Chapter 1
Introduction to Data Mining Methodology to Investigate Health Outcomes

Patricia Cerrito
University of Louisville, USA

ABSTRACT
In this case, we provide some basic information concerning the statistical methods used throughout this text. These methods include visualization by kernel density estimation, regression and predictive models, time series, and text analysis. This case is not intended to provide extensive information concerning these techniques; references are provided for the interested reader who wants to follow up on the methodology. Some of the methodology is fairly well known while the predictive modeling and text analysis are relatively recent. One of the primary visualization techniques used is that of kernel density estimation. Since the patient base is relatively heterogeneous, we cannot assume that any health outcome is normally distributed. Kernel density estimation can find the entire population distribution so that we no longer have to rely upon averages and can relate the results more to individuals.

INTRODUCTION
The “gold standard” for investigating medical treatments is the randomized, controlled, double-blind study. Volunteers are recruited and randomized into a control or treatment group; the physician is blinded as to which patients receive the treatment and which receive a placebo. The patients are also blinded and unaware of whether they receive the actual treatment or not. Because such studies are expensive, they can only be performed on a small number of potential treatments. They are done using a minimum number of subjects, usually on the patients at highest risk. Randomized studies are usually relatively short term and often use surrogate endpoints because it would take too long to examine real endpoints.

Because of informed consent requirements, such randomized, blinded studies are impossible to examine the effectiveness of a surgical procedure. Also, patients are usually excluded if they have multiple conditions that require more than one medication.
Until recently, young women with the potential to become pregnant were excluded from such trials for fear of damaging the woman’s baby; children were excluded as well. Therefore, there are many studies that cannot be performed using a randomized design. We cannot, for example, look to the randomization of patients into levels of cigarette use, alcohol use, or level of sexual activity.

In this book, we present a number of examples of using data mining techniques and healthcare databases to examine healthcare decision making. Because any dataset collected in the routine of treating patients, including billing data, is quite extensive, the more traditional statistical methods cannot be used to investigate the data. (Cerrito, 2009a) This chapter is not intended to be a complete discussion of data mining; for more information, we refer the reader to texts by Cerrito (2009b) and deVille (2006).

**DATA MINING TECHNIQUES**

There are many different methods that can be used to investigate the data. We will discuss them briefly here. For more information, we refer the reader to Cerrito. (Cerrito, 2009a) These techniques include data visualization, predictive modeling, market basket analysis, time series analysis, and text analysis. Data mining is a process as well as a collection of techniques. We will list some of the differences between data mining and traditional statistical techniques.

Data mining deals with heterogeneous data, sometimes with a complex internal structure such as multimedia; including images, video, and text. Because most data collected in healthcare related to the routine of patient treatment consists of heterogeneous populations, the techniques of data mining are ideal to use with the various healthcare datasets. Data mining starts with the assumption that the raw data set is not of sufficient quality to apply a statistical model directly without some appropriate preprocessing techniques such that the preprocessing will have as much or even more influence on the quality of the final results compared to the selected statistical technique. Data mining uses flexible predictive techniques that often are based on algorithmic foundations, but may have weak formal statistical justification. The data mining process often uses hidden variables as tools to perform a step-by-step compression of raw input data. Data mining attempts to find not only general, global models based on a data set, but also to find basic patterns in the data. Data mining is more concentrated on the aspects of data management and optimization of a search through the data while statistics is more oriented toward formalisms for final model representation, and score function formalization in the data space to perform inference. Data mining has been more focused on estimation, and the process generally ignores inferential models.

Data mining and statistics have generally developed in different domains. Statisticians are primarily interested in inference; data miners in exploratory data analysis. Nevertheless, there are some instances where data mining and statistics have blended. Many statisticians remain dubious about the data mining process. (Lee, 1995) Others are concerned with the lack of a theoretical framework similar to the one for inferential statistics, especially since data mining tends to be algorithmic-based. (Giudici & Passerone, 2002; Hand & Bolton, 2004; Sargan, 2001)

Statistics and data mining differ in the use of machine learning methods, the volume of data, and the role of computational complexity. Our need for analysis is exceeding our abilities to handle the complexity. (Hosking, Pednault, & Sudan, 1997; Keim, Mansmann, Schneidewind, & Ziegler, 2006) Preprocessing is far more important with large datasets, especially as we approach the petabyte level, although healthcare data have yet to approach that level of size, with gigabytes considered to be extensively large. (Mannila, 1996) However, there are indications that data mining is focused on the data mining process itself with little
Related Content

**Reporting Clinical Gait Analysis Data**
[www.igi-global.com/chapter/reporting-clinical-gait-analysis-data/53651?camid=4v1a](www.igi-global.com/chapter/reporting-clinical-gait-analysis-data/53651?camid=4v1a)

**The Graphic Display of Labor Events**
[www.igi-global.com/chapter/graphic-display-labor-events/53644?camid=4v1a](www.igi-global.com/chapter/graphic-display-labor-events/53644?camid=4v1a)

**Organizational Factors: Their Role in Health Informatics Implementation**
[www.igi-global.com/chapter/organizational-factors-their-role-health/26196?camid=4v1a](www.igi-global.com/chapter/organizational-factors-their-role-health/26196?camid=4v1a)

**The Extension of Theory and Methodology to B-Splines**
Carlo Ciulla (2009). *Improved Signal and Image Interpolation in Biomedical Applications: The Case of Magnetic Resonance Imaging (MRI)* (pp. 223-238).
[www.igi-global.com/chapter/extension-theory-methodology-splines/22500?camid=4v1a](www.igi-global.com/chapter/extension-theory-methodology-splines/22500?camid=4v1a)