Chapter 8.21
Mining Clinical Trial Data

Jose Ma. J. Alvir
Pfizer Inc., USA

Javier Cabrera
Rutgers University, USA

Frank Caridi
Pfizer Inc., USA

Ha Nguyen
Pfizer Inc., USA

ABSTRACT

Mining clinical trails is becoming an important tool for extracting information that might help design better clinical trials. One important objective is to identify characteristics of a subset of cases that responds substantially differently than the rest. For example, what are the characteristics of placebo respondents? Who have the best or worst response to a particular treatment? Are there subsets among the treated group who perform particularly well? In this chapter we give an overview of the processes of conducting clinical trials and the places where data mining might be of interest. We also introduce an algorithm for constructing data mining trees that are very useful for answering the above questions by detecting interesting features of the data. We illustrate the ARF method with an analysis of data from four placebo-controlled trials of ziprasidone in schizophrenia.

INTRODUCTION

Data mining is a broad area aimed at extracting relevant information from data. In the 1950s and 60s, J.W. Tukey (1952) introduced the concepts and methods of exploratory data analysis (EDA). Until the early 1980s, EDA methods focused mainly on the analysis of small to medium size datasets using data visualization, data computations and simulations. But the computer revolution created an explosion in data acquisition and in data processing capabilities that demanded the expansion of EDA methods into the new area of data mining. Data mining was created as a large umbrella including simple analysis and visualization of massive data together with more theoretical areas like machine learning or machine vision.

In the biopharmaceutical field, clinical repositories contain large amounts of information from many studies on individual subjects and their characteristics and outcomes. These include data collected to test the safety and efficacy of promising drug compounds, the bases on which a pharmaceutical company submits a new drug application (NDA) to the Food and Drug Administration in the United States. These data may also include data from postmarketing studies that are carried out after the drug has already been approved for marketing. However, in many circumstances, possible hidden relationship and patterns within these data are not fully explored due to the lack of an easy-to-use exploratory statistical tool. In the next sections, we will discuss the basic ideas behind clinical trials, introduce the active region finder methodology and apply it to a clinical problem.

CLINICAL TRIALS

Clinical trials collect large amounts of data ranging from patient demographics, medical history and clinical signs and symptoms of disease to measures of disease state, clinical outcomes and side effects.

Typically, it will take a pharmaceutical company 8-10 years and $800 million to $1.3 billion to develop a promising molecule discovered in the laboratory into a marketed drug (Girard, 2005). To have a medicine approved by the Food and Drug Administration, the sponsoring pharmaceutical company has to take the compound through many development stages (see Figure 1). To obtain final approval to market a drug, it's efficacy, compared to an appropriate control group (usually a placebo group), must be confirmed in at least two clinical trials (U.S. Department of Health and Human Services, FDA, CDER, CBER, 1998). These clinical trials are well-controlled, randomized and double-blind (Pocock, 1999). Briefly, the compound has to show in vitro efficacy and efficacy/safety in animal models. Once approved for testing in humans, the toxicity, pharmacokinetic properties and dosage have to be studied (phase I studies), then tested in a small number of patients for efficacy and tolerability (phase II studies) before running large clinical trials (phase III studies). After the drug is approved for marketing, additional trials are conducted to monitor adverse events, to study the morbidity and mortality, and to market the product (phase IV studies).

In a typical clinical trial, primary and secondary objectives, in terms of the clinical endpoints that measure safety and/or efficacy, are clearly defined in the protocol. Case report forms are developed to collect a large number of observations on each patient for the safety and efficacy endpoints (variables) that are relevant to the primary and secondary objectives of the trial. The statistical methodology and hypothesis to be tested have to be clearly specified in a statistical analysis plan (SAP) prior to the unblinding of the randomization code. The results are summarized in a clinical study report (CSR) in a structured format in the sense that only primary and secondary hypotheses specified in the protocol and SAP will be presented in detail. The discussion section might include some post hoc analyses but generally these additional results do not carry the same weight as the primary/secondary analyses.

The plan of analysis is usually defined in terms of the primary clinical endpoints and statistical analyses that address the primary objective of the trial. Secondary analyses are defined in an analogous manner. *The primary analysis* defines the clinical measurements of disease state along with the appropriate statistical hypotheses (or estimations) and statistical criteria that are *necessary* to demonstrate the primary scientific hypothesis of the trial. *Secondary analyses* are similarly defined to address the secondary scientific hypotheses under study, or to support the primary analysis in the sense of elucidating the primary result or