ACTA MÉDICA PORTUGUESA
A Revista Científica da Ordem dos Médicos

Kaplan-Meier Survival Analysis: Practical Insights for Clinicians

Análise de Sobrevida de Kaplan-Meier: Fundamentos Práticos para Clínicos

António Pedro GOMES✉[1], Bruna COSTA[2], Rita MARQUES[3], Vitor NUNES[1], Constança COELHO[4]
Acta Med Port 2024 Apr;37(4):280-285 ▪ https://doi.org/10.20344/amp.21080

ABSTRACT

This article aims to provide a guide that will help healthcare professionals and clinical researchers from all fields that deal with Kaplan-Meier curves. Survival analysis methods are among the most frequently used in the medical sciences and in clinical research. Overall survival, progression free survival, time to recurrence, or any other clinically relevant parameter represented by a Kaplan-Meier curve will be discussed. We will present a practical and straightforward interpretation of these curves, setting aside intricate mathematical considerations. Our focus will be on essential concepts that interface with biological sciences and medicine in order to guarantee proficiency in one of the most popular yet frequently misunderstood methods in clinical research. Being familiar with these concepts is not only essential for designing new clinical studies but also for critically assessing and interpreting published data.
Keywords: Kaplan-Meier Estimate; Survival Analysis

RESUMO

Este artigo tem como objetivo funcionar como um guia que ajudará profissionais de saúde e investigadores clínicos de todas as áreas que lidam com curvas de sobrevida. Os métodos de análise de sobrevida estão entre os mais usados nas ciências médicas e na investigação clínica. Serão discutidos os conceitos de sobrevida global, sobrevida livre de progressão, tempo de recidiva e todos os parâmetros de interesse clínico que possam ser representados por curvas de Kaplan-Meier, com uma interpretação prática e direta dessas curvas. Deixaremos de lado todas as considerações no campo da matemática. Referiremos apenas conceitos essenciais que interagem com as ciências biológicas e a medicina, de forma a garantir a proficiência de um dos métodos mais populares e frequentemente incompreendidos na investigação clínica. Estar familiarizado com esses conceitos é essencial não apenas para projetar novos estudos clínicos, mas também para avaliar e interpretar estudos publicados de forma crítica.
Palavras-chave: Análise de Sobrevivência; Estimativa de Kaplan-Meier

INTRODUCTION

Survival analysis provides an essential tool for understanding results in clinical research. It is of the utmost importance to know how to interpret these methods, as well as to be aware of alternatives and their limitations.

This paper was specifically designed to elucidate the basic principles of the analysis, interpretation, design, and execution of the most widespread and standard method in survival analysis, according to the typical needs arising in clinical practice. It does not explain alternative and more accurate methods adapted to specific scenarios. Also, it does not address multivariable regression analysis.

The core of survival analysis is time and how the factors under study may affect the time until the event, rather than the event itself. Although the term 'survival' intuitively makes us think of the event as death, it may be any other outcome. A straightforward example is given by José de la Mata *et al*[1] when they analyzed the time to "treatment termination because of toxicity and lack of response"; Zhang *et al*, while studying the association of levels of metalloproteinase-7 and the risk of renal survival and fibrosis, defined one of the endpoints as the time between the diagnosis of nephropathy and the event which was defined as a 50% decline in glomerular filtration rate[2]; Conden *et al* studied the influence of type D personality and the time to the event of recurrent myocardial infarction.[3]

Kaplan-Meier curves

Kaplan-Meier (KM) curves are the most traditional method to show survival data. Its success is due to the amount of information that can be obtained from one single chart. Fig. 1 depicts an example of a KM curve. The horizontal x-line measures time, while the vertical y-axis refers to the survival rate. Using this figure as an example, in the beginning, 100% of the study sample did not experience the event. The first step-down means that the first event occurred at four months. Each step-down represents one or more events. The line immediately after the step-down represents the proportion of patients who did not experience the event after that time. In contrast, the line immediately before the step-down represents the proportion of patients who did not experience the event until that time. The vertical difference between both lines represents the number of events that

1. Surgery Department. Hospital de Vila Franca de Xira. Vila Franca de Xira. Portugal.
2. Champalimaud Centre for the Unknown. Lisbon. Portugal.
3. Surgery Department. Unidade de Caldas da Rainha. Centro Hospitalar do Oeste. Caldas da Rainha. Portugal.
4. Genetics Laboratory. Faculdade de Medicina. Universidade de Lisboa. Lisbon. Portugal.
✉ **Autor correspondente:** António Pedro Gomes. antonio.gomes2@hvfx.min-saude.pt
Recebido/*Received*: 10/12/2023 - **Aceite/*Accepted*:** 06/02/2024 - **Publicado/*Published*:** 01/04/2024

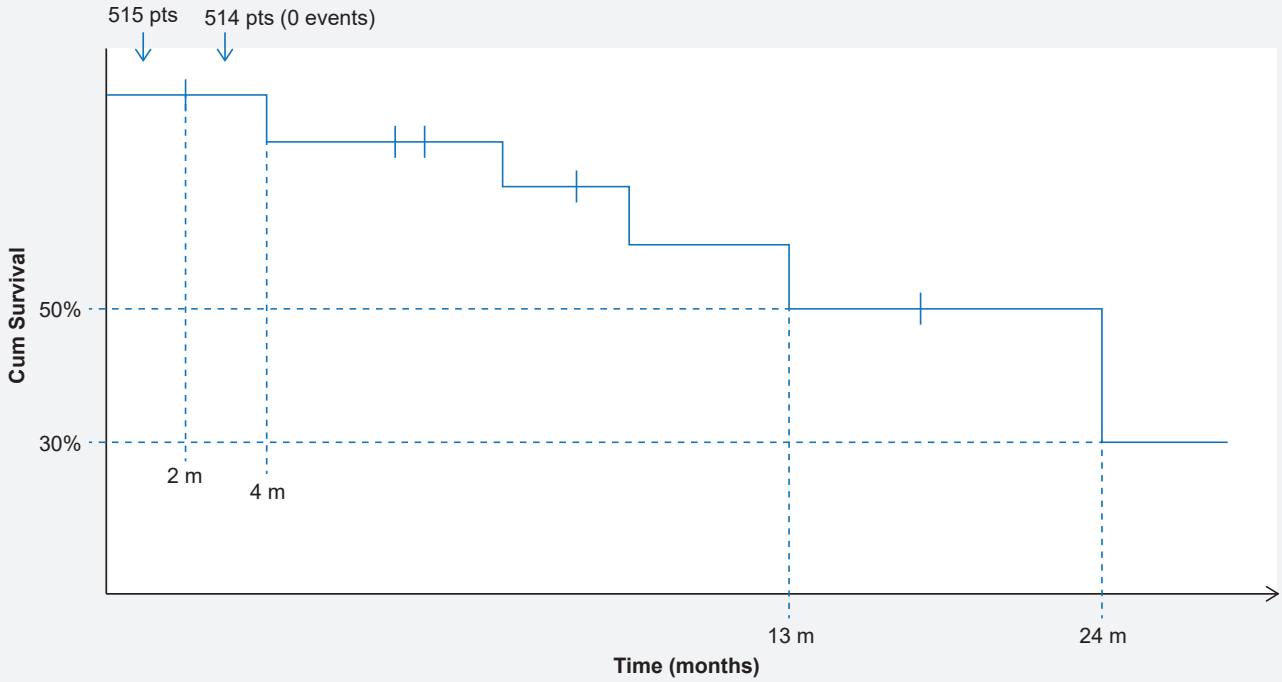

Figura 1 – Example of a Kaplan-Meier curve

occurred at that time point. The vertical step-down is as high as the number of events. However, the number of right-censored patients up to that time-point (the number of patients who were lost to follow-up without experiencing an event in the last time interval) will also increase the step-down.

In survival analysis, we usually refer to the median survival time (and not the average time): a median survival time of 13 months means that at 13 months after treatment (or any other variable), 50% of all subjects have experienced the event.

In Fig. 1, survival at two years is 30%. Two-year survival, five-year survival, or survival at any specific time point corresponds to the proportion of patients who, after two years, five years, or X years of follow-up, have not experienced the event. This is computed by dividing the number of patients who did not experience the event after the analyzed time by the number of patients included.

The vertical lines (|) in the figure indicate censoring, i.e., at any given time, a patient leaves the study without experiencing an event. In other words, censoring represents patients who exit the study, either due to being lost to follow-up or reaching the end of the study time period without having experienced the event under investigation. Censoring in a study occurs when there is incomplete information regarding the event in a given participant. Right censoring happens specifically when a subject leaves the study before experiencing the event. In such cases, we do not know whether the patient experienced the event or when the event occurred. However, we do know that during the period in which a given patient was included, no event occurred. That incomplete information is of utmost importance for the study. This is the case of lost-to-follow-up (or when the study time ends without any event occurring) and is referred to as type I right censoring. On the other hand, type II right censoring refers to studies that ended when a predefined number of individuals have experienced the event.

Left censoring is observed when the subject is already at risk for the event being studied before the study initiation, a scenario that is unusual in most clinical studies.

Interval censoring, on the other hand, occurs when there is no information regarding the occurrence of the event for a given patient during a specific time period of the study. This situation is also not commonly observed in 'standard' clinical studies.[4]

After censoring, the proportion of patients with no event is lower (censored subjects leave the study), although the number of events is the same.

Still in Fig. 1, 515 patients were included and reached two months with no event. However, only 514 patients remained after two months without an event. The patient represented by the vertical line was a 'dropout' at that time, but no event had occurred.

Comparing different groups and different outcomes with Kaplan-Meier curves

Lee EC *et al* studied the time until death for all causes and the time until recurrence.[5] The authors could draw a KM curve regarding time to recurrence according to T-stage, N-stage, gender, and other variables. However, no information could be collected regarding the influence of all variables together since KM is not a multivariable analysis.

Each variable must be a categorical/ordinal variable: it is possible to compare T1 *vs* T2 *vs* T3; N0 *vs* N+; male *vs* female. However, it is not possible to study continuous variables such as age, unless the continuous variable can be grouped in classes.

Figure 2A shows KM curves, which can be read as time to event according to a given variable, where the variable is represented as group 1 and 2. Immediately, the reader assumes that group 1 has a better prognosis with a time to event lower than group 2 and, consequently, a higher survival at any given time point. However, to infer population results based on data from a sample (statistical inference), a hypothesis test has to be performed. The hypothesis test will determine if the result is statistically significant.

This may be a misinterpretation because to know if that difference is statistically significant, a statistical test must be performed. The most widely used test is the log-rank test, although there are others for specific situations. The log-rank test compares the difference between the real-life observed curves with a hypothetic expected curve where all the events would be equally distributed between groups. The larger the difference between the observed and expected curves, the higher the likelihood of the different curves being statistically different. The log-rank test is the most powerful rank test when considering the proportional hazard model.

This model assumes that the effect of exposure on the hazard rate remains constant over time. In simpler terms, it implies that the risk of death, recurrence, or any other event remains the same after one week of the treatment (or other intervention) or after one year. This is the opposite of accelerated models, where the risk of an event is higher or lower over time. For instance, in accelerated models, the risk of death or cancer after one week of quitting smoking is higher than the risk of death or cancer after two years. Whenever the hazard rate remains constant over time, it is called a proportional hazards model, and the log-rank test is the appropriate statistical method.

The Gehan-Breslow (Wilcoxon) test assigns more weight to early time events, making it particularly useful when studying early differences in two or more survival curves (Fig. 2A). While the log-rank test provides an outline between observed survival times in both groups and how they deviate from the expected values, the Gehan-Breslow-Wilcoxon test is based on the number of participants at risk in both groups at a given time. That is why it assigns greater

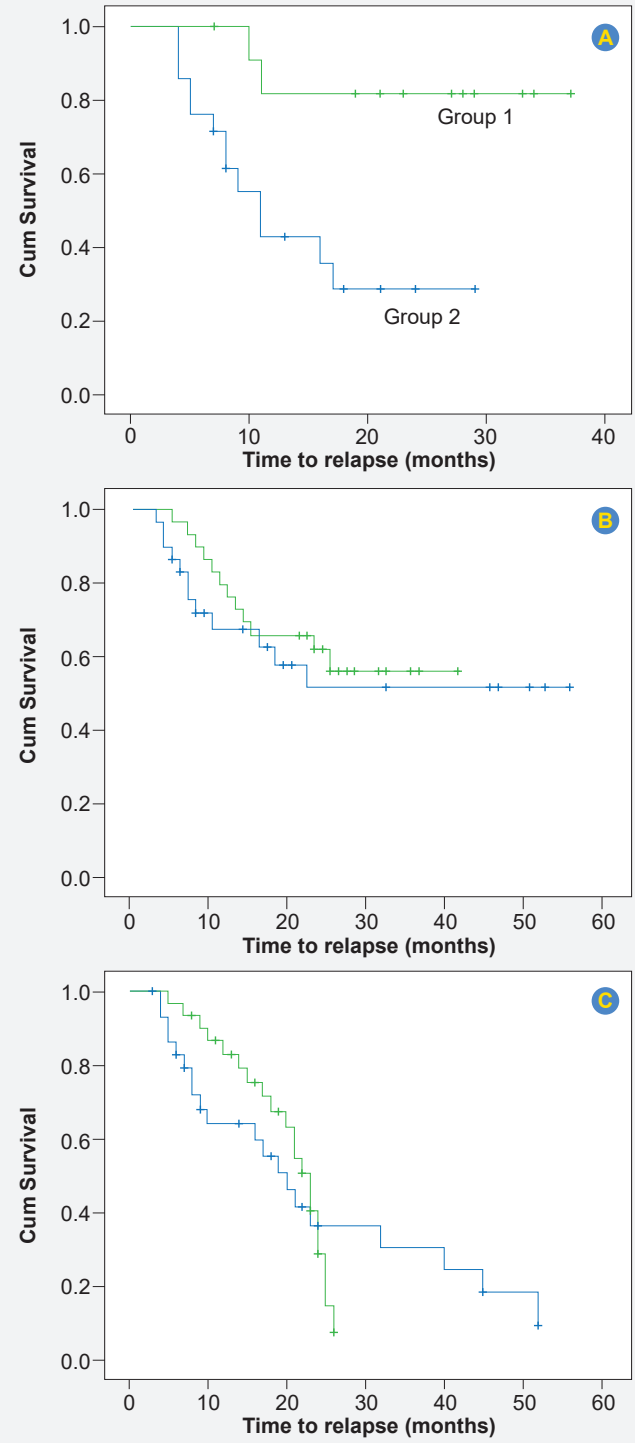

Figura 2 – Examples of a Kaplan-Meier curves comparing two groups of patients/subjects

weight to early differences. For example, it is often applied in certain types of cancer studies where death or recurrence tends to occur within the first or second years, even when patients are followed up to 10 or more years. However, this test is more sensitive to early censoring.

When two or more survival curves cross each other (Fig. 2B), it indicates a change in the hazard rate over time, implying non-proportional hazards. In such cases, Tarone-Ware, Peto-Peto, and Fleming-Harrington family tests may be performed. However, addressing these cases poses statistical challenges, and there is no correct answer regarding the most powerful test to apply. Most widely available software packages offer a variety of tests specifically tailored to each of these situations.

In clinical research, a 95% confidence level is usually the standard. This means there is a 95% probability that the 'true' value of the population falls within that range. If we were to repeat the experiment 100 times under the same conditions, in approximately 95 of those instances, the true value of the population would fall within the assumed confidence interval. It is important to note that the width of the confidence interval decreases as the sample size increases; it is inversely proportional to the square root of the sample size).[6]

Verifying Kaplan-Meier assumptions

Statistical models aim to represent reality and to make predictions as accurately as possible. To do so, statistical models need to rely on certain principles (assumptions). By violating the assumptions, we are withdrawing predictive acuity from these models. To obtain valid conclusions from these graphs, it is necessary to keep in mind the following principles:

1 – Independent samples: groups are based on random selection (the selection of one patient does not influence another patient's selection);
2 – Censoring is independent of the outcome and independent of the study group – the chance of a subject 'drop out' before the event should be the same in each group;
3 – If a subject is included today, the probability of having the event in the next year must be the same for a subject that will be included in the following months;
4 – An event can occur at any given time. However, in clinical sciences, most of the time we can only know that the event occurred between two time points. For example, a KM curve from a sample where patients are followed-up through 18 months with scheduled visits every six months has much less information than the same sample followed-up through 56 months with scheduled visits every three months.

The previously mentioned assumptions are intimately associated with methodology, data collection, and database design and not so much with statistical issues. Even though most of the times it is impossible to absolutely verify all the assumptions, the way we measure their impact on our results is very important for external validity and reproducibility of an experiment.

How to collect data to compute KM curves

It is somewhat intuitive to compute survival analysis statistics and KM curves with most of the available statistical packages. However, this relies on data collection and database design.

For each patient/subject it is necessary to clearly state:

1 – Follow-up period (when a patient enters the study and when the patient leaves the study);
2 – The reason for the patient leaving the study (whether it is due to an event or censoring).

How to present descriptive data in survival analysis

a) Study time follow-up;
b) Median follow-up time (FUP) for the whole sample and for study groups: typically used to ascertain how long the subjects were under study. This concept is relevant since in survival analysis each patient may enter and leave the study at different times. A median FUP time of 52 weeks means that 50% of the patients/subjects were studied for at least (or for no more than) 52 weeks;
c) Sample size for the whole sample and for study groups;
d) Number and proportion of events in a given time for the whole sample and for each group. This is usually expressed as two-year survival, five-year survival, etc.;
e) Median survival time (explained above).

This is the minimum standard set of descriptive data needed to present or read a study with survival analysis. As an example, we selected the work of Ryu *et al*[6]: the authors explain that 34 patients were treated for gastric cancer liver metastasis with three different techniques (n = 14, n = 13 and n = 7). They clearly show the median FUP time: 29.4 months (ranging between 2.2 and 170.4 months) and the one-year, three-year and five-year overall survival (84.4%, 38.6%, and 34.7% respectively). Also, censored patients are clearly presented with a dot in the KM curves (Fig. 3).

Most frequently used concepts in survival analysis

Overall survival

Overall survival (OS) measures the time between the diagnosis or treatment initiation of a disease (or other factor clearly defined as the beginning of the study period) and death (or other event) for all causes.

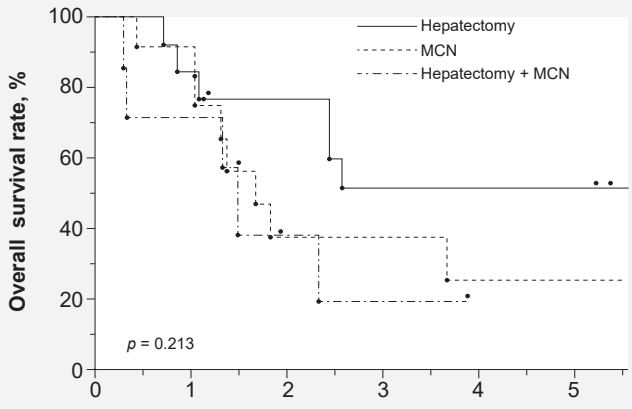

Figura 3 – Kaplan-Meier analysis of overall survival. Three groups, representing three different approaches to liver metastasis from gastric cancer are compared with long-rank statistic. Dots represent censured patients. Reproduced with permission from Tomoki Ryu.[7]

Progression free survival/recurrence free survival/ disease free survival/time to recurrence

Progression free survival (PFS) / Recurrence free survival (RFS)/ Time to recurrence (TTR) and disease-free survival (DFS) represent similar concepts, although arriving at a clear definition is challenging and not consensual.

The primary aim of these measures is to study the time interval between the diagnosis (or treatment) and diagnosis of recurrence. However, it is worth noting that some authors consider death as an event, while others consider it as right-censoring (similar to a loss to follow-up). Furthermore, determining whether recurrence was the cause of death or not can often be a complex task, introducing potential biases into the results.

Given the lack of a universally accepted definition, we strongly recommend that research methods clearly state whether death is considered an event or not, and how the cause of death is determined.

Providing this information allows readers to better understand the limitations and the direction and magnitude of potential associated biases.[8,9] Progression free survival is more frequently used than TTR and is more 'unfavorable' to the drug being studied: if a patient does not have a recurrence and dies, we assume that the patient dies because of the disease. In TTR, if the same patient dies, it would not be considered for the number of events (it would be censored), even if the death is most likely to be due to recurrence.

In Fig. 3 both patients would be included to assess OS (patient 1 at point B); both patients would be included to assess PFS (patient 1 at point A); and only patient 1 at point A would be included to assess TTR (patient 2 would be censored at time of death).

Hazard ratio

The hazard ratio (HR) refers to the ratio between the proportion of patients with the event with the condition/treatment under study and the proportion of patients with the event in the reference condition for a given period.

As an example, HR = 2 for lymph node metastasis (N1) to recurrence. This means that, for any given interval of time, the probability of having recurrence is twice as high for the N1 patients compared to N0 patients.

HR = (N1 with recurrence/N1 with no recurrence)/(N0 with recurrence/N0 with no recurrence) = 2.

Another example of the application of this concept is illustrated by Zhang *et al*: "Follow-up analyses revealed that increased serum MMP-7 levels were linked with a greater risk of poor renal outcome with a hazard ratio of 1.898 per doubling MMP-7 concentration".[2] Here, the reader would conclude that the proportion of patients with poor renal outcome is 1.898 times the proportion of patients with poor renal outcome with half of the MMP-7 concentration.

CONCLUSION

Clinicians should understand statistics well enough to conduct and evaluate studies which provide evidence-based data for clinical practice. Being familiar with these concepts is essential to critically assess and interpret published data.

This is an approach specifically designed to elucidate the basic principles in the analysis, interpretation, design, and execution of the most widespread and standard method in survival analysis, according to the typical needs of clinical practice. It does not explain alternative and more accurate methods adapted to specific scenarios. Also, it does not address multivariable regression analysis.

AUTHOR CONTRIBUTIONS

APG, RM: Literature review, writing of the manuscript.
BC: Writing and critical review of the manuscript.
VN, CC: Critical review of the manuscript.

PROTECTION OF HUMANS AND ANIMALS

The authors declare that the procedures were followed according to the regulations established by the Clinical Research and Ethics Committee and to the Helsinki Declaration of the World Medical Association updated in 2013.

DATA CONFIDENTIALITY

The authors declare having followed the protocols in use at their working center regarding patients' data publication.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

FUNDING SOURCES

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. De La Mata J, Blanco FJ, Gomez-Reino JJ. Survival analysis of disease modifying antirheumatic drugs in Spanish rheumatoid arthritis patients. Ann Rheum Dis. 1995;54:881-5.
2. Zhang J, Ren P, Wang Y, Feng S, Wang C, Shen X, et al. Serum matrix metalloproteinase-7 level is associated with fibrosis and renal survival in patients with IgA nephropathy. Kidney Blood Press Res. 2017;42:541-52.
3. Conden E, Rosenblad A, Wagner P, Leppert J, Ekselius L, Aslund C. Is type D personality an independent risk factor for recurrent myocardial infarction or all-cause mortality in post-acute myocardial infarction patients? Eur J Prev Cardiol. 2017;24:522-33.
4. Turkson AJ, Ayiah-Mensah F, Nimoh V. Handling censoring and censored data in survival analysis: a standalone systematic literature review. Int J Math Math Sci. 2021;2021:1-16.
5. Lee EC, Kim SH, Park H, Lee SD, Lee SA, Park SJ. Survival analysis after liver resection for hepatocellular carcinoma: a consecutive cohort of 1002 patients. J Gastro Hepatol. 2017;32:1055-63.
6. Zwiener I, Blettner M, Hommel G. Survival analysis: part 15 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2011;108:163-9.
7. Ryu T, Takami Y, Wada Y, Tateishi M, Matsushima H, Yoshitomi M, et al. Oncological outcomes after hepatic resection and/or surgical microwave ablation for liver metastasis from gastric cancer. Asian J Surg. 2019;42:100-5.
8. National Cancer Institute. NCI Dictionary of Cancer Terms. [cited 2022 Jun 28]. Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/progression-free-survival.
9. Yan WT, Quan B, Xing H, Wu MC, Yang T. Time to recurrence, but not recurrence-free survival, should be the endpoint used to predict early recurrence after HCC resection. J Hepatol. 2019;70:570-1.