Survival analysis, more than meets the eye

Marko Lucijanic*[1], Marko Skelin[2], Tomo Lucijanic[3]

[1]Hematology Department, University Hospital Dubrava, Zagreb, Croatia
[2]Pharmacy Department, General Hospital Šibenik, Šibenik, Croatia
[3]Endocrinology, diabetes and metabolism disorders Department, University Hospital Dubrava, Zagreb, Croatia

*Corresponding author: markolucijanic@yahoo.com

Abstract

The log-rank test is a cornerstone of phase III oncology clinical trials. However, there are at least three different mathematical procedures that can be named the log-rank test and two of them are widely used by commercial statistical programs. Consequently, different P values can be obtained. In the case of a borderline statistical significance, this can mean the difference between the evidence (significant P value) and merely an observation. Since all three methods can be reported under the same name, space for possible data manipulation occurs. This should be of a particular concern in a drug regulatory context. Randomized clinical trials with borderline significant results should perhaps be required to report P values calculated by all three methods, in order to properly evaluate drug efficacy. An interactive MS Excel spreadsheet that uses all three logrank test variants is prepared as a supplementary file accompanying this article. Association of high grade of bone marrow fibrosis with poor outcome in patients with myelofibrosis is used as an example.

Key words: statistics; clinical trial; survival analysis; primary myelofibrosis; software

Received: October 10, 2016 *Accepted: January 08, 2017*

Introduction

Survival analysis based on the method by Kaplan and Meier is a cornerstone of phase III oncology clinical trials (1). The log-rank test is a statistical test of choice to compare survival/time to event of interest between two or more groups of patients. However, one name (the log-rank test) can be used for three related but different mathematical procedures. Two of them are widely employed inside different commercial statistical programs. "Behind the scenes" mathematics are not the same and thus different results can be obtained. In the case of a borderline statistical significance, this can mean the difference between the evidence (significant P value) and merely an observation. In other words, two persons analysing the same data set with two different statistical programs can "unknowingly" reach a different conclusion. Since all three methods can be reported under the same name, space for possible data manipulation occurs.

The log-rank test variants

Mathematical overview of all three methods is shown in Table 1 and the supplementary file. The first method that was proposed by Mantel in 1966 represents an extension of the Mantel-Haenszel procedure for comparing 2 x 2 tables (2). Most commercial statistical programs provide it under the name of the log-rank test (e.g. STATA, StataCorp v. 14 and MedCalc, MedCalc Software v 16). The second method that was developed by Peto and Peto in 1972 uses alternative computational approach to produce the same test statistic but different variance (3). It is computationally simpler and therefore easier to calculate by hand/table calculator. Although developed later, it was originally named the log-rank test by the authors and the name was thereafter generalized for both procedures. This method is provided by e.g. Statistica StatSoft v. 13 under the name of the log-rank test. It should be noted that this program also provides the first method proposed by Mantel, but under a

Table 1. Mathematical overview of three different log-rank test variants.

Key steps in data analysis	The Cox-Mantel test	The simple χ^2 log-rank test	The Peto log-rank test
Step 1 (Data sorting)*	Data are sorted in a time ascending order. At the time of each death, a new interval is created.		
Step 2 (Preliminary calculations)†	‡$O_{Group1} = \sum_{j=1}^{I} O_{jGroup1}$		‖$\Lambda_I = \sum_{j=1}^{I} O_j / R_j$
	‡$E_{Group1} = \sum_{j=1}^{I} O_j \times R_{jGroup1} / R_j$		‖$W_i = \begin{cases} 1 - \Lambda_{Ii} & \text{if death} \\ -\Lambda_{Ii} & \text{if censoring} \end{cases}$
	§$T_{Mantel} = \sum_{j=1}^{I} (O_{jGroup1} - O_j \times R_{jGroup1} / R_j)$		§$T_{Peto} = \sum_{i=1}^{N\,group1} W_i$
	$T_{Mantel} = O_{Group1} - E_{Group1}$		$T_{Peto} = O_{Group1} - E_{Group1}$
	§$V_{Mantel} = \sum_{j=1}^{I} \frac{R_{jGroup1} \times R_{jGroup2} \times O_j \times (R_j - O_j)}{R_j^2 \times (R_j - 1)}$		§$V_{Peto} = \frac{N_{Group1} \times N_{Group2} \times \sum_{i=1}^{N} W_i^2}{N \times (N - 1)}$
Step 3 (Calculation of χ^2 value)	$\chi^2_{Mantel} = \frac{T_{Mantel}^2}{V_{Mantel}}$	$\chi^2 = \frac{(O_{Group1} - E_{Group1})^2}{E_{Group1}} + \frac{(O_{Group2} - E_{Group2})^2}{E_{Group2}}$	$\chi^2_{Peto} = \frac{T_{Peto}^2}{V_{Peto}}$
Step 4 (P value)	P values that correspond to calculated χ^2 values are found using one degree of freedom χ^2 distribution table.		

*Step 1 (Sorting data) is same for all three methods. Intervals are necessary if we want to obtain correct calculations when more than one death occurs at the same time (tied observations). All concurrent deaths are considered to happen in the same interval. Central calculations for all three methods are interval specific (i.e. occur at death times).
†Step 2 (Preliminary calculations) is necessary for later calculations of χ^2 value. The Cox-Mantel test and the simple χ^2 test share calculation of observed and expected number of deaths.
‡O, E, R and I represent observed number of deaths, expected number of deaths, number of patients at risk and number of intervals, respectively. O_j, E_j and R_j (j=1,..,I) represent aforementioned parameters at the time of the j-ordered interval.
§T and V represent test statistic and variance for particular test variant, respectively. The Cox-Mantel test and the Peto log-rank test produce the same test statistic. Calculations are done in one of the groups only. Variance for both methods is calculated on a whole data-set.
‖N, Λ_I and W represent number of observations, the Nelson-Aalen estimator for a particular interval and a specific Peto log-rank test score, respectively. N_i, Λ_{Ii} and W_i (i=1,..,N) represent aforementioned parameters at the time of the i-ordered observation.

different name (the Cox-Mantel test). The third method is based on the simple χ^2 (chi squared) principle of analysing observed and expected number of events. This method is rarely used by commercial statistical programs but deserves to be mentioned because it is widely accepted as an explanation to the logic behind the test (4). We refer to particular methods throughout our manuscript and supplementary file as the Cox-Mantel test, the Peto log-rank test and the simple χ^2 log-rank test, respectively. All three methods produce a one degree of freedom χ^2 statistic that is used to obtain the corresponding P value. These tests should not be confused with weighted two-sample tests for survival data (Gehan generalization of the Wilcoxon test, Peto and Peto generalization of the Wilcoxon test, the Tarone-Ware test, the Fleming-Harrington test, etc.) (5).

It is hard to recommend which method should be favoured over the other. Variance of the Peto log-rank test is calculated under the assumption of equal censoring and other log-rank tests might

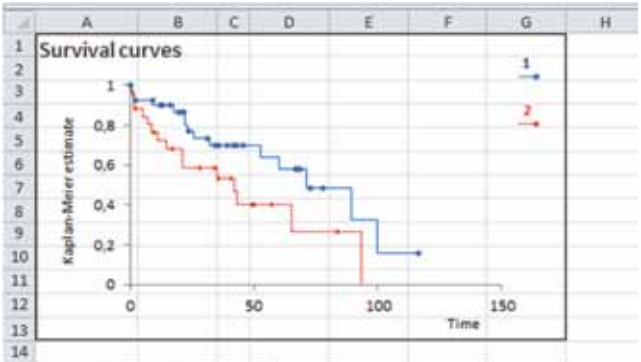
perform better if censoring does not occur at random with respect to group membership (e.g. if withdrawals due to side-effects occur mainly in one treatment group) (3). However, it is unclear how important in practice unequal censoring is. On the other hand, it was suggested that the Cox-Mantel test tends to underestimate true variance (and therefore produce unrealistically lower P value in comparison to the Peto log-rank test) when the test statistic is large in absolute value (6). As we have observed in multiple data sets, these two tests exchange in providing more significant P value in different clinical situations. It should be noted that if the assumption of proportional hazards is violated (e.g. survival curves cross) neither of the log-rank test methods should be used. Alternative statistical methods were developed for such situations (7).

All three log-rank test variants are considered to be the log-rank test and are named as such on different occasions. Actually, medical researchers are mostly unaware of the method used and currently, there is no discrimination between the log-rank test variants in most of published medical literature. Some statistical programs do not clearly report their method of choice either, and sometimes it is almost impossible to know how the P value was obtained unless data are recalculated in a known manner. Therefore, an interactive MS Excel spreadsheet that uses all three methods is prepared as a supplementary file accompanying this article. Users are encouraged to experiment with the provided data set or test their own, and become more acquainted with the problem. Spreadsheet can analyse up to 200 entries that can be copy-pasted inside corresponding columns and can serve as a standalone statistical program. It should be noted that it is unethical to "fish" for significant P value and to report only one most significant result. Such "P value hacking" is strongly discouraged by the authors.

Application of three methods to example data set

Primary myelofibrosis (PMF) is a Philadelphia chromosome negative chronic myeloproliferative neoplasm (Ph- MPN) originating from transformed hematopoietic stem cell (8). Secondary myelofibrosis (SMF) can develop from PMF biologically related Ph- MPNs and it clinically resembles PMF. Typical feature of these diseases is scarring of the bone marrow (i.e. myelofibrosis) that can be graded according to the current European consensus (9).

In our example, we have evaluated impact of highly advanced (grade 3) bone marrow fibrosis present at the time of diagnosis on overall survival in a cohort of 67 patients with PMF and SMF. Data were acquired in a retrospective manner and represent single centre experience. One might have the feeling that there is a real effect in place by observing the Kaplan-Meier curves (Figure 1). But is there statistical evidence to support it? Stated in other words, can inferences about population be made from these results (based on a sample)? We performed necessary calculations for all three log-rank test variants. Calculations for first ten observations are shown in Table 2. Step by step procedure for each approach is shown in the supplementary file. After obtaining corresponding P values, we encounter a controversial situation. When



Log-rank test variants:	χ^2 value	P value
Cox-Mantel	3,96	0,04664
Peto log-rank	3,78	0,05185
simple χ^2 log-rank	3,84	0,04993

Summary:	Group 1	Group 2
Number of cases	41	26
Observed events	15	15
Expected events	20,06	9,94

Figure 1. Survival curves of myelofibrosis patients with grade < 3 (blue line marked as 1) and grade 3 (dashed red line marked as 2) bone marrow fibrosis and P values obtained with three different long-rank tests.

Table 2. Parameters needed for log rank test calculations obtained in an example data set of myelofibrosis patients

N	Time (months)	Status	Group	I	O	R	R_{Gr1}	R_{Gr2}	E_{Gr1}	E_{Gr2}	V_{Mantel}*	O/R	Λ†	W‡	W²
1	0.10	0	1	1	0	67	41	26					0.00	0.00	0.00
2	0.57	1	1	2	1	66	40	26	0.61	0.39	0.24	0.02	0.02	0.98	0.97
3	0.60	1	2	3	1	65	39	26	0.60	0.40	0.24	0.02	0.03	0.97	0.94
4	1.17	1	1	4	1	64	39	25	0.61	0.39	0.24	0.02	0.05	0.95	0.91
5	1.47	1	1	5	1	63	38	25	0.60	0.40	0.24	0.02	0.06	0.94	0.88
6	1.57	1	2	6	1	62	37	25	0.60	0.40	0.24	0.02	0.08	0.92	0.85
7	1.70	1	2	7	1	61	37	24	0.61	0.39	0.24	0.02	0.09	0.91	0.82
8	2.17	0	1	7	1	61	37	24					0.09	0.09	0.01
9	2.23	0	2	7	1	61	37	24					0.09	0.09	0.01
10	4.90	1	2	8	1	58	36	22	0.62	0.38	0.24	0.02	0.11	0.89	0.79

The first ten observations are shown. For specific calculations please see Table 1 and the supplementary file. N – number of observation. Time - duration of follow-up, recorded in months in our example. Status - censoring variable, 1 for death and 0 for alive or lost to follow-up. Group - 1 for grade < 3 myelofibrosis and 2 for grade 3 myelofibrosis patients. I – number of interval. O – observed number of deaths per interval. R – overall number at risk per interval. R_{Gr1} and R_{Gr2} – overall number at risk in a specific group per interval. E_{Gr1} and E_{Gr2} – expected number of deaths in a specific group per interval. * Variance of the log-rank test calculated by Mantel method. †Λ – the Nelson-Aalen estimator. ‡W – score for the Peto log-rank test.

P values are reported to three decimal places, the Cox-Mantel test suggests that the result is significant (P = 0.047), the Peto log-rank test suggest that the result is insignificant (P = 0.052), and the simple χ^2 test suggests that the result is of borderline statistical significance (P = 0.050). None of the methods used is currently considered the gold standard, and all three P values can be reported as results of the log-rank test. According to our interpretation, our result seems to be truly of borderline statistical significance as suggested by inhomogeneity of obtained P values. Significant association of higher grade of bone marrow fibrosis with inferior overall survival was previously reported in multiple cohorts of myelofibrosis patients (10-12), although this finding was not universal (13). Therefore, we conclude that our data are in line with most of previously published results and are in support of adverse prognostic significance of highly advanced bone marrow fibrosis in these patients. However, we cannot consider our result alone to represent high level of evidence due to borderline statistical significance and retrospective study design.

Are P values that important?

A recent statement by the American Statistical Association discussed that no single index should substitute for scientific reasoning, and proper inference requires full reporting and transparency; e.g. patient selection, contextual factors, number of hypotheses explored, measures of effect size, etc. (14). Although there are many valid arguments against a blind use of specific threshold P values to determine statistical significance (and we agree they should not be used in that way), P values remain an important landmark in scientific decision-making. Medical literature is overladen with borderline significant results regarding survival benefit of a new drug or a new procedure. Our example adds a new dimension to an issue of their appropriate interpretation. The statement "the log-rank test was used" is not unequivocal as it seems at first and this should be of a particular concern in a drug regulatory context. Things get especially suspicious if different statistical programs are used for survival analyses and data analysis in general.

As we previously stated, it would be unethical to "fish for significant P value" and to report only one most significant result. We would like to point out that this danger will exist in borderline significant situations until scientific and professional authorities establish the consensus about the log-rank test method of choice. Until then, there is probably no need to insist on the least significant result in analysis of retrospective data sets and researchers should be adhering to their standard practice/statistical program of choice. This is because retrospective studies are biased by numerous factors. Their results do not provide high strength of evidence and usually do not have direct effects on clinical practice. However, in a drug regulatory context, one must insist on the clear evidence of improved survival because randomized clinical trials are taken as a very high level of evidence that bears clinical-practice-related and financial implications. This would perhaps be the situation in which all three log-rank test variants should be tested, in order to properly evaluate drug efficacy. In our opinion, firm result should be significant irrespective of the method used. If the result of a randomized clinical trial is of borderline statistical significance (not consistent among three variants of the log-rank test) then it should not be taken as the clear evidence of a drug/procedure benefit. Regulatory and clinical reasoning should be based on the least significant result as the relevant one. Definite conclusion would require replicating results in new independent samples.

Acknowledgements

Data used for presentation are part of PhD thesis of the first author. Authors would like to thank Mladen Petrovecki, Rajko Kusec, Ljiljana Miletic and Ivo Veletic for support.

Potential conflict of interest

None declared.

References

1. *Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81. https://doi.org/10.1080/01621459.1958.10501452.*
2. *Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 1966;50:163-70.*
3. *Peto R, Peto J. Asymptotically Efficient Rank Invariant Test Procedures. J R Statist Soc A 1972;135:185-207. https://doi.org/10.2307/2344317.*
4. *Bland JM, Altman DG. The logrank test. BMJ 2004;328:1073. https://doi.org/10.1136/bmj.328.7447.1073.*
5. *Klein JP, Moeschberger ML, eds. Survival analysis: techniques for censored and truncated data. 2nd ed. New York: Springer; 2003. p. 536.*
6. *Brown M. On the choice of variance for the log rank test. Biometrika 1984;71:65-74. https://doi.org/10.1093/biomet/71.1.65.*
7. *Li H, Han D, Hou Y, Chen H, Chen Z. Statistical inference methods for two crossing survival curves: a comparison of methods. PloS ONE 2015;10:e0116774. https://doi.org/10.1371/journal.pone.0116774.*
8. *Vardiman JW, Thiele J, Arber DA, Brunning RD, Borowitz MJ, Porwit A, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. Blood 2009;114:937-51. https://doi.org/10.1182/blood-2009-03-209262.*
9. *Thiele J, Kvasnicka HM, Facchetti F, Franco V, van der Walt J, Orazi A. European consensus on grading bone marrow fibrosis and assessment of cellularity. Haematologica 2005;90:1128-32.*
10. *Gianelli U, Vener C, Bossi A, Cortinovis I, Iurlo A, Fracchiolla NS, et al. The European Consensus on grading of bone marrow fibrosis allows a better prognostication of patients with primary myelofibrosis. Mod Pathol 2012;25:1193-202. https://doi.org/10.1038/modpathol.2012.87.*
11. *Lekovic D, Gotic M, Perunicic-Jovanovic M, Vidovic A, Bogdanovic A, Jankovic G, et al. Contribution of comorbidities and grade of bone marrow fibrosis to the prognosis of survival in patients with primary myelofibrosis. Med Oncol 2014;31:869. https://doi.org/10.1007/s12032-014-0869-8.*
12. *Barosi G, Rosti V, Bonetti E, Campanelli R, Carolei A, Catarsi P, et al. Evidence that prefibrotic myelofibrosis is aligned along a clinical and biological continuum featuring primary myelofibrosis. PloS ONE 2012;7:e35631. https://doi.org/10.1371/journal.pone.0035631*
13. *Nazha A, Estrov Z, Cortes J, Bueso-Ramos CE, Kantarjian H, Verstovsek S. Prognostic implications and clinical characteristics associated with bone marrow fibrosis in patients with myelofibrosis. Leuk Lymphoma 2013;54:2537-9. https://doi.org/10.3109/10428194.2013.769537.*
14. *Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician 2016;70:129-33. https://doi.org/10.1080/00031305.201.1154108.*