Neuro-Oncology Practice
4(2), 77–81, 2017 | doi:10.1093/nop/npw012 | Advance Access date 9 December 2016

Essence of survival analysis†

Stephanie L. Pugh

Department of Statistics, American College of Radiology, 1818 Market Street, Suite 1720, Philadelphia, PA 19103, USA (S.L.P.)

Corresponding Author: Stephanie L. Pugh, PhD, 1818 Market Street, Suite 1720, Philadelphia, PA 19103. 215-717-0850 (spugh@acr.org).

†This paper is part of the statistics series.

Abstract

Many clinical trials are designed based on a time-to-event endpoint. Overall survival and progression-free survival are commonly used, especially in Phase II and III clinical trials. Overall survival measures the time to death from any cause, while progression-free survival measures the time to progression of the disease or death from any cause. The key distinguishing factor is that the event of interest, such as death, may not occur in all individuals, making their time to this event unknown. Survival analysis comprises of the methods used to estimate the rates associated with time-to-an-event data, compare the rates between groups, and assess how other factors impact these rates.

Key words

log-rank test | survival analysis | time to event

Many clinical trials are designed based on a time-to-event endpoint. Overall survival and progression-free survival are commonly used, especially in Phase 2 and 3 clinical trials. Overall survival measures the time to death from any cause while progression-free survival measures the time to progression of the disease or death from any cause. The key distinguishing factor is that the event of interest, such as death, may not occur in all individuals, making their time to this event unknown. Survival analysis comprises of the methods used to estimate the rates associated with time-to-an-event data, compare the rates between groups, and assess how other factors impact these rates.

A patient diagnosed with a grade 2 glioma has just been referred to you for treatment. After discussing her treatment, her first question is how much time she has left. Recalling the results of Buckner et al, the median survival time for this patient population receiving radiotherapy (RT) followed by 6 cycles of procarbazine, lomustine, and vincristine (PCV) is 13.3 years, which was shown to be significantly longer than RT alone (7.8 years).[1]The median time to progression is 10.4 years for patients receiving RT + PCV compared with 4.0 years for patients receiving RT alone.[1] You inform her that with this particular treatment, approximately 50% of patients progress before 10.5 years and 50% survive just over 13 years when treated with RT + PCV.

Median survival time (MST) is a commonly used statistic in survival analysis, which is an analytic approach when time to an event is of interest, noting that not all patients may experience the event. For example, time to death is commonly used in cancer research but is usually called survival time. Not all patients may die during the study so their survival time is unknown. Another example is progression-free survival, which is the time to disease progression or death, whichever occurs first. If a patient does not progress or die during the study, then this patient's progression-free-survival time is unknown. Since the time to an event could be unknown and tends to have a skewed distribution, the median survival time, rather than the mean, is used.

Censoring

How are the patients with unknown survival time handled in the analysis? Censoring occurs when some information on a patient's survival time exists, but the exact time is unknown since the event has not occurred within the time frame under

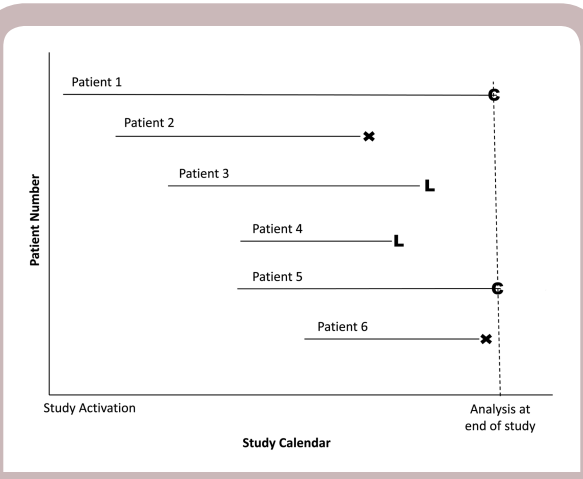

Fig. 1 Types of censoring: Patients were enrolled in a study at different calendar times and followed until the end of the study when the analysis took place. Patients 1 and 5 were right censored at the end of the study since an event did not occur. Patients 2 and 6 experienced an event on study. Patients 3 and 4 were lost to follow-up and left censored at the last time of follow-up. C = censored, L = lost to follow-up, X = event.

investigation.[2] Using time to death in a clinical trial as an example, a patient who is alive at the end of the trial or time of analysis would be censored at that time. The patient's information, that he or she was alive at the end of the study, is used in the analysis even though an event did not occur. If a patient is lost to follow-up, the information until the last date the patient was seen can still be used in the analysis and the patient would be censored at the last date seen. Censoring can also occur if the patient experiences a different event that makes it impossible to experience the event of interest.[3] These examples of censoring typically occur prior to the end of the time frame under investigation and are thus known as left censoring. Another type of censoring that is quite common is right censoring, which occurs when the patient does experience the event of interest, but this occurs after the end of the time frame under investigation. For both types of censoring, the observed survival time is shorter than the actual survival time, which is unknown. The observed survival time is then used to draw inferences about the actual survival time.[2] Fig. 1 depicts examples of these types of censoring.

Estimation of Survival Curves

Survival data are modeled in terms of two functions, survival and hazard.[3] The survival function models the probability that an individual survives past a specified time. The survival function is nonincreasing and will eventually go to 0 as all patients eventually die. The hazard function provides the instantaneous event rate, the rate measured at that instant, at a specified time given that an individual has survived up to that time.[3] In other words, the survival function focuses on not failing and is cumulative, while the hazard function focuses on failing and is tied to a time point.

Survival curves are typically plotted to allow visualization of survival across time. The Kaplan-Meier approach is commonly used.[4] This is a nonparametric approach, meaning it does not depend on a probability distribution such as the exponential function. It is assumed that events occur independently of each other. The times associated with each event are ordered from the smallest to the largest; if multiple patients have the same event time, that time is only listed once. Each event time creates a new time interval. Within each interval, the number of patients with an event is calculated along with the number at risk. If a patient is censored, that patient is removed from the number at risk in the subsequent time interval. Thus the number at risk can decrease even when no events have occurred. A Kaplan-Meier plot will then look like a step function (Fig. 2). The more events there are, the smoother the lines will look. The MST can be found from a Kaplan-Meier plot by finding the point on the line that corresponds to 50% survival (on the y axis). The corresponding time (on the x axis) is the MST. Looking at Fig. 2, the MST for the RT-alone arm is about 8 years, while the MST for the RT + PCV arm occurs after 12 years, which is as far as the figure goes.

Testing Differences in Survival Distributions

After plotting survival curves for two different treatments, for example, determining whether they are different is of interest. The log-rank test is the most commonly used method to determine statistically if survival curves are the same or not. The null hypothesis is that the survival curves are the same and can be extended to more than two curves.[5,6] Specifically, it calculates the number of expected events since the previous event, assuming there is no difference between the groups, at each time in each group. The observed number of events for each group are compared with the expected using a test statistic that is then compared with the chi-square distribution to determine the *P* value. A stratified log-rank test will test the difference between groups created by multiple categorical factors, such as treatment (RT alone vs RT + PCV) and age (<40 vs ≥40 years old), by grouping patients based on these variables. In this situation, the null hypothesis being tested is that the survival distributions for all of the groups are the same. If it is rejected, it does not specify which groups are different.

Recall the example of a patient with a grade 2 glioma. In Fig. 2, the survival curves for each treatment arm cross early and then a large separation of the curves occurs. Is this difference between treatment arms statistically significant? The answer to this question is yes because the *P* value from the log-rank test is .001, which is less than the prespecified .05 significance.[1] Therefore, we can conclude that the treatment arms had different survival distributions. Examining the figure shows us that the RT + PCV arm is better since the RT-alone arm declines faster.

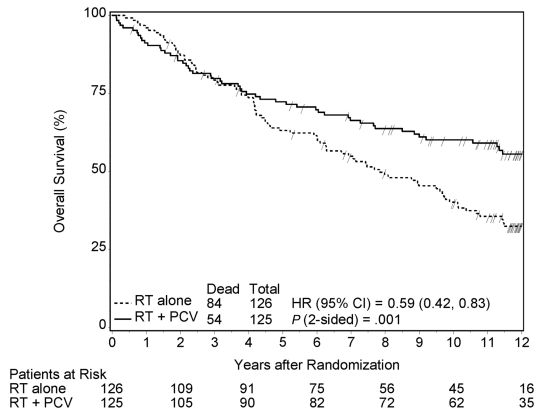

Fig. 2 Overall survival results using the Kaplan-Meier method from NRG Oncology's RTOG 9802 trial, a randomized phase 3 trial of radiation therapy with or without PCV chemotherapy in unfavorable low-grade glioma.[1] The hash marks on the survival curves indicate when a patient was censored.

There are other approaches to testing the differences between survival curves. Briefly, one example is the class of tests developed by Harrington and Fleming.[7] This class of tests uses weights on each death to emphasize differences at different time points. The log-rank test, contained within this class of tests, assigns equal weight to all events. The Wilcoxon test, on the other hand, assigns more weight to early events when the number of patients at risk is larger and is therefore more sensitive to early differences.[2,8] The paper by Buckner et al focuses on the long-term follow-up of patients on NRG Oncology's trial RTOG 9802, a randomized phase 3 trial of RT with or without PCV chemotherapy in unfavorable low-grade glioma.[1] The initial reporting of this study by Shaw et al reported the Wilcoxon test.[9] However, in the long-term follow-up analysis, early differences were not the focus and thus the log-rank test was used to test the survival distribution differences between treatment arms.[1] The Tarone-Ware test also applies more weight to early event times and is suitable to be used in the case of crossed curves as seen in Fig. 2.[10]

When conducting a survival analysis, it is crucial to analyze the data after all patients have been followed for a sufficient amount of time. For example, if the MST of the patient population is 13.3 years, as was the case in the example used, only following patients for 5 years will not provide enough events to yield sufficient statistical power when testing differences between treatments.[1,3] However, in a trial for patients with newly diagnosed glioblastoma, where the MST is only about 16 months, 5 years of follow-up would be more than sufficient.[11]

Modeling Survival Distributions

The log-rank test applies only to groups; meaning that the effect of a continuous variable on a time-to-event outcome cannot be tested using the log-rank unless the continuous variable was categorized into groups. This can create a loss of statistical power and sensitivity when conducting the test. Since the log-rank test functions by separating the cohort into groups based on the categorical predictors, the sample size within each group becomes much smaller, thus decreasing the statistical power of the test. The log-rank test also gives the significance of the difference, but not the magnitude of the difference. Is there a way to test the difference between groups without these limitations?

Consider the example of the patient with a grade 2 oligodendroglioma. Does histology, as well as the type of treatment, impact the survival distribution? Buckner et al reported that this particular histology is a favorable prognostic variable indicating that patients with an oligodendroglioma have significantly improved survival when compared with patients with oligoastrocytoma or astrocytoma, while adjusting for treatment, the patient's age, and whether or not the patient has a specific mutation.[1] This analysis was performed using a statistical regression model, the Cox proportional hazards model, which examines the association between survival time and one or more variables and provides an estimate of the strength of effect for each variable.[12,13] Although there are other types of statistical models for survival data, such as the accelerated failure time model, the Cox proportional hazards model is the most common and will be described here.

The hazard ratio, the ratio of the hazard rates from two groups, is an important measure in survival analysis and an integral component of Cox proportional hazards models. If there is no difference in survival, the hazard ratio is equal to 1. If the ratio is less than 1 then the effect is considered protective and if it is greater than 1, the effect is considered a risk factor. As an example, the hazard ratio for the survival curves presented in Fig. 2 is 0.59 and the 95% confidence interval is 0.42–0.83. In this example, since there are only 2 groups, the log-rank test is testing whether this ratio is equal to 1. The confidence interval does not include 1 which indicates, along with the log-rank test, that the RT + PCV arm is superior.

The Cox proportional hazards model provides the hazard ratio and a statistical test and corresponding *P* value for each variable included in the model. The time to event is the hazard function which is dependent on a set of variables called covariates. The covariates in the model can be categorical or continuous, one advantage over the log-rank test. Using the example above, the covariates would be treatment arm (RT alone vs RT + PCV), histology (oligodendroglioma vs oligoastrocytoma vs astrocytoma), age (categorized as < 40 years old vs ≥ 40 years old), and presence vs absence of a mutation. Since histology has more than two levels, in order to calculate a hazard ratio it must be split into two variables each with two levels: oligodendroglioma vs oligoastrocytoma and oligodendroglioma vs astrocytoma. The level that appears in both variables is called the reference level. As seen in Table 1, oligodendroglioma tumors have lower hazard ratios than oligoastrocytoma and astrocytoma tumors, meaning they are associated with a longer survival time.

A basic premise of the Cox proportional hazards model appears in its name: the hazards are assumed to be proportional. There are various ways to check the proportional hazards assumption.[14] A proportional hazard occurs

Table 1 Cox proportional hazards model from NRG Oncology's RTOG 9802 trial, a randomized phase 3 trial of radiation therapy with or without PCV chemotherapy in unfavorable low-grade glioma[1]

Variable (**Bolded** value has **favorable** outcome)	*P* value	Hazard Ratio (95% CI)
Assigned treatment: first 1-year follow-up[a] (RT + PCV vs **RT alone**)	.839	1.15 (0.30-4.33)
>1-year follow-up[a] (**RT + PCV** vs RT alone)	.001	0.35 (0.19-0.66)
IDH1-R132H Mutation (Absent vs **Present**)	.124	0.66 (0.39-1.12)
Histology (Astrocytoma vs **Oligdendroglioma**)	.012	0.38 (0.18-0.81)
(Oligoastrocytoma vs **Oligdendroglioma**)	.001	0.35 (0.19-0.66)
Age (**<40** vs >=40)	.014	0.50 (0.29-0.87)

From *New England Journal of Medicine*, Buckner J, Shaw EG, Pugh SL, et al., Radiation plus Procarbazine, CCNU, and Vincristine in Low-Grade Glioma, Volume No. 374, Page No. S3. Copyright © (2016) Massachusetts Medical Society. Reprinted with permission.
[a]1 year was the optimal survival time which yielded the largest log partial likelihood.
RT, radiation therapy; PCV, procarbazine, lomustine, and vincristine.

when the difference between groups in the hazard rate is the same across time; in other words, the hazard of one group is simply the hazard of the other group multiplied by some constant. Visually, one can see that the data portrayed in Fig. 2 do not have proportional hazards since the survival curves for treatment arm cross at around 3 years. Other models, such as an accelerated failure time model, could be used and are described elsewhere, or the Cox model could be modified to correct the nonproportional hazards.[2,11,15–17] A stratified Cox model, where subjects are divided into strata with distinct baseline hazard functions, is one example of a modification.[2] Another is the extended Cox model in which an interaction or time-dependent covariate is added to the model. When conducting the Cox model for the data in Fig. 2, treatment arm was divided into two variables based on time (Table 1). This extension allowed the use of the Cox model in the setting of nonproportional hazards. The Tarone-Ware test, mentioned in the previous section, is also a more powerful test than the log-rank test in the case of nonproportional hazards.

The assumptions associated with the Cox proportional hazards model should always be confirmed when using it in an analysis. The proportional hazard assumption can be checked visually, using a goodness of fit test or implementing an extended Cox model with time-dependent variables. More information on these approaches can be found elsewhere.[2] The functional form of the covariates, usually continuous, in the model must be assessed, as a nonlinear covariate may cause the hazards to appear nonproportional. A linear relationship between an outcome variable and a covariate can be visualized with a straight diagonal line.[18] As is the case with any regression model, idenitifying outliers, specifically ones that disproportionately influence analysis, is also critical when using Cox proportional hazards models.

Discussion

Survival analysis is a widely used and well-studied method of data analysis in statistics. It allows for calculation of both the failure and survival rates in the presence of censoring. The Kaplan-Meier method is commonly used to estimate the survival and hazard functions and depict these functions in a graphical form. The log-rank test is typically used to test the difference between survival distributions in at least two groups. In the case of two groups, a hazard ratio is calculated that describes the hazard rate in one group as compared with the other. In the case where there are multiple covariates that may impact the occurrence of the event, a Cox proportional hazards model can be used to assess these associations simultaneously.

Survival analysis has been extended to other types of data, such as longitudinal. In that particular case, the event may occur when a certain score or level is met. Brown et al conducted an analysis based on time to neurocognitive failure, where a score below a certain threshold on at least one test in a battery of neurocognitive tests, indicates failure.[19] Chinot et al assessed deterioration-free survival, which is a composite endpoint of time to a ≥10-point decline from baseline without a subsequent ≥10-point improvement from baseline in selected quality-of-life and brain-tumor-specific questionnaires or death, whichever occurred first.[20]

Another extension is competing risks, which occurs when a patient cannot experience the event of interest because another event has occurred. The analysis performed by Brown et al took the competing risk of death into account since a person cannot decline cognitively postmortem.[19] Therefore, there were three types of patient status: event (neurocognitive failure), censored, competing risk (death). The methods of analysis are different in the presence of a competing risk than those described here.[21–24]

The type of data and the hypothesis of interest will direct the type of analysis and choice of statistical tests. The tests described in this paper are robust, allowing them to be commonly applied. There are modifications of these tests that allow them to be applied even when certain assumptions are not met. Even in the presence of nonproportional hazards, the Cox model can be extended to allow its use. Most statistical software packages have these tests and models as built-in functions and also allow for extensions into competing risks analysis.

Funding

None.

Conflict of interest statement. None declared.

References

1. Buckner J, Shaw EG, Pugh SL et al. Radiation plus procarbazine, CCNU, and vincristine in low-grade glioma. *N Engl J Med*. 2016;374(14):1344–1135.
2. Kleinbaum DG, Klein M. *Survival analysis: A self-learning text*. 2nd ed. New York, NY: Springer; 2005.
3. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. *Br J Cancer*. 2003;89(2):232–238.
4. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457–481.
5. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*. 1966;50(3):163–170.
6. Peto R, Pike MC, Armitage P et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *Br J Cancer*. 1977;35(1):1–39.
7. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika*. 1982;69(3):553–566.
8. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *J R Stat Soc Ser A*. 1972;135(2):185–207.
9. Shaw EG, Wang M, Coons SW et al. Randomized trial of radiation therapy plus procarbazine, lomustine, and vincristine chemotherapy for supratentorial adult low-grade glioma: Initial results of RTOG 9802. *J Clin Oncol*. 2012;30(25):3065–3070.
10. Tarone RE, Ware JH. On distribution-free tests for equality for survival distributions. *Biometrika*. 1977;64(1):156–160.
11. Gilbert MR, Dignam JJ, Armstrong TS et al. A randomized trial of bevacizumab for newly diagnosed glioblastoma. *N Engl J Med*. 2014;370(8):699–708.
12. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol*. 1972;34(2):187–220.
13. 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):431–436.
14. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *J R Stat Soc Series B*. 2000;62(4):711–730.
15. Ata N, Sozer MT. Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe J Math Stat*. 2007;36(2):157–167.
16. Klein JP, Moeschberger ML. *Survival analysis: Techniques for censored and truncated data*. 2nd ed. New York, NY: Springer; 2003.
17. Therneau TM, Grambsch PM. *Modeling survival data: Extending the Cox model*. 1st ed. New York, NY: Springer; 2000.
18. Gerds TA, Schumacher M. On functional misspecification of covariates in the Cox regression model. *Biometrika*. 2001;88(2):572–580.
19. Brown PD, Pugh S, Laack NN et al. Memantine for the prevention of cognitive dysfunction in patients receiving whole-brain radiotherapy: a randomized, double-blind, placebo-controlled trial. *Neuro Oncol*. 2013;15(10):1429–1437.
20. Chinot OL, Wick W, Mason W et al. Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. *N Engl J Med*. 2014;370(8):709–722.
21. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. *Br J Cancer*. 2004;91(7):1229–1235.
22. Gray RJ. A class of K-sample test for comparing the cumulative incidence of a competing risk. *Ann Stat*. 1988;16(3):1141–1154.
23. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496–509.
24. Freidlin B, Korn EL. Testing treatment effects in the presence of competing risks. *Stat Med*. 2005;24(11):1703–1712.