

생존분석시 주의깊게 고려하여야 할 점은 무엇인가?

곽상규¹ · 최은진²

대구가톨릭대학교 의과대학 ¹의학통계학교실, ²소아과학교실

What Should We Consider Carefully When Performing Survival Analysis?

Sang Gyu Kwak¹ and Eun Jin Choi²

Departments of ¹Medical Statistics and ²Pediatrics, School of Medicine, Catholic University of Daegu, Daegu, Korea

The survival data and the survival analysis are the data and analysis methods used to study the probability of survival. The survival data consist of a period from the juncture of a start event to the juncture of the end event (occurrence event). The period is called the survival period or survival time. In this way, the method of analysing the survival time of subjects and appropriately summarizing the degree of survival is called survival analysis. To understand and analyse survival analysis methods, researchers must be aware of some concepts. Concepts to be aware of in the survival analysis include events, censored data, survival period, survival function, survival curve and so on. This review focuses on the terms and concepts used in the survival analysis. It will also cover the types of survival data that should be collected and prepared when using actual survival analysis method and how to prepare them.

pISSN 2233-5250 / eISSN 2233-4580
<https://doi.org/10.15264/cpho.2019.26.1.1>
Clin Pediatr Hematol Oncol
2019;26:1~5

Received on March 9, 2019

Revised on March 26, 2019

Accepted on April 3, 2019

Corresponding Author: Eun Jin Choi
Department of Pediatrics, School of
Medicine, Catholic University of
Daegu, 33 Duryugongwon-ro
17-gil, Nam-gu, Daegu 42472, Korea
Tel: +82-53-650-4248
Fax: +82-53-621-4106
E-mail: ejchoi2@cu.ac.kr
ORCID ID: orcid.org/0000-0001-7116-8538

Key Words: Censored data, Hazard ratio, Kaplan-Meier method, Survival analysis

Introduction

When patients are diagnosed with cancer or incurable disease for the first time in a hospital, they feel the fear of death very much. They may have a lot of questions about their illness, but among them, the question of whether they can live for years or even months is the most representative. In response to this question, the medical team cannot assert that the patient can live for exactly years or months, but based on the patient's health status or test results, the probability of surviving for three or five years

can be answered as a percentage. The rationale for this answer is based on the results of studies and papers published so far. The survival data and the survival analysis are the data and analysis methods used to study the probability of survival.

The survival data consist of a period from the juncture of a start event to the juncture of the end event (occurrence event). The period is called the survival period or survival time. For example, if surgery is the start event and death is the end event, the period from surgery to death is the survival time. It is judged according to the survival period, that is, whether it is short or long, and the curative effect

on the operation is judged. In this way, the method of analysing the survival time of subjects and appropriately summarizing the degree of survival is called survival analysis.

There are several original articles used survival analysis method. For example, Park et al. [1] used the survival analysis method to compare disease free survival rate and overall survival rate according to the NK cell percentages in newly diagnosed acute myeloid leukemia. Kang et al. [2] used the survival analysis method to compare overall survival rate and relapse-free survival rate according to 4 groups (high-risk, high-intermediate, low and low-intermediate) for classic IPI (International Prognostic Index), NCCN-IPI (National Comprehensive Cancer Network-International Prognostic Index) and modified IPI. Jung et al. [3] used the survival analysis method to compare progression-free survival rate and overall survival rate between STAT3-positive and STAT3-negative in non-elderly adult patients with newly diagnosed multiple myeloma. The three examples are the ones that calculate the overall survival rate and the event free survival rate and compare them by group. However, the survival analysis is not a method of statistical analysis that is limited to events related to living and dying. Because of the word “survival,” there has been a misconception that it will only be used for data on the death of a patient or the success or failure of treatment of a particular disease, and so the survival analysis is not widely used. If only data on a specific event and a period from a specific juncture to an event occurrence date are collected, the survival rate can be analysed for each situation using a survival analysis method.

To understand and analyse survival analysis methods, researchers must be aware of some concepts. Concepts to be aware of in the survival analysis include events, censored data, survival period, survival function, survival curve and so on.

This review focuses on the terms and concepts used in the survival analysis. It will be also covered the types of survival data that should be collected and prepared when using actual survival analysis method and how to prepare them.

Events

Events can be defined according to the topics and circumstances of each study, such as surgery, treatment, death, survival, relapse, and re-operation. However, events must be defined from binary data. The binary data are data that can be represented only in two categories or states. In other words, if death is defined as an event, it is defined by a binary data consisted death or survival. Data with more than three categories cannot be defined as an event, but can be defined as an event by modifying the data into two categories. For example, hazard classification data with three categories of high hazard, low hazard, and normal should be reconstructed into hazard data, which is high hazard and low hazard, together with hazard data. It is then possible to define hazard as an event from the hazard data. Researchers should also make clear the definition of the event. If death is defined as an event, the cause of death should be specifically mentioned. It should be clear whether the event includes all causes of death, such as accidental death, death due to illness not related to research, death due to research-related diseases, or an event involving only a specific cause of death. As described above, in the case of the event having the same type, although the cause is different, in this review, it is defined as “events of the same type with different causes”.

When creating the survival data, it should be recorded as the number 1 if the event occurred and the number 0 if the event did not occur. In most studies using survival analysis methods, death is defined as an event and death is recorded as 1, survival as 0 in the survival data, and the survival rate is presented as a result. Conversely, if survival is defined as an event, the survival should be recorded as 1 and death as 0, and the results presented using the survival analysis method should present the mortality rate rather than the survival rate.

The main events covered in the survival analysis are death and recurrence. Death and recurrence are clearly different concepts in clinical field. Therefore, when performing survival analysis for death and recurrence, definition of event should be precise. If only the definition of the event

is clear, there is no particular difference in mortality and recurrence when performing survival analysis [4].

Censored Data

The censored data are data that cannot be confirmed whether the event occurred between the date of the start event and the ended date of the last follow up. The most ideal data for using the survival analysis method is that all study subjects are observed until a defined event occurs. However, it is practically impossible for all subjects to be monitored until an event occurs. Therefore, censored data must exist in survival data. For example, if a researcher has set a tracking observation period of several years from the juncture of a start event, the end-of-study event may not occur and the end event may occur after the tracking observation period. This case is called right censoring data. In addition, if the subject of the study experiences a dropout or events of the same type with different causes before the end of the last follow up, the survival period may not be as long as the specified period set by the researcher. This case is called left censoring data [5].

There is no special consideration for censored data when creating survival data. However, the date of the start event, the occurrence date of the end event, the occurrence date of the dropout event, the occurrence date of the events of the same type with different causes, and the ended date of the last follow up should be clearly specified and re-

corded.

Survival Period

The survival period is the period from the juncture of the start event to the juncture of the end event (occurrence event). The period is called the survival period or survival time. For example, if surgery is the start event and death is the end event, the period from surgery to death is the survival period. When creating survival data, the survival data should include the censored data, so that the survival period should be calculated with different calculations at each period, as shown in Table 1.

Also, the unit of survival period is often used for years. This is because in studies using survival analysis methods, the survival rate of an annual unit, such as a survival rate of three years or a survival rate of five years, is concerned. Therefore, when the unit of survival period is converted, accurate calculation must be performed and at this time, there should be no distortion of the data value. For example, when converting the unit of survival period from day to month, data of the day should not be converted into data of month by dividing by 30. Since the number of days in one month is 28 days, 30 days, and 31 days in 12 months of the year, the survival period unit conversion should be performed in consideration of this. In this review, it is recommended to use the following two formulas to change the daily survival period data into monthly survival data,

Table 1. Calculation for survival period

No.	Lapse of time					Survival period
1	Occurrence of the start event	→	Occurrence of the end event	→	End of last follow up	Survival period= occurrence date of the end event—occurrence date of the start event
2	Occurrence of the start event	→	No end event occurs	→	End of last follow up	Survival period= ended date of last follow up—occurrence date of the start event
3	Occurrence of the start event	→	Occurrence of dropout	→	End of last follow up	Survival period= occurrence date of a dropout—occurrence date of the start event
4	Occurrence of the start event	→	Occurrence of the end event of the same type with different causes	→	End of last follow up	Survival period= occurrence date of the end event of the same type with different causes—occurrence date of the start event

Survival period (months)

$$= \frac{\text{Survival period (days)}}{365.25/12} \text{ month} \quad (1)$$

$$= \frac{\text{Survival period (days)}}{365/12} \text{ month} \quad (2)$$

It is recommended that the survival period be calculated differently depending on whether the leap month is included once every four years. First, if the leap month is included in the survival period, the survival period unit conversion formula is 365 days in a year, considering that there is a month (29 days in February) every 4 years, the number of days is assumed to be 365.25, and this is divided by 12 to divide the daily survival period data, and the monthly survival period data is calculated. Also, if the leap month is not included in the survival period, the survival period unit conversion formula is such that the number of days of the year is 365, the number of days is divided by 12, and the survival period data per month is calculated.

Survival Function

The survival function is a function indicating the probability that the end event will not occur until a specific juncture including the start event [6]. For example, if the end event is death, it indicates the probability of not dying up to a certain juncture, such as one year or two years. If the juncture in the survival function is 0, the survival rate is 1.0 (100%), and every time the end event occurs, the survival rate tends to decrease gradually. The number of end events occurring during the entire study period divided by the total number of subjects is the overall incidence rate for the end event. Therefore, it is different from the survival rate calculated by the survival function. If the planned study is to calculate the incidence rate for the event, the simple rate can be calculated, but if the study is to calculate the survival rate up to a certain juncture, the survival analysis method should be used.

Survival Curve

The survival curve is a curve in which the survival rate gradually decreases in a stepwise manner as the survival time increases and the end event occurs. As the survival time increases, the number of subjects included in the juncture decreases. A typical method for calculating the survival curve is the Kaplan-Meier method. This method is a method of calculating the survival rate in the juncture in which the end event occurs by arranging the data in the shortest survival period to longest. Also, Kaplan-Meier survival curves can be compared according to two or more groups. For example, this is the case in which the survival curves are compared according to the surgical method and the case in which survival curves are compared according to the type of the anticancer drug. There are many methods comparing survival curves according to factors such as the Mantel-Haenszel method, the Log-rank test method, the Breslow test, and the Likelihood ratio, but the Log-rank test method is used in many studies. The null hypothesis of the Log-rank test method is that “the survival curves of the two groups are not different overall” and the observed and expected counts are compared by the chi-square testing method. The Log-rank method is known to have excellent power when the survival curves of the comparison groups are significantly different. When comparing three or more groups of survival curves, multiple comparisons are possible, such as the Bonferroni test or the Holm-Sidak method [7].

To use the log-rank test method, the proportional hazards assumption that the hazard ratios of the two groups are constant over time should be satisfied. The simplest way to confirm the proportional hazards assumption is to check the Kaplan-Meier survival curves for both groups. If the two survival curves intersect and the slope trends of the two survival curves are largely different, it can be judged that the proportional hazards assumption is not satisfied. If the proportional hazards assumption is not met, the Breslow test or other methods that use the number of subjects at risk in the juncture as a weight should be applied.

Discussion

The method of analysing the survival time of subjects and appropriately summarizing the degree of survival is called survival analysis. In this review, the terms and concepts used in survival analysis were covered, and also types and instructions for survival data to be collected and created were covered when using the practical survival analysis method. It is expected that the concept of event, survival period, censored data, survival function, survival curve, comparison of two or more survival curves are understood, and studies applying survival analysis are actively performed.

References

1. Park Y, Lim J, Kim S, et al. The prognostic impact of lymphocyte subsets in newly diagnosed acute myeloid leukemia. *Blood Res* 2018;53:198-204.
2. Kang J, Yoon S, Suh C. Relevance of prognostic index with β 2-microglobulin for patients with diffuse large B-cell lymphoma in the rituximab era. *Blood Res* 2017;52:276-84.
3. Jung SH, Ahn SY, Choi HW, et al. STAT3 expression is associated with poor survival in non-elderly adult patients with newly diagnosed multiple myeloma. *Blood Res* 2017;52:293-9.
4. Mouracade P. Key concepts of survival analysis: Checking appropriateness. *Prog Urol* 2017;27:331-3.
5. Canales RA, Wilson AM, Pearce-Walker JL, Verhougstraete MP, Reynolds KA. Methods for handling left-censored data in quantitative microbial risk assessment. *Appl Environ Microbiol* 2018;84.
6. George B, Seals S, Aban I. Survival analysis and regression models. *J Nucl Cardiol* 2014;21:686-94.
7. Delgado J, Pereira A, Villamor N, López-Guillermo A, Rozman C. Survival analysis in hematologic malignancies: recommendations for clinicians. *Haematologica* 2014;99:1410-20.