IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA ITB10358
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

Chapter VII

Data Mining and Knowledge Discovery in Healthcare Organizations: A Decision-Tree Approach

Murat Caner Testik, Cukurova University, Turkey

George C. Runger, Arizona State University, USA

Bradford Kirkman-Liff, Arizona State University, USA

Edward A. Smith, University of Arizona and Translational Genomics Research Institute, USA

ABSTRACT

Health care organizations are struggling to find new ways to cut healthcare utilization and costs while improving quality and outcomes. Predictive models that have been developed to predict global utilization for a healthcare organization cannot be used to predict the behavior of individuals. On the other hand, massive amounts of healthcare data are available in databases that can be used for exploring patterns and therefore knowledge discovery. Diversity and complexity of the healthcare data requires attention to the use of statistical methods. By nature, healthcare data are multivariate, making the analysis difficult as well as interesting. In this chapter, our intention is to classify individuals that are future high-utilizers of healthcare. In particular, we answer the question of whether a mathematical model can be generated utilizing a large claims

database that will predict which individuals who are not using a service in a yet untested database will be high utilizers of that health service in the future. For this purpose, an integrated dataset from enrollment, medical claims, and pharmacy databases containing more than 150 million medical and pharmacy claim line items and for over four million patients is analyzed for knowledge discovery. A modern data-mining tool, namely decision trees, which may have a broad range of applications in healthcare organizations, was used in our analyses and a discussion of this valuable tool is provided. The results and managerial aspects are discussed. Several approaches are proposed for the use of this technique depending on the health plan.

INTRODUCTION

Many predictive models have been developed in healthcare in the past (Ash, 1999; Dunn, 1998; Dunn et al., 1995; Epstein & Cumella, 1988; Newhouse, 1986, 1995, 1998; van Vliet & Lamers, 1998; Weiner et al., 1995). However, for the most part, these models focused on how manipulation of plan design (deductibles, pays, etc.) will influence utilization behavior and to adjust for case-mix and risk for the purpose of predicting global costs and setting capitated reimbursement rates. Until recently, there has been little interest in applying prediction tools to individuals for the purpose of reducing costs and improving individual care. This lack of interest was mostly due to the absence of tools that can accurately predict future individual patient utilization, especially for patients who have had no current utilization. In general terms, current utilization of a particular kind of health service is the best predictor of future utilization of a particular kind of health service. Methods to predict future utilization of a particular service when there is no current utilization of the same service tend to produce results that are not meaningful for program managers.

Today, with the rapid increase in the generation and collection of data, researchers are able to explore patterns hidden in large databases. Massive amounts of healthcare data are also available in databases that can be used for knowledge discovery. Diversity and complexity of the healthcare data requires attention to the use of statistical methods. By nature, healthcare data are multivariate, making the analysis difficult as well as interesting.

The main objective of this research is to answer the question of whether a mathematical model can be generated utilizing a large claims database that will predict which individuals who are not using a service in a yet untested database will be high utilizers of that health service in the future. For this purpose we used a massive dataset containing more than 150 million medical and pharmacy claim line items and for over four million patients.

This research differs from previous related studies in a number of ways: (a) The focus is on identifying individuals for targeted interventions who currently have no use of the service under study; (b) An integrated dataset from commonly available data found in enrollment, medical claims, and pharmacy databases is used; (c) A model that is built on more advanced "episodes of care" cost groupings rather than merely raw claims data; and (d) A modern data mining technique, namely a decision tree, is used for knowledge discovery.