Survival Analysis: Where, Why, What and How?

ABHAYA INDRAYAN,[1] CHANDRA BHUSHAN TRIPATHI[2]
From [1]Max Healthcare and [2]Institute of Human Behaviour and Allied Sciences, Delhi.
Correspondence to: Dr A Indrayan, A-037 Telecom City, B-9/6 Sector 62, NOIDA 201 309, Uttar Pradesh. a.indrayan@gmail.com

Durations of any event, such as duration of hospitalization, is usually found to have a highly skewed distribution and incomplete values due to dropouts and limited follow-up. The usual methods of statistical analysis are, therefore, not applicable. The method of survival analysis is a nonparametric method and is designed to overcome these problems. Survival is a generic term and is used for any time-to-event data. The entire survival pattern at different points in time is studied by the Kaplan-Meier method under certain conditions. Log-rank method is used to compare survival pattern in two or more groups. Hazard is the rate of occurrence of an event per unit of time and studied by Cox method. The concept of survival and all these methods of survival analysis are briefly discussed in this short note in a non-mathematical format for medical audience.

Keywords: *Hazard rate, Kaplan-Meier method, Log-rank test.*

Published online: May 28, 2021; ***PII:*** S097475591600337

Survival analysis is an important concept in biostatistics with extensive usage in medical research. This is a multi-faceted technique and has evolved as a full subject. For medical researchers, we herein discuss, in brief, where and why survival analysis is needed, what it is, and how it is done in practical settings.

WHERE AND WHY?

A special quantitative measurement in medicine is the time taken for the occurrence of an event. This requires that the time of beginning and the time of the end, both, are well defined. Duration of hospitalization is a commonly studied duration, measured from the time of admission to the time of discharge or death. The effect of oxygen saturation targets on the duration of respiratory support, oxygen therapy, and hospitalization in extremely preterm infants [1] and continuation duration after contraception insertion among adolescents and young adults [2] have been recently studied using the method of survival analysis. Such durations have two special features that make them distinct from other quantitative measurements and render the usual statistical methods inapplicable.

Censoring

Quite often, the full duration is not observed such as duration of hospitalization in the case of a patient who left against medical advice (LAMA) or when a patient is transferred. Such incomplete observations more commonly happen when a follow-up of patients is planned for, say, two years, but the patient becomes untraceable or uncooperative after one year, or when the event under consideration, such as death, does not occur within the two-year period. The latter would mean that the survival is known for at least two years, but the full duration is not known. Such truncated durations are called 'censored', meaning thereby that the observation is incomplete.

This censoring is of three types. The example we have cited is for 'right censoring' where the endpoint has not reached at the time of the last observation. In rare cases, it could be 'left censoring' also where the beginning point is not known but the end point is known. This can happen when the interest is in the duration of disease, but the day of onset is not known or cannot be assessed. This has happened for many COVID-19 cases. The third could be 'interval censoring' where the exact duration is not known but only the interval is known. If surgery patients are followed quarterly for any complication, the only information available in the case of complication is that it occurred somewhere between, say, 9 and 12 months when no complication was reported at 9-month follow-up but reported at 12-month follow-up. Thus, the exact time of complication is not known. The present communication is restricted to only the right-censored data because that is the predominant type of censoring in medical research.

Skewed Distribution of the Durations

The second distinctive feature of durations is that the duration is generally relatively small for most cases but long or very long in some cases. For example, the duration of hospitalization in most heart surgery cases could be around 7 days but some develop complications and stay for 30, 40, or even 90 days. Thus, the statistical distribution of most

durations is highly skewed to the right. This makes the usual Gaussian based parametric analysis inapplicable and either logarithmic transformation or non-parametric methods are needed for analysis. In the case of highly skewed distribution, mean and standard deviation are misleading, and recourse to median and inter-quartile range (IQR) is taken.

Due to these two special features of data - censored values and highly skewed distribution - the usual statistical methods for quantitative measurements are not applicable to durations and a special method, called survival analysis is required. This is a nonparametric method and can be used for any kind of duration data. Yet, the method requires a reasonable sample size, at least 30 in each group, preferably 50 or more.

WHAT IS SURVIVAL ANALYSIS?

The method of survival analysis is needed when the study requires survival experience at different points in time. The term 'survival' in this case is generic and could mean any event such as discharge from the hospital or occurrence of a complication. Survival analysis tells us the percentage survived at different points in time – one year, two years, etc. - even when some of the durations are censored. It plots the survival pattern over time, called the survival curve, which can be used to draw inferences. An example is given later that would fix the idea.

If the interest is not in survival pattern but only in the percentage of patients who survived for, say, at least 3 months or at least 1 year, and if the duration is available for all (no censored values), there is no need of the method of the survival analysis. This percentage can be calculated in the usual manner and the inference regarding it being a prespecified value or for difference between two groups (such as one with treatment A and the other with treatment B) can be drawn as usual for proportions. This can be done even when the distribution of the duration is highly skewed. However, if some values are censored, survival analysis would be needed even for this proportion. Similarly, median survival time can be directly obtained without going through the process of survival analysis when the duration for all the subjects is known but this too would require the method of survival analysis if some durations are truncated (censored). Survival analysis is also needed if the entire survival pattern at different points in time is under study, whether the observations are censored or not.

Survival analysis requires that the duration for the occurrence of the event under investigation is recorded for each subject and the censored durations are marked. Consider the survival of children living with human immunodeficiency virus (HIV) [3]. Suppose these are followed up for 12 months after start of the antiretroviral therapy (ART). The duration of survival for 10 patients may be as shown in **Table I**.

The plus sign is for censored values. The third patient migrated and lost to follow-up after 4 months but was alive at that time. Patient numbers 5, 8, and 9 were alive at 12-month follow-up visit but were not followed-up thereafter. Their duration of survival is known at least 12 months but not beyond. Note in this case that mean or median duration of survival and percentage surviving 12 months or at any other time cannot be directly computed because of censored values.

Two extensions of survival analysis are commonly used. First, to compare the survival curves of two or more groups such as duration of hospitalization in mild and severe cases, and second, to find risk factors that affect the survival pattern such as the effect of age and time the antibodies are administered on the duration of survival in pediatric septic shock patients. The first requires log-rank test (discussed later) and the second is Cox regression, which also is briefly discussed later in the context of hazard rates. The Cox regression is also used to estimate hazard ratio relative to a reference base. Other aspects are not discussed here and available in the literature [4].

HOW SURVIVAL ANALYSIS IS DONE?

Kaplan-Meier Method

The method of survival analysis is quite mathematical, but we explain it here in simple terms. The primary method of survival analysis is the Kaplan-Meier (K-M) method. This method gives the proportion surviving at different points in time such as at 6 months and one year. Survival at time *t* in this case is the proportion surviving longer than *t* and the method considers the censored observations only till the time the subjects were last seen alive. After that they are ignored. This proportion is an estimate of the chance of survival at time *t* in the target population after due consideration of censored values. This can be easily obtained for each time-point with the help of an appropriate statistical package.

Table I Duration of Survival of Children With HIV on ART

Patient number	*Duration (mo)*
1	7
2	3
3	4+
4	10
5	12+
6	5
7	9
8	12+
9	12+

The estimated survival proportion can be plotted for different time-points *t*. This plot is called the survival function or survival curve. Only the unique time points are considered. If a time point is applicable to two or more subjects, it is counted only once. This method requires the calculation of as many survival rates as there are events unless several events occur at the same time. The more the number of time-points, the smoother the survival curve. This means that a smooth survival function is obtained when many time points are observed. In case the survival durations are not arranged in increasing order in your data, these should be arranged in ascending order before the calculations. The survival function can be used to find the mean and median duration of survival along with their confidence interval, which can easily be done using a computer program. This is explained with the help of an example.

Example – Wainstock, et al. [5] studied neurological morbidity in children born to severely anemic (Hb <7 mg/dL) women. The follow-up time varied in their study from child to child, but we assume it to be fixed 10 years for our example. The incidence of neurological morbidity is generally low, but suppose the following duration was observed in 20 children for developing neurological morbidity.

Duration (y) elapsed for occurrence of neurological morbidity in children born to severely anemic women:

10+	7	3	9	10+	10+	8	7	7	5+
7	9	5	4+	2	1	9	4	10+	10+

where + denotes that the observation is censored. Five children with 10+ did not develop any neurological morbidity till the age of 10 years, and two children with values 5+ and 4+ were lost to follow-up when their age was 5 years and 4 years, respectively. These durations are arranged in increasing order in **Table II** along with the other details of the K-M procedure.

The K-M method requires that the time, such as 'years' in this example, is considered as a continuous variable. This can be in decimals. Yet, for clarity, first column in **Table II** is the beginning time when, in our example, the start is from 20 cases. One case developed morbidity at 1 year so that only 19 remain at risk for developing morbidity at the beginning of the second year. Similarly, the number at risk at other time points is obtained. The K-M method calculates the proportion surviving at each point in time based on cases dying at each point in time after excluding the censored values. Subsequent proportion is obtained by multiplying with the previous proportion. The method is explained in the calculations shown in **Table II**. The event under study is developing neurological morbidity in our example in place of death.

The other important requirements for valid results from the K-M methods are: *i*) The censored values do not belong to special subjects so that the distribution of the censored duration is the same as of complete values; *ii*) The subjects are independent. That is, the survival duration of one does not affect the survival duration of the other; *iii*) The rate of the event is the same in early recruiters as in the late recruiters; *iv*) In general, the number of censored values should not exceed the number of complete values.

The plot of the survival function against time gives the survival curve (**Fig. 1**). This is based on the data in **Table II** on time for the occurrence of neurological morbidity in children born to severely anemic women. This shows how the numbers at risk are declining with time. The censored values are shown by a "+" sign (there are five overlapping plus signs at 10 years). The plot can be used to find median time to develop neurological morbidity by drawing a horizontal line at proportion = 0.50 and projecting it down from the point of intersection to the x-axis. This is the time when half of the patients remained at risk and the other half had the event. In this figure, the median duration of developing neurological morbidity is nearly 8 years. The survival table (**Table II**) gives a more exact median value, which is between 7 and 8 years where the survival proportion is 0.5.

The complimentary of survival (1–proportion survived) is called the hazard since it depicts the proportion of deaths at different points in time, where 'hazard' again is a generic term for any outcome of interest.

Table II illustrates that the computations for K-M survival function are tedious and a statistical package is usually used. The software also calculates the confidence intervals for each survival probability. The estimates for the

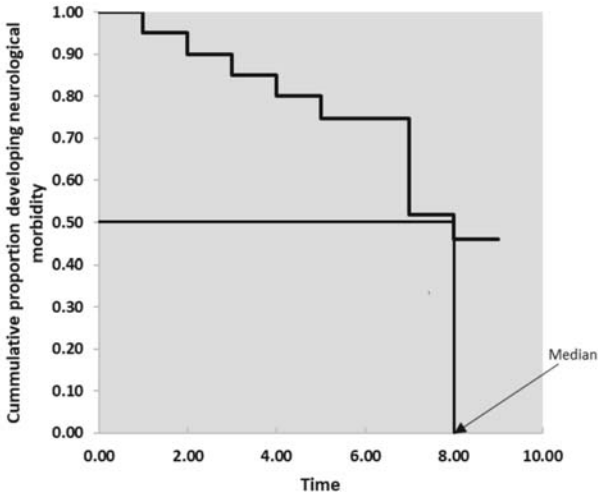

Fig. 1 Survival curve and the median survival time.

Table II Survival Analysis by K-M Method

Beginning time (year) t	Number at risk $n_{t+1} = n_t - d_t - c_t$, $n_0 = 20$	Developed morbidity at time t d_t	Censored c_t	Survival function (proportion with no morbidity) at time t $s_0 = 1$, $s_t = s_{t-1}(\frac{n_t - d_t}{n_t})$
1	20	1	0	$1 * \frac{20-1}{20} = 0.950$
2	19	1	0	$\frac{19}{20} * \frac{19-1}{19} = 0.900$
3	18	1	0	$\frac{18}{20} * \frac{18-1}{18} = 0.850$
4	17	1	1	$\frac{17}{20} * \frac{17-1}{17} = 0.800$
5	15	1	1	$\frac{16}{20} * \frac{15-1}{15} = 0.747$
6	13	0	0	$\frac{16}{20} * \frac{14}{15} * \frac{13-0}{13} = 0.747$
7	13	4	0	$\frac{16}{20} * \frac{14}{15} * \frac{13-4}{13} = 0.517$
8	9	1	0	$\frac{16}{20} * \frac{14}{15} * \frac{9}{13} * \frac{9-1}{9} = 0.459$
9	8	3	0	$\frac{16}{20} * \frac{14}{15} * \frac{8}{13} * \frac{8-3}{8} = 0.287$
10	5	0	5	

probability of survival are relatively unreliable for the time points towards the end because of fewer surviving subjects at those time points. The area under the curve is the same as mean surviving time but that is rarely used.

Estimation of the survival function or median survival time by itself may not be of much value unless it is compared with another group to find where the survival is longer. A plot of survival curves of different groups gives an indication of which group has better survival experience and the statistical significance can be checked with log-rank test.

Log-Rank Test

In place of comparing just the median survival time or the proportion surviving at specific time *t*, the log-rank test is used to compare the overall survival pattern of one group with other groups. The median may be nearly the same, but the pattern could be different (**Fig. 2**). In this figure, survival is better in one group at initial time points but worse at later time points, with survival curves shown as smooth curves for illustration. The right kind of survival pattern for comparison by log-rank test is shown in **Fig. 3**.

For example, the interest might be in comparing the pattern of duration of hospitalization of cases on a new regimen with those on the existing regimen. Another example is the time to return to school by students with and without attention-deficit/hyperactivity disorder (ADHD) following concussion [6]. Each of these can be compared using log-rank test.

Log-rank test also is a non-parametric procedure based on chi-square and used to check whether two or more survival curves are statistically significantly different. The null hypothesis is that the difference between the survival curves is due to sampling of cases and not real. For simplicity, we describe this test for two groups. The test requires that all the conditions mentioned earlier for the K-M method are met. That is, the sample size must be large and representative, censored values are random and not related to survival, early recruiters have the same survival as the late recruiters, and the survival time is recorded exactly and not in intervals. In addition, the survival curves should not cross each other – one should be lower than the other at most time points. Thus, the curves of the type shown in **Fig. 2** cannot be compared with log-rank test but those in **Fig. 3** can be compared. Also, the two groups must be independent – the survival of the patients in one group should not have any relationship with the survival of patients in the other group.

The null hypothesis of identical survival curves in the two populations implies that the probability of survival (or of death) at *each* point of time is the same in one group as in the other. Under this null hypothesis, the expected number of

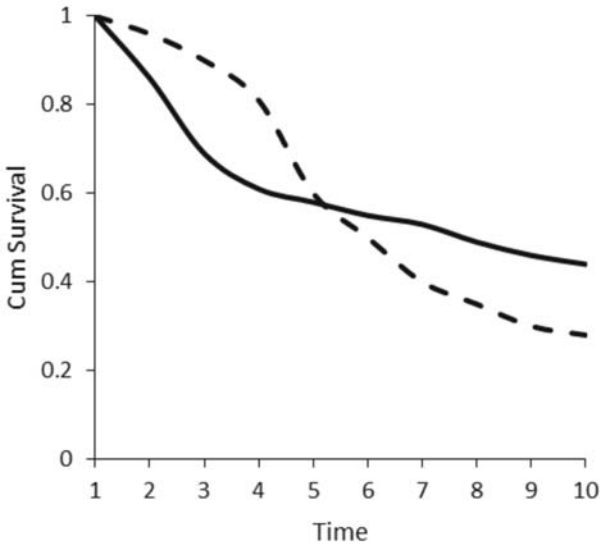

Fig. 2 Survival curves crossing each other.

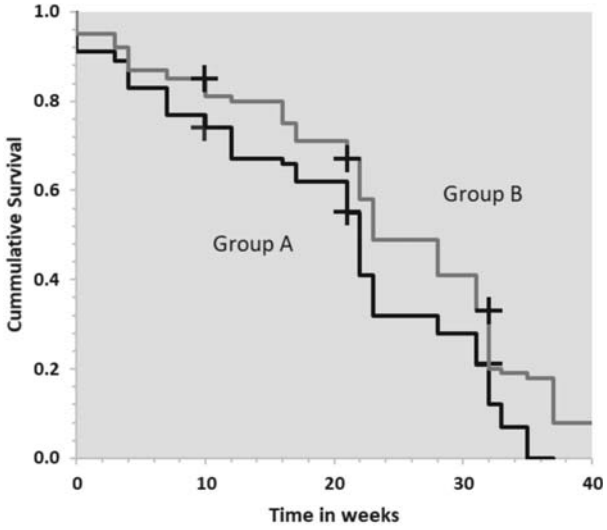

Fig. 3 Survival curves in two groups. The decline in survival in group A is steeper than in group B.

deaths is calculated for each time using the combined experience in the two groups.

For example, if there are 5 deaths in group-I at time *t* when 200 people are still at risk (after deaths and dropouts before time *t*), and 7 deaths in group-II where 100 people are at risk, the expected deaths under the null hypothesis in group-I are $\frac{100}{300}[5+7]=4$ and in group-II are $\frac{200}{300}[5+7]=8$. The sum of these kinds of numbers over different time-points are used to calculate the log-rank test.

The *P* value can be obtained in the usual manner using chi-square. This will only tell you that the difference is statistically significant or not, but would not tell you about the magnitude of the difference. For magnitude, hazard ratios at different points of time are calculated.

One subtle and difficult to understand limitation of the log-rank method is that statistical significance is mostly driven by the difference in survival at initial time points where the number of subjects is high. The number of available subjects declines at subsequent time points due to deaths and dropouts. To overcome this problem, Breslow test or Tarone-Ware test is used for comparing two survival curves when the number of subjects is large in the beginning but quickly declines and becomes small towards the end. That is, when the mortality and dropouts are high [7]. **Box I** lists when to use and when not to use survival analyses.

Hazard Ratio

Risk is the probability of occurrence of an event and is generally calculated at the end of the study irrespective of the time; whereas, the hazard rate is the risk per unit of time such as hazard of recurrence of papillary thyroid cancer per year in a high-risk population [8]. This can be different from year-to-year such as a low rate in the first year and progressively increasing each year. Hazard rate is especially used in survival studies because time is an important factor in these studies.

Box I Where to Use Survival Analysis

- For analyzing any duration (time-to-event) data if it has highly skewed distribution, particularly if some durations could not be fully ascertained (censored), but the subjects with censored values should not have special survival pattern.
- When the interest is in studying the complete survival pattern at different points in time even if no duration is incomplete.
- For comparing survival pattern in two or more groups (log-rank test)

Do not use

- When the number of censored values is more than the complete values because then the median survival duration and the area under the survival curve have poor reliability.
- The log-rank test can give misleading result when the two survival curves cross each other. Also, the groups under comparison must be independent.
- When the deaths or dropouts are fast, the numbers towards the tail become too small and the log-rank test loses its power.

The ratio of the two hazards called hazard ratio is used to compare hazard rate in one group with another, such as hazard of developing anemia per year in adolescent girls of low versus high socioeconomic (SE) status. A hazard ratio of 1.25 says that the hazard of developing anemia per year in girls of low SE status is 1.25 times (or 25% higher) of such a hazard in girls of high SE status. If this ratio remains the same over the period of the study, called proportional hazards, Cox [9] has shown that the factors affecting this ratio can be easily studied by a regression model. This analysis tells which factor is contributing how much and its significance towards difference in the hazard rates in the two groups under study. A more detailed explanation has been given by Kleinbaum and Klein [10].

CONCLUSIONS

A separate method of analysis is required for duration (time-to-event) data because durations generally have censored values and a highly skewed distribution. The method of survival analysis is nonparametric and takes care of both these 'aberrations' in the data. Survival is a generic term, and the method is applicable to any duration data. Kaplan-Meier is the method of choice to study the complete survival pattern when the duration is measured on continuous scale and when censored values are not related to the survival pattern. This can be used to estimate the median survival time despite censored values. Comparison of survival pattern in two or more independent groups is done by log-rank test.

REFERENCES

1. Hanna Y, Laliberté C, Ben Fadel N, et al. Effect of oxygen saturation targets on the incidence of bronchopulmonary dysplasia and duration of respiratory supports in extremely preterm infants. Pediatr Child Health. 2020;25:173-9.
2. Bryson AE, Cabral HJ, Coles MS. Attendance of an initial follow-up visit after long-acting reversible contraception insertion and method continuation among adolescents and young adults: a retrospective study. J Pediatr Adolesc Gynecol. 2021:S1083-3188(21)00004-8.
3. Jha, UM, Dhingra N, Raj Y, et al. Survival of children living with human immunodeficiency virus on antiretroviral therapy in Andhra Pradesh, India. Indian Pediatr. 2018;55;301-5.
4. Indrayan A, Malhotra RK. Medical Biostatistics, Fourth Edition. CRC Press; 2018.
5. Wainstock T, Walfisch A, Sergienko R, Sheiner E. Maternal anemia and pediatric neurological morbidity in the offspring - Results from a population based cohort study. Early Hum Dev. 2019;128:15-20.
6. Cook NE, Iverson GL, Maxwell B, Zafonte R, Berkner PD. Adolescents with ADHD do not take longer to recover from concussion. Front Pediatr. 2021;8:606879.
7. Cleves M, Gould W, Marchenko Y. An introduction to survival analysis using stata. Third edition. Stata Press; 2016.
8. Siraj AK, Parvathareddy SK, Qadri Z, et al. Annual hazard rate of recurrence in Middle Eastern papillary thyroid cancer over a long-term follow-up. Cancers (Basel). 2020;12:3624.
9. Cox DR. Regression models and life tables. J Royal Stat Soc B 1972;34:187-220.
10. Kleinbaum DG, Klein M. Survival Analysis: A Self learning text. Second edition. Springer; 2005.