Applied statistics in vascular surgery Part V: The use of Kaplan-Meier and Cox proportional hazard regression model

Constantine N. Antonopoulos[1], Efthymios D. Avgerinos[2], John, Kakisis[1]

[1]Department of Vascular Surgery, Athens University Medical School, "Attikon" Hospital, Athens, Greece
[2]Department of Vascular Surgery, University of Pittsburgh, UPMC Presbyterian Hospital, Pittsburgh, PA

Abstract:

Survival data analysis is used when the time until the event is of interest. The significance of survival analysis is that it takes into consideration that the event will probably not have occurred for all patients at the end of the follow-up period. One of the most commonly used approaches in survival analysis is the Kaplan-Meier estimator, which splits the estimation of survival probability into a series of intervals and calculates the probability of surviving until the end of each interval. However, it cannot control for multiple covariates. Multivariate analysis, using the Cox proportional hazard regression model, is applied when there are multiple, potentially interacting covariates and provides an estimate of the Hazard Ratio and its Confidence Intervals. A brief description of these basic survival data analyses is presented.

INTRODUCTION

Survival or time-to-event analysis is the process of analyzing data measuring the time until a specific event occurs in the population under investigation[1-5]. Such event may be adverse, (e.g. death), positive, (e.g. discharge from hospital) or neutral (e.g. walking a distance). The unique characteristic of this type of analysis is that the event will probably not have occurred for all patients at the end of the follow-up period, or patients may have been lost to follow-up, or they may have experienced another event, which will make further follow-up impossible[1-5]. In this case, we only know the time period (eg. the total number of days) within which the event did not occur; these observations (subjects who did not experience the event) are called “censored” as they drop off the analysis of the subsequent follow up. It is evident that survival time has two components which must be clearly defined; a beginning point and an endpoint that is reached either when the event occurs or when the follow-up time has ended. A basic assumption of the survival analysis is that censored individuals have the same probability to experience a subsequent event as individuals that remain in the study and there is sufficient follow-up time and number of events for adequate statistical power[1-5].

TYPES OF APPROACHES FOR SURVIVAL ANALYSIS

Based on the research hypothesis, three main types of time-to-event analysis can be used, either alone or in conjunction; non-parametric, parametric and semi-parametric. Parametric implies that the model comes from a known distribution (e.g. normal distribution), non-parametric makes no assumptions about the distribution, while in semi-parametric some components, even from unknown distributions, can be added to a parametric model. The most common non-parametric approaches are the Kaplan-Meier (or product limit) estimator and the life table estimator of the survival function[2, 5-8]. In a non-parametric approach, a univariable analysis for categorical factors of interest is conducted. Semi- and fully-parametric models are used when we want to investigate the relationship between several covariates and the time-to-event. For that reason, we usually use non-parametric approaches as the first step in our analysis to generate the descriptive statistics and we thereafter continue with semi-parametric or parametric approaches in case of multivariate models[2, 5-8].

COMMON APPROACHES FOR SURVIVAL ANALYSIS: THE KAPLAN-MEIER ANALYSIS

Kaplan-Meier analysis was first introduced in 1958 by Edward L. Kaplan and Paul Meier[9] who collaborated to publish a study on how to deal with incomplete observations. The Kaplan-Meier method is used to estimate the probability of survival past given time points and it calculates a survival distribution. Furthermore, the survival distributions of two or more groups can be compared for equality[1-6, 8-11]. When preparing a Kaplan-Meier analysis the researcher needs to construct a raw dataset table, in which each subject is characterized by three variables: 1) **time (t)**, which is the survival or censoring time 2) **status** of the event at time t (1=event or 0=no event/censored), and 3) the study **group** in which the participant belongs (eg. 1=treatment group or 2=control group). The table is then sorted by ascending serial time beginning with the shortest time for each group. The technique is to divide the follow-up period into a number of small-time

Author for correspondence:

Constantine N. Antonopoulos

Department of Vascular Surgery, Athens University Medical School, Attikon University Hospital, Athens, Greece
E-mail: kostas.antonopoulos@gmail.com
ISSN 1106-7237/ 2020 Hellenic Society of Vascular and Endovascular Surgery Published by Rotonda Publications

intervals, determining for each interval the number of cases followed up over that interval and the number of events of interest (e.g. deaths) during each period. The **survival probability** (which is also called the survivor function) S(t) is the probability that an individual survives (in case of death) from the start time to a specified time (t), which means that the event of interest has not yet occurred by time (t), or in other words, the probability that the time of the event (eg. death) is later than some specified time (t). The **hazard function** h(t) or λ(t), which is also named as "*force of mortality*" [μ (t)], is the probability that a subject experienced the event of interest (death, relapse, etc.) during a small-time interval, given that the individual had survived up to the beginning of that interval. The survival probability can be plotted against time, using the **Kaplan-Meier curve**[1-6, 8-11]. In a Kaplan-Meier plot (Figure 1), the *x axis* represents time, from start to the last observed time point, while the *y axis* is the proportion of subjects surviving, in a way that all subjects are alive without an event at time zero. The Kaplan-Meier curve is usually drawn as a solid line (similar to a staircase), which shows the progression of event occurrences. In such a curve, a vertical drop indicates an event, while a vertical line indicates that a patient was censored at this time. Kaplan-Meier analysis can estimate median time, which is the time at which, in 50% of cases, an event of interest has occurred and mean time, which is the average time for the event.

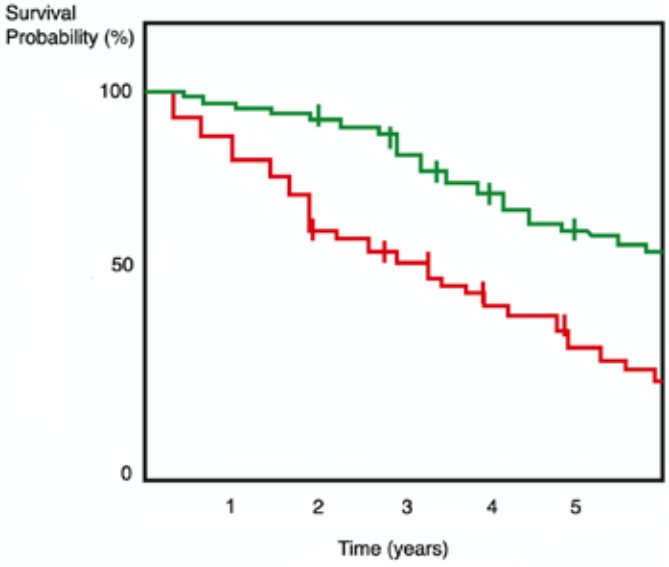

Figure 1. A Kaplan-Meier plot: The *x axis* represents time in years, from start to the last observed time point, while the *y axis* is the (%) proportion of subjects surviving. Two Kaplan-Meier curves are drawn as solid lines (similar to a staircase) with green (treatment A) and red (treatment B) color. The vertical drops indicate events, while the vertical lines indicate that a patient was censored at this time

After plotting of survival data, a **life table** is usually used to depict the number of events and the proportion surviving at each event time point. In a life table, the reader usually finds data concerning the time at which an events occurs, the number of subjects who experience an event at that time, the number of subjects who did not have an event or who were not censored before that time (*patients at risk*), the proportion of surviving at that time and its standard error with lower and upper 95% Confidence Intervals (CIs). In case we want to compare the survival times of two or more groups, the **log-rank test** is used, which is based on chi-square statistic and checks if the observed number of events in each group is significantly different from the expected number. In log-rank test all time points are weighted equally. Other tests include the **weighted log-rank test** in case we want to compare groups, but with more importance ("weight") to certain events, the **Breslow test**, in which the time points are weighted by the number of cases at risk at each time point, the **Tarone-Ware test**, in which the time points are weighted by the square root of the number of cases at risk at each time point and others.[1, 3, 7].

THE COX PROPORTIONAL HAZARD REGRESSION MODEL

The Kaplan-Meier estimator is one of the most commonly used methods to illustrate survival curves. However, the disadvantage of Kaplan-Meier estimator is the lack of controlling for other covariates. In that case, a **Cox proportional hazard regression model** should be used. The latter is a semi-parametric model, which can act as a multiple regression model investigating the association between the survival time with one or more predictor variables and provides an estimate of the hazard ratio (HR) and its CIs[1, 4, 8, 12] . It is similar to multiple regression analysis, except that the dependent variable is the hazard function at a given time. Furthermore, while Kaplan-Meier analysis requires categorical variables, Cox regression can also work with continuous variables. In Cox models, the baseline or underlying hazard function corresponds to the probability of dying (or reaching an event) when all the explanatory variables are zero. A basic assumption in Cox models is the "*proportional hazards*" assumption[11, 13], which means that the hazard functions for any two individuals at any point in time are proportional. So, if the risk of death at some initial point in time in an individual is twice as high as that of another individual, then at all later times the risk of death should remain twice as high.

CONCLUSION

As scientific literature frequently deals with survival time data, censorings cannot be overlooked, as they carry important information. Survival analysis and comparisons using the log-rank test are important in that case. However, when multiple confounders should be taken into consideration, multivariable analyses can be performed using Cox proportional hazard regression model and the results can be interpreted using hazard ratios. Understanding of survival analysis is of paramount importance for both researchers and clinicians.

No conflict of interest.

REFERENCES

1 Fink SA, Brown RS, Jr. Survival Analysis. Gastroenterol Hepatol (N Y). 2006;2(5):380-3.

2 Singh R, Mukhopadhyay K. Survival analysis in clinical trials: Basics and must know areas. Perspect Clin Res. 2011;2(4):145-8.

3 Lee ET, Go OT. Survival analysis in public health research. Annu Rev Public Health. 1997;18:105-34.

4 Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003;89(2):232-8.

5 Antonopoulos C, Avgerinos E, Kakisis J. Applied statistics in vascular surgery Part IV: Introduction to survival analysis. Heljves. 2019;1(4):180-1.

6 Sedgwick P. How to read a Kaplan-Meier survival plot. BMJ. 2014;349:g5608.

7 Rich JT, Neely JG, Paniello RC, Voelker CC, Nussenbaum B, Wang EW. A practical guide to understanding Kaplan-Meier curves. Otolaryngol Head Neck Surg. 2010;143(3):331-6.

8 Altman DG, Bland JM. Time to event (survival) data. BMJ. 1998;317(7156):468-9.

9 Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958;53(282):457-81.

10 Shaw P, Johnson L, Proschan M. Chapter 27 - Intermediate Topics in Biostatistics. In: Elsevier, editor. Principles and Practice of Clinical Research. 4th ed2018. p. 383-409.

11 Kleinbaum D.G., Klein M. Evaluating the Proportional Hazards Assumption. Survival Analysis Statistics for Biology and Health. New York, NY: Springer; 2012.

12 Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. Br J Cancer. 2003;89(3):431-6.

13 Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part IV: further concepts and methods in survival analysis. Br J Cancer. 2003;89(5):781-6.