery for orthopedic complaints. This chapter also examines the preprocessing necessary to investigate healthcare data. This study, too, relies upon the definition of an episode, and also on the definition of the zero time point. In this example, the zero point starts at physical therapy and the survival model ends with surgery. Patients are censored if they do not undergo the surgery.

Chapter 8 examines the relationship of patient procedures to inpatient care. It demonstrates that the compliance of patients in testing blood glucose reduces the cost of treatment. The importance of this monitoring cannot be understated. Similarly, chapter 9 looks at patient compliance and the patient condition in dental care. It shows that patients with the worst dental problems have the least compliance with treatment. Do they have such problems because of a lack of compliance, or are the most compliant the ones who have the best dental outcomes?

The final three chapters in section one examine the treatment of gastrointestinal problems and their relationship to mental disorders, the condition of hydrocephalus in infants, and common problems in childhood and adolescence. In particular, this chapter examines the issue of adolescent obesity and also some issues with vaccines in childhood and adolescents.

The second section of this book is related to case studies in healthcare delivery. Two of the studies examine healthcare delivery in the hospital emergency department. The first examines the scheduling of personnel; the second examines the patients who present at the emergency department. The objective of the first study concerning the emergency department is to use time series techniques to predict the need for personnel throughout the day. The second study looked at the detailed demographic information of patients presenting to the emergency department to determine the relationship between the demographics and the type of visit, non-urgent, urgent, or emergency conditions. The goal was to determine which patients should be referred to a no-cost clinic that treats patients with chronic conditions at no charge. It introduces another type of analysis, that of spatial data and spatial analysis using geographic information systems (GIS).

A third case study in the section examines time trends in physician prescribing of antibiotics and a fourth looks at the current process of reimbursing hospital providers by negotiated amount for a specific DRG code. One additional paper in this section relates to the information contained within the voluntary reporting of adverse events as supported by the Centers for Disease Control, or CDC. It looks at some standard issues in the treatment of pediatric patients, including the issue of obesity and exercise. Many of the studies in this section rely upon the use of time series methods to investigate health and treatment trends.

The third section in the book looks at the use of data mining techniques to model the relationship between brain activity and cognitive functioning. It is possible that some children are treated for learning disabilities when they should be treated instead for sleep apnea. This can have considerable impact on the type and amount of medication that is typically prescribed for problems such as ADHD. The tech-

niques shown in this final chapter can also be used in microbiology research related to gene networks and interactions.

All of these examples can give the reader some excellent concepts of how data mining techniques can be used to investigate these datasets to enhance decision making. These large databases are invaluable in investigating general trends, and also to provide individual results.

Patricia Cerrito
Editor