Chapter 9
More Survival Data Mining of Multiple Time of Endpoints

INTRODUCTION

Survival analysis is almost always reserved for an endpoint of mortality or recurrence. (Mantel, 1966) However, it can be used for many different types of endpoints as the survival distribution is defined as the time to an event. That event can be any endpoint of interest. For patients with chronic diseases, there are many endpoints to examine. For example, patients with diabetes want to avoid organ failure as well as death, and treatments that can prolong the time to organ failure will be beneficial. For patients with resistant infections, treatments that prevent one or multiple recurrences should be explored.

Survival data mining differs from survival analysis in that multiple patient events can occur in sequence. The first step in survival data mining is to define an episode of treatment so that the patient events can be found for analysis. It can be thought of as a sequence of survival functions. In this chapter, we will look at the time to a switch in medications, and contrast how prescriptions are given to patients, either following or disregarding treatment guidelines.

Another issue in survival data mining is that there is no real starting time point in observational data. Survival analysis measures the time to an event starting with time 0. Without a specific starting point, there is no time 0. In other words, the data can be left-censored as well as right-censored, and that is more difficult to examine. We will discuss some techniques to define the zero point.

DOI: 10.4018/978-1-61520-905-7.ch009
BACKGROUND

There are not many resources for survival data mining, although many exist for survival analysis, or the time to an event. (Aalen, Borgan, & Gjessing, 2008; Allison, 1995; Potts, 2007) We will not produce all of that material here. We will just give a brief introduction. This statistical model is common to examine the time to death or the time to recurrence of disease. It is more general than that, examining the time to a specified event, whatever that event is defined to be.

There are two major survival analysis statistics: log rank statistics and Cox Regression. (Grambsch & Grambsch, 2001) They are based largely upon the exponential distribution rather than the normal. The log rank statistic is nonparametric. However, it has limitations in that it is not possible to do more than compare survival across specified groups. Cox regression is more detailed in that independent variables, even variables that are time dependent can be used to predict the time to the endpoint. However, it also requires the assumption of proportional hazards. The assumption requires that the probability of the time to the event is not related to the point in time but only to the difference in time. In other words, a patient with cancer remission is not less likely to have recurrence at 4.5 years after remission compared to just 1 year distant from treatment. As many chronic diseases are progressive, such an assumption may not always be realistic. Therefore, we have to look carefully at any situation before we apply Cox regression.

Survival data mining can be thought of as multiple survival analyses computed in sequence. In addition, it can look at different types of events in that sequence. One that is of particular interest to patients with chronic diseases is the time to disease progression, however that progression is defined. For example, a patient with diabetes can examine the time to the need for insulin, the time to heart failure, the time to renal failure, and the time to death. Many of these progressions are identified during inpatient or outpatient treatment, so they can be identified in claims databases.

LOG RANK STATISTICS

The logrank test statistic compares estimates of the hazard functions of the two groups at each observed event time. (Peto & Peto, 1972) It is constructed by computing the observed and expected number of events in one of the groups at each observed event time and then adding these to obtain an overall summary across all time points where there is an event.

The hazard rate is defined from the failure rate, \( \lambda(t) \), which can be thought of as the probability that a failure occurs in a specified interval, given that there is no failure before time \( t \). It can be defined with the aid of the survival function (also called the reliability function), \( R(t) \), the probability of no failure before time \( t \), such that

\[
\lambda(t) = \frac{R(t_1) - R(t_2)}{(t_2 - t_1) R(t_1)} = \frac{R(t) - R(t + \Delta t)}{\Delta t R(t)}
\]

where \( \Delta t = t_2 - t_1 \). This is a conditional probability. By taking the limit, we define the hazard function, which is the instantaneous failure rate at any point in time,

\[
h(t) = \lim_{\Delta t \to 0} \frac{R(t) - R(t + \Delta t)}{\Delta t \cdot R(t)}
\]

The continuous failure rate depends on a failure distribution, \( F(t) \), which is the cumulative distribution function that defines the probability of failure prior to time \( t \), \( P(T \leq t) = F(t) = 1 - R(t) \) for non-negative values of \( t \) where \( T \) is the failure time. The failure distribution function is the integral of the failure density function \( f(x) \) such that
Related Content

Generation of Scaffold Free 3-D Cartilage-Like Microtissues from Human Chondrocytes
www.igi-global.com/chapter/generation-scaffold-free-cartilage-like/71981?camid=4v1a

Revisiting the Feature and Content Gap for Landmark-Based and Image-to-Image Retrieval in Medical CBIR
www.igi-global.com/chapter/revisiting-feature-content-gap-landmark/55138?camid=4v1a

Finite Element Analysis and its Application in Dental Implant Research
www.igi-global.com/chapter/finite-element-analysis-its-application/8092?camid=4v1a

Analysis and Quantification of Motion within the Cardiovascular System: Implications for the Mechanical Strain of Cardiovascular Structures
www.igi-global.com/chapter/analysis-quantification-motion-within-cardiovascular/19586?camid=4v1a