Quaderni di Statistica
Vol. 12, 2010

Combination-based permutation testing in survival analysis

Rosa Arboretti Giancristofaro
Department of Territory and Agri-Forestal Systems, University of Padova
E-mail: rosa.arboretti@unipd.it

Mario Bolzan
Department of Statistics, University of Padova
E-mail: mario.bolzan@unipd.it

Federico Campigotto Livio Corain Luigi Salmaso
Department of Management and Engineering, University of Padova
E-mail: federico.campigotto@gmail.com, livio.corain@unipd.it,
luigi.salmaso@unipd.it

Summary: In survival analysis, it is very common to test whether two survival time distributions are equal. In the framework of random censorship model, two of the most frequently used procedures in order to compare samples of right-censored survival data are the asymptotic weighted log-rank test (WLR; Mantel, 1966) and the weighted Kaplan-Meier test (WKM; Kaplan and Meier, 1958). In this work we present a novel permutation combination-based testing approach for survival analysis. Within permutation methodology, censored data problems can be thought within a missing data setting. In this sense, it is possible to take into consideration a Multidimensional Permutation Test based on the theory of permutation testing with missing data. A comparative Monte Carlo simulation study has been performed, along with an application to a real case study, in order to evaluate the behaviour of the permutation procedures, with respect to some other asymptotic nonparametric methods proposed in the recent literature. The aim of this work is to point out a possible flexible and robust procedure, in terms of power, among the investigated methods. In general, the achieved results mainly suggest the use of a multidimensional permutation methodology in case of equal censoring.

Keywords: Survival analysis, Right and unequal censoring data, Permutation tests.

1. Introduction and notation

This work is about permutation methodologies for hypothesis testing problems within the context of survival analysis. The theme of survival analysis embraces methods related to the analysis of data on events observed over a certain period of time, and related to the investigation of predictors associated with the occurrence rates of this endpoint. In this framework, a failure (a death, for instance) is considered the event of interest in the statistical analysis. Thus, the statistical comparison is focused on failure time data which come up when statistical units are exposed to the hazard of failure under different experimental conditions. In this work we make the assumption that an individual can have an event at most once.

Now, let us assume that the experiment have is based on the comparison of independent samples of size n_1 and n_2, respectively, where $n = n_1 + n_2 \in \mathbb{N}$ is the pooled sample size of the study. Thus, we define $(\Omega^{(n)}, \mathcal{B}^{(n)}, P^{(n)})$ as a given sequence of fixed probability spaces, and $\left\{\mathcal{B}_t^{(n)} : 0 \leq t < \infty\right\}$ as a family of right-continuous, non-decreasing complete sub-σ-algebras , which correspond to the history of survival up to and including time t.

Usually in the type II censoring model, T_{mj}, $m = 1, \ldots, n_j$, $j = 1, 2$, designates the true survival times under testing (i.e. the length of time to event), and C_{mj} represents the censoring variable for the longest time subject m can be observed. Given this notation, F_j indicates the cumulative distribution function (c.d.f.) of T_{mj} (with survival function $S_j(t) = \Pr(T_{mj} > t) = 1 - F_j(t)$), and K_j represents the cumulative distribution function of C_{mj} (with censoring survival function $G_j(t) = \Pr(C_{mj} > t) = 1 - K_j(t)$). Notice that the cumulative distribution function $F(t) = \Pr(T \leq t)$ of T is of primary interest, while the one related to C, $G(t) = \Pr(C \leq t)$, is supposed to be an unknown nuisance entity.

In this setting, a typical right-censored survival dataset comprises n independent realizations of the random pair (X, Δ). Hence, the random vector (X_{mj}, Δ_{mj}) belongs to subject m in the j^{th} sample, for $j = 1, 2$ and $m = 1, \ldots, n_j$. Here, variable X_{mj} indicates the failure time of subject m in the j^{th} group. X_{mj} is called uncensored if the event of interest occurs before the end of the observation period; otherwise the observation is called censored. Thus, $B_j(t) = \Pr(X_{mj} \leq t)$ is the cumulative distribution function of X_{mj}. On the other hand, variable Δ_{mj} indicates the censoring marker, so that it is equal to 1 if the observation is uncensored, and is equal to 0 otherwise. Concisely,

$$X_{mj} = \min(T_{mj}, C_{mj}) \qquad \text{and} \qquad \Delta_m = I(T_{mj} \leq C_{mj})\,,$$

where $I(A)$ denotes the indicator function of event A.

In addition, let us denote with $t_1 < \cdots < t_n$ the distinct ordered pooled times (where $t_m{}'s$ consist of either event or censoring times); and with $t_1 < \cdots < t_D$ the distinct event times in the pooled sample (i.e. in this case t_i represents only the event times). Finally, let us define τ as the largest of the observed event times.

The entire collection of observed data is captured by the pair of associated matrices $(\mathbf{X}, \delta)$:

$$\begin{aligned}(\mathbf{X}, \delta) = {} & ((\mathbf{X}_j, \delta_j), j = 1, 2) = \{(X_{mj}, \delta_{mj}), m = 1, \ldots, n_j\,, j = 1, 2\,\} \\ = {} & \{(X_{mji}, \delta_{mji}), m = 1, \ldots, n_j\,, j = 1, 2\,, i = 1, \ldots, D\}\,.\end{aligned}$$

Hence, let us make the assumption that the data from a random array $(\mathbf{X}, \delta) = \{(\mathbf{X}_1, \delta_1) \uplus (\mathbf{X}_2, \delta_2)\}$ on n subjects are splitted into two samples of n_1 and n_2 units, respectively, and related to two levels of a treatment. Let us also make the assumption that the outcome variables in the two samples have unknown distributions $P_1 = P_{1\Delta} \cdot P_{1X|\Delta}$ and $P_2 = P_{2\Delta} \cdot P_{2X|\Delta}$, respectively, (with $P_j \in \mathcal{P}$, where $\mathcal{P}$ is a, potentially not specified, nonparametric family of non-degenerate distributions), both defined on the same probability space $(\Omega, \mathcal{B})$, where $\Omega = (\mathcal{X}, \Delta)$ is the sample space and $\mathcal{B}$ is a σ-algebra of events. Thus, let us define $\Omega_{/(\mathbf{X},\mathbf{\Delta})}$ as the permutation sample space given $(\mathbf{X}, \mathbf{\Delta})$**.** In this terms, $(\mathcal{X}, \Delta)_{/(\mathbf{X},\mathbf{\Delta})}$ represents the orbit associated with the data $(\mathbf{X}, \mathbf{\Delta})$, as the set containing all permutations $(\mathbf{X}^*, \mathbf{\Delta}^*)$. of the observed data set $(\mathbf{X}, \mathbf{\Delta})$.

In the permutation context, let us indicate $(\mathbf{X}_j, \delta_j)$ as the observed data set of n_j units corresponding to the j^{th} group or sample, with $j = 1, 2$. On the other hand, $\left\{(\mathbf{X}_j^{*b}, \delta_j^{*b}), j = 1, 2\,, b = 1, \ldots, B\right\}$ is a random sample from the permutation sample space.

2. Comparison of survival curves

In survival analysis, it is very common to test whether or not two survival time distributions are equal. In this context, data are collected in order to study the failure time of a group of sample size n, assuming that observations are independent. For simplicity, and without loss of generality, let us consider the case of two independent samples. In this case, the hypotheses testing system is commonly focused on studying the null hypothesis H_0: $F_1(t) = F_2(t), \forall\, t \in \mathbb{R}^+$.

Explicitly, we are interested in the following hypotheses system:

$$\begin{aligned} H_0 \quad &: \quad \{P_1(t) = P_2(t) = P(t),\ \forall t \leq \tau\} \\ &= \quad \{[S_1(t) = S_2(t)] \text{ and } [K_1(t) = K_2(t)]\,,\ \forall t \leq \tau\}\,, \end{aligned}$$

against:

$$H_1 : \{P_1(t) <\neq> P_2(t) \text{ for some } t \leq \tau\}\,,$$

in case of equal censoring, or:

$$H_1 : \{[S_1(t) <\neq> S_2(t)] \text{ or } [K_1(t) \neq K_2(t)] \text{ for some } t \leq \tau\}\,,$$

in case of unequal censoring, where with the symbol ”$<\neq>$” we denote the specific kind of the alternative, i.e. one between the one-sided stochastic dominance (”$<$” or ”$>$”) or the two-sided stochastic difference (”$\neq$”) and where the censoring distributions K_j (with $j = 1, 2$) may vary between treatments.

A critical peculiarity of survival analysis is of course due to censored observations when time to event data is collected. It is of particular interest the case of right-censored data, a problem that happens when the unknown and unobserved time to event is greater than the observed time for which a subject was followed-up during the study. In a clinical trial for example, drop-outs can occur because some patients choose to stop the therapy or because the study ends before all of the subjects have experienced the event under study.

This work deal with complicated censoring patterns and, most important, with the type of censoring. In the framework of right-censored survival data, it is frequently made the assumption that censored data come from an underlying random process, which might or might not be associated with treatment levels or with event occurrences. If we suppose that the probability that an observation is censored is not affected by its unobserved value, then we can ignore this process and therefore there is no need to specify it.

In the literature related to right-censored survival data, most of the statistical procedures make the assumption that censoring effects are, in a very specific sense, noninformative in terms of distribution of survival time, i.e. unaffected by treatment levels. In the event of equal censoring, the censoring process is not affected by treatments, and observed data can be thought as a random sub-sample of the complete data set. Hence, in this case, the process that affects censored data can be disregarded, without making any impact on the inferences on $\mathbf{X}$. Thus, if the censoring distributions are equal, the censoring data process is called ignorable, and statistical tests might be performed conditionally on the actual observed data.

On the other hand, in case of unequal censoring, the observation pairs from the first sample do not have the same distribution of those from the second sample, and that is valid even if the null hypothesis on survival times is true. If the censoring distributions are not equal, then the censored data process must be properly specified, in order to make valid inferences on $\mathbf{X}$. As a result, dealing with unequal censoring data is much more complicated than when we make the assumption that censoring distributions are equal, because inferences ought to be based on the whole data set, and most important it is necessary to specify a suitable model for the underlying censoring trend.

If we assume that, under the null hypothesis, both event and censoring times are jointly exchangeable with respect to samples, then it is possible to solve these multivariate testing problems by using the nonparametric combination of dependent permutation tests. In order to do that, the overall hypotheses can be broken down into a set of sub-hypotheses, assuming that the associated partial tests are significant for large values, marginally unbiased, and consistent.

Based on this reasoning, two articles from Callegaro, Pesarin and Salmaso (2003), and Bonnini, Salmaso and Solari (2005) presented several solutions for permutation analysis of survival data.

This paper introduces some procedures which are exact and one which is approximated.

3. A novel permutation combination-based testing procedure for survival analysis

In the framework of permutation methodologies, it is possible to think about an approach of analysis performed in two stages: a first phase focused on the D observed distinct event times in the pooled sample, where each of those tests can be considered partial aspect of the whole hypothesis testing problem; and then a second phase focused on the combination of these partial aspects into a global one.

Hence, in the right-censored survival data framework, one can thought censored data as a missing data problem. In this case, it is possible to take into consideration a multidimensional permutation test based on the theory of permutation testing with missing data. In fact, if we make the assumption that, after having fixed an observed time t_i, $i = 1, ..., D$, the data already censored might be considered as missing data, then it is reasonable to extend the theory of permutation methods of missing values proposed in Pesarin and Salmaso (2010) to the right-censored survival analysis.

Therefore, the nonparametric combination procedure for dependent tests can be thought as a two-phase (or multi-phase) testing method. In the first stage, let us define $\Gamma_i : (\mathcal{X}^{(n)}, \Delta^{(n)}) \longrightarrow \mathbb{R}^1$, $i = 1, \dots, D$, as an appropriate univariate partial test statistic for the i^{th} sub-hypothesis H_{0i} against H_{1i} as defined later. Now, we assume, without loss of generality, that Γ_i is non-degenerate, marginally unbiased, consistent and that large values of Γ_i are significant, i.e. large values are stochastically larger in H_{1i} than in H_{0i}. Therefore, in the second stage, we define the global test statistic $\Gamma'' = \psi(\widehat{\lambda}_1, \dots, \widehat{\lambda}_i, \dots, \widehat{\lambda}_D)$, as the combination of the permutation p-values $\widehat{\lambda}_i = \widehat{\lambda}_{\Gamma_i}$ associated with the D partial tests, using a proper combining function ψ. Thus, the second-level combined test is a function of D dependent partial tests. Again, we assume, without loss of generality, that $\psi : [0,1]^D \rightarrow \mathbb{R}^1$ is a continuous, non-increasing, non-degenerate univariate combining function, and that large values of ψ are significant.

In general, when we have to deal with a more complex data pattern (for instance, in the framework of hypothesis testing with repeated measures, and stratification variables, or multi-aspect testing, and closed-testing), the nonparametric combination can be viewed as a multi-phase methodology based on several intermediate combinations, where for instance it is at first necessary to combine partial tests with respect to variables within each of the s strata, $s = 1, \ldots, S$, and then to combine the obtained second-order tests with respect to strata using a single third-order combined test.

Here, we are interested in the overall (or global) null hypothesis which states that the two samples have the same underlying distribution:

$$H_0^G : \left\{(X_1, \Delta_1) \overset{d}{=} (X_2,\Delta_2)\right\} ,$$

against a one-sided (stochastic dominance) or a two-sided (inequality in distribution) global alternative hypothesis:

$$H_1^G : \left\{(X_1, \Delta_1) \overset{d}{<\neq>} (X_2,\Delta_2)\right\} .$$

Under the null hypothesis, let us make the assumption that the data $(\mathbf{X}, \mathbf{\Delta})$ are jointly exchangeable with respect to the two samples on both $\mathbf{X}$ and $\mathbf{\Delta}$ variables. It is crucial to emphasize that the whole set of observed data $(\mathbf{x}, \delta)$ under the null hypothesis, is a set of jointly *sufficient statistics* for the underlying observed and censoring data pattern. Likewise, H_0^G entails the exchangeability of the vectors of observations with respect to samples, i.e. the permutation multivariate testing principle can be properly applied.

In a contest of such complexity, it is unlikely to come out with a single overall test statistic. Thus, this hypothesis problem might be handled by using the nonparametric combination of a set of dependent permutation tests (Pesarin and Salmaso, 2010).

Therefore, this method includes a set of D partial tests and, subsequently their nonparametric combination. In this context, the overall null hypothesis can be equivalently written as follow:

$$H_0^G : \left\{ \bigcap_{i=1}^{D} \left[(X_{i1}, O_{i1}) \overset{d}{=} (X_{i2}, O_{i2}) \right] \right\} = \left\{ \bigcap_{i=1}^{D} H_{0i} \right\},$$

which is equivalent to

$$H_0^G : \left\{ \left[\bigcap_{i=1}^{D} \left(O_{1i} \overset{d}{=} O_{2i} \right) \right] \bigcap \left[\bigcap_{i=1}^{D} \left(X_{1i} \overset{d}{=} X_{2i} \right) |\mathbf{O} \right] \right\} = H_0^{\mathbf{O}} \bigcap H_0^{\mathbf{X}|\mathbf{O}}.$$

These formulas serve to emphasize the opportunity of breaking down the global null hypothesis H_0^G. The global alternative hypothesis can be written as:

$$\begin{aligned} H_1^G \quad : \quad & \left\{ \bigcup_{i=1}^{D} [(X_{1i}, O_{1i}) < \overset{d}{\dot{\neq}} > (X_{2i}, \mathbf{O}_{2i})] \right\} = \left\{ \bigcup_{i=1}^{D} H_{1i} \right\} \\ = \quad & \left\{ \left[\bigcup_{i=1}^{D} \left(O_{1i} \overset{d}{=} O_{2i} \right) \right] \bigcup \left[\bigcup_{i=1}^{D} \left(X_{1i} \overset{d}{=} X_{2i} \right) |\mathbf{O} \right] \right\} \\ = \quad & H_1^{\mathbf{O}} \bigcup H_1^{\mathbf{X}|\mathbf{O}}. \end{aligned}$$

At this point, the overall hypotheses system, H_0^G against H_1^G, is broken down into D systems of sub-hypotheses, H_{0i} against H_{1i}, $i = 1, ..., D$, in such a way that H_0^G is true if all the H_{0i} are jointly true. H_1^G is stated in order to reject H_0 when at least one partial null hypothesis is false.

The testing problem H_0^G against H_1^G, is based on a D-dimensional vector of real-valued test statistics $\mathbf{\Gamma} = \{\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_i \ldots, \mathbf{\Gamma}_D\}$, where Γ_i is the univariate partial test for the i^{th} sub-hypothesis H_{0i} against H_{1i}. Without loss of generality, we make the assumption that partial tests Γ_i are marginally unbiased, non-degenerate, consistent and significant for large values. Thus, the second-level combined test is a function of D dependent partial tests. It is important to highlight that, the combination need to be nonparametric, especially with respect to the underlying dependence relation pattern.

4. The structure of the multidimensional permutation test

Let us denote $t_{(1)} < ... < t_{(D)}$, $i = 1, ..., D$, as the ordered and distinct observed time of the event of interest.

Thus, for each statistical unit m within the j^{th} sample, $m = 1, \dots, n_j$, $j = 1, 2$, and each time t_i^o we compute $V_{mji}(t_i)$ as

$$V_{mji}(t_i) = \begin{cases} 1 & if & X_{mj} > t_i \\ 0 & if & X_{mj} \leq t_i \text{ and } X_{mj} = T_{mj} \\ C & if & X_{mj} \leq t_i \text{ and } X_{mj} = C_{mj}; \end{cases} ,$$

and $O_{mji}(t_i)$ as

$$O_{mji}(t_i) = \begin{cases} 0 & \text{if } V_{mji}(t_i) = C \\ 1 & otherwise \end{cases} .$$

Hence, let us consider $u_{ij}(t_i) = \sum_{m=1}^{n_j} O_{mj}(t_i)$ the number of subjects that have not already been censored at time t_i in the j^{th} sample, and $u_i(t_i) = \sum_{j=1}^{2} u_j(t_i)$ the number of subjects that have not already been censored at time t_i in the pooled sample.

4.1. Multidimensional permutation test in case of equal censoring (i.e. $\Delta_1 \stackrel{d}{=} \Delta_2$)

The following procedure, proposed by Callegaro, Pesarin and Salmaso (2003), assumes that censoring effect are non-informative with regard to the distribution of survival time.

Under this assumption, the hypotheses system is focused on the comparison between the global null hypothesis:

$$\begin{aligned} H_0^G \quad &: \quad \left[\{S_1(t_i) = S_2(t_i)\ \forall t_i\ , i = 1, \dots, D\} \text{ and } \left\{\Delta_1 \stackrel{d}{=} \Delta_2\right\}\right] \\ &= \quad \left[\left\{X_1 \stackrel{d}{=} X_2\right\} \text{and} \left\{\Delta_1 \stackrel{d}{=} \Delta_2\right\}\right] , \end{aligned}$$

and the overall alternative hypothesis H_1^G:

$$\begin{aligned} H_1^G \quad &: \quad \{S_1(t_i) <\neq> S_2(t_i)\, \forall t_i, \exists t_i : S_1(t_i) <\neq> S_2(t_i)\} \\ &= \quad \left\{X_1 \overset{d}{<\neq>} X_2\right\}. \end{aligned}$$

In the event of equal censoring, the analysis can be conducted conditionally with regard to the observed censoring variable $\mathbf{O}$ and, most importantly, it can ignore $H_0^{\mathbf{O}}$, because in this setting it is stated that $\mathbf{O}$ does not add any information about the effects of the variable under testing. Therefore, $H_0^{\mathbf{O}} : \{\mathbf{O}_1 \overset{d}{=} \mathbf{O}_2\}$ may be ignored because we made the assumption that the sub-hypotheses on $\mathbf{O}$ are true. On the other hand, the global null and alternative hypotheses $H_0^{\mathbf{X}|\mathbf{O}}$ and $H_1^{\mathbf{X}|\mathbf{O}}$are broken down in D sub-hypotheses, one for each D observed times $t_{(1)} < ... < t_{(D)}$. In conclusion, the overall null hypothesis can be stated in the following simpler fashion:

$$H_0^G = H_0^{\mathbf{X}|\mathbf{O}} : \left\{\bigcap_{i=1}^{D} \left[\left(X_{i1} \overset{d}{=} X_{i2}\right)|\mathbf{O}\right]\right\} = \left\{\bigcap\nolimits_i H_{0i}^{\mathbf{X}|\mathbf{O}}\right\},$$

against

$$H_1^G : \left\{\bigcup\nolimits_i H_{1i}^{\mathbf{X}|\mathbf{O}}\right\}.$$

Now, each of the partial permutation test statistics for testing the sub-hypothesis $H_{0i}^{\mathbf{X}|\mathbf{O}}$ against the sub-alternative $H_{1i}^{\mathbf{X}|\mathbf{O}}$is defined as:

$$\Gamma_i^{*\mathbf{X}|\mathbf{O}}(t_i) = \overline{S}_2^*(t_i)\sqrt{\frac{u_{i1}^*(t_i)}{u_{i2}^*(t_i)}} - \overline{S}_1^*(t_i)\sqrt{\frac{u_{i2}^*(t_i)}{u_{i1}^*(t_i)}},$$

where $\overline{S}_j^*(t_i) = \sum_{m=1}^{n_j} V_{mj}^*(t_i) \cdot O_{mj}^*(t_i)$ is a suitable function of the univariate sampling totals of valid data, and $u_{ij}^*(t_i) = \sum_m O_{mj}^*(t_i)$ is the number of non-censored data permutations at the observed time t_i.

It is important to emphasize that, each test statistic $\Gamma_i^{\mathbf{X}|\mathbf{O}}(t_i)$ is permutationally invariant, in mean value and variance, with regard to the permutation actual sample size $u_j^* = \sum_{m=1}^{n_j} O_{mj}^*$, which changes based on the random assignment of observations to the two samples, because in this case all the observations play a part in the permutation process, even those observations with censoring data. In addition, when $u_j^* = n_j$, $j = 1, 2$, i.e. when no censoring values are observed, the procedure is permutationally equivalent to the standard two-sample permutation test in case of comparison between locations.

Given that the test statistic Γ^* is approximately exact, then it is also approximately unbiased. Another important property of Γ^* is that the test statistic is consistent.

This solution is properly defined, if we make the assumption that u_1^* and u_2^* are jointly positive. In general, this entails that it is necessary to discard from the analysis all those points of the permutation sample space $(\mathcal{X}, \mathcal{O})_{/(\mathbf{X},\mathbf{O})}$ where even a single component of the permutation sample $\mathbf{u}^*$, of actual sample sizes of valid data, is zero. This restriction does not affect the inferential conclusions.

In the proposed solution, the survival analysis is computed using the nonparametric combination $\Gamma''(t_i) = \Gamma''^{\mathbf{X}|\mathbf{O}}(t_i) = \psi_X\left(\widehat{\lambda}_1^{\mathbf{X}|\mathbf{O}}, \ldots, \widehat{\lambda}_D^{\mathbf{X}|\mathbf{O}}\right)$.

Here,

$$\widehat{\lambda}_i^{\mathbf{X}|\mathbf{O}} = \frac{\frac{1}{2} + \sum_{b=1}^{B} \mathbf{I}\left\{\Gamma_i^{*b}(t_i) \geq \Gamma_i^o\right\}}{B+1},$$

represents the estimated p-value related to each t_i, where ψ_X is a proper combining function ψ and B represents the number of independent permutations.

According to Rubin (1976), it is possible to ignore the variable Δ because it is assumed that Δ does not add any information about the effect of the variable under test (because we are considering the event of equal censoring). Therefore, the process which affects the censoring data can be ignored, and the survival analysis can be computed conditionally on the actual observed data.

4.2. Multivariate permutation test in case of unequal censoring (i.e. $\Delta_1 \overset{d}{\neq} \Delta_2$ *)*

This solution, introduced by Bonnini, Salmaso and Solari (2005), is a conditional test based on the probability of failure and on the distribution of observed time to failure conditional upon censoring data.

In the current framework, the hypothesis system is focused on the comparison between the global null hypothesis H_0^G:

$$\begin{aligned} H_0^G \quad &: \quad \left\{ [S_1(t_i) = S_2(t_i)\ \forall t_i\ , i = 1, \ldots, D] \text{ and } \left[\Delta_1 \overset{d}{=} \Delta_2\right] \right\} \\ &= \quad \left\{ \left[X_1 \overset{d}{=} X_2\right] \text{ and } \left[\Delta_1 \overset{d}{=} \Delta_2\right] \right\} , \end{aligned}$$

and the overall alternative hypothesis H_1^G:

$$\begin{aligned} H_1^G : &\left\{ [S_1(t_i) <\neq> S_2(t_i)\ \forall t_i, \exists t_i : S_1(t_i) <\neq> S_2(t_i)] \text{ or } \left[\Delta_1 \overset{d}{\neq} \Delta_2\right] \right\} \\ &= \left\{ \left[X_1 \overset{d}{<\neq>} X_2\right] \text{ or } \left[\Delta_1 \overset{d}{\neq} \Delta_2\right] \right\} . \end{aligned}$$

In case of unequal censoring, H_0 considers the homogeneity in distribution, with regard to the two groups, of the effective observed and collected data $\mathbf{X}$, in conjunction with that associated to the censored data process $\mathbf{O}$. Under unequal censoring data, the null hypothesis must take into consideration the joint distributional equality of the censored data process in the two samples, conditional to $\mathbf{O}$, and of outcome $\mathbf{X}$ conditional to $\mathbf{O}$. Thus, the overall null hypothesis can be written as:

$$H_0^G : \left\{ \left[\mathbf{O}_1 \overset{d}{=} \mathbf{O}_2\right] \bigcap \left[\left(\mathbf{X}_1 \overset{d}{=} \mathbf{X}_2\right) |\mathbf{O} \right] \right\} .$$

Under the null hypothesis, we assume exchangeability among the n individual data vectors in $(\mathbf{X}, \mathbf{O})$, with regard to the two samples, which entails that the effects of the treatment are null on *all* observed and unobserved variables. This means that we are making the assumption that there is no difference in distribution with respect to the multivariate censoring indicator variables $\mathbf{O}_j$, $j = 1, 2$, and, conditionally with regard to $\mathbf{O}$, with respect to the effective observed variables $\mathbf{X}$. Therefore, there is no need to specify both the censored data process and the data distribution, as long as marginally unbiased permutation procedures are available. This is particularly helpful, because in this way we do not have to specify the dependence relation model in $(\mathbf{X}, \mathbf{O})$, as this inferential component is nonparametrically processed.

Consequently, the global hypothesis system is broken down into the $2 \times D$ sub-hypotheses

$$\begin{aligned} H_0^G \quad &: \quad \left\{ \left[\bigcap_{i=1}^{D} \left(O_{i1} \overset{d}{=} O_{i2} \right) \right] \bigcap \left[\bigcap_{i=1}^{D} \left(X_{i1} \overset{d}{=} X_{i2} \right) |\mathbf{O} \right] \right\} \\ &= \quad \left\{ H_0^{\mathbf{O}} \bigcap H_0^{\mathbf{X}|\mathbf{O}} \right\} = \left\{ \left(\bigcap_{i=1}^{D} H_{0i}^{\mathbf{O}} \right) \bigcap \left(\bigcap_{i=1}^{D} H_{0i}^{\mathbf{X}|\mathbf{O}} \right) \right\}, \end{aligned}$$

against

$$H_1^G : \left\{ \left(\bigcup_{i=1}^{D} H_{1i}^{\mathbf{O}} \right) \bigcup \left(\bigcup_{i=1}^{D} H_{1i}^{\mathbf{X}|\mathbf{O}} \right) \right\}.$$

Here, the null sub-hypothesis $H_{0i}^{\mathbf{O}}$ points out the equality in distribution between the two levels of the i^{th} marginal component of the censoring indicator pattern; and the null sub-hypothesis $H_{0i}^{\mathbf{X}|\mathbf{O}}$ points out the equality in distribution of the i^{th} component of $\mathbf{X}$, conditional on $\mathbf{O}$. Now, for each of the D sub-hypotheses $H_{0i}^{\mathbf{O}}$ versus $H_{1i}^{\mathbf{O}}$, it is possible to choose a permutation test statistic like for instance the Fisher's exact probability test or any other appropriate test statistics for suitable testing for binary data.

The partial permutation test statistic for testing the sub-hypothesis $H_{0i}^{\mathbf{O}}$ against the sub-alternative hypothesis $H_{1i}^{\mathbf{O}}$ is defined as:

$$\Gamma_i^{\mathbf{O}}(t_i) = \Gamma_i^{\mathbf{O}}(\mathbf{X}, \mathbf{O}) = \sum_{i=1}^{n_2} O_{2i}\,.$$

This partial test is permutationally equivalent to Fisher's exact probability test.

On the other hand, for each of the D sub-hypotheses $H_{0i}^{\mathbf{X}|\mathbf{O}}$, $\mathbf{O}$ is considered fixed at its observed value, and therefore it is possible to work conditionally.

In the framework of unequal censoring, it is also necessary to combine the D test statistics related to the components of the censoring variable $\mathbf{O}$, assuming that all partial tests are marginally unbiased.

In conclusion, in order to test $H_0 : \{[\bigcap_i H_{0l}^{\mathbf{O}}]\bigcap[\bigcap_i H_{0l}^{\mathbf{X}|\mathbf{O}}]\}$ against $H_1 : \{[\bigcup_i H_{1l}^{\mathbf{O}}]\bigcup[\bigcup_i H_{1l}^{\mathbf{X}|\mathbf{O}}]\}$ the final step involves the combination of the D tests $\Gamma_i^{\mathbf{O}*}$ and the D tests $\Gamma_i^{\mathbf{X}|\mathbf{O}*}$, $i = 1, \ldots, D$. Therefore

$$\Gamma'' = \psi(\widehat{\lambda}_1^{\mathbf{O}}, \ldots, \widehat{\lambda}_D^{\mathbf{O}}; \widehat{\lambda}_1^{\mathbf{X}|\mathbf{O}}, \ldots, \widehat{\lambda}_D^{\mathbf{X}|\mathbf{O}})\,.$$

The nonparametric combination of $2 \times D$ partial tests may be executed in at least three different methods:
(i) we consider one single combining function on all $2 \times D$ partial tests such as: $\Gamma_a'' = \psi(\lambda_1^{\mathbf{O}}, \ldots, \lambda_D^{\mathbf{O}}; \lambda_1^{\mathbf{X}|\mathbf{O}}, \ldots, \lambda_D^{\mathbf{X}|\mathbf{O}})$;
(ii) we can consider D second-order combinations, one for each component variable, $\Gamma_{bl}'' = \psi_i(\lambda_i^{\mathbf{O}}; \lambda_i^{\mathbf{X}|\mathbf{O}})$, $i = 1, \ldots, D$, followed by a third order combination $T_b''' = \psi(\lambda_{b1}'', \ldots, \lambda_{bL}'')$;
or as another alternative, (iii) we might consider two second-order combinations, $\Gamma_{c\mathbf{O}}'' = \psi_O(\lambda_1^{\mathbf{O}}, \ldots, \lambda_D^{\mathbf{O}})$ and $\Gamma_{c\mathbf{X}|\mathbf{O}}'' = \psi_X(\lambda_1^{\mathbf{X}|\mathbf{O}}, \ldots, \lambda_D^{\mathbf{X}|\mathbf{O}})$, respectively, on the censoring indicator $\mathbf{O}$ and on the effective observed $(\mathbf{X}|\mathbf{O})$, followed by a third order combination $\Gamma_c''' = \psi(\lambda_{c\mathbf{O}}''; \lambda_{c\mathbf{X}|\mathbf{O}}'')$.

If we use the same combining function ψ in all the phases and in each of the three methods of combination, then Γ_a'', Γ_b''' and Γ_c''' are almost permutationally equivalent, except for approximations due to the Monte Carlo procedure and non linearity of combining functions. In addition, due to assumptions on partial tests, the second-level partial test $\Gamma_{c\mathbf{X}|\mathbf{O}}''$ is marginally unbiased for $H_0^{\mathbf{X}|\mathbf{O}} : \left\{\left[\left(X_1 \overset{d}{=} \ldots \overset{d}{=} X_C\right)|\mathbf{O}\right]\right\}$, hence it allows for separate testing on effective observed data, conditional on $\mathbf{O}$, even under unequal censoring. This property is useful in many situations, in particular when the analysis is focused on effective observed data.

Partial tests $\Gamma_i^{*\mathbf{O}}$ on the components related to variable $\mathbf{O}$ are exact, unbiased and consistent, while $\Gamma_i^{*\mathbf{X}|\mathbf{O}}$ on the components related to variable $\mathbf{X}$ are unbiased, consistent, but approximately exact. This means that the combined test T'' is unbiased, consistent, and approximately exact for all $\psi \in C$ (see also Pesarin and Salmaso, 2010).

5. A comparative simulation study

The performance of the proposed solutions has been investigated by means of Monte Carlo (MC) simulations in order to estimate the size and power under different experimental situations. For each setting, we considered $MC = 1000$ simulations and $B = 1000$ Conditional Monte Carlo iterations (CMC) for the censoring proportion of 50% in both samples (in case of equal censoring) and 25% against 75% (in case of unequal censoring).

The different configurations are particularly interesting because they represent an extensive variety of situations. These configurations involve most of the challenges tackled in survival analysis. Therefore the aim of the simulation study was to find out whether or not there is a specific procedure able to handle all of these issues in a “robust” fashion. More specifically, the simulated configurations take into consideration: (i) the type of hazard model under alternative hypothesis (models with proportional hazard rates, early, middle, late and crossing hazard differences), (ii) the type of censoring model (equal or unequal censoring), (iii) the sample sizes between the two groups (balanced, left or right unbalanced sample sizes) and lastly, (iv) the type of alternative (one sided or two sided hypothesis testing).

The power of the analyzed tests has been compared under several hazard rate models, and specifically the situation of proportional hazard rates and the interesting circumstances of early, middle, late, and crossing hazard rate differences.

Now, in the simulation study, the survival (or failure) times T_{mj}, $m = 1, \ldots, n_j$, $j = 1, 2$, were generated from piecewise Weibull distributions with shape and scale parameters γ_j and β_j, respectively ($\gamma > 0, \beta > 0$), hazard function $\lambda(t) = \gamma\beta(\frac{t}{\beta})^{\gamma-1}I_{(0,\infty)}(t)$, probability density function:

$$f_T(t,\beta,\gamma) = \gamma\beta^{-\gamma}t^{\gamma-1}exp[-(\frac{t}{\beta})^{\gamma}]I_{(0,\infty)}(t)\,,$$

and survival function:

$$S_T(t,\beta,\gamma) = exp[-(\frac{t}{\beta})^{\gamma}]I_{(0,\infty)}(t)\,,$$

Therefore, the notation used is:

$$T_{m1} \sim \mathcal{W}ei(\beta_1,\gamma_1), T_{m2} \sim \mathcal{W}ei(\beta_2,\gamma_2)\,.$$

and these are the failure times related to the two groups, and generated from Weibull distributions, where, in all our simulations, $\gamma_1 = \gamma_2 = \gamma = 2$. The choice of the shape parameter is intended to take into consideration that in real situations event frequencies increase over time. On the other hand, the scale parameter β_j was properly chosen for each model in order to simulate the hazard rates configuration of interest in sample j.

Then, the censoring times C_{mj}, $m = 1,\ldots,n_j$, $j = 1,2$ were independently generated from the survival times using the uniform distribution. Therefore, C_{mj} has density function $f_C(c,0,\xi) = \frac{1}{\xi}I_{(0,\xi)}(c)$. In particular we denote with: $C_{m1} \sim \mathcal{U}(0,\xi_1)$ and $C_{m2} \sim \mathcal{U}(0,\xi_2)$ the censoring times related to the two samples.

In each configuration, the constants ξ_j were chosen to achieve the target censoring proportions in sample j. Thus, the parameters ξ_j have been selected such that $\xi_j : p_j(\xi_j) = \pi = \Pr(\Delta_{mj} = 1)$ with $0 \leq \pi \leq 1$.

In the simulation study, the multidimensional permutation tests in case of equal censoring (Callegaro, Salmaso and Pesarin, 2003) and in the event of unequal censoring (Bonnini, Salmaso and Solari, 2005), denoted with the acronyms ECNPC and UCNPC, respectively, represent the suitable permutation approach for survival analysis. These combination-based procedures are compared with some other permutation and asymptotic nonparametric techniques introduced in the recent literature. With respect to permutation methods, we considered: (i) the conditional (exact) log-rank test (PLR, Heimann and Neuhaus, 1998), and (ii) the conditional (exact) ”Renyi-type” test (RUPT; Callegaro, Salmaso and Pesarin, 2003).

On the other hand, with regard to asymptotic approaches, we took into consideration the two most frequently used methodologies for comparing groups related to right-censoring survival data. Those tests are (iii) the log-rank test (LR, Mantel, 1966), and (iv) the weighted Kaplan-Meier test (WKM; Pepe and Fleming, 1989). In addition, we considered other two recent works related to (v) an integrated version of the WKM (IWKM, Lee, Lee and Omolo, 2008), and (vi) a modified version of the Fleming-Harringhton test (MFH; Fleming and Harringhton 1991; Gaugler, Kim and Liao, 2007).

Tables 1-5 show the power behaviour of the tests in some of the investigated experimental settings. The results presented in Tables 1-5 show a general good behaviour of the nonparametric combination method in different settings. On the whole, the achieved results suggest the use of the multidimensional permutation procedure, especially in case of equal censoring (ECNPC).

Table 1. Power behaviour of the proposed tests with right-censored data. Models with proportional hazard rates under equal censoring. Two-sided alternative with right-unbalanced sample sizes ($n_1 = 10; n_2 = 50$).

	Permutation Tests				Asymptotic Tests			
α	ECNPC	UCNPC	PLR	RUPT	LR	WKM	IWKM	MFH
0.01	0.540	0.920	0.571	0.596	0.661	0.499	0.542	0.560
0.025	0.683	0.929	0.736	0.755	0.775	0.631	0.663	0.682
0.05	0.790	0.939	0.830	0.846	0.851	0.730	0.769	0.756
0.10	0.876	0.951	0.913	0.912	0.919	0.828	0.860	0.843
0.20	0.938	0.958	0.962	0.957	0.966	0.909	0.928	0.922
0.30	0.966	0.962	0.982	0.957	0.982	0.943	0.961	0.948
0.40	0.984	0.964	0.989	0.986	0.990	0.963	0.969	0.964
0.50	0.989	0.968	0.995	0.989	0.995	0.974	0.979	0.975
0.60	0.992	0.972	0.996	0.992	0.996	0.983	0.988	0.983
0.70	0.999	0.973	0.997	0.995	0.997	0.988	0.991	0.990
0.80	1.000	0.974	1.000	0.997	0.998	0.992	0.993	0.996
0.90	1.000	0.979	1.000	0.999	1.000	0.995	0.998	0.999
1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2. Models with early hazard differences, unequal censoring. Power behaviour of some tests with right-censored data. Two-sided alternative, left-unbalanced sample sizes ($n_1 = 50; n_2 = 10$).

	Permutation Tests				Asymptotic Tests			
α	ECNPC	UCNPC	PLR	RUPT	LR	WKM	IWKM	MFH
0.01	-	0.847	0.056	0.051	0.094	0.052	0.089	0.168
0.025	-	0.857	0.136	0.125	0.195	0.105	0.133	0.317
0.05	-	0.874	0.259	0.229	0.301	0.166	0.177	0.434
0.10	-	0.884	0.393	0.351	0.426	0.255	0.260	0.569
0.20	-	0.899	0.555	0.510	0.575	0.387	0.386	0.733
0.30	-	0.909	0.650	0.606	0.658	0.494	0.493	0.813
0.40	-	0.916	0.717	0.681	0.720	0.579	0.593	0.870
0.50	-	0.924	0.779	0.755	0.785	0.645	0.678	0.898
0.60	-	0.931	0.829	0.815	0.835	0.720	0.764	0.927
0.70	-	0.938	0.882	0.878	0.885	0.800	0.854	0.952
0.80	-	0.944	0.933	0.932	0.936	0.872	0.929	0.972
0.90	-	0.951	0.971	0.967	0.972	0.938	0.981	0.987
1.00	-	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 3. Models with middle hazard differences under equal censoring. Power behaviour of tests with right-censored data. One-sided alternative, right-unbalanced sample sizes ($n_1 = 10; n_2 = 50$).

	Permutation Tests				Asymptotic Tests			
α	ECNPC	UCNPC	PLR	RUPT	LR	WKM	IWKM	MFH
0.01	0.110	0.478	0.024	0.042	0.041	0.061	0.060	0.116
0.025	0.199	0.494	0.041	0.086	0.066	0.113	0.105	0.165
0.05	0.301	0.521	0.072	0.146	0.100	0.165	0.157	0.225
0.10	0.432	0.553	0.137	0.275	0.168	0.245	0.240	0.319
0.20	0.585	0.592	0.245	0.471	0.277	0.387	0.421	0.453
0.30	0.698	0.621	0.353	0.651	0.374	0.490	0.580	0.536
0.40	0.782	0.637	0.449	0.763	0.472	0.581	0.771	0.620
0.50	0.840	0.659	0.540	0.866	0.554	0.678	0.875	0.708
0.60	0.883	0.684	0.650	0.897	0.666	0.750	0.941	0.788
0.70	0.923	0.704	0.748	0.957	0.769	0.825	0.965	0.847
0.80	0.957	0.723	0.854	0.980	0.865	0.887	0.973	0.903
0.90	0.981	0.748	0.925	0.987	0.940	0.944	1.000	0.961
1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4. Models with late hazard differences under unequal censoring. Power behaviour of the proposed tests with right-censored data. Two-sided alternative with balanced sample sizes ($n_1 = 10; n_2 = 10$*).*

	Permutation Tests				Asymptotic Tests			
α	ECNPC	UCNPC	PLR	RUPT	LR	WKM	IWKM	MFH
0.01	-	0.890	0.079	0.138	0.102	0.190	0.426	0.091
0.025	-	0.906	0.149	0.254	0.186	0.322	0.520	0.156
0.05	-	0.919	0.248	0.411	0.294	0.455	0.596	0.255
0.10	-	0.928	0.397	0.550	0.421	0.581	0.676	0.370
0.20	-	0.942	0.578	0.704	0.600	0.710	0.776	0.547
0.30	-	0.955	0.690	0.800	0.699	0.792	0.839	0.646
0.40	-	0.964	0.767	0.867	0.781	0.857	0.893	0.724
0.50	-	0.972	0.842	0.908	0.846	0.891	0.924	0.784
0.60	-	0.980	0.882	0.941	0.887	0.933	0.958	0.834
0.70	-	0.986	0.915	0.964	0.918	0.957	0.979	0.868
0.80	-	0.991	0.946	0.982	0.949	0.973	0.986	0.911
0.90	-	0.992	0.973	0.990	0.975	0.986	0.998	0.954
1.00	-	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 5. Models with crossing hazard differences under equal censoring. Power behaviour of the proposed tests with right-censored data. Two-sided alternative with right-unbalanced sample sizes ($n_1 = 10; n_2 = 50$*).*

	Permutation Tests				Asymptotic Tests			
α	ECNPC	UCNPC	PLR	RUPT	LR	WKM	IWKM	MFH
0.01	0.730	0.974	0.238	0.428	0.289	0.609	0.514	0.088
0.025	0.834	0.978	0.313	0.575	0.347	0.712	0.637	0.114
0.05	0.896	0.981	0.391	0.695	0.426	0.776	0.723	0.134
0.10	0.946	0.983	0.484	0.804	0.502	0.846	0.813	0.188
0.20	0.970	0.983	0.595	0.887	0.607	0.895	0.903	0.298
0.30	0.979	0.985	0.659	0.924	0.666	0.922	0.939	0.390
0.40	0.991	0.989	0.717	0.953	0.728	0.940	0.967	0.473
0.50	0.993	0.992	0.774	0.974	0.778	0.954	0.985	0.575
0.60	0.995	0.993	0.822	0.984	0.830	0.962	0.992	0.648
0.70	0.997	0.995	0.871	0.989	0.872	0.982	0.997	0.733
0.80	0.999	0.995	0.919	0.997	0.923	0.987	0.999	0.823
0.90	1.000	0.996	0.961	0.999	0.965	0.993	0.999	0.917
1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

The combination-based solution appears to be reasonably effective, especially in case of equal censoring, and even in those configurations well-known in the literature as ”ideal” settings for other common permutation and asymptotic tests. It is important to emphasize that this solution is also effective in those configurations where other nonparametric procedures cannot properly be used, i.e. with nonproportional hazard rates models. The combination-based solution can be suggested in the event of equal censoring, particularly when the (small) sample sizes of the two groups are markedly unbalanced. The simulation study also underlines that the novel approach shows a good overall behaviour and the solution appears to be sensitive to most of all the investigated configurations.

6. An application to a biomedical study

In the specialized biomedical literature, there is evidence to believe that there is an association between Tricuspide Valve Replacement (TVR) and high mortality and morbidity; but current knowledge in long-term results related to TVR is still limited. In this section, it is shown an application of survival analysis based on data collected at the Department of cardiovascular disease, Cardiac Surgery Unit, Policlinico S.Donato Hospital (Garatti et al, 2009), which was focused on postoperative or in-hospital mortality and long-term survival of a high-risk population. We studied a well-known risk factor, and specifically the pre-surgery New York Heart Association functional classification (NYHA), which enables to grade the extent of heart failures. This is a seriousness measure which can classify patients in one of four classes with respect to their limitations and symptoms during physical activities related to normal breathing and changeable levels in shortness of breath and/or angina pain. In this particular group of patients we observed 16 patients with class-II functional capacity (28%), 30 with class III (54%) and 10 with class IV (18%). The group of patients in NYHA class III or IV was compared with the group of patients in class II (control group).

The aim of the study was to analyze the overall mortality (either perioperative mortality or long-term follow-up (FU) mortality). Perioperative mortality includes: i) post-operative mortality, ii) in-hospital mortality (all deaths observed at the hospital), and iii) 30-days mortality (all deaths observed within 30 days from surgery).

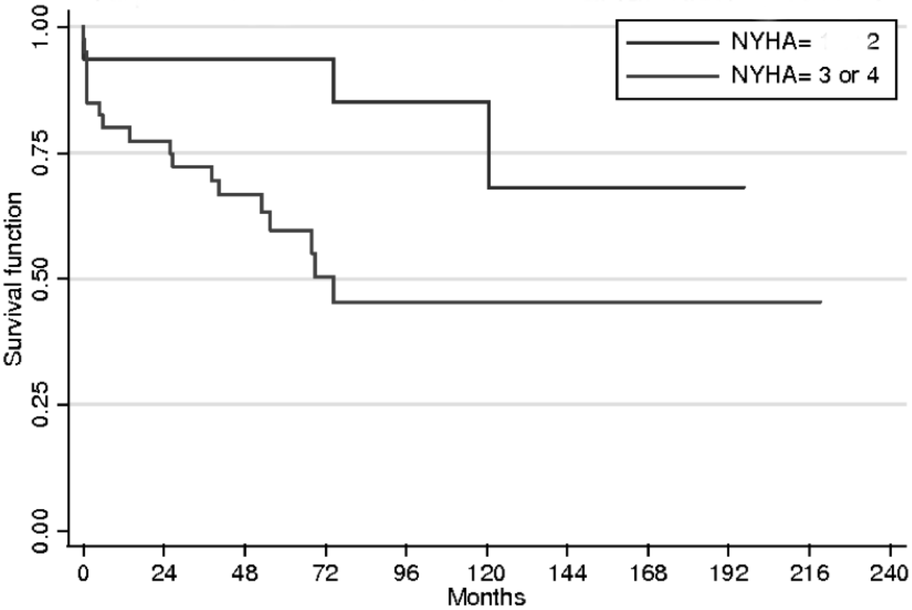

Figure 1. Kaplan-Meier survival estimates by NYHA.

Table 6. Results of survival analysis for NYHA. Department of Cardiovascular Disease, Policlinico S.Donato, Milan: June 1990 - December 2005

	Permutation Tests	Asymptotic Tests
Alternatives	ECNPC	WKN
Two-sided	0.020	0.027
One-sided (2)>(3,4)	0.939	0.987

This retrospective study followed a cohort of 56 patients who underwent TVR for a period of 15 years, from June 1990 to December 2005. 21

(37.5%) deaths was observed during the study period, 8 (38%) during the hospitalization, and 13 (62%) during the follow-up. Figure 1 illustrates the survival curves for the two groups.

The sample size of the study is relatively small, and the two groups are roundly balanced. This trial is particularly interesting because the hazard rates seem to be proportional, but there is a moderate to heavy unequal censoring pattern.

For this specific survival analysis, we computed the proposed permutation combination-based approach (ECNPC) and the traditional asymptotic Weighted Kaplan-Meier (WKM) test. In order to analyze the effect of NYHA on the survival of the two groups of patients. For the permutation solution, we used $B = 10000$ conditional Monte Carlo iterations (CMC). A p-value < 0.05 was considered statistically significant.

We refer the reader to Pesarin and Salmaso (2010) for details on the software code.

7. *Conclusions and final remarks*

This work has introduced a novel permutation combination-based testing technique for survival analysis when the researchers are interested in comparing whether or not two survival time distributions are equal. In light of the comparative Monte Carlo simulation study and then as it can be seen in an application from a real case study, the suggested permutation-based approach seems to be quite reliable and effective when compared with traditional asymptotic counterparts. Indeed, the achieved results show the overall good behaviour of the nonparametric combination method under different settings. On the whole, the suggested solution achieves a good performance and it appears to be sensitive to almost all the examined situations. The achieved results suggest the use of the multidimensional permutation procedure especially in case of equal censoring, and particularly when the small sample sizes of the two samples are substantially unbalanced.

Acknowledgements: Authors wish to thank the University of Padova (CPDA 088513/08 and CPDA092350/09) and the Italian Ministry for University and Research (2008 WKHJPK /002) for providing the financial support for this research.

References

Bonnini S., Salmaso L. and Solari A. (2005), Multivariate permutation tests for evaluating effectiveness of universities through the analysis of student dropouts, *Statistica & Applicazioni,* 3, 37–44.

Callegaro A., Pesarin F. and Salmaso L. (2003), Test di permutazione per il confronto di curve di sopravvivenza, *Statistica Applicata*, 15(2), 241–261.

Fleming T.R. and Harrington D.P. (1991), *Counting processes and survival analysis*, Wiley, New York.

Garatti A., Canziani A., Mossuto E., Gagliardotto P., Innocente F., Corain L., Frigiola A. and Menicanti L. (2010), Tricuspid valve replacement with mechanical prostheses: long-term results, *The Journal of Heart Valve Disease*, 19, 194–200.

Gaugler T., Kim D. and Liao S. (2007), Comparing two survival time distributions: an investigation of several weight functions for the weighted log-rank statistic, *Communications in Statistics - Sim. and Comp.*, 36, 423–435.

Heimann G. and Neahaus G. (1998), Permutational distribution of the log-rank statistic under random censorship with applications to carcinogenicity assays, *Biometrics*, 54, 168–184.

Kaplan E. and Meier P. (1958), Nonparametric estimation from incomplete observations *Journal of the American Statistical Association*, 53, 457–481.

Lee S., Lee H.E.J. and Omolo B.O. (2008), Using integrated weighted survival difference for the two-sample censored data problem, *Computational Statistics and Data Analysis*, 52, 4410–4416.

Mantel N. (1966), Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports*, 50, 163–170.

Pepe M.S. and Fleming T.R. (1989), Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data, *Biometrics*, 45, 497–507.

Pepe M.S., Fleming T.R. (1991), Weighted Kaplan-Meier statistics: large sample and optimality considerations, *Journal of the Royal Statistical Society*, 53, 341–352.

Pesarin F., Salmaso L. (2010), *Permutation tests for complex data: theory, applications and software*, J. Wiley, Chichester.

Rubin D.B. (1976), Inference and missing data, *Biometrika*, 63, 581–592.