STATISTICS FOR RESEARCHERS

Survival Analysis

NJ Gogtay, UM Thatte

Introduction

Often in research, we are not just interested in knowing the association of a risk factor/exposure with the presence [or absence] of an outcome, as seen in the article on Measures of Association,[1] but rather, in knowing how a risk factor/exposure affects *time to disease occurrence/recurrence or time to disease remission or time to some other outcome of interest.*

Survival analysis is defined as the set of methods used for analysis of data where **time to an event is the outcome of interest.** Originally, this analysis was concerned with time from treatment until death and hence the name. Survival analysis however can be applied to a wide variety of situations. Medical examples include time to metastases, time to tumor recurrence, time to discharge from the hospital, time to first exacerbation after a new drug treatment in patients with Chronic Obstructive Pulmonary Disease [COPD], time to dialysis in patients with renal dysfunction and so on. In the real world, *survival* could be time to a light bulb fusing, time to replacing the battery on the wall clock or time to the change the gas cylinder. The other terms used for survival analysis are "failure-time analysis", "reliability analysis", and "event history" analysis.

Why Survival Analysis is Important

Studies of how patients respond to treatment *over time* are fundamentally important to understanding how treatments influence both disease progression and quality of life. Such studies can last for weeks, months or even years. When we capture *not just the event, but also the time frame over which the event/s occurs*, this becomes a much more powerful tool, than simply looking at events alone.

An additional advantage with this type of analysis is the use of the technique of "censoring" [described below], whereby each patient contributes data even if he/she does not achieve the desired outcome of interest or drops out during the course of the study for any reason.

Key Concepts in Survival Analysis

In order to understand survival analysis certain concepts need to be understood before doing survival analysis such as: Time or survival time, time of entry into the study, censoring, cumulative probability, hazards and hazard ratio and survival and hazard functions. We discuss these briefly below.

Time or survival time

The time variable in a survival analysis is called as "survival time", while the event of interest itself is called "failure".

Time of entry into the study- the concept of zero time

Let us understand this with an example of a study evaluating a new medical treatment for lung cancer, which has a follow up duration of 5 years. The first 2 years of this study are used for enrolling or accruing patients. These patients are followed up for the remaining 3 years of the study duration. Patients can enter the study at any point during the first 2 years. For example, they can enter the study right at the beginning [Month 1] or at end of the accrual period [Month 24]. The first patient will undergo the full 5 year follow up, while the patient who came into the study at month 24, will only have a 3 year follow up as the cut off that we have defined for this study is 5 years. However, when this data is analyzed, regardless of the time of entry into the study, every patient would be analyzed from his/her point of entry into the study. This is called "zero time" and is the time when the patient is enrolled.

Censoring

In the above example, any one of the following could happen to any of the patients. Thus, the patient

i. Would actually achieve the outcome of interest [death in this case]
ii. Does not achieve the outcome of interest although the study ends (i.e. 5 years are over)
iii. Is lost to follow up [so we do not know whether the outcome has or has not occurred]
iv. Withdraws consent
v. Dies of some cause other than the disease under investigation, in this case, lung cancer [for example a patient enrolled in the trial dies of myocardial infarction rather than lung cancer. This is called "competing risk"].

Department of Clinical Pharmacology, Seth GS Medical College and KEM Hospital, Mumbai, Maharashtra
Received: 09.04.2017; Accepted: 11.04.2017

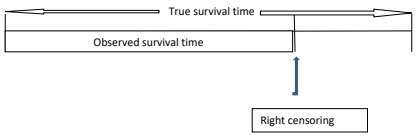

Fig. 1: Right Censoring

When a patient achieves the outcome of interest (i), it is useful to the researcher as it contributes valuable data. But what if the patient experiences any of the other situations (ii to v)?

Survival analysis is unique in that it "allows" the researcher to use data from such patients up until the point of their last follow up by using a method called as *censoring*. There are three main types of censoring: right, left, and interval. The most commonly used one is the "right censoring" where censoring occurs *after* the patient has entered the study (Figure 1) because the participant has left the study for any of the reasons mentioned above.

Let us take another example of a breast-feeding survey done monthly. Two types of mothers can enter the study: those who are breast feeding at the time of entering the study and those who have stopped breast feeding at the point of entry into the study. For the former, right censoring can be done, while for the latter as we do not know exactly when they stopped breast feeding before entering the study, we need to do "left censoring" for the point where they stopped breast-feeding. Interval censoring occurs if the breast-feeding ended between two successive surveys since one can only say that breast feeding ended somewhere between the two surveys.[2]

For censoring to have validity, when a patient is censored, the risk for achieving the outcome for the remainder of the patients who continue on the study, should be unchanged. In addition, censoring should be randomly distributed over time as its main assumption is that it is independent of time and the intervention/treatment under evaluation.

Cumulative probability of survival

Probability is the chance of a single event occurring whereas if you want to calculate the chance of two, three, or more events happening, we measure the "Cumulative probability". Two caveats need to be fulfilled for calculating Cumulative probability: 1) each event needs to be independent of the other and 2) outcome of the first event should not influence the probability of occurrence of the second.

This concept can be best understood with an unbiased coin toss experiment. Let us say we want to answer the question "When a coin is flipped twice, what is the probability of getting heads on both occasions?" Each toss will have two outcomes (H or T) and two consecutive tosses will have four possible outcomes:

- HH
- TT
- HT
- TH

And the answer to our question is therefore one in four or 25%. When the coin is tossed the first time, the probability of getting heads is ½ or 50%. When it is tossed the second time, the probability of heads is again ½ as this outcome is independent of the first and its probability is not influenced by the probability of the first. Thus, the cumulative probability of getting two consecutive heads is calculated as the product of the probabilities of each event – i.e. ½ x ½ or ¼ or 25%. We will see how this is applied to calculate the cumulative survival and survival and draw the survival curve [see below].

Hazard and hazard ratio

A "hazard" is simply the rate at which a particular event occurs. If we were to take the example of lung cancer presented earlier, and we are interested in death as the outcome, then the hazard would be the rate at which patients die during the course of the study [or the time course of their death]. Mathematically, it is expressed as the *hazard function* [described below]. Since studies on survival analysis involve the comparison of two or more groups, the hazard in one group is compared with the other group and expressed as a ratio called the *hazard ratio*. This is defined as the ratio of the hazard in the experimental to the hazard in the control arm. The distinction between hazard ratio and odds ratio [or relative risk] lies in the fact that the latter are simply ratios of proportions, while hazard, which incorporates time, is a ratio of incidence rates.

Survival and hazard functions

Both of these are crucial to the analysis of survival data and are related to each other. They describe the distribution of event times. The survival function s(t) is the probability that an individual survives from a specified time point (e.g., the diagnosis of cancer) to a specified future time t. It directly summarizes time to event experience of a group of patients and is crucial to analyzing time to event data. Hazard function is denoted as h(t) and represents the probability that an individual who is under observation at a time t, has an event at that time t. It can range from o to infinity. In other words, it represents the instantaneous event rate [at time t], given that the individual has already survived upto that time t. The distinction between the two functions lies in the fact that the survival function relates to *not* having the event, while the hazard function relates to the *probability of the event occurring per unit time*.[3] Mathematical relationships between the two functions have been defined and computer software can return the value of one function, given the value of the other.

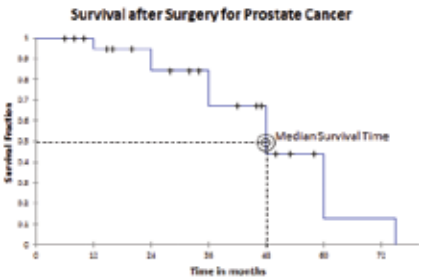

Fig. 2: Data of n =100 patients depicting Survival after surgery for prostate cancer

Types of survival analysis

A number of models are available to analyse the relationship of a set of predictor variables with the survival time. The methods for doing a survival analysis fall into three broad categories

- Non-parametric
- Semi-parametric, and
- Parametric.

The difference between these methods lies in the assumptions that we make regarding the distribution of survival data. In nonparametric analysis, there is absolutely no assumption made and these methods can be used for all types of data. The Kaplan Meier method is a widely used non-parametric method in medicine for survival analysis. It can be used for study of a variable in a single group over time (e.g. group of smokers developing lung cancer over time), but it also serves the purpose when you want to compare two or more groups over time (e.g. progression free survival with Trastuzumab innovator *vs.* Trastuzumab biosimilar in metastatic breast cancer over time). This method plots a curve (Kaplan Meier curve; see Figure 2) of cumulative probability against time and can be used to obtain univariate descriptive statistics for survival data, e.g. median survival time. Inferential statistics can be applied using several tests and the log-rank test is the most popular (see below).

Semi-parametric tests are equally commonly used in medicine to analyze survival data and the Cox proportional hazards regression (see below) is representative of these methods. When we assume that survival times conform to some specific statistical distribution (e.g. exponential, Weibull, and lognormal distributions) we use parametric survival analysis, less commonly seen in medicine.

Table 1: Hypothetical survival data for n = 100 patients with prostate cancer following surgery

Time interval (years)	No. at risk	No. censored (Data not available beyond a point)	No. who achieved the outcome of interest (died)	No. survived
1	100	3	5	95
2	92	3	10	82
3	79	3	15	64
4	61	3	20	41
5	38	3	25	13

Table 2: Kaplan Meier cumulative survival estimates in hypothetical cases of prostate cancer (n=100 at start)

Time interval (years)	No. at risk	No. censored (Data not available beyond a point)	No. who achieved the outcome of interest (died)	No. survived	Kaplan-Meier cumulative survival estimate
1	100	3	5	95	(95/100)=0.95
2	92	3	10	82	0.95 x(82/92)=0.8467
3	79	3	15	64	0.8467 x (64/79)=0.70
4	61	3	20	41	0.7 x (41/61)=0.4611
5	38	3	25	13	0.4611 x (13/38)=0.1577
6	10	2	8	0	0

Analysis of Survival Data

Several techniques exist for analyzing survival data. In the Life table technique [also called actuarial analysis] the data is divided into fixed time intervals and for each time interval, we assess cumulative probability with patients who have achieved the outcome of interest and those who have been censored. The Kaplan-Meier technique described above, on the other hand, uses the time intervals that are data driven and not necessarily fixed as in Life Table Methods (although they can be fixed as well).[4] Another technique is the Cox Proportional Hazard Method, used when there are several predictor variables impacting the event of interest.

Before we can do any survival analysis, we need to make sure that our data is structured in a manner that makes analysis easy. This needs to be done for each and every patient in the study. Thus, the first thing to do is to organize the data. Each patient as discussed earlier would enter the study at time zero. Then, we look at each patient to see whether he/she has achieved the outcome of interest or needs to be censored. Let us understand this with an example.

A hypothetical research question we are trying to answer could be "How many years do patients with prostate cancer survive after they undergo surgery"? We lay down the following unequivocal boundaries for the study right at the beginning: 1) all patients who undergo surgery in the first 1 year after the start of the study would be included i.e., 1 year would be the total accrual period 2) The study would be done for a total of 5 years regardless of when they enter the study. The data for 100 patients The data for 100 patients is given in Table 1.

The Kaplan-Meier procedure

then calculates the *cumulative probability* (calculated as the number of subjects surviving divided by the number of patients at risk) for each of the t time periods, except the first (Table 2).

Kaplan Meier / Product-Limit Curve

The Kaplan-Meier survival curve is the curve that results at the end of the survival analysis and after the calculation of the survival probabilities. The generation of the curve makes certain assumptions

- At any time, patients who are censored have the same survival prospects as those who continue to follow up on the study (non-informative).
- Survival probabilities are the same for patients regardless of the time point at which they enter the study
- The event of interest happens at the time or time interval specified.

A Kaplan Meier curve generated on the hypothetical data is described in Figure 2. Time is given on the x axis and cumulative probability on the y axis. Censored patients are shown as vertical lines or "ticks" on the curve while the deaths/outcome of interest are the dips in the curve

Pitfalls of the Kaplan Meier curve

- Curves that have many small steps usually have a higher number of participating subjects, whereas curves with large steps usually have a limited number of subjects and tend to be less accurate.
- Most studies have a *minimum duration of follow-up based on knowledge of disease biology and overall survival*. At this minimum duration of follow up, the status of each patient is known. The survival rate at this point becomes the most accurate reflection of the survival rate of the group. At the end of the survival curve, there are far fewer patients remaining and thus survival estimates at the end of the curve are less accurate.
- When patients are censored, we do not know whether they have actually experienced the outcome of interest. Thus, more the number of patients that are censored, the less reliable is the Kaplan Meier curve.

Cox Regression or Proportional Hazards Regression

It is intuitive that time to an event/outcome is influenced by not one by multiple predictor variables. For example, coronary artery disease [CAD] we know is associated with a high Body mass index [BMI], gender, hypertension, and smoking. Cox regression also called proportional hazards regression [semi-parametric] helps investigate the effect and impact of several predictor variables on the time to event or the variable of interest. It has three critical assumptions – 1) The effects of the predictor variables upon survival remain constant over time, the concept of proportional hazards (the parametric component). In other words, in this example, the impact of BMI, gender and hypertension all remain constant or do not change over the duration of the study. 2) It makes no assumptions regarding the baseline hazard (the non-parametric component) 3) It assumes that the censoring is non-informative.

The Log Rank Test – Comparing Two Survival Curves

This is a popular test that is closely related to the chi-square test and tests the null hypothesis of no difference in survival between two or more independent groups. The null hypothesis would be that there is no difference between the population survival curves (i.e. the probability of an event occurring at any time point is the same for each of the two groups being studied). Survival curves are estimated separately for each group, using the Kaplan-Meier method and compared statistically using this test to give a chi-square value.

Assumptions of the log rank test

- Survival times are ordinal or continuous.
- The risk of an event in one group relative to the other does not change with time

The Mantel Cox test is also another commonly used test.

Conclusions

The fundamental reason why survival analysis is done is that time to an event is a far more powerful tool than simply looking at events alone. Several examples of these can be found in medical literature. One recent example is that of association of atypical femoral shaft fractures with long term bisphosphonate use. We would thus be answering the question "What is the time to developing atypical femoral fractures after the use of bisphosphonates for several years" and how would it change the way we precribe bisphosphonates? This would then drive evidence based practice. Thus, the recommendation of a task force of the American Society for Bone and Mineral Research is to give a "bisphosphonate holiday" of 2-3 years in postmenopausal women at low risk of fractures who have received 3-5 years of treatment with bisphosphonates.[6]

Censoring is an integral component of survival analysis. However, this must not mean that the researcher should relax with respect to follow up as completeness of follow-up is crucial so as not to miss events and lose patients as this can lead to biased results. Unequal follow-up between different treatment groups may also produce biased results. Finally, it must be ensured that when Cox regression is used, the

proportional hazards assumption is met with.

Acknowledgments

The authors are grateful to Dr. L. Jeyaseelan, Dr. Girish Chinnaswamy and Dr. Manju Sengar for helpful inputs in refining the manuscript.

References

1. Gogtay NJ, Deshpande S, Thatte UM. Measures of Association. *J Assoc Phy Ind* 2016; 64:70-3.
2. https://www.cscu.cornell.edu/news/statnews/stnews67.pdf, accessed on 7th April 2017.
3. George B, Seals S, Aban I. Survival analysis and regression models. Journal of Nuclear Cardiology: official publication of the *American Society of Nuclear Cardiology* 2014; 21:686-694.
4. Sambath AK, Ramanujam R, Chinnaiyan P. Survival analysis: Kaplan Meier and life table estimates for time to event clinical trial tuberculosis data. *Concepts in Pure and Applied Science* 2013; 1:17-21.
5. Edwards BJ, Bunta AD, Lane J, et al. Bisphosphonates and Nonhealing Femoral Fractures: Analysis of the FDA Adverse Event Reporting System (FAERS) and International Safety Efforts: A Systematic Review from the Research on Adverse Drug Events And Reports (RADAR) Project. *The Journal of Bone and Joint Surgery American volume* 2013; 95:297-307.
6. https://www.asbmr.org/About/detail.aspx?cid=9caf8b31-8157-407e-a61d-e9b34065e3b4, accessed on 7th April 2017.