Comparing Multiple Survival Functions with Crossing Hazards in R

by Hsin-wen Chang, Pei-Yuan Tsai, Jen-Tse Kao and Guo-You Lan

Abstract It is frequently of interest in time-to-event analysis to compare multiple survival functions nonparametrically. However, when the hazard functions cross, tests in existing R packages do not perform well. To address the issue, we introduce the package **survELtest**, which provides tests for comparing multiple survival functions with possibly crossing hazards. Due to its powerful likelihood ratio formulation, this is the only R package to date that works when the hazard functions cross. We illustrate the use of the procedures in **survELtest** by applying them to data from randomized clinical trials and simulated datasets. We show that these methods lead to more significant results than those obtained by existing R packages.

Introduction

The nonparametric comparison of multiple survival functions is of interest in numerous biomedical settings, such as clinical trials (Robert et al., 2015), preclinical studies (Liebl et al., 2015) and observational studies (Loupy et al., 2013) with right-censored time-to-event endpoints. It has been implemented in existing R packages using log-rank-type statistics. However, these log-rank-type tests can fail to detect differences among survival curves when the hazard functions cross. For example, consider the Kaplan–Meier (KM) estimated survival functions in Figure 1 for the treatment and control groups of patients in a randomized clinical trial. There is a clear gap between the survival curves, which we would expect to be detected by a reasonable statistical test. Nevertheless, the log-rank test provided in the **survival** (Therneau et al., 2020) package returns a p-value of 0.07, indicating no significant difference between the two survival functions at $\alpha = 0.05$. Note that in this case the gap between the survival curves shrinks in the middle of the follow-up period, suggesting that the estimated hazard functions may cross at some time point.

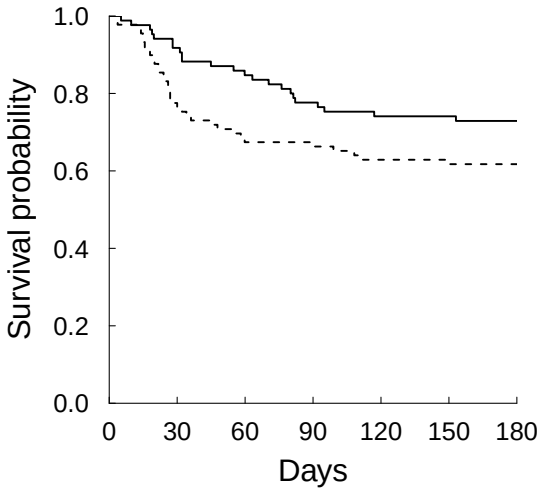

Figure 1: Estimated survival functions for treatment (solid line) versus control (dashed line) groups, based on a randomized clinical trial for treatment of severe alcoholic hepatitis (Nguyen-Khac et al., 2011).

The need to compare survival functions with crossing hazards has been documented in the statistics literature (see, e.g., Pepe and Fleming, 1989; Yang and Prentice, 2010). There are many practical situations in which the hazard functions cross, indicating that the instantaneous treatment effect changes direction. For example, some treatments are initially harmful due to toxicity or other complications, but may be beneficial later on. Other treatments can have short-term benefits but produce side effects in the long run. Despite the varying instantaneous treatment effect, the cumulative treatment effect can still be positive, as reflected by a positive difference between the treatment and control survival functions throughout the follow-up period. It is important to be able to detect such a difference, as the treatment would be worth considering in this case. To this end, an adaptive weighted

log-rank test was implemented in the R package **YPmodel** (Sun and Yang, 2015), but this test involves a parametric assumption on the hazard functions, is limited to two-sample comparisons, and cannot deal with the general k-sample situation. Another method, the restricted mean survival time (RMST) approach, was implemented in the R package **survRM2** (Uno et al., 2020). However, to our knowledge the RMST method can only deal with two-sample comparisons nonparametrically. For the k-sample case, certain model-based assumptions still need to be made (see, e.g., Cronin et al., 2016).

To address this issue, the package **survELtest** (Chang, 2020) provides nonparametric tests that can deal with general k-sample comparisons while accounting for possibly crossing hazards. It avoids the pitfalls of log-rank-type statistics, in which negative and positive differences between the estimated hazard functions cancel each other in a weighted sum (see Section Existing test statistics in R and their pitfalls for more details). Further, the statistics are constructed using empirical likelihood (EL), which has been shown to produce tests with optimal power (see, e.g., Kitamura et al., 2012). EL is a nonparametric likelihood which does not assume that the data come from a particular parametric family of distributions. It serves as the basis for constructing a nonparametric likelihood ratio (i.e., the EL ratio), which leads to more efficient inference than Wald-type procedures such as log-rank-type tests, as seen in the literature on parametric (see, e.g., Mukerjee, 1994) and nonparametric inference (Bravo, 2003; Kitamura et al., 2012, p. 116). There are R packages available for survival analysis using EL, namely **emplik** (Zhou, 2020), **emplik2** (Barton, 2018) and **ELYP** (Zhou, 2018), but they are limited to inference regarding finite-dimensional parameters, whereas our package handles survival functions, an infinite dimensional problem.

Our approach computes EL ratios at each observed uncensored time point, then summarizes them into maximal-deviation-type and integral-type statistics. The statistical theory of this approach and the empirical levels and powers in various simulation scenarios have been studied in Chang and McKeague (2016; 2019), but these authors focused on the technical development of one-sided tests, and did not provide a software package or an accessible guide for implementing the method. In this paper we provide a general framework for both two-sided and one-sided testing, an accessible guide to the R package **survELtest**, and a comparison with existing R packages (reviewed briefly in Section Existing test statistics in R and their pitfalls) in applications to data from clinical trials and simulated datasets.

For k-sample nonparametric testing under right censorship, to our knowledge all existing R packages use log-rank-type statistics (see the CRAN Task View *Survival*), often referred to as the weighted log-rank statistics. The package **FHtest** (Oller and Langohr, 2017) and the `survdiff` function in the package **survival** consider the Fleming-Harrington G^ρ family, which belongs to the class of weighted log-rank statistics. The package **clinfun** (Seshan, 2018) adopts a permutation version of the log-rank test. The package **LogrankA** (Richter-Dumke and Rau, 2013) provides a log-rank test based on aggregated survival data. `SurvivalTests` in the **coin** (Hothorn et al., 2019) package implements a reformulated weighted log-rank test as a linear rank test. The **maxstat** (Hothorn, 2017) package performs tests using maximally selected log-rank statistics.

This paper is organized as follows. In the next section, we provide a brief review of k-sample nonparametric methods used in existing R packages, their pitfalls, and the use of EL tests as a solution. Section Program description describes our package functions, along with a flow chart showing the procedure for using those functions. In Sections Application of supELtest to threearm data and Application of intELtest to hepatitis data, we apply the proposed routines to datasets from clinical trials, and obtain more significant results than the log-rank-type tests. Some concluding remarks are made in Section Discussion. The availability of the program is given in Section Availability. In the Appendix, we compare our procedures with more existing methods, including those in the aforementioned packages **YPmodel** and **survRM2**, which cannot deal with the general k-sample case nonparametrically.

Theoretical background

Existing test statistics in R and their pitfalls

This section briefly reviews the log-rank-type statistics in existing R packages, for testing whether the k survival functions are the same. The null and alternative hypotheses are $H_0 \colon S_1 = \ldots = S_k$ and $H_1 \colon H_0$ is not true, respectively, where S_j is the survival function of the j-th sample. To quantify the discrepancy between the j-th sample ($j = 1, \ldots, k-1$) and other samples, a weighted sum of differences between the estimated hazard function of the j-th sample and that of the pooled sample is computed. The $k-1$ weighted sums are then summarized using a quadratic form to obtain the final log-rank-type statistc. Different choices of the weight lead to different log-rank-type statistics, of which the commonly used log-rank test is a special case.

To illustrate the pitfalls of this formulation under crossing hazards, we restrict our attention to $k = 2$ for simplicity. When $k = 2$, there is only one weighted sum involved, which can be expressed as

$$\sum_{i=1}^{m} v_i \left(\hat{h}_1(t_i) - \hat{h}_2(t_i) \right), \tag{1}$$

where $0 < t_1 < \ldots < t_m < \infty$ are the (ordered) observed uncensored times, $\hat{h}_j(t_i)$ are the estimated hazard functions at time t_i, and v_i is the corresponding weight at t_i. When the survival functions are different, the hazard functions can cross each other. In this case, there are both positive differences (i.e., when $\hat{h}_1(\cdot) > \hat{h}_2(\cdot)$) and negative differences (i.e., when $\hat{h}_1(\cdot) < \hat{h}_2(\cdot)$) in (1). These differences between the estimated hazard functions cancel out, leading to a smaller value of the statistic and hence a less significant result. Consequently, the formulation can fail to detect the difference between the survival curves.

EL ratio and test statistics

In the proposed package **survELtest**, we use a likelihood ratio statistic, namely a pointwise EL statistic, to replace the estimated hazard difference in (1). This pointwise EL statistic quantifies, at each time point, the difference in the multiple survival functions. It is always positive, as are all typical likelihood ratio statistics, which prevents the problematic cancellation described in the previous section. In the rest of Section Theoretical background, we provide a brief description of this approach. More details can be found in Chang and McKeague (2016, 2019).

The pointwise EL statistic is constructed from the following likelihood ratio:

$$\mathcal{R}(t) = \frac{\sup \{L(S_1, S_2, \ldots, S_k) \colon S_1(t) = S_2(t) = \ldots = S_k(t)\}}{\sup \{L(S_1, S_2, \ldots, S_k)\}}, \tag{2}$$

where $L(S_1, S_2, \ldots, S_k)$ is a nonparametric likelihood which does not assume that the data come from a particular parametric family of distributions (Thomas and Grunkemeier, 1975). As in a usual (parametric) likelihood ratio, the numerator of (2) maximizes the likelihood subject to the constraint under H_0, whereas the denominator maximizes the likelihood globally, as it corresponds to the union of H_0 and H_1. We then use $-2 \log \mathcal{R}(t)$ as our pointwise EL statistic; such transformation of likelihood ratios has been widely used in the literature. A larger $-2 \log \mathcal{R}(t)$ gives less evidence for $S_1(t) = S_2(t) = \ldots = S_k(t)$.

For the desired simultaneous inference, we summarize the pointwise statistics in two ways. The first, provided by the routine `intELtest`, takes a weighted sum:

$$I = \sum_{i=1}^{m} w_i \{-2 \log \mathcal{R}(t_i)\}, \tag{3}$$

where w_i is the weight at each t_i. This is an integral-type statistic because the summation can be written into a stochastic integral. The form of a weighted sum is similar to the components of the log-rank-type statistics shown in the previous section. Here we avoid ad hoc choices of the weight w_i by setting equal weight for data with no ties. We do this because the EL statistic $-2 \log \mathcal{R}(t_i)$ implicitly provides optimal (i.e., nonparametric-likelihood-optimized) weighting for contrasting the survival functions. More details regarding the weighting schemes are given in Section Weight.

Another way to summarize the pointwise statistics is to take a maximum $K = \sup_{i=1,\ldots,m} \{-2 \log \mathcal{R}(t_i)\}$, which is provided by the function `supELtest`. Such maximal-deviation-type statistics have been used in the classical Kolmogorov–Smirnov test, and are more sensitive to local differences amongst survival curves (i.e., differences among survival curves that appear only in a short period of time). In contrast, the integral-type statistic I in (3) is designed to detect moderate differences spread over a sizable portion of the follow-up period. The choice between the two statistics should be guided by prior knowledge and practical considerations. In particular, if prior knowledge does not suggest the presence of local differences, we recommend I for general use. Otherwise K can be implemented to exploit the additional knowledge of the existence of a local difference. For example, a local difference is present when medical knowledge suggests that a treatment has a benefit in some localized time interval, or when a pilot study shows that the difference among the KM estimated survival curves appears only over a short period of time. In the latter case, the evidence would be even stronger if a significant result was obtained from a statistical test that is sensitive to such local differences, such as the maximal-deviation-type statistics described above.

Two-step procedure for one-sided testing

So far, we have focused on two-sided testing. For one-sided testing, we consider the alternative H_1^o: $S_1 \succ S_2 \succ \ldots \succ S_k$ that there is an ordering among the survival functions, where $f \succ g$ for functions $f(t)$ and $g(t)$ of t means $f(t) \geq g(t)$ for all t with a strict inequality for some t. The EL statistics are the same as the ones in the previous section, except that now we put an additional constraint $S_1(t) \geq S_2(t) \geq \ldots \geq S_k(t)$ in the denominator of $\mathcal{R}(t)$ in (2).

The resulting EL test will be preceded by an initial test that excludes the possibility of crossings or alternative orderings among the survival functions. The reason is due to the fact that for functional parameters (e.g., survival functions), testing a one-sided alternative hypothesis usually involves certain assumptions, such as that the functions are not crossed and that their ordering is as hypothesized. These assumptions may be checked using the initial test, with the null hypothesis that the assumptions are not satisfied, versus the alternative that they are. If the null hypothesis of the initial test is rejected, we conclude that the assumptions are satisfied and proceed to the EL test. Rejection of the null hypothesis H_0 of the EL test then gives support for H_1. On the other hand, if the null hypothesis of the initial test is not rejected, we conclude that the assumptions are not satisfied and do not proceed to the EL test. The family-wise error rate of this two-step procedure has been shown to be asymptotically controlled at the same α-level as the individual tests.

Weight

As mentioned in Section EL ratio and test statistics, we need to specify w_i in (3). This can be done by setting the value of the argument `wt` in the routine `intELtest`. The default is an objective weight $w_i = d_i/n$, where d_i denotes the number of events at each time point t_i and n is the total sample size. This simplifies to equal weight $w_i = 1/n$ when there are no ties (i.e., $d_i = 1$) in the data. This default weight is specified by the option `wt = "p.event"`.

Despite the default weight, we provide in `intELtest` two alternative options that have been used for integral-type statistics in the literature. One option is $w_i = \hat{F}(t_i) - \hat{F}(t_{i-1})$ for $i = 1, \ldots, m$, where $\hat{F}(t) = 1 - \hat{S}(t)$, $\hat{S}(t)$ is the pooled KM estimator, and $t_0 \equiv 0$. This w_i reduces to the objective weight $w_i = d_i/n$ when there is no censoring (see, e.g., El Barmi and McKeague, 2013). The resulting I can be seen as an empirical version of the expected negative two EL ratio under H_0. This weight can be chosen via the option `wt = "dF"`.

Another weight, proposed by Pepe and Fleming (1989), is $w_i = t_{i+1} - t_i$ for $i = 1, \ldots, m$, where $t_{m+1} \equiv t_m$. This approach gives more weight to the time intervals when there are fewer observed uncensored times, but can be affected by extreme observations. This weight can be chosen via the option `wt = "dt"`.

Bootstrap critical values

Having computed the statistics, we need to calibrate the tests. Possible methods include bootstrapping or simulating the limiting distributions. We choose the former for the following two reasons: (a) in small samples, calibration using the bootstrap can perform better than using the limiting distribution (Heller and Venkatraman, 1996). (b) in our experience the bootstrap can be more computationally efficient, since the limiting distributions of I and K involve stochastic processes that depend on unknown parameters.

Here we adopt a Gaussian multiplier bootstrap approach, which is commonly used instead of the nonparametric bootstrap in survival analysis to avoid producing tied data in the bootstrap samples. To form the bootstrap samples, the original data are perturbed using independent standard Gaussian random variables, termed Gaussian multipliers (see, e.g., Parzen et al., 1997). We denote the number of bootstrap samples as B, which is specified by the `nboot` argument (default is 1000). In cases when m is too large for the computation to be handled with reasonable speed, we split the calculation of the B bootstrap replications into `nsplit` parts, where `nsplit` = $\lceil m$ `/ nlimit` $\rceil$ (default `nlimit = 200`). Here and in the sequel, if we do not specify which R function an option or argument is applied to, the option or argument applies to all the functions provided by the **survELtest** package.

Since the bootstrap involves random sampling, the critical values will differ based on different sets of bootstrap samples. To make the critical values reproducible, we set a seed for random number generation via the `seed` option in our routines.

User guide and numerical examples

Program description

The **survELtest** package can be installed along with **survELtest** using the following R code:

```
>   install.package("survELtest")
```

The following code loads the package:

```
>   library(survELtest)
```

The main routines in **survELtest** are `intELtest`, `supELtest`, `nocrossings`, and `ptwiseELtest`. The `intELtest` routine conducts testing based on the integrated EL statistics I in (3) that can detect moderate differences among the survival curves over time. The `supELtest` routine conducts testing based on the maximally selected EL statistics K that is more sensitive to differences locally in time. Both routines give two- or one-sided test statistics, the critical value based on bootstrap, and the p-value of the test. As mentioned in Section EL ratio and test statistics, the choice between the two routines should be guided by prior knowledge and practical considerations regarding whether there is a local difference among the survival curves. The choice between two-sided and one-sided testing should be determined a priori as well, depending on the research question of interest. One-sided testing can be specified by the option `sided = 1` in both `intELtest` and `supELtest`, but should be preceded by the initial test in `nocrossings` to exclude the possibility of crossings or alternative orderings among the survival functions. While the first three routines provide simultaneous testing, `ptwiseELtest` conducts pointwise testing to compare the survival curves at each time point. It can be used to identify periods of local differences, after `intELtest` or `supELtest` test gives a significant result. A flow chart of the procedure for using the **survELtest** package is given in Figure 2. Methods defined for the resulting objects of the main routines are provided for `print` and `summary`. In addition to the aforementioned routines, **survELtest** contains four datasets: `hepatitis`, `threearm`, `hazardcross` and `hazardcross_Weibull`, which will be analyzed in Sections Application of supELtest to threearm data, Application of intELtest to hepatitis data, and the Appendix to illustrate the use of the routines.

A summary of the R code and the input arguments of the routines is given as follows. Among the input arguments below, only the `formula` input is compulsory. The rest of the arguments can be omitted if the default settings are used.

```
>   intELtest(formula, data = NULL, group_order = NULL, t1 = 0, t2 = Inf, sided = 2,
+   nboot = 1000, wt = "p.event", alpha = 0.05, seed = 1011, nlimit = 200)

>   supELtest(formula, data = NULL, group_order = NULL, t1 = 0, t2 = Inf, sided = 2,
+   nboot = 1000, alpha = 0.05, seed = 1011, nlimit = 200)

>   nocrossings(formula, data = NULL, group_order = NULL, t1 = 0, t2 = Inf, sided = 2,
+   nboot = 1000, alpha = 0.05, seed = 1011, nlimit = 200)

>   ptwiseELtest(formula, data = NULL, group_order = NULL, t1 = 0, t2 = Inf, sided = 2,
+   nboot = 1000, alpha = 0.05, seed = 1011, nlimit = 200)
```

The time needed to run these functions depend on the total number n of observations, the number k of samples, the speed of the processor and the amount of PC memory. For example, to run `intELtest` with the default settings, it takes about 0.32 seconds on the dataset `hepatitis` with $n = 174$ and $k = 2$, and 1.79 minutes on the dataset `threearm` with $n = 664$ and $k = 3$, on a desktop computer with Intel i7-7700 CPU @ 3.60 GHz and 64 GB RAM.

- `formula`: a formula object with a Surv object as the response on the left of the $\sim$ operator and the grouping variable as the term on the right. The Surv object involves two variables: the observed survival and censoring times, and the censoring indicator, which takes a value of 1 if the observed time is uncensored and 0 otherwise. The grouping variable takes different values for different groups.
- `data`: an optional data frame containing the variables in the formula: the observed survival and censoring times, the censoring indicator, and the grouping variable. If not found in `data`, the variables in the formula should be already defined by the user or in attached R objects. The default is the data frame with three columns of variables taken from the formula: column 1 contains the observed survival and censoring times, column 2 the censoring indicator, and column 3 the grouping variable.

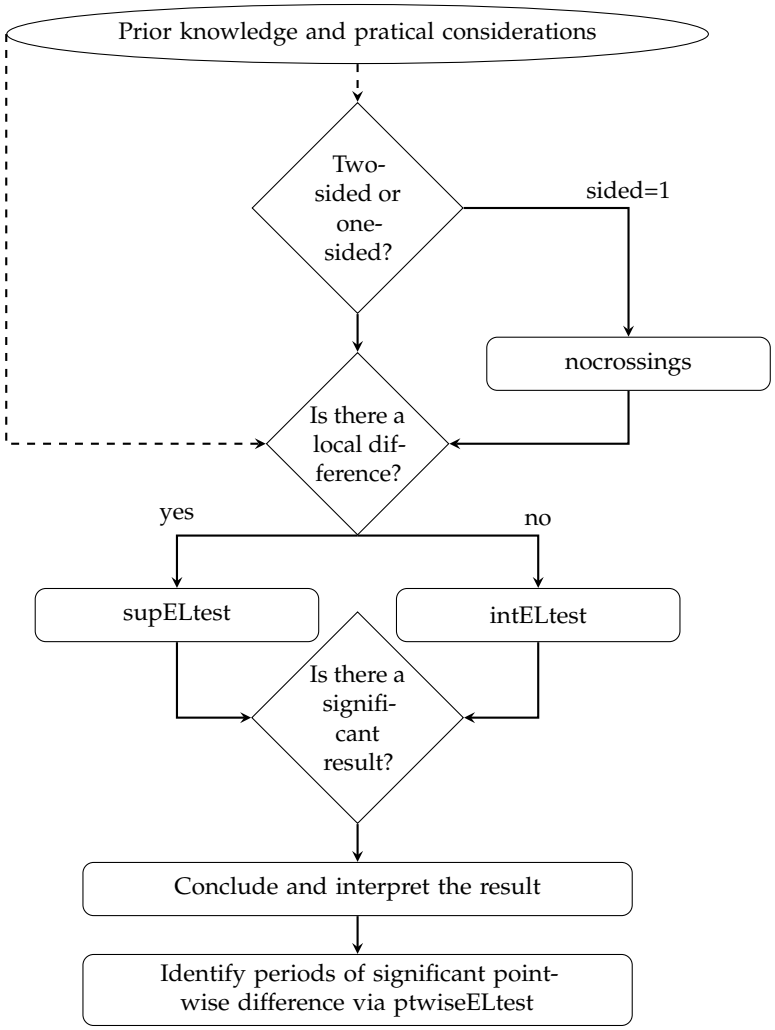

Figure 2: Flow chart of the procedure for using the routines in the **survELtest** package.

- `group_order`: a k-vector containing the values of the grouping variable, with the j-th element being the group hypothesized to have the j-th highest survival rates, $j = 1, \ldots, k$. The default is the vector of sorted grouping variables.
- `t1`: the first endpoint of a prespecified time interval, if any, to which the comparison of the survival functions is restricted. The default value is 0.
- `t2`: the second endpoint of a prespecified time interval, if any, to which the comparison of the survival functions is restricted. The default value is ∞.
- `sided`: 2 if two-sided test, and 1 if one-sided test. The default value is 2.
- `nboot`: the number of bootstrap replications in calculating critical values for the tests. The default value is 1000.
- `wt`: the name of the weight for the integrated EL statistics in `intELtest`: `"p.event"`, `"dF"`, or `"dt"`. The default is `"p.event"`.
- `alpha`: the pre-specified significance level of the tests. The default value is 0.05.
- `seed`: the seed for the random number generator in R, for generating bootstrap samples needed to calculate the critical values for the tests. The default value is 1011.
- `nlimit`: a number used to calculate `nsplit =` $\lceil m$ `/ nlimit`$\rceil$, the number of parts into which the calculation of the `nboot` bootstrap replications is split. The use of this variable can make computation faster when the number of time points m is large. The default value for `nlimit` is 200.

Application of supELtest to threearm data

In this section we apply the routines `supELtest` and `ptwiseELtest` to the dataset `threearm` provided in the **survELtest** package, and compare the results with the log-rank-type tests for trend. The

dataset is obtained by resampling from a perturbed dataset of patients from a randomized clinical trial for the treatment of major depression, where the perturbation is achieved by adding a random $U(-0.01\ell, 0.01\ell)$ variable to existing observations, ℓ is the minimal observation in the original data, and the resampling is done by conditional bootstrapping with stratified survival and censoring distributions using the `censboot` function in the package **boot** (Canty and Ripley, 2020). The original data were analyzed by Chang and McKeague (2019), who observed a local difference among the survival functions.

The purpose of analyzing the `threearm` dataset is to assess whether the survival functions of the three arms are ordered: that is, whether the experimental treatment group ($n_1 = 262$) is better than the standard treatment group ($n_2 = 267$), which is in turn superior to the placebo group ($n_3 = 135$). This question can be answered using the one-sided tests described in Section Two-step procedure for one-sided testing . Since prior knowledge suggests there is a local difference among the survival functions, here we conduct the maximally selected EL test via `supELtest`.

The endpoint of the clinical trial is time (in days) to first remission. Because a shorter time to first remission is desirable, a lower value of the survival function indicates a better treatment in this dataset. Based on this information, from the KM estimated survival curves in the left panel of Figure 3, it seems that the three groups are similar initially but become ordered for the rest of the follow-up period.

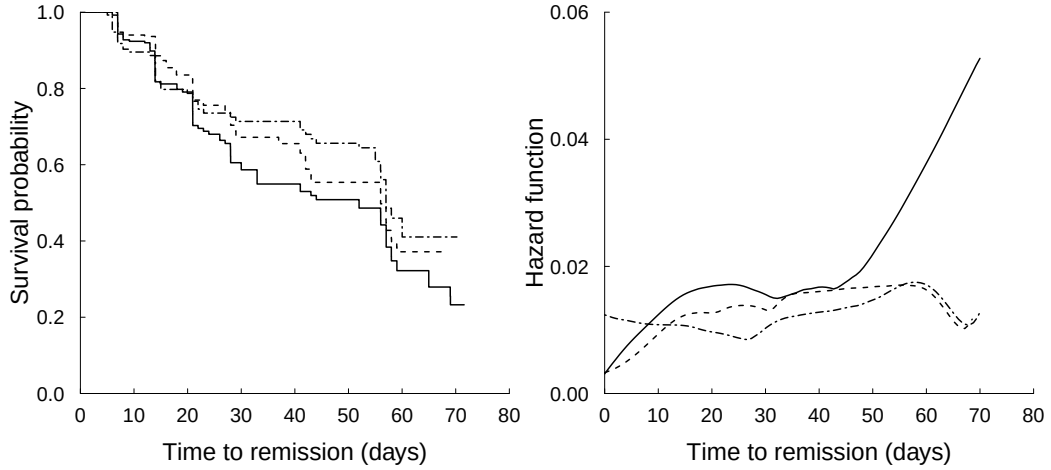

Figure 3: The KM estimated survival curves (left) and the estimated hazard functions (right) in the `threearm` dataset: experimental treatment group (solid), standard treatment group (dashed) and placebo group (two-dashed).

To see if the curves are statistically significantly ordered, we start with conducting the commonly used log-rank-type tests. The trend test is needed for the one-sided research question. Using the common choice $c(3, 2, 1)$ for the score vector (see, e.g. Andersen et al., 1993, page 388), the log-rank test for trend is implemented as follows:

```
>   library(survival)
>   dat = Surv(threearm[, 1], threearm[, 2])
>   logrank = survdiff(dat ~ threearm[, 3])
>   score_vec = 3 : 1
>   logrankteststat = matrix(score_vec, nrow = 1, ncol = 3)
+   %*% (logrank$obs - logrank$exp) / sqrt(matrix(score_vec, nrow = 1, ncol = 3)
+   %*% (logrank$var) %*% matrix(score_vec, nrow = 3, ncol = 1))
>   if(logrankteststat < 0){
+     pval = 2 * pnorm(logrankteststat)
+   }else{
+     pval = 2 * (1 - pnorm(logrankteststat))
+   }
>   round(pval, 2)

     [,1]
[1,] 0.04
```

As the log-rank test for trend gives a p-value of 0.04, we conclude that the three survival functions are ordered at $\alpha = 0.05$. The other extreme in the G^ρ family can be implemented by setting `survdiff(dat`

$\sim$ `threearm[, 3], rho = 1)` in the above code, which leads to a p-value of 0.08. These results mean the weighted log-rank statistics in the entire G^ρ family give a p-value that ranges from 0.04 to 0.08 for the trend test.

Now we conduct the proposed one-sided testing for the `threearm` data. We anticipate a more significant result than the log-rank-type tests, as there seems to be crossing among the estimated hazard functions in the right panel of Figure 3, created using the function `muhaz` in the package **muhaz** (Hess and Gentleman, 2019) with the default settings. The initial test and the maximally selected EL test are implemented by the routines `nocrossings` and `supELtest`, respectively. (Note that if two-sided testing is conducted instead, then the initial test is not needed.) To use the routines, we need to specify two options: `sided = 1` for the one-sided test, and `group_order = c(3, 2, 1)`, since the hypothesized order of the survival rates from the largest to the smallest is the placebo (coded as 3 in our data matrix), standard treatment (coded as 2), and experiment treatment (coded as 1). The rest of the options are kept at their default values. The R code for performing the initial test is as follows:

```
>   nocrossings(Surv(threearm$time, threearm$censor) ~ threearm$group,
+   group_order = c(3, 2, 1), sided = 1)

Call:
nocrossings(formula = Surv(threearm$time, threearm$censor) ~ threearm$group,
group = c(3, 2, 1), sided = 1)

Decision = 1
```

A decision value of 1 means there is no crossing or alternative orderings among the survival functions. Thus, we can proceed to the main (maximally selected EL) test in the second step:

```
>   supELtest(Surv(threearm$time, threearm$censor) ~ threearm$group,
+   group_order = c(3, 2, 1), sided = 1)

Call:
supELtest(formula = Surv(threearm$time, threearm$censor) ~ threearm$group,
group = c(3, 2, 1), sided = 1)

One-sided maximally selected EL test statistic = 14.23, p = 0.004
```

As the maximally selected EL test gives a p-value < 0.01, we obtain the same conclusion—that the three survival functions are significantly ordered—as the log-rank-type tests for trend, but with a statistically more significant result. This finding is as we anticipated after seeing the crossing estimated hazard functions in the right panel of Figure 3.

Since our procedure leads to the conclusion that the survival functions are ordered, it can be of interest to identify periods of local differences for further clinical investigation. To this end, we can use the routine `ptwiseELtest` for pointwise testing at each observed uncensored time point:

```
>   ptwise = ptwiseELtest(Surv(threearm$time, threearm$censor) ~ threearm$group,
+   group_order = c(3, 2, 1), sided = 1)
```

The list of the time points at which the survival functions are ordered (i.e., `decision == 1`) is obtained by

```
>   round(ptwise$result_dataframe$time_pts[ptwise$result_dataframe$decision == 1], 2)

 [1] 13.91 13.91 13.91 13.92 13.92 13.92 13.92 13.93 13.98 13.99 13.99 14.00 14.00
[14] 14.00 14.01 14.01 20.96 20.96 27.98 27.99 28.00 28.00 28.00 28.02 28.02 28.98
[27] 28.99 29.01 30.00 32.96 36.97 40.97 40.98 40.99 41.02 41.98 41.98 41.99 42.00
[40] 42.01 42.02 42.99 43.01 43.02 43.02 44.00 44.00 51.97 51.97 55.00 56.01 56.01
[53] 56.02 59.03 59.04 64.97 68.98 69.01
```

From the result, we see there are local differences occuring near the time points 14, 21, 30, 40, 52, 56, 59, 65 and 69 days.

Application of intELtest to hepatitis data

Now we turn to our motivating example in the Introduction and demonstrate the use of `intELtest` and its benefit over the log-rank-type tests. The corresponding dataset `hepatitis` is provided in the **survELtest** package. The dataset was obtained by reconstructing survival and censoring information (Guyot et al., 2012) based on digitizing the KM curves presented in Nguyen-Khac et al. (2011). It contains survival data (in days, rounded to one decimal place) from patients in a randomized clinical trial for the treatment of severe alcoholic hepatitis. The purpose of the clinical trial was to assess if the treatment group ($n_1 = 85$) had a significantly different survival rate than the control group ($n_2 = 89$).

From the KM estimated survival curves in Figure 1, the survival rate of the treatment group seems to be greater than that of the control group over the entire follow-up period. To see whether the difference between the survival functions are statistically significant, we start with conducting the commonly used two-sided log-rank test:

```
>   library(survival)
>   dat = Surv(hepatitis[, 1], hepatitis[, 2])
>   logrank = survdiff(dat ~ hepatitis[, 3])
>   round(1 - pchisq(logrank$chisq, df = 1), 2)

[1] 0.07
```

The log-rank test gives a p-value of 0.07, failing to detect a difference between the survival curves at $\alpha = 0.05$. The reason may be due to the crossing estimated hazard functions in Figure 4 (created using the function `muhaz` in the package **muhaz** with the default settings). We also conduct another log-rank-type test—the Peto and Peto's modification of the Gehan-Wilcoxon test—by setting `survdiff(dat ~ hepatitis[, 3], rho = 1)` in the above code, which leads to a p-value of 0.05. Since this test and the log-rank test are the two extremes in the G^ρ family, these results mean the weighted log-rank statistics in the entire G^ρ family give either insignificant or borderline significant conclusions.

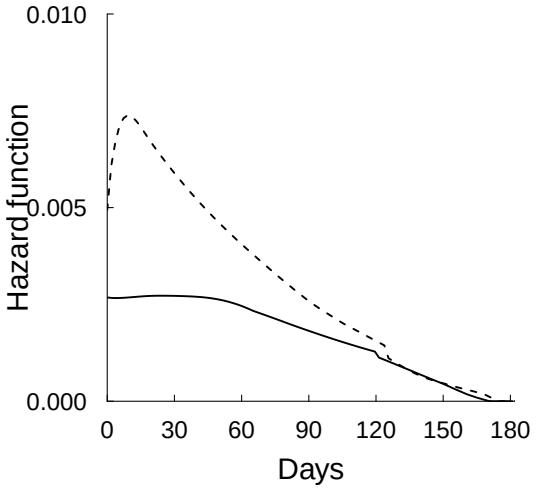

Figure 4: Estimated hazard functions for treatment (solid line) versus control (dashed line) groups.

Now we apply the proposed two-sided integrated EL test to the `hepatitis` data to see if we can better detect a difference between the survival functions. The default options are used and the R code is as simple as

```
>   intELtest(Surv(hepatitis$time, hepatitis$censor) ~ hepatitis$group)

Call:
intELtest(formula = Surv(hepatitis$time, hepatitis$censor) ~ hepatitis$group)

Two-sided integrated EL test statistic = 1.42, p = 0.007
```

As the integrated EL test gives a p-value of 0.01, we conclude there is a significant difference between the two survival functions at $\alpha = 0.05$. The p-value is much smaller than those given by the previous

log-rank-type tests, which indicates that the integrated EL test is better at detecting the difference between the survival curves.

Note the decision as to whether there is a significant discrepancy between the two survival functions is totally different for the log-rank and the integrated EL tests at $\alpha = 0.05$. It may be tempting to pick the most significant result, but this practice is data snooping and has been shown to be problematic. Instead, we recommend setting a primary method prior to the data analysis and making the decision based on that method. Any other methods are treated as secondary, and their results can serve an exploratory purpose for future work.

Discussion

In this paper we introduce the R package **survELtest** for comparing two or more survival functions nonparametrically based on right-censored data. It is the only R package to date that utilizes the powerful likelihood ratio formulation instead of log-rank-type statistics, thereby performing well when the hazard functions cross. We provide both maximal-deviation-type and integral-type statistics, for detecting local and cumulative differences among the survival functions, respectively.

The use of the software is illustrated using two data sets from randomized clinical trials, where the estimated survival functions seem to be ordered, but the estimated hazard functions cross. In these cases, our procedures lead to more significant results than the results obtained from the log-rank-type tests. Specifically, in one of the examples, the original clinical trial concludes that there is no significant difference between the treatment and the control groups (log-rank $p = 0.07$), whereas our test suggests otherwise, based on a much smaller p-value of 0.01. We envision the **survELtest** package will be valuable for finding more significant results in numerous biomedical settings involving the comparison of multiple survival functions, especially in the presence of crossing hazards.

Availability

The package is available from the Comprehensive R Archive Network at https://CRAN.R-project.org/package=survELtest. The development website is available at https://github.com/news11/survELtest.

Acknowledgements

The research of Hsin-wen Chang was partially supported by Ministry of Science and Technology of Taiwan under grants 106-2118-M-001-015-MY3 and MOST 109-2118-M-001-005-. The authors thank Yu-Ju Wang for computational support and Shih-Hao Huang for helpful comments. The authors declare that they have no conflict of interest.

Bibliography

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. New York: Springer, 1993. URL https://doi.org/10.1007/978-1-4612-4348-9. [p7]

W. H. Barton. *emplik2: Empirical Likelihood Ratio Test for Two Samples with Censored Data*, 2018. URL https://cran.r-project.org/package=emplik2. R package version: 1.21. [p2]

F. Bravo. Second-order power comparisons for a class of nonparametric likelihood-based tests. *Biometrika*, 90(4):881–890, 2003. URL https://doi.org/10.1093/biomet/90.4.881. [p2]

A. Canty and B. Ripley. *boot: Bootstrap Functions (Originally by Angelo Canty for S)*, 2020. URL https://CRAN.R-project.org/package=boot. R package version: 1.3-25. [p7]

H.-w. Chang. *survELtest: Comparing Multiple Survival Functions with Crossing Hazards*, 2020. URL https://CRAN.R-project.org/package=surELtest. R package version: 2.0.1. [p2]

H.-w. Chang and I. W. McKeague. Empirical likelihood based tests for stochastic ordering under right censorship. *Electronic Journal of Statistics*, 10(2):2511–2536, 2016. URL https://doi.org/10.1214/16-EJS1180. [p2, 3]

H.-w. Chang and I. W. McKeague. Nonparametric testing for multiple survival functions with non-inferiority margins. *Annals of Statistics*, 47(1):205–232, 2019. URL https://doi.org/10.1214/18-AOS1686. [p2, 3, 7]

A. Cronin, L. Tian, and H. Uno. strmst2 and strmst2pw: New commands to compare survival curves using the restricted mean survival time. *Stata Journal*, 16(3):702–716, 2016. URL https://doi.org/10.1177/1536867X1601600310. [p2]

H. El Barmi and I. W. McKeague. Empirical likelihood based tests for stochastic ordering. *Bernoulli*, 19: 295–307, 2013. URL https://doi.org/10.3150/11-BEJ393. [p4]

P. Guyot, A. E. Ades, M. J. N. M. Ouwens, and N. J. Welton. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan–Meier survival curves. *BMC Medical Research Methodology*, 12(1):1–13, 2012. URL https://doi.org/10.1186/1471-2288-12-9. [p9]

G. Heller and E. S. Venkatraman. Resampling procedures to compare two survival distributions in the presence of right-censored data. *Biometrics*, 52(4):1204–1213, 1996. URL https://doi.org/10.2307/2532836. [p4]

K. Hess and R. Gentleman. *muhaz: Hazard Function Estimation in Survival Analysis*, 2019. URL https://CRAN.R-project.org/package=muhaz. R package version: 1.2.6.1. [p8]

T. Hothorn. *maxstat: Maximally Selected Rank Statistics*, 2017. URL https://cran.r-project.org/package=maxstat. R package version: 0.7-25. [p2]

T. Hothorn, H. Winell, K. Hornik, M. A. van de Wiel, and A. Zeileis. *coin: Conditional Inference Procedures in a Permutation Test Framework*, 2019. URL https://cran.r-project.org/package=coin. R package version: 1.3-1. [p2]

Y. Kitamura, A. Santos, and A. M. Shaikh. On the asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*, 80(1):413–423, 2012. URL https://doi.org/10.3982/ECTA8773. [p2]

M. Liebl, J. Windschmitt, A. S Besemer, A.-K. Schäfer, H. Reber, C. Behl, and A. Clement. Low-frequency magnetic fields do not aggravate disease in mouse models of Alzheimer's disease and amyotrophic lateral sclerosis. *Scientific Reports*, 5:8585, 2015. URL https://doi.org/10.1038/srep08585. [p1]

A. Loupy, C. Lefaucheur, D. Vernerey, C. Prugger, J.-P. D. van Huyen, N. Mooney, C. Suberbielle, V. Frémeaux-Bacchi, A. Méjean, F. Desgrandchamps, D. Anglicheau, D. Nochy, D. Charron, J.-P. Empana, M. Delahousse, C. Legendre, D. Glotz, G. S. Hill, A. Zeevi, and X. Jouven. Complement-binding anti-HLA antibodies and kidney-allograft survival. *New England Journal of Medicine*, 369 (13):1215–1226, 2013. URL https://doi.org/10.1056/NEJMoa1302506. [p1]

R. Mukerjee. Comparison of tests in their original forms. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 56(1):118–127, 1994. URL www.jstor.org/stable/25050974. [p2]

E. Nguyen-Khac, T. Thevenot, M.-A. Piquet, S. Benferhat, O. Goria, D. Chatelain, B. Tramier, F. Dewaele, S. Ghrib, M. Rudler, N. Carbonell, H. Tossou, A. Bental, B. Bernard-Chabert, and J.-L. Dupas. Glucocorticoids plus N-acetylcysteine in severe alcoholic hepatitis. *New England Journal of Medicine*, 365(19):1781–1789, 2011. URL https://doi.org/10.1056/NEJMoa1101214. [p1, 9]

R. Oller and K. Langohr. *FHtest: Tests for Right and Interval-Censored Survival Data Based on the Fleming-Harrington Class*, 2017. URL https://cran.r-project.org/package=FHtest. R package version: 1.4. [p2]

M. I. Parzen, L. J. Wei, and Z. Ying. Simultaneous confidence intervals for the difference of two survival functions. *Scandinavian Journal of Statistics*, 24(3):309–314, 1997. URL https://doi.org/10.1111/1467-9469.t01-1-00065. [p4]

M. S. Pepe and T. R. Fleming. Weighted Kaplan–Meier statistics: a class of distance tests for censored survival data. *Biometrics*, 45(2):497–507, 1989. URL https://doi.org/10.2307/2531492. [p1, 4]

J. Richter-Dumke and R. Rau. *LogrankA: Logrank Test for Aggregated Survival Data*, 2013. URL https://cran.r-project.org/package=LogrankA. R package version:1.0. [p2]

C. Robert, B. Karaszewska, J. Schachter, P. Rutkowski, A. Mackiewicz, D. Stroiakovski, M. Lichinitser, R. Dummer, F. Grange, L. Mortier, V. Chiarion-Sileni, K. Drucis, I. Krajsova, A. Hauschild, P. Lorigan, P. Wolter, G. V. Long, K. Flaherty, P. Nathan, A. Ribas, A.-M. Martin, P. Sun, W. Crist, J. Legos, S. D. Rubin, S. M. Little, and D. Schadendorf. Improved overall survival in melanoma with combined dabrafenib and trametinib. *New England Journal of Medicine*, 372(1):30–39, 2015. URL https://doi.org/10.1056/NEJMoa1412690. [p1]

V. E. Seshan. *clinfun: Clinical Trial Design and Data Analysis Functions*, 2018. URL https://cran.r-project.org/package=clinfun. R package version: 1.0.15. [p2]

J. Sun and S. Yang. *YPmodel: The Short-Term and Long-Term Hazard Ratio Model for Survival Data*, 2015. URL https://CRAN.R-project.org/package=YPmodel. R package version: 1.3. [p2]

T. M. Therneau, T. Lumley, E. Atkinson, and C. Crowson. *survival: Survival Analysis*, 2020. URL https://CRAN.R-project.org/package=survival. R package version: 3.2-3. [p1]

D. R. Thomas and G. L. Grunkemeier. Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70:865–871, 1975. URL https://doi.org/10.1080/01621459.1975.10480315. [p3]

H. Uno, L. Tian, A. Cronin, C. Battioui, and M. Horiguchi. *survRM2: Comparing Restricted Mean Survival Time*, 2020. URL https://CRAN.R-project.org/package=survRM2. R package version: 1.0-3. [p2]

S. Yang and R. Prentice. Improved logrank-type tests for survival data using adaptive weights. *Biometrics*, 66(1):30–38, 2010. URL https://doi.org/10.1111/j.1541-0420.2009.01243.x. [p1]

M. Zhou. *ELYP: Empirical Likelihood Analysis for the Cox Model and Yang-Prentice (2005) Model*, 2018. URL https://cran.r-project.org/package=ELYP. R package version: 0.7-5. [p2]

M. Zhou. *emplik: Empirical Likelihood Ratio for Censored/Truncated Data*, 2020. URL https://cran.r-project.org/package=emplik. R package version: 1.1-1. [p2]

Hsin-wen Chang
Institute of Statistical Science
Academia Sinica
128 Academia Road, Section 2,
Nankang, Taipei 11529, Taiwan (R.O.C)
ORCiD: 0000-0003-4566-7047
hwchang@stat.sinica.edu.tw

Pei-Yuan Tsai
Institute for Information Industry
Taipei, Taiwan (R.O.C)

Jen-Tse Kao
Institute of Statistical Science
Academia Sinica
Taipei, Taiwan (R.O.C)

Guo-You Lan
Department of Economics
National Chengchi University
Taipei, Taiwan (R.O.C)

Appendix: Comparison of survELtest with other existing tests in two simulated datasets

Here we provide two more examples for comparing our procedures with other existing tests in the literature, namely the log-rank test, the Peto and Peto's modification of the Gehan-Wilcoxon test, the adaptive weighted log-rank test implemented in the R package **YPmodel**, and the RMST method implemented in the R package **survRM2**. Since the latter two methods cannot deal with the general k-sample case nonparametrically, the examples provided here are restricted to the two-sample case. For the first dataset `hazardcross`, the survival time is generated from the piecewise exponential model displayed in the left panel of Figure 5. Since the difference between the true survival curves appears only during $[0, 6]$ but not later on, we use `supELtest` to detect such local differences. For the second dataset `hazardcross_Weibull`, the survival time is generated from the Weibull model displayed in the right panel of Figure 5. We use `intELtest` because the difference between the true survival curves is spread over the entire follow-up period. For both datasets, the true hazard functions cross, but there is an obvious gap between the survival curves. The censoring distributions are specified to be the same in each arm, and uniform with administrative censoring at t = 10 and a censoring rate of 25% in the first group. In implementing the tests, we use the default settings given in the aforementioned two packages.

The results are given in Table 1. Our tests provide more significant results in detecting the gap between the survival curves than any of the other tests for both datasets.

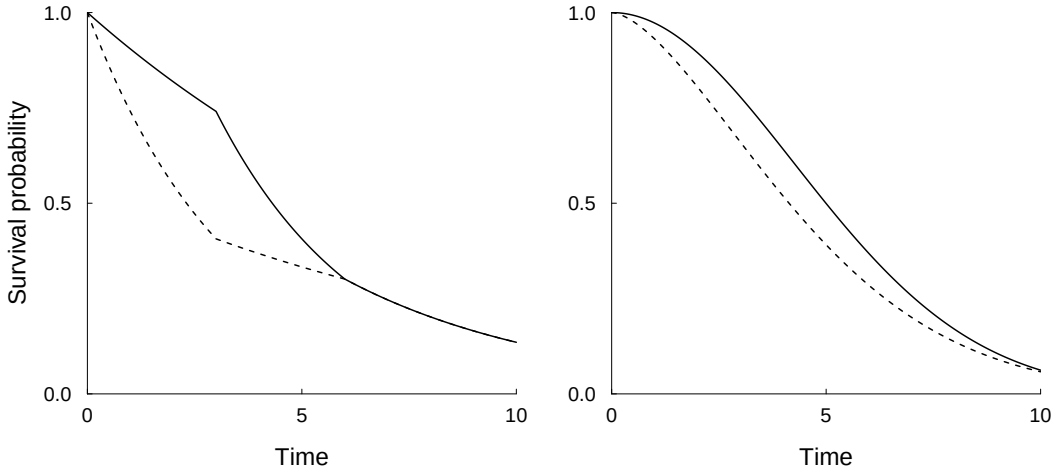

Figure 5: The true survival curves for generating `hazardcross` (left) and `hazardcross_Weibull` (right) datasets: the first (solid) and second (dashed) group.

Table 1: p-values from various tests for comparing the survival curves of the `hazardcross` and `hazardcross_Weibull` datasets. EL denotes the suitable EL test implemented in the R package **survELtest**, PP denotes the Peto and Peto's modification of the Gehan-Wilcoxon test, YP denotes the adaptive weighted log-rank test implemented in the R package **YPmodel**, and dRMST and rRMST denote the results in the R package **survRM2** for difference in and ratio of RMST, respectively.

Datasets	EL	log-rank	PP	YP	dRMST	rRMST
`hazardcross`	0.037	0.106	0.060	0.096	0.126	0.130
`hazardcross_Weibull`	0.005	0.080	0.006	0.009	0.014	0.018