Novel Test for Survival Data Analysis of Cancer Patients

Dmitriy Klyushin[1] and Pavel Yakovlev[2]

[1]*Taras Shevchenko National University of Kyiv, Ukraine, Akademika Glushkova Avenue 4D, Kyiv, 03680, Ukraine*
[2]*Feofaniya Clinical Hospital, Akademika Zabolotnogo 21, Kyiv, 03143, Ukraine*

Abstract

Modern medical information systems necessarily include functions for assessing the effectiveness of treatment provided to patients. As a rule, this problem is solved by calculating the survival functions for estimation of the risk of death. Traditionally, three nonparametric tests are used to analyze survival: the Cochran–Mantel–Hansel log-rank test, the Wilcoxon test for censored data, and the Tarone–Ware test. In these tests, testing statistical hypotheses about the equivalence of survival functions, as a rule, is reduced to calculating the critical value of the standard normal distribution. These tests give reliable results only if the samples are large enough and additional conditions are met. Consequently, for the development of effective medical information systems that perform survival analysis, nonparametric tests are required that use a minimum of preliminary assumptions and allow the use of small samples. The paper proposes a test for testing the hypothesis of the equivalence of the survival functions, which does not depend on the sample size and does not use additional preconditions, except for the condition of the continuity of the distribution functions of the initial data.

Keywords

Survival analysis, risk of death, Kaplan-Mayer curve, Log-rank test, Wilcoxon test, Tarone–Ware test

1. Introduction

To assess the effectiveness of the treatment provided to patients and the risk of death during a given period, many cancer healthcare facilities design information systems that analyze data and assess patient survival using the Kaplan–Meier curve [1]. Three nonparametric tests are usually used in the survival analysis based on the Kaplan–Meier estimator: the Cochran–Mantel–Hansel log-rank test [2], the Wilcoxon test [3], and the Tarone–Ware test [4]. To test statistical hypotheses about the identity of the survival functions, these tests mainly calculate the values of the standard normal distribution. However, these tests give reliable results only if the samples are large enough and additional conditions are met. The most popular is the log rank test, which gives the maximum power under the alternatives with proportional hazards [5]. However, these tests give reliable results only if the samples are large enough and additional conditions are met. For example, the Wilcoxon test is preferable when deaths at early time points have more weights [6], and the Tarone–Ware test also places more heavy weight on hazards at the early time [7].

CITRisk'2021: 2nd International Workshop on Computational & Information Technologies for Risk-Informed Systems, September 16–17, 2021, Kherson, Ukraine
EMAIL: dokmed5@gmail.com (D.Klyushin); pavel_3@hotmail.com (P.Yakovlev)
ORCID: 0000-0003-4554-1049 (D.Klyushin); 0000-0002-1767-3231 (P.Yakovlev)


The nonparametric Kaplan-Meier estimate measures the survival time of patients, i.e. the interval of time between a certain date (for example, the date of surgery) and the moment of death or censuring. It allows the construction of survival functions based on data on the life expectancy of patients and estimates the risk of death during a given time period. Similarly, it can be used to estimate the time to equipment failure or other significant event. Thus, it can be used for assessment of the risk of a specific event (death, failure, etc.) based on observations (censored and uncensored).

The aim of this paper is to describe an alternative nonparametric test that does not use any assumption excepting the most general (continuity of the distribution) and allow using small samples (size less than 50). This test use the p-statistics investigated in [8–11] and base on the $A_{(n)}$ Hill's assumption [12]. The theoretical background of the *p*-statistics is developed by Matveichuk and Petunin [8, 9] and later by Johnson and Kotz [10], and Klyushin and Petunin [11]. The high sensitivity and specificity of the nonparametric test for homogeneity of two samples based on the *p*-statistics is demonstrated in [11]. Here we propose new application of this test for comparison of two survival curves.

2. Theoretical background

Consider samples $x=(x_1,x_2,...,x_n)\in G_1$ and $y=(y_1,y_2,...,y_n)\in G_2$ from absolutely continuous distributions F_1 and F_2. The Hill's assumption $A_{(n)}$ [12] states that for exchangeable random values $x_1,x_2,...,x_n\in G$ following to an absolutely continuous distribution function

$$P\left(x\in\left(x_{(i)},x_{(j)}\right)\right)=\frac{j-i}{n+1},\ \ j<i, \tag{1}$$

where $x_{(i)}$ and $x_{(j)}$ are the *i*-th and *j*-th order statistics. Find the relative frequency h_{ij} of the event $y_m\in\left(x_{(i)},x_{(j)}\right)$ for the elements of *y* and estimate the deviation of h_{ij} from the expected probability $\frac{j-i}{n+1}$ using the Wilson confidence interval $I_{ij}^{(n)}=\left(p_{ij}^{(1)},p_{ij}^{(2)}\right)$ where

$$\begin{aligned} p_{ij}^{(1)}&=\frac{h_{ij}n+g^2/2-g\sqrt{h_{ij}(1-h_{ij})n+g^2/4}}{n+g^2},\\ p_{ij}^{(2)}&=\frac{h_{ij}n+g^2/2+g\sqrt{h_{ij}(1-h_{ij})n+g^2/4}}{n+g^2}. \end{aligned} \tag{2}$$

The significance level of this interval is the function of *g*. When *g* = 3 the significance level of $I_{ij}^{(n)}$ does not exceed 0.05 [11]. *P*-statistics, estimating the homogeneity of samples *x* and *y*, is

$$h=\#\left\{p_{ij}=\frac{j-i}{n+1}\in I_{ij}^{(n)}\right\}\Big/\left(\frac{n(n-1)}{2}\right), \tag{3}$$

It is the relative frequency of the event $\left\{ p_{ij} = \frac{j-i}{n+1} \in I_{ij}^{(n)} \right\}$. Therefore, using (2) and (3) we may construct the Wilson interval I for the *p*-statistics an formulate the following test: the null hypothesis on identity of the survival functions is accepted if the upper bound of I is greater than 0.95, else it is rejected.

For the true null hypothesis is true, the events $\left\{ p_{ij} = \frac{j-i}{n+1} \in I_{ij}^{(n)} \right\}$ form a generalized Bernoulli scheme [8, 9]. For the false null hypothesis they form a modified Bernoulli scheme. If the null hypothesis may be either true or false, they form the Matveichuk–Petunin scheme [10]. If the null hypothesis is true, $\lim\limits_{n\to\infty} \frac{j-i}{n+1} \in (0,1)$, and $\lim\limits_{n\to\infty} \frac{i}{n+1} \in (0,1)$, then the asymptotic significance level β of a sequence of confidence intervals $I_{ij}^{(n)}$ is less than 0.05 [11].

3. Experiments and results

To confirm the high sensitivity and specificity of the proposed test, we considered two groups of patients with a nondifferential diagnosis of bladder cancer of stages T2 and T3, who in 1998–2016 received special surgical care (radical and salvage cystectomy) at the urology department of the Kiev City Clinical Oncological Dispensary. For the analysis, patients were taken who had a complete history and an accurate survival result (uncensored). Characterization of the prevalence of the malignant process was carried out according to the clinical classification TNM 7th ed. (2010).

The first group (stage T2) consists of 38 patients, among them 22 patients were underwent to radical cystectomy (17 died and 5 are alive), and 16 were underwent to the salvage cystectomy (7 died and 9 are alive). The second group (stage T3) consists of 51 patients, among them 33 patients were underwent to radical cystectomy (24 died and 9 are alive), and 18 were underwent to the salvage cystectomy (10 died and 8 are alive). The survival curves for the first and second groups are demonstrated in Fig. 1 and Fig. 2. Here the mark 1 means the radical cystectomy and 0 means the salvage cystectomy, Tables 1–4 contain the mean survival times and results of testing identity of the survival curves using four tests: log-rank, Wilcoxon, Tarone–Ware, and p-statistics,

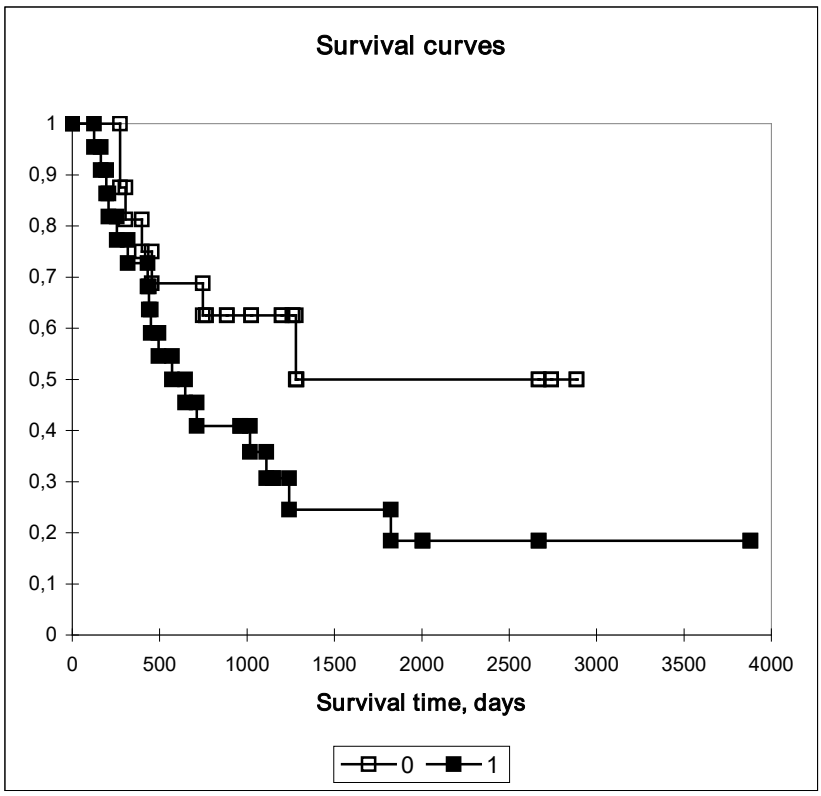

Figure 1: Survival curves in the first group of patients (stage T2)

As we see, in the first group (stage T2) the survival curve of the patients who were underwent to radical cystectomy goes above the survival curve of the patients who were underwent to salvage cystectomy. Therefore, intuitively, the risk of death for the former patients is less than for the latter ones and the salvage cystectomy prolongs life of patients better than the radical cystectomy. However, this hypothesis must be rigorously tested using statistical tests. Traditionally, to estimate the significance of the deviation between to survival curves the log-rank test, the Wilcoxon test, and the Tarone–Ware are used. Their p-values are the critical values of these tests.

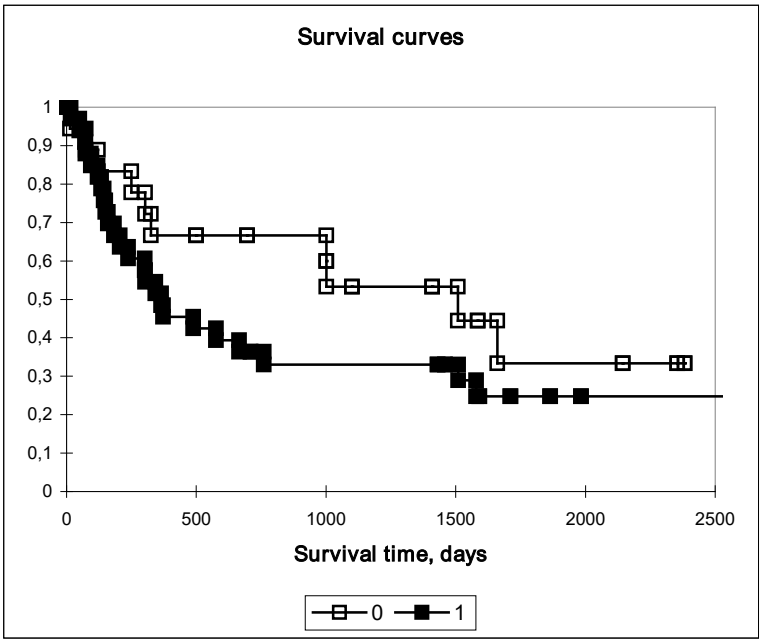

Figure 2: Survival curves in the second group of patients (stage T3)

In the second group (stage T3) the survival curve of the patients who were underwent to radical cystectomy also goes above the survival curve of the patients who were underwent to salvage cystectomy. We again may suppose that the risk of death for the former patients is less than for the latter ones. Note, that since the stage T3 is harder that T2, the survival interval became mush shorter. The maximum survival time in the first group is avout 4000 days (almost 11 years) but in second group it is about 2500 days (almost 7 years). Thus, the effectiveness of the cytectomy in this group is compensated by the stage of tumors. To estimate the significance of the deviation between to survival curves we again used the log-rank test, the Wilcoxon test, and the Tarone–Ware and their p-values.

In both cases we completed the traditional analysis by computing the p-statistics as an alternative to the three above tests. Descriptive statistics of the data are provided in Tables 1–3

Table 1

Mean survival time in the first group (stage T2)

Cystectomy	Mean survival time	Standard deviation	Lower bound (95%)	Upper bound (95%)
Radical	1015,720	202,769	618,300	1413,141
Salvage	1647,688	309,949	1040,198	2255,177

Table 2

Results of survival analysis in the first group of patients (stage T2) at significance level 0.05

Test	Observed value	Critical value	p-value
Log-rank	3.239	3.841	0,072
Wilcoxon	2.533	3.841	0,111

Tarone-Ware	2.893	3.841	0,089
P-statistics	0.997	0.950	0.003

Table 3

Mean survival time in the first group (stage T3)

Cystectomy	Mean survival time	Standard deviation	Lower bound (95%)	Upper bound (95%)
Radical	1015.720	202.769	618.300	1413.141
Salvage	1647.688	309.949	1040.198	2255.177

Table 4 contains the observed values, critical values and *p*-values of the log-rank test, the Wilcoxon test, the Tarone–Ware test, and the p-statistics.

Table 4

Results of survival analysis in the second group of patients (stage T3) at significance level 0.05

Test	Observed value	Critical value	p-value
Log-rank	1.718	3.841	0.190
Wilcoxon	2.083	3.841	0.149
Tarone-Ware	2.046	3.841	0.153
P-statistics	0.981	0.950	0.019

The hypothesis of the identity of the two survival functions (0 — the salvage cystectomy and 1 —the radical cystectomy) in the first and second groups (stages T2 and T3, respectively) was tested using four tests at a significance level of 0.05. In all the results, there were no statistically significant differences between the survival curves, since the observed values did not exceed the critical value and the upper confidence bound for the p-statistics exceeds 0.95. The log-rank test, the Wilcoxon test and the Tarone–Ware test acceps the null hypothesis is the corresponding p-values are less than 0.05, and the test based on the p-statistics, in opposite, accepts the null hypothesis if its p-value is greater than 0.05.

Noteworthy is the fact that the observed *p*-value (the probability of rejecting the null hypothesis, provided that it is true) in the *p*-statistics test is an order of magnitude less than in the three traditional nonparametric tests used in the analysis of survival. This is the evidence of high sensitivity and specificity of the proposed test.

4. Conclusions

Mathematical basis of modern medical information systems for assessing the effectiveness of treatment and the risk of death during a given time period must be more rigorously justified. Traditional nonparametric tests used in survival analysis (the log-rank test, the Wilcoxon test, and the Tarone–Ware test) assume conditions that not always are met in practice. These tests reduce the verification of statistical hypotheses about the equivalence of survival functions to calculating the critical value of the standard normal distribution. This is justified only when samples are large enough and additional conditions are met. Thus, to develop an effective medical information system for survival analysis, we need in nonparametric tests with minimal preliminary assumptions and minimal requirements to the size of samples.

In paper, we described a test for verification of the hypothesis of the equivalence of the survival functions and risk of death during a given time period, which does not depend on the sample size and does not use additional preconditions, except for the condition that the samples have not ties.

We have provided the strong mathematical background and demonstrated high sensitivity and specificity of testing homogeneity of two samples of random samples from continuous distributions in comparison with three traditional tests. We have shown the practical application of this test in survival analysis of the patient with bladder cancer and demonstrated its high performance. This test may be used for the development of effective medical information systems that perform survival analysis of cancer patients. Note, that the scheme described in the paper is easily expanded on much wider spectrum of problems connected with the assessment of the risk of device failure or risk of some significant event based on the censored and uncensored observations.

Future work will be directed to the improvement of computational complexity of the proposed test and its expanding to the various problem of the risk assessment.

References

[1] M.Morris, S.Landon, I.Reguilon, J.Butler, M.McKee, E.Nolte, Understanding the link between health systems and cancer survival: A novel methodological approach using a system-level conceptual model, Journal of Cancer Policy, 25, 202, 100233. doi: 10.1111/codi.15622

[2] J.M.Bland, D.G.Altman, The logrank test. British Medical Journal, 328, 2004, 1073. doi: 10.1136/bmj.328.7447.1073

[3] M.A.Proschan, L.E.Dodd, Re-randomization tests in clinical trials, Statistics in medicine, 38, 2019, pp. 2292-2302. doi: 10.1002/sim.8093

[4] R.E.Tarone, J.Ware, On distribution-free tests for equality of survival distributions, Biometrika, 64, 1977, pp. 156–160. doi: 10.1093/biomet/64.1.156

[5] T.G.Karrison, Versatile tests for comparing survival curves based on weighted log-rank statistics, The Stata Journal, 16, 2016, pp. 678–690

[6] A.Hazra, N.Gogtay, Biostatistics Series Module 9: Survival Analysis, Indian Journal of Dermatology, 62, 2017, pp.: 251–257. doi: 10.4103/ijd.IJD_201_17

[7] P.G.Karadeniz, I.Ercan, Examining tests for comparing survival curves with right censored data, Statistics in Transition New Series, 18, 2017, pp. 311–328. doi: 10.21307/stattrans-2016-072

[8] S.A.Matveichuk, Yu.I.Petunin, Generalization of Bernoulli schemes that arise in order statistics, I. Ukrainian Mathematical Journal, 42, 1990, pp. 459–466. doi: 10.1007/BF01058940

[9] S.A.Matveichuk, Yu.I Petunin, Generalization of Bernoulli schemes that arise in order statistics, II. Ukrainian Mathematical Journal, 43, 1991, pp. 728–734. doi: 10.1007/BF01058940

[10] N.Johnson, S.Kotz, Some generalizations of Bernoulli and Polya-Eggenberger contagion models, Statist Paper, 32, 1991, pp. 1–17. doi: 10.1007/BF02925473

[11] D.A.Klyushin, Yu.I.Petunin, A Nonparametric Test for the Equivalence of Populations Based on a Measure of Proximity of Samples, Ukrainian Mathematical Journal, 55, 2003, pp. 181–198. doi: 10.1023/A:1025495727612

[12] B.M.Hill, Posterior distribution of percentiles: Bayes' theorem for sampling from a population, Journal of American Statistical Association, 63, 1968, pp. 677–691