SURVIVAL ANALYSIS

D.G. Kleinbaum and D.L. Christensen
Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia, U.S.A

S.Y. Rowe
US Centers for Disease Control and Prevention, Atlanta, GA, USA

Keywords: Survival analysis, failure, censoring, survival function, hazard function, ordered failure times, average survival time, average hazard rate, hazard ratio, Kaplan-Meier method, log-rank test, Wilcoxon test, Cox proportional hazards model, Wald test, likelihood ratio test, log-log survival curves, goodness-of-fit test, Heaviside function, stratified Cox model, extended Cox model

Contents

Summary

Survival analysis refers to statistical procedures used to analyze data where the outcome of interest is time to an event. Examples of events include death and recurrence of illness. Events are designated as "failures" and time to the event is designated "survival time." Study subjects who do not experience an event during the study period are designated as "censored" and are included in the analysis by counting the follow-up time they contributed during the study. Common methods for conducting survival analysis include calculating survival probabilities, plotting survival curves, and using mathematical models. The Kaplan-Meier method is a technique to empirically estimate survival probabilities and plot survival curves. Survival curves can be compared visually or by statistical tests such as the log-rank and Wilcoxon tests. The Cox proportional hazards model is a popular mathematical model used to estimate regression coefficients for variables predicting survival time. This model estimates a hazard rate for a set of predictor variables. Hazard rates can be divided to give a hazard ratio, a comparison of the hazard rates at different levels of predictor variables. An important assumption of the Cox proportional hazards model is that the hazard for an individual is proportional to that for another individual, regardless of time. Alternate models are available to analyze data when this assumption is not satisfied, including the stratified Cox model and the extended Cox model. The stratified Cox model estimates regression coefficients for variables that do satisfy the proportional hazards assumption, stratifying on variables that do not satisfy the assumption. The extended Cox model estimates regression coefficients for variables that do and do not satisfy the proportional hazards assumption.

1. Introduction

In certain types of epidemiologic studies, the outcome variable of interest is *time until an event occurs*. An example of this type of study is one that follows leukemia patients in remission over several weeks to see how long these patients stay in remission. Another example is a study that follows patients who received a heart transplant to find out how long these patients survive after receiving the transplant. A third example is a study that follows a group of persons who are initially disease-free over several years to see who develops heart disease. To analyze data from these types of studies, a collection of statistical procedures called **survival analysis** is needed.

2. Basic Concepts of Survival Analysis

In survival analysis, the occurrence of an event is often called a **failure,** and the time variable is often referred to as **survival time** because it designates the amount of time an

individual "survived" without having an event during a specific follow-up period. In this introductory description of survival analysis, only one event per individual is of interest. Though an individual might have more than one event during a specific follow-up period, the procedures required to address this statistical complexity (known as "competing risks procedures") are beyond the scope of this chapter.

2.1. Censoring

Another important aspect of survival analysis is that it can address **censoring**, which occurs when some information about an individual's survival time is known, but the exact survival time is unknown. Data can be right or left censored. The most common form of censoring is right censoring.

A person's survival time can be **right-censored** if the actual survival time is at least as long as the time observed by the investigator. Three common situations where an individual's survival time is right-censored are the following: an individual does not experience the event before the study ends, an individual is lost to follow-up during the study period, or an individual withdraws from the study. Figure 1 illustrates these situations.

Left censoring occurs when the actual survival time is less than what is observed by the investigator. This can occur when an event has occurred by the time of first examination, and all that is known is an individual's survival time is less than a certain value. For example, if a new disease were recognized at a certain time, individuals diagnosed at that time might have just developed disease, or they might have had existing disease that had not been previously recognized.

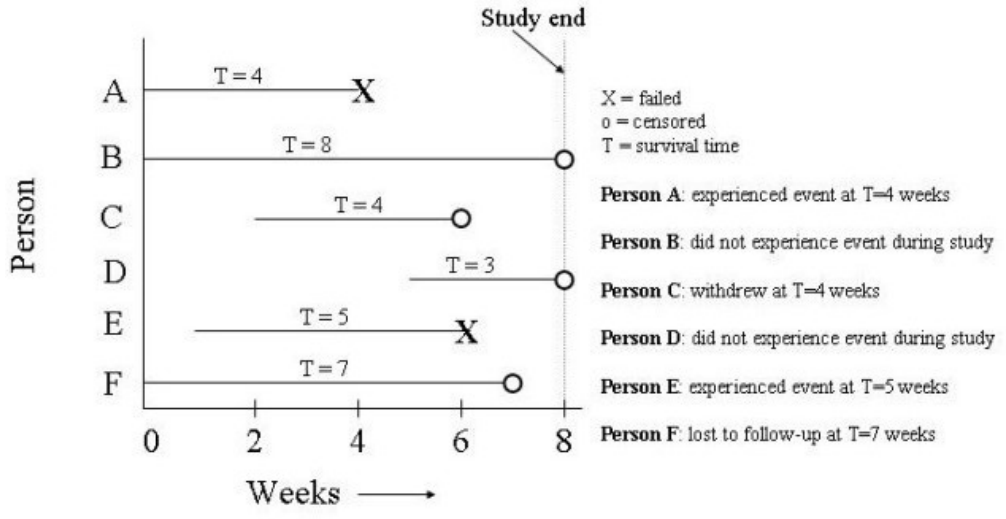

Figure 1: Three common situations where survival time is right-censored

An individual's survival time can be left-truncated if it is incomplete at the left side of the individual's follow-up period. For example, in a study where patients with HIV infection are followed to find out how long these patients survive after exposure to the

virus, follow-up might begin when a person first tests positive for the HIV virus. A person's time of exposure to the virus, which evidently is sometime before the positive HIV test, might be unknown. The survival time for this individual is left-truncated, since the follow-up time from the first exposure to the virus up to the time of first positive HIV test is unknown. Nevertheless, left-truncated data is always right censored since the actual survival time is longer than the observed survival time.

Since most censoring that occurs in survival data is right-censoring, this chapter only will consider right-censored data. In this text, an individual who experiences an event (i.e., fails) will be considered **non-censored**, while an individual whose exact survival time is unknown will be considered **censored**.

Although censored observations are incomplete, their survival time up to the point of censorship can provide useful information. Utilization of censored survival time in analysis will be addressed in Section 2.4.

2.2. Terminology and Notation

The random variable for an individual's survival time is denoted by T, for which values can range from zero to infinity. A specific time of interest for survival time T is denoted by t.

An important quantitative term considered in survival analysis is the **survival function**, denoted by $S(t)$, which directly describes the survival experience of a study cohort. The survival function summarizes information from survival data by giving survival probabilities for different values of time. A survival probability is the probability a person survives longer than specified time t, or

$$S(t) = P(T > t). \qquad (1)$$

Theoretically, all survival functions have the following characteristics (see Figure 2):

- As time t increases, $S(t)$ decreases.
- $S(0) = 1$, since at the beginning of the study, no one has experienced an event, and the probability of surviving past time 0 is unity.
- $S(\infty) = 0$, since if the study period were limitless, presumably everyone eventually would experience the event, and the probability of surviving would ultimately fall to zero.

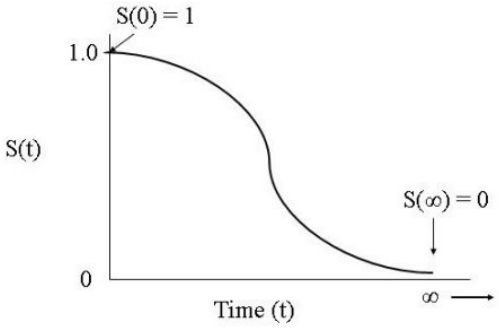

Figure 2: Theoretical survival function, $S(t)$, versus time

When using actual data, the plot of $S(t)$ versus time t usually results in a step function, as shown in Figure 3, rather than a smooth curve.

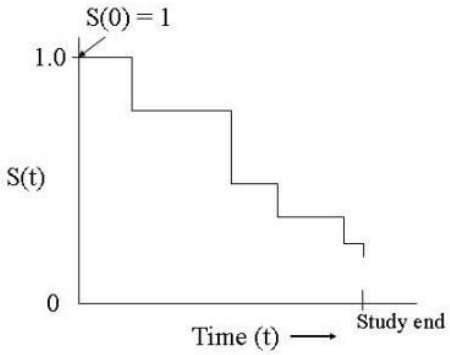

Figure 3: $S(t)$ versus time as a step function

Also, since a study period has a finite length and not everyone necessarily experiences an event by the end of the study period, the survival function curve does not always decrease to zero.

Another important quantitative term considered in survival analysis is the **hazard function**, denoted by $h(t)$. In contrast to the survival function, the hazard function summarizes survival data by focusing on failures. The hazard function gives the instantaneous potential per unit time for an event to occur, given that the individual has survived up to time t. Also, while the survival function is an expression for a survival probability, the hazard function is an expression for a failure rate. The mathematical formula for the hazard function is

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \qquad (2)$$

This expression defines $h(t)$ as the limit, as the time interval Δt approaches zero, of the ratio of two quantities: 1) the probability that the event will occur between time t and $t + \Delta t$, given that the survival time T is greater than or equal to t, and 2) the time

interval Δt.

Sometimes called the *conditional failure rate*, the hazard function gives the conditional probability of failure per unit time. For any specified value of t, $h(t)$ is always nonnegative and has no upper bound (see Figure 4). Furthermore, as will be shown later in this chapter, the hazard function is of particular interest because the mathematical model used to describe survival data is usually written in terms of the hazard function (see Sections 4, 6, and 7).

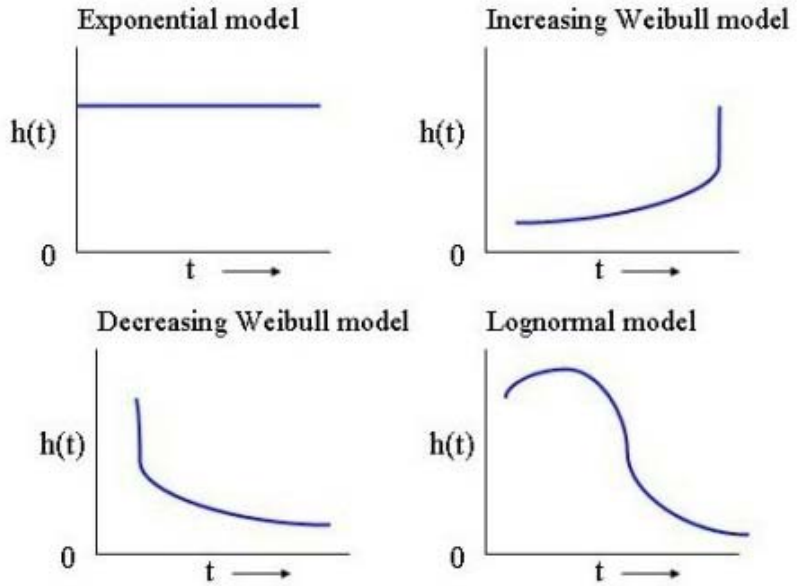

Figure 4: Examples of hazard functions, $h(t)$

The hazard and survival function are related such that if the form of $h(t)$ is known, $S(t)$ can be derived, and vice versa. The relationship between $h(t)$ and $S(t)$ can be shown in two mathematical formulae:

$$S(t) = \exp\left(-\int_0^t h(u)du\right), \tag{3}$$

$$h(t) = -\frac{1}{S(t)}\frac{dS(t)}{dt}. \tag{4}$$

These formulae indicate that for a given value of t, a high $S(t)$ corresponds to a small $h(t)$ and a low $S(t)$ corresponds to a high $h(t)$.

2.3. Goals of Survival Analysis

In survival analysis, the following three basic goals exist:

♦ **Goal #1:** *Estimate and interpret survival and/or hazard functions from survival data*

To address this goal, survival function values could be plotted versus time ($S(t)$ on the y-axis and time t on the x-axis). For example, Figure 5 shows the survival experience for a study cohort A. Cohort A's survival probabilities quickly dropped to 15% early in the follow-up period, remained at that level for about 4 weeks, then gradually decreased close to zero by the end of the study period.

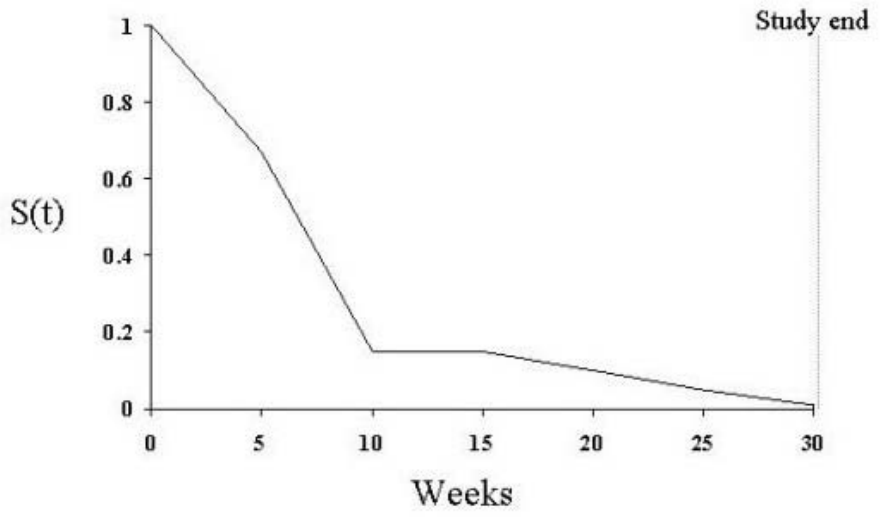

Figure 5: Survival function for study cohort A

♦ **Goal #2:** *Compare survival and/or hazard functions*

To accomplish this, survival functions for more than one study cohort could be plotted against the same time axis. For example, Figure 6 shows the survival experience for study cohorts A (placebo group) and B (treatment group). As noted earlier, Cohort A's survival probabilities dropped early in the follow-up period. In contrast, Cohort B's survival probabilities dropped sharply much later in the study period. Also, the $S(t)$ curve for Cohort B lay above that for Cohort A consistently for the entire follow-up period, indicating that the treatment was more effective than the placebo for this time period.

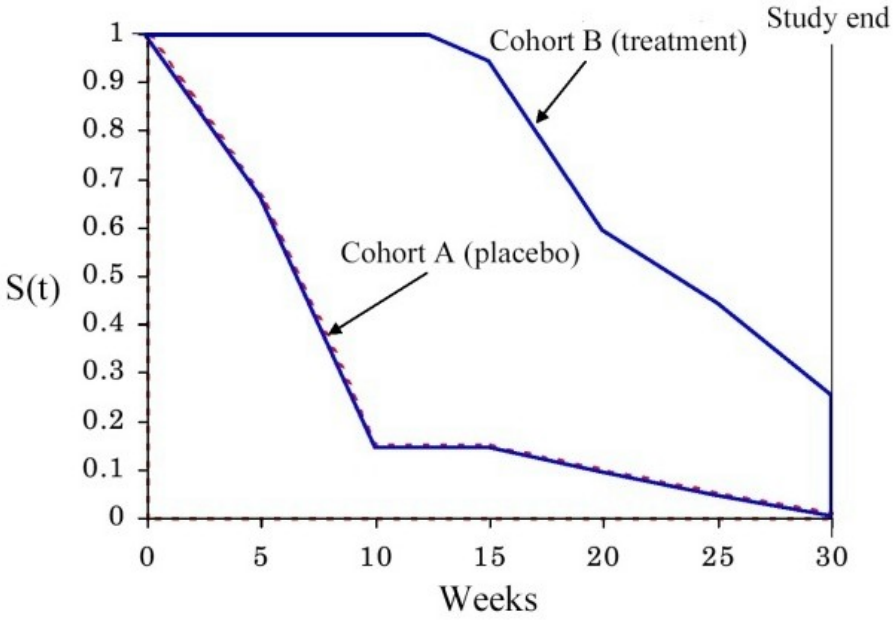

Figure 6: Survival functions for cohort A (placebo) and B (treatment)

- **Goal #3:** *Assess the relationship of explanatory variables to survival time*

Assessing this relationship usually requires some form of mathematical modeling. Examples of this modeling, the Cox proportional hazards (PH) model, stratified Cox (SC) model, and extended Cox model, will be discussed in subsequent sections of this chapter (see Sections 4, 6, and 7, respectively). Analogous to linear and logistic regression modeling, mathematical modeling in survival analysis typically is used to describe the relationship between an exposure and survival time, controlling for the possible confounding and interaction effects of additional factors. In survival analysis, the measure of the effect of an exposure on survival time is called the hazard ratio, which is the ratio of failure rates for the unexposed and exposed groups. Similar to the odds ratio in the logistic regression model, the hazard ratio is expressed in terms of an exponential of a regression coefficient in the model for survival data. Also, the analysis strategy to find the most appropriate mathematical model for survival data is analogous to the strategy used when fitting logistic regression.

-
-
-

Bibliography

Hennekens C.H. and Buring J.E. (1987). *Epidemiology in Medicine*. First Edition. (ed. S.L. Mayrent), p. 33. Boston, Toronto: Little, Brown and Company. [This book provides an overview of basic principles

and methods used in epidemiologic research, discussion of study designs, and addresses issues in the analysis and interpretation of epidemiologic data.]

Rothman K.J. and Greenland, S. (1998). Modern Epidemiology. First Edition. Philadelphia, Pennsylvania: Lippincott-Raven Publishers. [This text discusses analytic techniques and their applications in scientific research.]

Kalbfleisch J.D. and Prentice R.L. (1980). The Statistical Analysis of Failure Time Data. New York: John Wiley and Sons. [This text features theory on the Cox proportional hazards model and describes methods for making inferences about regression model parameters.]

Kleinbaum D.G., Kupper L.L., and Muller K.E. (1988). *Applied Regression Analysis and Other Multivariable Methods*. Second Edition. p. 24, 314-340. Belmont, California: Duxbury Press. [This text features details about regression analysis, analysis of variance, use of linear models, multiple regression analysis, maximum likelihood estimation methods, and procedures for fine-tuning a regression analysis.]

Kleinbaum D.G. (1994). *Logistic Regression, A Self-Learning Text*. First Edition. p. 161-226. New York: Springer-Verlag New York, Inc. [This book presents basic information about logistic regression, including the logistic model, computation of the odds ratio from the logistic model, maximum likelihood estimation techniques, and modeling strategy guidelines].

Kleinbaum D.G. (1996). *Survival Analysis, A Self-Learning Text*. First Edition. New York: Springer-Verlag New York, Inc. [This book presents basic information about survival analysis, including the Kaplan-Meier method, the log-rank test, mathematical models for survival analysis, computation of the hazard ratio from these models, and maximum likelihood estimation methods.]SAS Procedures Guide, Version 8, Volumes 1 and 2. Cary, NC: SAS Institute, Inc.SPSS 10.0 Advanced Models Manual. Chicago: SPSS, Inc.STATA Reference Manual. College Station, TX: Stata Press, 2001.

Biographical Sketches

David G. Kleinbaum holds a Ph.D. in Mathematical Statistics, University of North Carolina at Chapel Hill. Dr. Kleinbaum's primary experience for thirty three years has concerned biostatistical applications to epidemiologic research as well as principles and methods of epidemiology. He is first author of two widely acclaimed textbooks: Applied Regression Analysis and Other Multivariable Methods, Duxbury Press, Third Edition, 1998, and Epidemiologic Research: Principles and Quantitative Methods, Van Nostrand Reinhold, 1982. He is sole author of two other recent texts, Logistic Regression- A Self-Learning Text, Springer-Verlag, 1994, and Survival Analysis-A Self-Learning Text, Springer-Verlag,1996. The second edition of Logistic Regression- A Self-Learning Text (now co-authored by Mitch Klein) was recently published in August 2002; this second edition includes 5 new chapters covering polytomous and ordinal logistic regression and methods for analyzing correlated binary data using logistic regression.

Dr. Kleinbaum has just completed an electronic textbook, ActivEpi, a multi-media learner-interactive CD ROM course on fundamentals of epidemiology, in collaboration with the Centers for Disease Control and Data Description, Inc. (Ithaca, NY). This work has just been published by Springer-Verlag in September 2002, along with The ActivEpi Companion Text (authors: David G. Kleinbaum, Kevin M. Sullivan, and Nancy D. Barker), in December 2002.

Dr. Kleinbaum is considered an outstanding teacher of biostatistical and epidemiological concepts and methods at all levels, particularly to the non-mathematically-sophisticated student. He has had over 30 years of experience teaching more than 120 short courses on statistical and epidemiologic methods to a variety of audiences, nationally and internationally.

Deborah L. Christensen, AB, BSN, MPH is a Ph.D. candidate in the Department of Epidemiology at Emory University in Atlanta, Georgia. She is currently a Project Coordinator in the Department of Epidemiology at Emory University. Her research interests include prenatal factors related to chronic disease in adulthood, breast cancer etiology, and cancer prevention programs.

Samantha Y. Rowe earned a Ph.D. in Epidemiology from Emory University in Atlanta, Georgia. She is currently a researcher in the Division of Parasitic Diseases at the U.S. Centers for Disease Control and Prevention in Atlanta, Georgia. Her research interests include pediatric infectious diseases in developing countries, child survival programs and health services research.