DOES LOG-RANK TEST GIVE SATISFACTORY RESULTS?

Qamruz ZAMAN
PhD Associate Professor,
Department of Statistics, University of Peshawar, Pakistan

E-mail: ayanqamar@gmail.com

Karl Peter PFEIFFER
PhD Professor,
Department of Medical Statistics, Informatics and Health Economics,
Medical University Innsbruck, Austria

E-mail: peter.pfeiffer@i-med.ac.at

Abstract:

Comparison of effects of two treatments by log-rank test is a very common phenomenon in medical research. Researchers prefer to use log-rank test with out carrying about the assumptions of test, which sometimes not only destroy the effects of study but also misguides the readers. The idea of this article is to review some aspects of log-rank test and to provide some rules of thumb.

Key words: *Log-rank test, Proportional hazards assumption, Kaplan-Meier survival curve*

INTRODUCTION

Medical researchers are often interested in comparing the survival experience of two groups of individuals. They usually prefer to use simple and straight forward tests. For this purpose several methods are available for comparing survival distributions, out of which the most commonly used rank based test, is the log-rank test [1]. Log-rank test is the first choice of researchers due to its easy concept as well as easily availability of software. The test is based on different assumptions and performs better in a specific situation, but some researchers do not care about these. Like the Kaplan-Meier survival function [2], log-rank test is also based on the assumption of non-informative censoring. It is cited so many times in the literature that the log-rank test is more appropriate, powerful and reliable as compared to other tests in a situation where two or more survival curves do not cross i.e. whose hazard functions are proportional (Figure 1).

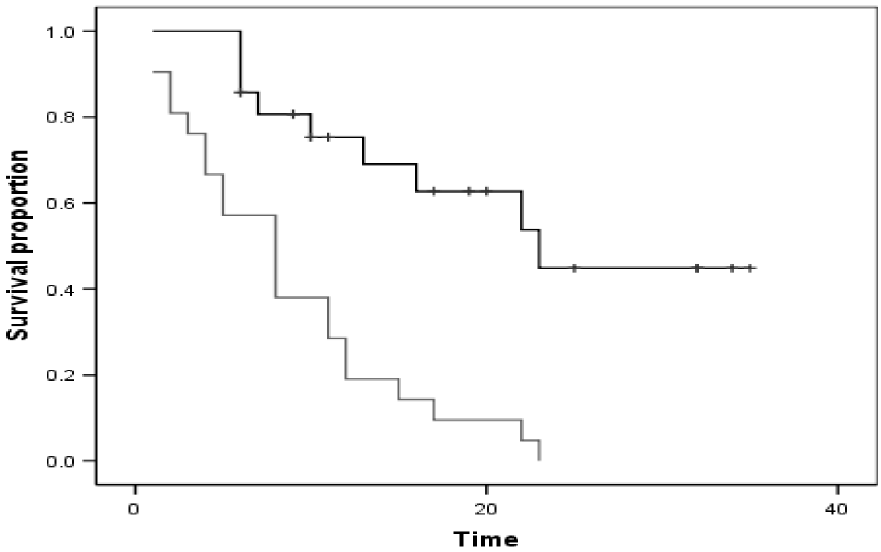

Figure 1. Survival curves of two groups

This assumption raises an important practical question that if any data set fulfills the proportional hazards assumption, log-rank test gives satisfactory results? We do not think so and we illustrate our point by considering two different published data sets which satisfied the proportional hazards assumption. Furthermore, we will try to provide some common rules of thumb on the use of log-rank test.

First data set consists of survival times of 30 cervical cancer patients, recruited to a randomized trial of the addition of a radiosensitiser to radiotherapy (Group-II) versus radiotherapy alone [3]. Group-I consists of 16 patients (5 censored and 11 events) and Group-II 14 patients (5 events and 9 censored). For further detail about the data concern the book. Table 1 summarized the data.

Table 1. Treatment group of 30 patients recruited to a cervical cancer trial.

Group-I	Group-II
1037	1476*
1429	827
680	519*
291	1100*
1577*	
90	1307
1090*	1360*
142	919*
1297	373
1113*	563*
1153	978*
150	650*
837	362
890*	383*
269	272
468*	

* Censored survival time

Survival comparison of two groups was made by using SPSS. This gave the value of log-rank test 1.682 with corresponding p-value 0.195, indicates the difference between groups is not statistically significant. This contradicts the fact which is shown in Figure 2. Except the log-rank test, we tried also different weighted tests, but every test gave the result in favour of null hypothesis that the two groups have the same survival probability. So log-rank test failed to detect difference.

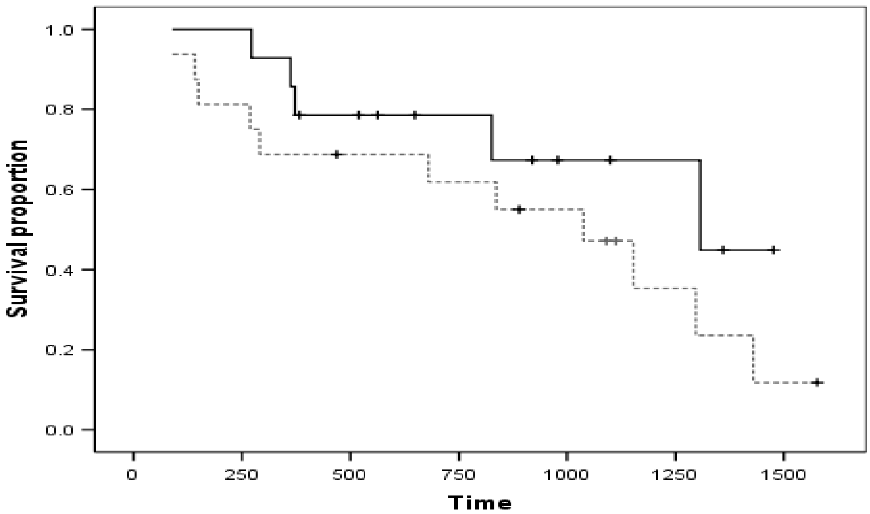

Figure 2. Kaplan-Meier survival curves of two treatments (radiosensitiser to radiotherapy denoted by dark line and radiotherapy denoted by dotted line).

For further verification, second data set is considered from Collet [4] (a brief introduction about the data set is given on page 7 of the book). The data set consists of survival times in months of women with tumours, which were classified negatively or positively stained with Helix pomatia agglutinin (HPA). There were 13 women in the negative stained group out of which 8 were censored. Positive group composed of 32 patients. Out of 32, 11 were censored. Data set is given in Table 2.The value of log-rank test along with some weighted tests values are summarized in the Table 3.

In comparison with $\alpha = 0.05$, p-value of the log-rank test does not support the fact (Figure 3), although two groups satisfied the proportional hazards assumption. While the two weighted tests, which are considered to be more appropriate for crossing curves, in this case give satisfactory results.

Table 2. Survival times in month of tumours women

Negative staining		Positive staining	
23		5	68
47		8	71
69		1	76*
70*	0		105*
71*		1	107*

JAQM

Vol. 7 No. 1 Spring 2012

100*	3		109*
101*		1	113
148	8		116*
181		2	118
198*	4		143
208*		2	154*
212*	6		162*
224*		2	188*
	6		212*
		3	217*
	1		225*
		3	
	5		
		4	
	0		
		4	
	1		
		4	
	8		
		5	
	0		
		5	
	9		
		6	
	1		

* Censored survival time

Table 3. Chi-Square statistics and p-values from the application of the log-rank, Wilcoxon and Tarone-Ware tests for tumours data

Statistical test	Chi-Square	p-value
Log-rank	3.515	0.061
Wilcoxon	4.180	0.041
Tarone-Ware	4.050	0.044

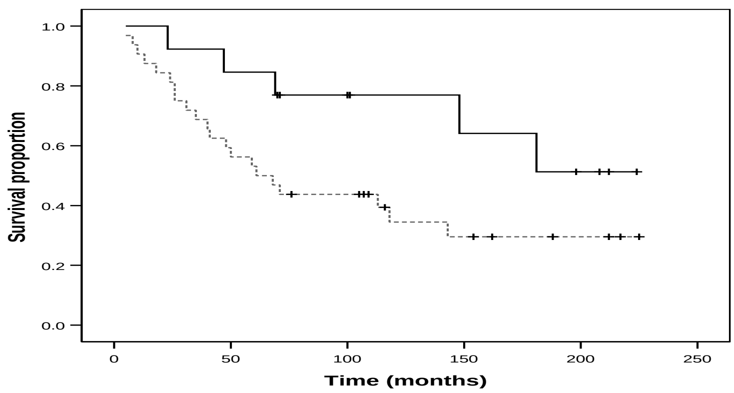

Figure 3. Kaplan-Meier survival curves of two treatments (positive stained by dark line and negative stained by dotted line).

Except these two data sets, there may be many more data sets which come across the same situation. In some cases weighted tests may be helpful but not always. It is also observed that in most cases in which log-rank test gives satisfactory results, weighted tests also and vice versa. Example is that of the famous leukaemia data set [5].

Therefore, the log-rank test which is considered to be the best choice, if the groups satisfied the proportional hazards assumption is not always true. Sometimes weighted tests are also helpful as in example 2. In some cases it may happen that no available test is able to detect the differences correctly as in example 1 and sometimes all tests give correct results.

Now the question is why the log-rank test fails in ideal situations?

The test is more suitable, if the risk of an event is considerably greater for one group [6]; although this proved in given examples still log-rank test fails. This means that the condition is not sufficient; there must be some other factors which influence the performance of log-rank test. The factor may be number of events ≤ 5, may be the range of data and may also be the difference between sizes of two groups.

We may face the same problem for crossing survival groups, on which no available weighted test fits well. This fact opens the door for new research and development. On the basis of these realities, one can not say any thing about an ideal test which is suitable in each and every situation; one test may be suitable in one situation and fails in other.

Conclusion

We conclude our discussion by mentioning the following rules

- The best way is to first check the proportionality assumption by plotting survival curves of groups or by hazard plotting of groups or by any other available method.
- It is not always true that if proportional hazards assumption satisfies, log-rank test gives a satisfactory answer.
- Do not restrict yourself to log-rank test, also apply the weighted tests. Sometimes weighted tests give more satisfactory results than log-rank test.
- If the existence tests are not able to produce appropriate results, develop a more powerful test.

References

1. Mantel N., **Evaluation of survival data and two new rank order statistics arising in its consideration**. Cancer Chemotherapy Report 1966; 50: 163-170.
2. Kaplan EL and Meier P., **Nonparametric estimation from incomplete observations. Journal of American Statistical Association** 1958; 53: 457-81.
3. Parmar MKB and Machin D., **Survival Analysis- A Practical Approach**. John Wiley & Sons Chichester 1995.
4. Collet D., **Modelling Survival Data in Medical Research**, (1st edn.) Chapman & Hall/CRC: Boca Raton 1994.

5. Freireich EJ, Gehan E, Frei III E, Schroeder LR, Wolman IJ, Anbari R, Burgert EO, Mills Sd, Pinkel D, Selwry OS, Moon JH, Gendel BR, Spurr, CL, Storrs R, Haurani F, Hoogstraten B, Lee S. **The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukaemia: a model for evaluation of other potentially useful therapy**. Blood, 1963; 21: 699-716.

6. Bland JM and Altman DG. **The log-rank test**. British Medical Journal 2004; 328: 1073.