

# Analysis of asymptotic time complexity of an assumption-free alternative to the log-rank test

Lubomír Štěpánek

Department of Statistics and Probability  
Faculty of Informatics and Statistics  
University of Economics

nám. W. Churchilla 4, 130 67 Prague, Czech Republic  
lubomir.stepanek@vse.cz

&

Institute of Biophysics and Informatics  
First Faculty of Medicine  
Charles University

Salmovská 1, Prague, Czech Republic  
lubomir.stepanek@lf1.cuni.cz

Ivana Malá

Department of Statistics and Probability  
Faculty of Informatics and Statistics  
University of Economics

nám. W. Churchilla 4, 130 67 Prague, Czech Republic  
malai@vse.cz

Filip Habarta

Department of Statistics and Probability  
Faculty of Informatics and Statistics  
University of Economics

nám. W. Churchilla 4, 130 67 Prague, Czech Republic  
filip.habarta@vse.cz

Luboš Marek

Department of Statistics and Probability  
Faculty of Informatics and Statistics  
University of Economics

nám. W. Churchilla 4, 130 67 Prague, Czech Republic  
marek@vse.cz

**Abstract**—Comparison of two time-event survival curves representing two groups of individuals' evolution in time is relatively usual in applied biostatistics. Although the log-rank test is the suggested tool how to face the above-mentioned problem, there is a rich statistical toolbox used to overcome some of the properties of the log-rank test. However, all of these methods are limited by relatively rigorous statistical assumptions.

In this study, we introduce a new robust method for comparing two time-event survival curves. We briefly discuss selected issues of the robustness of the log-rank test and analyse a bit more some of the properties and mostly asymptotic time complexity of the proposed method. The new method models individual time-event survival curves in a discrete combinatorial way as orthogonal monotonic paths, which enables direct estimation of the  $p$ -value as it was originally defined. We also gently investigate how the surface of an area, bounded by two survival curves plotted onto a plane chart, is related to the test's  $p$ -value. Finally, using simulated time-event data, we check the robustness of the introduced method in comparison with the log-rank test.

Based on the theoretical analysis and simulations, the introduced method seems to be a promising and valid alternative to the log-rank test, particularly in case on how to compare two time-event curves regardless of any statistical assumptions.

## I. INTRODUCTION

**I**N SURVIVAL analysis, the response variable is usually two-dimensional, since it takes into account both the time of the event of our interest and whether the event (or the censoring) even occurred. More than intuitively, such a target variable suggests being plotted in a two-dimensional plot. As usual, while a number of subjects who do not evinced the

event of interest to all subjects is plotted on a vertical axis at a given time point, the time points where the event occurred are aligned with the horizontal axis, see also Fig. 1 That is the way how Kaplan-Meier estimators are commonly illustrated [1]. Therefore, the survival curve as a response variable could be represented as a monotonic orthogonal path, i. e., a polygonal path of a finite number of horizontal and vertical segments, in the Cartesian two-dimensional chart. Since such a variable deals both with the events of interest and their times, it is ordinarily called the time-event (survival) curve.

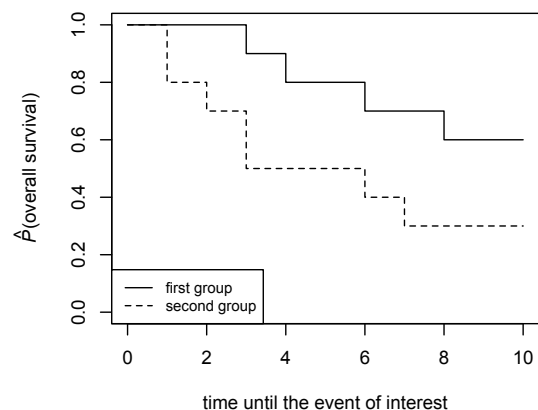


Fig. 1. Two time-event survival curves in a survival plot.

Whenever two or more time-event survival curves, describing evolution of events in time within two groups of individuals, are to be compared, several well-established methods could be used. A classical log-rank test could solve the problem, when there are only two groups supposed to be compared [2]. Assuming some special settings, particularly when the time-event survival curves are constructed using data that are not censored, i. e. the data fully describe all events of interest occurred in the groups, then a simple Wilcoxon rank-sum test might be applied. If more than two groups are to be compared, the problem could be battled using either a score-rank test, or even a Cox proportional hazards model [3]. All the approaches mentioned above may be performed in various software, including R language and environment [4], such that a pure R package *stats* or a package *survival* [5] could be employed to do the job. Nevertheless, each one of the described methods has its limitations, and its application is determined by meeting relatively rigorous statistical assumptions.

Application of the log-rank test that compares two non-crossing time-event survival curves (similar as plotted in Fig. 1) is limited mostly by assuming the fact that censoring (i) should not induce anyhow the observed events and (ii) is equally likely to occur in both the groups. What is more, the counts of events of interest should be large enough to satisfy asymptotic properties of  $\chi^2$  distribution and fulfill the central limit theorem to let the log-rank test statistic follow an asymptotically normal distribution. That means an incidence of the events of interest in each group across all the time points should be neither too small nor too large.

To overcome the limitations of the classical log-rank test, several diverse modifications of the log-rank test were published to either increase efficiency of the test, or its robustness against violation of the statistical assumptions, or both. Whereas Kong (1997) in [6] adjusted the log-rank test efficiency by improving the hazard functions, i. e. functions of rates of events based on fixed proportions of the events in the past, Song et al. (2008) dived deeper into covariate matrix decomposition, by which they derived formulas estimating minimal sample sizes that enables a valid usage of the log-rank test [7]. Several authors such as Peto and Peto (1972) [8], Yang and Prentice (2010) [9], and Li (2018) [10], suggested the usage of weights of individual observations, usually lower weights for later events when there the numbers of observations tend to be not so high; by this they improve the validity of the log-rank test outputs.

There are also articles handling with exact discrete calculations when compare two survival curves which is more similar to our proposed approach. Thomas (1975) simplified the computations by fixing total numbers in the compared groups [11]. The algorithm was improved a bit computationally by Mehta et al. (1985) [12]. Finally, Heinze et al. (2003), similarly asymptotic approaches above, incorporated a weighting scheme into the calculations to increase significance of earlier observations [13].

Studies that go deeper into asymptotic complexities of the statistical inference test, particularly the exact ones that ex-

haustively compute over a polynomial universe, are generally missing. Some significant pieces of related knowledge focused on complexity of classic but robust and computationally-hard inference tests are discussed by Mosler (2002) [14], Smolinski et al. in (2008) [15], and Kulikov et al. (2014) [16].

Vast majority of the papers listed above work with a hazard function, which is event of interest rate in a given time point conditional on overall survival rate until the time point or they assume constant total numbers of subjects in all the compared groups. Unlike them, in this proceeding, besides a brief discussion on limitations of the log-rank test, we model the time-event survival curves using a discrete combinatorial approach, considering the survival curves to be orthogonal monotonic paths on a plane of two-dimensional plot (as shown in Fig. 1 and Fig. 2), and taking into account their mutual "Manhattan" grid distances. That indicates how easily the  $p$ -value of this modified log-rank test could be calculated using its original statistical definition as a conditional probability of observing data of given properties. Then we analyse asymptotic time complexity of algorithmic approaches behind the proposed method. We also briefly discuss the possible relationship between the two-dimensional surface bounded by two non-crossing survival curves in the plot and the test's  $p$ -value. Finally, using simulations of artificial survival curves, the first type errors as rates of detection the cases, when similar curves are supposed to be different, are estimated for both the log-rank test and our proposed alternative, mutually compared and discussed within the frame of the robustness of the methods.

## II. PRINCIPLES, ASSUMPTIONS AND LIMITATIONS OF THE LOG-RANK TEST

Firstly, we gently introduce principles of the log-rank test, by which we can better understand its assumptions and limitations.

### A. Principles of the log-rank test

Let's assume two groups of individuals (marked by indices 1, and 2, respectively) and  $k \in \mathbb{N}$  distinct event times. At each event time, we can construct a  $2 \times 2$  contingency table and compare the event rates between the two groups. Let the  $(t_1, t_2, \dots, t_k)^T$  be an ordered tuple of the event time points, then for the  $j$ -th event time  $t_j$ , such that  $j \in \{1, 2, 3, \dots, k\}$ , we can construct the (contingency) table Tab. I. At  $j$ -th event time, there are  $d_{1,j}$  and  $d_{2,j}$  individuals who experienced the events in the group 1 and 2, respectively, and  $r_{1,j}$  and  $r_{2,j}$  subjects at risk (who have not yet had the event or been censored) in the groups 1 and 2, respectively, see Tab. I.

TABLE I  
NUMBERS OF THE EVENTS OF INTEREST IN BOTH GROUPS AT TIME  $t_j$ .

group	event of interest at the event time $t_j$		total
	yes	no	
1	$d_{1,j}$	$r_{1,j} - d_{1,j}$	$r_{1,j}$
2	$d_{2,j}$	$r_{2,j} - d_{2,j}$	$r_{2,j}$
total	$d_j$	$r_j - d_j$	$r_j$

The log-rank test checks the null hypothesis  $H_0$  that both groups have identical hazard functions, i. e. that rates of the events of interest in time conditional on fixed rates in the past are the same. Under the null hypothesis  $H_0$ , the observed numbers of the events could be considered as random variables  $D_{1,j}$  and  $D_{2,j}$  following a hypergeometric distribution with parameters  $(r_j, r_{i,j}, d_j)$  for both  $i \in \{1, 2\}$ . Thus, the expected value of the variable  $D_{i,j}$  is  $\mathbb{E}(D_{i,j}) = r_{i,j} \frac{d_j}{r_j}$  and variance is  $\text{var}(D_{i,j}) = \frac{r_{1,j} r_{2,j} d_j}{r_j^2} \left( \frac{r_j - d_j}{r_j - 1} \right)$  for both  $i \in \{1, 2\}$ . For all  $j \in \{1, 2, 3, \dots, k\}$  we can compare the observed numbers of events of interest,  $d_{i,j}$ , to their expected values  $\mathbb{E}(D_{i,j}) = r_{i,j} \frac{d_j}{r_j}$ , under  $H_0$ . So, the test statistic for both  $i \in \{1, 2\}$  is finally

$$\begin{aligned} \chi_{\log\text{-rank}}^2 &= \frac{\left( \sum_{j=1}^k d_{i,j} - \mathbb{E}(D_{i,j}) \right)^2}{\sum_{j=1}^k \text{var}(D_{i,j})} = \\ &= \frac{\left( \sum_{j=1}^k d_{i,j} - r_{i,j} \frac{d_j}{r_j} \right)^2}{\sum_{j=1}^k \frac{r_{1,j} r_{2,j} d_j}{r_j^2} \left( \frac{r_j - d_j}{r_j - 1} \right)}, \end{aligned} \quad (1)$$

which follows under  $H_0$  a  $\chi^2$  distribution with 1 degree of freedom,  $\chi_{\log\text{-rank}}^2 \sim \chi^2(1)$ . For feasible large  $r_j$ , at least  $r_j \geq 30$ , a square root of  $\chi_{\log\text{-rank}}^2$  follows a standard normal distribution,  $\sqrt{\chi_{\log\text{-rank}}^2} \sim \mathcal{N}(0, 1^2)$ .

#### B. Some of the assumptions and limitations of the log-rank test

Firstly, censoring is assumed not to affect anyhow the occurrence of event of interest, and the proportion of censored data are supposed being of nearly equal size in both the groups, as well. Otherwise, the test statistic  $\chi_{\log\text{-rank}}^2$  calculated using (1) either for  $i = 1$ , or for  $i = 2$ , respectively, could be biased and therefore mutually different. That may affect the interpretability, i. e. the robustness of the log-rank test applied on such data.

Then, since the test statistic  $\chi_{\log\text{-rank}}^2$  follows a  $\chi^2$  distribution, the initial total number of individuals  $r_0$  and the number of all event times  $k$  should be large enough. Analogously but inversely, whenever the numbers of individuals  $d_j$  experiencing the event of interest are generally large (relatively to  $r_j$ ), than both the numerator and denominator of the fraction in the formula (1) is relatively small, too, and, consequently, one could expect that the  $\chi_{\log\text{-rank}}^2$  statistic (or the derived  $\sqrt{\chi_{\log\text{-rank}}^2}$  statistic) does not fulfil its assumed asymptotic properties, and its estimate could be thus biased. That might influence both the robustness and the power of the log-rank test when applied to data of such limitations.

By researching the denominator of the equation (1) a bit deeper, we can realize the test statistic  $\chi_{\log\text{-rank}}^2$  is the highest when the denominator  $\sum_{j=1}^k \text{var}(D_{i,j})$  is as low as possible given the values  $d_{i,j}$  and  $r_{i,j}$  for all  $i \in \{1, 2\}$  and  $j \in \{1, 2, 3, \dots, k\}$ . It is worth mentioning this holds just when the proportions  $\frac{r_{1,j}}{r_j} = \frac{r_{1,j}}{r_{1,j} + r_{2,j}}$  and  $\frac{r_{2,j}}{r_j} = \frac{r_{2,j}}{r_{1,j} + r_{2,j}}$  are

both constant (and mutually different enough) across all the time points  $(t_1, t_2, \dots, t_k)^T$ , and then the log-rank test is the most powerful; i. e. in other words, its ability to reject the null hypothesis  $H_0$ , claiming the survival curves are equivalent, when they are in fact different, is maximal possible. That used to be the most usual issue that may decrease the power of the log-rank test. The mentioned proportions are typically not constant when the time event curves change a lot their mutual distance across the time points or when they even cross themselves one or more times. Consequently, the power of the log-rank test may be decreased by any deviations from the constant values of the proportions  $\frac{r_{1,j}}{r_j}$ , and  $\frac{r_{2,j}}{r_j}$ , respectively.

### III. INTRODUCTION OF AN ASSUMPTION-FREE ALTERNATIVE TO THE LOG-RANK TEST

Within this section, we introduce an assumption-free alternative to the log-rank test. The alternative algorithm for two time-event curves comparison is based on a discrete combinatorial calculation of possible states (i. e. all possible time-event curves) that would be theoretically obtained and that are at least as extreme as the original two survival curves. This approach corresponds to an original definition of a  $p$ -value as a probability of obtaining data at least as extreme as the data currently observed, assuming that the null hypothesis is true (i. e. the observed survival curves are not statistically different).

All the possible states could be considered as monotonic orthogonal paths in the two-dimensional chart of two original survival curves, excluding (for simplicity) the crossing curves. By calculating (or estimating) the numbers of all the paths at least as extreme as the plotted two curves, i. e. all the paths such that one is above the first observed one and the other is below the second observed one, we get a point estimate of the  $p$ -value as a proportion of all pairs of orthogonal paths contradicting the same way or even more to the observed survival curves. Or in other words, as a proportion of all pairs of orthogonal paths that are at least as distant one from the other than the original two time-event curves.

#### A. Principle of the proposed assumption-free alternative to the log-rank test

Again, let two groups of individuals (marked by indices 1, and 2, respectively) to be compared and  $k \in \mathbb{N}$  distinct event times when events of interest could occur. Let the  $(t_1, t_2, \dots, t_k)^T$  be an ordered tuple of the event times. At each event time, we can compute the number of individuals who experienced the event at the  $j$ -th event time  $t_j$  for both groups, similarly to the construction of contingency tables, as shown in table Tab. I. By repeating this approach  $k$  times, consequently, once we get the proportions of subjects at risk,  $\frac{r_{1,j}}{r_j}$ , and  $\frac{r_{2,j}}{r_j}$ , respectively, for each event time  $t_j$ , we could plot the time-event survival curves based on the proportions of individual in risk  $\frac{r_{1,j}}{r_j}$ , and  $\frac{r_{2,j}}{r_j}$  similarly to Fig. 1.

For simplicity, the survival curves are assumed not to cross themselves. More technically spoken, it for each  $j$ -th event time  $t_j$  holds

$$\frac{r_{1,j}}{r_j} \geq \frac{r_{2,j}}{r_j}, \quad (2)$$

as illustrated in Fig. 1. By adding a grid into the Fig. 1, we get Fig. 2, which is a bit closer to an idea of calculating (or estimating) a number of monotonic orthogonal paths starting at the proportion of subjects at risk  $\frac{r_{i,0}}{r_0} = 1$  and ending — after  $k$  event times — at the proportion of subjects at risk  $\geq \frac{r_{i,k}}{r_k}$  (one of such possible paths is the blue line for  $i = 1$  in Fig. 2)  $\leq \frac{r_{i,k}}{r_k}$  (similarly to the red line for  $i = 2$  in Fig. 2).

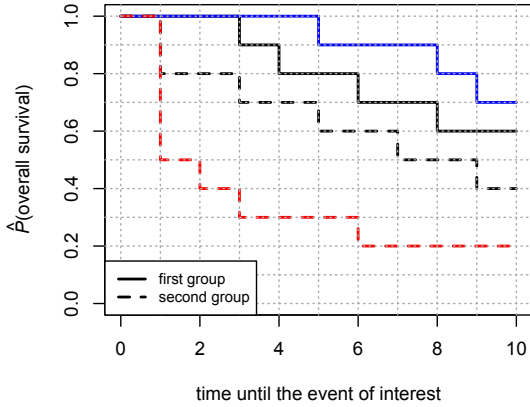


Fig. 2. Two original time-event survival curves in a survival plot (black lines) and an example of a pair of monotonic orthogonal paths such that one is above (blue solid line,  $i = 1$ ) the upper original survival curve and the second one is below (red dashed line,  $i = 2$ ) the lower original survival curve.

Let  $N_{(1,k,u,v)}$  stands for the number of all orthogonal paths (respecting the grid, i. e. all segments of such a path are parallel to horizontal or vertical lines of the grid and its edges are aligned to grid points) starting at the proportion 1 (left upper corner of the Fig. 2) and ending after  $k$  event times at the proportion of subjects at risk  $\frac{u}{v}$  (a point with coordinates  $[k, \frac{u}{v}]$  in Fig. 2). Eventually, let  $N_{(1,k,u,v)}^+$  be a number of all orthogonal paths starting at the proportion 1, going above the 1-st survival curve or tangentially meeting it (without crossing it) and ending at the proportion of subjects at risk  $\geq \frac{u}{v}$ . Analogously, let  $N_{(1,k,u,v)}^-$  be a number of all orthogonal paths starting at the proportion 1, going below the 2-nd survival curve or tangentially meeting it (without crossing it) and ending at the proportion of subjects at risk  $\leq \frac{u}{v}$  after  $k$  event times. The numbers  $N_{(1,k,u,v)}^+$  and  $N_{(1,k,u,v)}^-$  could be computed perhaps exhaustively in a combinatorial way (this is an open problem) or could definitely be estimated by numerical simulations.

Let us define a null hypothesis  $H_0$  that claims the original (observed) survival curves are not significantly different. One of the tricky part on the proposed method is that, since we do not need any more initial assumptions for this testing, we also do not require modelling a null distribution. The  $p$ -value, as mentioned above, is the probability of obtaining data (expected

survival curves described as monotonic orthogonal paths in the survival plot) at least as extreme as the data currently observed (the two original survival curves), assuming that the null hypothesis  $H_0$  is correct. Following the definition of the  $p$ -value and marking it as  $p$ , we get

$$p = p\text{-value}$$

$$p = P(\text{getting data at least as extreme as the observed} | H_0)$$

$$p = P \left( \frac{N_{1,k,r_{1,k},r_k}^+ \cdot N_{2,k,r_{2,k},r_k}^-}{\left( \sum_{j=0}^{r_k} N_{k,j,r_k} \right)^2 - N_{cc}} \right) \quad (3)$$

where  $N_{cc}$  is a number of pairs of survival curves crossing each other. Again, the number  $N_{cc}$  can be calculated probably either using a discrete combinatorial analysis, or be numerically simulated (which is far easier).

In comparison with the term in the denominator of the equation (3), the number of pairs of survival curves depicted by the numerator can not include any crossing curves. Since we assume all curves ending in the proportion  $\frac{r_{1,k}}{r_k}$  or greater, and all curves ending in the proportion  $\frac{r_{2,k}}{r_k}$  or lower, considering that  $\frac{r_{1,j}}{r_j} \geq \frac{r_{2,j}}{r_j}$  for each  $j \in \{1, 2, 3, \dots, k\}$  as stated in (2), thus, since  $\frac{r_{1,k}}{r_k} \geq \frac{r_{2,j}}{r_j}$  for all time points, there are no pairs of crossing curves taken into account in the numerator of (3). The curves could tangentially meet themselves (in case of  $=$ ) or run one above the other (in case of  $>$ ), but could not cross each other.

### B. A brief analysis of surface bounded by two non-crossing survival curves and the test's $p$ -value

Surfaces above the first, upper survival curve (let us mark it as  $S_1^+$ ) and below the second, bottom curve (let us mark it as  $S_2^-$ ) in Fig. 2 suggest investigating on how are the surfaces related to the  $p$ -value of the test.

By following the first impression, when  $S$  stands for a surface of the whole canvas of the chart in Fig. 2, it seems that  $p$ -value is proportional to the term  $\frac{S_1^+ + S_2^-}{S}$ . However, the relationship between the  $p$ -value and the surfaces is more complex and not so straightforward. The numbers of all orthogonal paths in some dedicated surface, let us assume e. g.  $N_{(1,k,u,v)}^+$ , is not proportional to the size of the surface. As a sketch of a proof by contradiction, let us suppose we are to calculate the number of  $N_{(1,k,0,v)}^-$  curves below a horizontal curve crossing the point  $[k, v]$ . Then, simply using combinatorial rules, we realize that  $N_{(1,k,0,v)}^- = \binom{k+v}{k}$ . However, if we now want to calculate the number of  $N_{(1,k,0,2v)}^-$  curves below a horizontal curve crossing the point  $[k, 2v]$ , we get that  $N_{(1,k,0,2v)}^- = \binom{k+2v}{k}$ . Whereas the proportion of the surfaces below the two lines crossing the points  $[k, 2v]$  and  $[k, v]$  is equal to 2, the proportion of the numbers of the paths is in general much greater than 2, since generally  $\frac{\binom{k+2v}{k}}{\binom{k+v}{k}} \gg 2$ .

Thus,  $\frac{N_{1,k,r_{1,k},r_k}^+}{N_{2,k,r_{2,k},r_k}^-} \neq \frac{S_1^+}{S_2^-}$  in general and the  $p$ -value is not (!) proportional to the term  $\frac{S_1^+ + S_2^-}{S}$ .

*C. Approaches on calculation the  $p$ -value of the proposed alternative to the log-rank test*

The terms such as  $N_{1,k,r_{1,k},r_k}^+$ ,  $N_{2,k,r_{2,k},r_k}^-$ ,  $\sum_{j=0}^{r_k} N_{k,j,r_k}$ , and  $N_{cc}$ , respectively, in equation (3) could be estimated either numerically by re-sampling, or calculated exhaustively. A fully analytical approach is under current research.

*Numerical estimation.* All the terms such as  $N_{1,k,r_{1,k},r_k}^+$ ,  $N_{2,k,r_{2,k},r_k}^-$ ,  $\sum_{j=0}^{r_k} N_{k,j,r_k}$ , and  $N_{cc}$ , respectively, in equation (3) could be numerically estimated by *re-sampling* approach. Let us assume we got two non-crossing survival curves similarly to the plot in Fig. 1, so that we know the values  $k$ ,  $r_{1,j}$ ,  $r_{2,j}$ ,  $d_{1,j}$ , and  $d_{2,j}$  for all  $j \in \{1, 2, 3, \dots, k\}$ . Let us suppose we generate  $n$  pairs of survival curves. Then, let  $N(\forall+, \forall-)(n)$  be the number of all pairs such that one of the curves is completely above the first original curve and the other is completely below the second original curve (and, thus, they do not cross each other), in all  $n$  generated pairs. Let  $N(\text{non-crossing})(n)$  be the number of all pairs such that the curves of the pair do not cross each other, in all  $n$  generated pairs. Then we can simply derive that

$$N_{1,k,r_{1,k},r_k}^+ \cdot N_{2,k,r_{2,k},r_k}^- = \lim_{n \rightarrow \infty} \frac{N(\forall+, \forall-)(n)}{\left( \sum_{j=0}^{r_k} N_{k,j,r_k} \right)^2} - N_{cc} = N(\text{non-crossing})(n),$$

and, consequently, by replacing in equation (3)

$$\hat{p} = \lim_{n \rightarrow \infty} \frac{N(\forall+, \forall-)(n)}{N(\text{non-crossing})(n)}.$$

By this *re-sampling* approach, we can obtain for reasonably large  $n \in \mathbb{N}$  an unbiased estimate of  $p$ -value in equation (3) of the proposed alternative to the log-rank test. The algorithm is also described in Algorithm 1

*Exhaustive approach.* Let us again assume we got two non-crossing survival curves similarly to the plot in Fig. 1, so that we know the values  $k$ ,  $r_{1,j}$ ,  $r_{2,j}$ ,  $d_{1,j}$ , and  $d_{2,j}$  for all  $j \in \{1, 2, 3, \dots, k\}$ . The exhaustive, greedy approach is based on grid search for all possible pairs of survival curves such that one of the curves is completely above the first original curve and the other is completely below the second original curve (and, thus, they do not cross each other). In case the exhaustive approach is finished successfully, one could obtain more confident estimate of  $p$ -value in equation (3) of the proposed alternative to the log-rank test than in case of the numerical re-sampling.

Since the exhaustive approach is greedy, so that one could expect a large asymptotic time complexity, we enumerated worst-case scenarios estimates of all terms such as  $N_{1,k,r_{1,k},r_k}^+$ ,  $N_{2,k,r_{2,k},r_k}^-$ ,  $\sum_{j=0}^{r_k} N_{k,j,r_k}$ , and  $N_{cc}$ , respectively, in equation (3).

Since  $N_{(1,k,r_{1,k},r_k)}^+$  (or  $N_{(2,k,r_{2,k},r_k)}^-$ ) is a number of all orthogonal paths starting at the proportion 1, going above (or below) the 1-st (or the 2-nd) survival curve or tangentially meeting it (without crossing it) and ending at the proportion of subjects at risk  $\geq \frac{u}{v}$ , number of such paths could not be

---

**Algorithm 1:** Re-sampling approach on how to obtain for reasonably large  $n \in \mathbb{N}$  an unbiased estimate of  $p$ -value in equation (3) of the proposed alternative to the log-rank test.

---

**Data:** two non-crossing survival curves  
**Result:** an unbiased estimate of  $p$ -value in equation (3) of the proposed alternative to the log-rank test

```

1  $k, r_{1,j}, r_{2,j}, d_{1,j}, d_{2,j}$  // parameters of
  the original two survival curves ;
2  $n$  // number of
  repetitions;
3  $N(\forall+, \forall-)(0) = 0$  // number of all
  pairs such that one of the curves is
  completely above the first original
  curve and the other is completely
  below the second original curve;
4  $N(\text{non-crossing})(0) = 0$  // number of all
  pairs such that the curves of the
  pair do not cross each other;
5 for  $j = 1 : n$  do
6   generate a pair of two survival curves;
7   if the curves of the pair do not cross each other
   then
8      $N(\text{non-crossing})(j) = N(\text{non-crossing})(j) + 1$ ;
9     if one of the curves is completely above the
   first original curve and the other is completely
   below the second original curve then
10     $N(\forall+, \forall-)(j) = N(\forall+, \forall-)(j) + 1$ ;
11 end
12 calculate an estimate of  $p$ -value as
    $\hat{p} = \frac{N(\forall+, \forall-)(n)}{N(\text{non-crossing})(n)}$  ;
```

---

larger than a number of all monotonic orthogonal paths in a rectangle of size  $k \times d_{1,k}$  (or  $k \times d_{2,k}$ ). Then,

$$N_{1,k,r_{1,k},r_k}^+ \leq \sum_{j=0}^{d_{1,k}} \binom{k+j}{k} = \binom{d_{1,k} + k + 1}{k + 1}$$

$$N_{2,k,r_{2,k},r_k}^- \leq \sum_{j=0}^{r_{2,k}} \binom{k+j}{k} - \sum_{j=0}^{d_{2,k}} \binom{k+j}{k} =$$

$$= \binom{r_{2,k} + k + 1}{k + 1} - \binom{d_{2,k} + k + 1}{k + 1}.$$

Number of all monotonic orthogonal paths in the grid,  $\sum_{j=0}^{r_k} N_{k,j,r_k}$ , is by assuming (for simplicity)  $r_{1,k} = r_{2,k}$  similarly

$$\sum_{j=0}^{r_k} N_{k,j,r_k} \leq \sum_{j=0}^{r_{2,k}} \binom{k+j}{k} = \binom{r_{2,k} + k + 1}{k + 1}.$$

Since parts of crossing curves in a pair could be rearranged such that the crossing segments could be "re-coloured" eventually, i. e. switched so that the curves only tangentially

meet each other keeping them monotonic, we can assume that  $N_{cc} \ll \sum_{j=0}^{r_k} N_{k,j,r_k}$ .

Putting all the derivations together, we can estimate an upper estimate  $\Theta(\bullet)$  of all the monotonic orthogonal paths' grid searching by the formula

$$\begin{aligned} \Theta(\bullet) &= \Theta\left(N_{1,k,r_{1,k},r_k}^+\right) + \Theta\left(N_{2,k,r_{2,k},r_k}^-\right) + \\ &+ \Theta\left(\sum_{j=0}^{r_k} N_{k,j,r_k}\right) + \Theta(N_{cc}) = \\ &= \Theta\left(\binom{d_{1,k} + k + 1}{k + 1}\right) + \\ &+ \Theta\left(\binom{r_{2,k} + k + 1}{k + 1} - \binom{d_{2,k} + k + 1}{k + 1}\right) + \\ &+ \Theta\left(\binom{r_{2,k} + k + 1}{k + 1}\right) + \\ &+ \Theta(0) = \\ &= \Theta\left((k + d_{1,k}/2)^{d_{1,k}}\right) + \\ &+ \Theta\left((k + r_{2,k}/2)^{r_{2,k}} - (k + d_{2,k}/2)^{d_{2,k}}\right) + \\ &+ \Theta\left((k + r_{2,k}/2)^{r_{2,k}}\right). \end{aligned}$$

Since we assume  $r_k = r_{1,k} = r_{2,k}$ , then there is  $r_{1,k} \geq d_{1,k}$  and the final worst-case scenario's asymptotic time complexity of  $p$ -value exhaustive calculation using the proposed method alternative to the log-rank test is equal to  $\Theta(\bullet) = \Theta((k + r_k/2)^{r_k})$ .

In comparison to the novel method's time complexity, the log-rank test's  $\chi_{\log\text{-rank}}^2$  statistic based on equation (1) is significantly simpler, considering its asymptotic time complexity. Both by inspecting the numerator and denominator of the fraction in (1), we can see the calculation is based only on two summations of  $k$  terms, so the asymptotic time complexity of the log-rank test's  $\chi_{\log\text{-rank}}^2$  statistic as about  $\Theta(k)$ , where  $k$  is number of time points where the event of interest may occur.

*Analytical approach.* At the current moment, a fully analytical approach on how to calculate the term  $N_{1,k,r_{1,k},r_k}^+$ ,  $N_{2,k,r_{2,k},r_k}^-$ ,  $\sum_{j=0}^{r_k} N_{k,j,r_k}$ , and  $N_{cc}$ , respectively, in equation (3) is an open problem and requires authors' ongoing research.

#### IV. SIMULATION STUDY

We compared the log-rank test and the assumption-free method proposed above by simulating many pairs of random non-crossing curves, assuming the curves in the pairs are not significantly different. Then, we calculated the first type errors rates, i. e. rates of the situations, when the inference test (either the log-rank test, or the new method) claims that two statistically similar survival curves are (falsely) detected as different. Finally, we assume that more robust the method is, the lower value of the first type error it should return.

The simulation study was performed using R programming language and environment [4]. There is more on numerical applications of R programming language to various fields in [17]–[21].

When generating the pairs of survival curves, we applied the following negatively exponential survival function,

$$s(t) = \sigma\left(e^{-\frac{10+\varepsilon}{10000}t}\right)$$

where  $\varepsilon$  is a random noise term that follows a standard normal distribution, i. e.  $\varepsilon \sim \mathcal{N}(0, 1^2)$ , and  $\sigma(\bullet)$  is a function rounding its argument to the nearest integer, e. g.  $\sigma(4.3) = 4$ ,  $\sigma(4.5) = 5$  or  $\sigma(5.8) = 6$ .

There were  $n = 1000$  pairs of significantly non-different survival curves generated in total and within each pair, the curves were compared using both the log rank test, and the above-proposed method. By summing up numbers of cases where  $p$ -value was lower than or equal to 0.05 regardless of the method, we got the point estimates of the first type error frequencies as illustrated in Tab. II.

TABLE II  
POINT ESTIMATES OF THE FIRST TYPE ERROR RATES FOR THE LOG-RANK TEST AND THE PROPOSED METHOD, BASED ON THE SIMULATION DESCRIBED ABOVE.

	method	
	the log-rank test	the proposed method
# of simulated cases in total	1000	1000
# of cases $p$ -value $\leq 0.05$	54	15
first type error rate estimate	0.054	0.015

Whereas the log-rank test output a point estimate of the first type error rate about 0.054, the method introduced above returned a point estimate of the first type error rate about 0.015, therefore lower than the one for the log-rank test. Thus, the proposed method seems to be more robust than the log-rank test, based on the simulation described above. The first type error settings follows the common value of the alpha level equal to 0.050, as usual in applied sciences.

#### V. CONCLUSION REMARKS

By calculation of monotonic orthogonal paths in the grid of survival plot, we can get a ratio of the number of all pairs of the paths that are more distant one to each other, which opposes the null hypothesis, and the number of all non-crossing pairs of possible paths. This is a suggested point estimate of the  $p$ -value of the proposed alternative to the classical log-rank test.

Based on the simulation, the introduced method proved to be of higher robustness than the log-rank test. So, the assumption-free version of the log-rank test seems to be a valid alternative for the comparison of two time-event curves. However, while the numerical estimation of the  $p$ -value seems to be relatively simple-to-follow, exhaustive (greedy) calculation of exact values is of a high asymptotic time complexity and analytical derivations of the  $p$ -value formula requires following research.

Besides, the method and the computational aspects could also be a topic for a new R package development.

## VI. ACKNOWLEDGEMENT

This research was supported by the grant no. F/45/2020 provided by Internal grant agency of University of Economics, Prague.

## REFERENCES

- [1] E. L. Kaplan and Paul Meier. “Nonparametric Estimation from Incomplete Observations”. In: *Journal of the American Statistical Association* 53.282 (June 1958), pp. 457–481. DOI: 10.1080/01621459.1958.10501452. URL: <https://doi.org/10.1080/01621459.1958.10501452>.
- [2] Nathan Mantel. “Evaluation of survival data and two new rank order statistics arising in its consideration”. In: *Cancer chemotherapy reports* 3.50 (1966), pp. 163–170.
- [3] Huimin Li, Dong Han, Yawen Hou, et al. “Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods”. In: *PLOS ONE* 10.1 (Jan. 2015). Ed. by Zhongxue Chen, e0116774. DOI: 10.1371/journal.pone.0116774. URL: <https://doi.org/10.1371/journal.pone.0116774>.
- [4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: <https://www.R-project.org/>.
- [5] Therneau T. *survival: A Package for Survival Analysis in R*. Vienna, Austria, R package version 3.1-12. URL: <https://CRAN.R-project.org/package=survival/>.
- [6] F. Kong. “Robust covariate-adjusted logrank tests”. In: *Biometrika* 84.4 (Dec. 1997), pp. 847–862. DOI: 10.1093/biomet/84.4.847. URL: <https://doi.org/10.1093/biomet/84.4.847>.
- [7] Rui Song, Michael R. Kosorok, and Jianwen Cai. “Robust Covariate-Adjusted Log-Rank Statistics and Corresponding Sample Size Formula for Recurrent Events Data”. In: *Biometrics* 64.3 (Dec. 2007), pp. 741–750. DOI: 10.1111/j.1541-0420.2007.00948.x. URL: <https://doi.org/10.1111/j.1541-0420.2007.00948.x>.
- [8] Richard Peto and Julian Peto. “Asymptotically Efficient Rank Invariant Test Procedures”. In: *Journal of the Royal Statistical Society. Series A (General)* 135.2 (1972), p. 185. DOI: 10.2307/2344317. URL: <https://doi.org/10.2307/2344317>.
- [9] Song Yang and Ross Prentice. “Improved Logrank-Type Tests for Survival Data Using Adaptive Weights”. In: *Biometrics* 66.1 (Apr. 2009), pp. 30–38. DOI: 10.1111/j.1541-0420.2009.01243.x. URL: <https://doi.org/10.1111/j.1541-0420.2009.01243.x>.
- [10] Chenxi Li. “Doubly robust weighted log-rank tests and Renyi-type tests under non-random treatment assignment and dependent censoring”. In: *Statistical Methods in Medical Research* 28.9 (July 2018), pp. 2649–2664. DOI: 10.1177/0962280218785926. URL: <https://doi.org/10.1177/0962280218785926>.
- [11] Donald G. Thomas. “Exact and asymptotic methods for the combination of  $2 \times 2$  tables”. In: *Computers and Biomedical Research* 8.5 (Oct. 1975), pp. 423–446. DOI: 10.1016/0010-4809(75)90048-8. URL: [https://doi.org/10.1016/0010-4809\(75\)90048-8](https://doi.org/10.1016/0010-4809(75)90048-8).
- [12] Cyrus R. Mehta, Nitin R. Patel, and Robert Gray. “Computing an Exact Confidence Interval for the Common Odds Ratio in Several  $2 \times 2$  Contingency Tables”. In: *Journal of the American Statistical Association* 80.392 (Dec. 1985), p. 969. DOI: 10.2307/2288562. URL: <https://doi.org/10.2307/2288562>.
- [13] Georg Heinze, Michael Gnant, and Michael Schemper. “Exact Log-Rank Tests for Unequal Follow-Up”. In: *Biometrics* 59.4 (Dec. 2003), pp. 1151–1157. DOI: 10.1111/j.0006-341x.2003.00132.x. URL: <https://doi.org/10.1111/j.0006-341x.2003.00132.x>.
- [14] Karl Mosler. *Multivariate dispersion, central regions, and depth : the lift zonoid approach*. New York: Springer, 2002. ISBN: 0387954120.
- [15] Tomasz Smolinski. *Computational intelligence in biomedicine and bioinformatics : current trends and applications*. Berlin: Springer, 2008. ISBN: 978-3-540-70776-9.
- [16] Alexander Kulikov. *Combinatorial pattern matching : 25th annual symposium, CPM 2014 Moscow, Russia, June 16-18, 2014, proceedings*. Cham: Springer, 2014. ISBN: 978-3-319-07565-5.
- [17] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. “Evaluation of facial attractiveness for purposes of plastic surgery using machine-learning methods and image analysis”. In: *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, Sept. 2018. DOI: 10.1109/healthcom.2018.8531195. URL: <https://doi.org/10.1109/healthcom.2018.8531195>.
- [18] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. “Machine-learning at the service of plastic surgery: a case study evaluating facial attractiveness and emotions using R language”. In: *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2019. DOI: 10.15439/2019f264. URL: <https://doi.org/10.15439/2019f264>.
- [19] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. “Evaluation of Facial Attractiveness after Undergoing Rhinoplasty Using Tree-based and Regression Methods”. In: *2019 E-Health and Bioengineering Conference (EHB)*. IEEE, Nov. 2019. DOI: 10.1109/ehb47216.2019.8969932. URL: <https://doi.org/10.1109/ehb47216.2019.8969932>.
- [20] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. “Machine-Learning and R in Plastic Surgery – Evaluation of Facial Attractiveness and Classification of Facial Emotions”. In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, Sept. 2019, pp. 243–252. DOI: 10.1007/978-3-030-30604-

- 5\_22. URL: [https://doi.org/10.1007/978-3-030-30604-5\\_22](https://doi.org/10.1007/978-3-030-30604-5_22).
- [21] Patřicia Martinkova, Lubomır Štěpanek, Adela Drabinova, et al. “Semi-real-time analyses of item characteristics for medical school admission tests”. In: *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*. Ed. by M. Ganzha, L. Maciaszek, and M. Paprzycki. Vol. 11. Annals of Computer Science and Information Systems. IEEE, 2017, pp. 189–194. DOI: 10.15439/2017F380. URL: <http://dx.doi.org/10.15439/2017F380>.