Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–31

DOI: 10.1177/0962280217710836
journals.sagepub.com/home/smm
SAGE

Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring

Rosa Arboretti,[1] Roberto Fontana,[2] Fortunato Pesarin[3] and Luigi Salmaso[4]

Abstract

This paper looks at permutation methods used to deal with hypothesis testing within the survival analysis framework. In the literature, several attempts have been made to deal with the comparison of survival curves and, depending on the survival and hazard functions of two groups, they can be more or less efficient in detecting differences. Furthermore, in some situations, censoring can be informative in that it depends on treatment effect. Our proposal is based on the nonparametric combination approach and has proven to be very effective under different configurations of survival and hazard functions. It allows the practitioner to test jointly on primary and censoring events and, by using multiple testing methods, to assess the significance of the treatment effect separately on the survival and the censoring process.

Keywords

Permutation test, nonparametric combination test, weighted log-rank test

1 Motivation and overview from the literature

The main motivation for this paper comes from problems in applied research. Often a researcher's aim is to test the equality of two survival processes corresponding to two different therapies or treatments. A significant problem is that well-known methods, like the log-rank test, assume that censoring is non-informative. When informative censoring occurs, it is usually very difficult to model censoring patterns, which can be very complex. In addition, assuming non-informative censoring when it is not true, increases the risk of making wrong decisions. In this paper, we compare several methods which assume non-informative censoring with nonparametric combination tests that can work both with informative and non-informative censoring.

In this section, we briefly present the most widely used statistics for testing the equality of two survival processes based on independent randomly censored samples. We will refer to the survival process as the primary process and to the censoring process as the secondary process.

The different nonparametric approaches can be classified as asymptotic or, for finite sample sizes, as permutation procedures. In the framework of asymptotic nonparametric methods, several classes of tests may be recognized in the literature.

A first class is based on integrated-weighted comparisons of the estimated cumulative hazard functions in the two-sample design under the null and alternative hypotheses. These are based on the Nelson–Aalen estimator[1,2] within weighted log-rank or weighted Cox–Mantel statistics. In particular, these methods are based on the weighted differences between the observed and expected hazard rates. The test is based on weighted comparisons of the estimated hazard rates of the jth population, $j = 1, 2,$ under the null and alternative

[1]Department of Civil, Environmental and Architectural Engineering, Università di Padova, Italy
[2]Department of Mathematical Sciences, Politecnico di Torino, Italy
[3]Department of Statistics, Università di Padova, Italy
[4]Department of Management and Engineering, Università di Padova, Italy

Corresponding author:
Luigi Salmaso, Stradella San Nicola 3 Vicenza 36100 Italy.
Email: luigi.salmaso@unipd.it

hypotheses. An important consideration in applying this class of tests is the choice of weight function to be used. Weights are used to highlight certain parts of the survival curves under study. A variety of weight functions have been proposed in the literature and the related tests are more or less sensitive to early or late departures from the hypothesized relationship between the two hazard functions (as specified in the null hypothesis) according to the weight function used in the testing procedure. All these statistics are censored data generalizations of linear rank statistics. For instance, the most commonly used rank-based test statistic is the log-rank test proposed by Mantel,[3] Peto and Peto[4] and Cox;[5] it is a generalization of the exponential ordered score test of Savage[6] for censored data. This test statistic has good power when it comes to detecting differences in the hazard rates, when the ratio of hazard functions in the populations being compared is approximately constant. Gilbert,[7] Gehan[8] and Breslow[9] proposed a censored data generalization of the two-sample Wilcoxon–Mann–Whitney rank test (non-informative censoring). Peto and Peto[4] and Prentice et al.[10] proposed other generalizations of the Wilcoxon–Mann–Whitney test. They used a different estimate of the survival function based on the combined sample. Tarone and Ware[11] proposed a class of multi-sample statistics for right-censored survival data that includes the log-rank test and the censored data generalized Wilcoxon–Mann–Whitney procedures.

Fleming–Harrington[12] introduced a very general class of tests which includes, as special cases, the log-rank test and another version of the Wilcoxon–Mann–Whitney test. They use the Kaplan–Meier estimate of the survival function based on the combined sample at the previous event time. More recently Gaugler et al.[13] proposed a modified Fleming–Harrington test very similar to the original version which includes as a special case the Peto–Peto and Kalbfleisch–Prentice tests. Here, the Peto–Peto and Kalbfleisch–Prentice estimate of the survival function is based on the pooled sample at the current event time. Jones and Crowley[14] introduced a more general class of single-covariate nonparametric tests for right-censored survival data that includes the Tarone–Ware two-sample class, the Cox[5] score test, the Tarone[15] and Jonckheere[16] C-sample trend statistics, the Brown et al.[17] modification of the Kendall rank statistic, Prentice's linear rank statistics,[10] O'Brien's logit rank statistic (1978)[18] and several new procedures. This class can be generalized to include the Tarone–Ware C-sample class.

The statistical properties of the aforementioned test statistics have been studied by many authors. Here we mention, among others, works by Gill,[19] Fleming and Harrington,[20] Breslow et al.,[21] Fleming et al.,[22] Lee,[23] Kosorok and Lin.[24]

They are based on the weighted log-rank statistics

$$\sum_{i=1}^{D} W(t_i)\left\{d_{ij} - \hat{E}\left[d_{ij}\right]\right\} \quad j = 1, 2$$

where $j = 1, 2$ denote the groups, $0 < t_1 < \cdots < t_D$ are D different observation times in the interval $(0, \tau]$ ($t_D = \tau$), d_{ij} is the number of primary events that occurred in the interval $(t_{i-1}, t_i]$, $t_0 \equiv 0$ and $\hat{E}\left[d_{ij}\right]$ is an estimate of its expected value. A variety of weight functions $W(t_i)$ have been proposed in the literature (see Table 1 where r_i is the number of individuals at risk at $t_i, i = 1, \ldots, D$ and $\hat{S}(.)$ and $\tilde{S}(.)$ are different estimates of the survival functions as shall be clarified in Section 4). We observe that the symbol τ is usually used for the terminal time, see e.g. the monograph of Andersen et al.,[25] while in this work τ will be smaller than or equal to the terminal time.

These tests are sensitive to alternatives of ordered hazard functions.[20] When they are applied to samples from populations where the hazard rates cross, they have little power because early positive differences in favour of one group are compensated by late differences in favour of the other treatment.

Table 1. Different types of weight function.

Weight function	Test's name
$W(t_i) = 1 \quad \forall t_i$	Log-rank test
$W(t_i) = r_i$	Gehan-Breslow test
$W(t_i) = \sqrt{r_i}$	Tarone–Ware test
$W(t_i) = \tilde{S}(t_i)$	Peto–Peto and Kalbfleisch–Prentice tests
$W(t_i) = \frac{\tilde{S}(t_i) r_i}{r_i + 1}$	Modified Peto–Peto and Kalbfleisch–Prentice test[25]
$W_{p,q}(t_i) = (\hat{S}(t_{i-1})^p (1 - \hat{S}(t_{i-1}))^q, p \geq 0; q \geq 0$	Fleming–Harrington test
$W_{p,q}(t_i) = (\tilde{S}(t_i)^p (1 - \tilde{S}(t_i))^q, p \geq 0; q \geq 0$	Modified Fleming-Harrington test[26]

A second class of procedures is based on the maximum of the sequential evaluation of the weighted log-rank tests at each event time. These procedures are known as ''Renyi-type'' statistics. Such tests are presumed to have 'good' power behaviour to detect crossing hazards. These supremum versions of the weighted log-rank tests were proposed by Gill[19] and are generalizations of the Kolmogorov–Smirnov statistic for comparing two censored data samples (non-informative censoring).

A third class of procedures is the weighted Kaplan–Meier (WKM) statistic based directly on integrated weighted comparisons of survival functions (rather than on ranks) in the two samples under the null and alternative hypotheses, based on the Kaplan–Meier estimator. This class of tests was initially proposed by Pepe and Fleming;[26–28] more recently Lee et al.[23] studied an integrated version of the WKM. Note that two other versions of these two statistics could be obtained by replacing the Kaplan–Meier estimators with the Peto–Peto and the Kalbfleisch–Prentice estimators. WKM statistics provide censored data generalizations of the two-sample z- or t-test statistic (non-informative censoring). Asymptotic distribution properties of the WKM statistics can be found in Pepe and Fleming[28], Pepe and Fleming,[29] and Lee et al.[23]

A fourth class of procedures is a censored data version of the Cramer–Von Mises statistics, based on the integrated squared difference between the two estimated cumulative hazard rates, based on the Nelson–Aalen estimator (non-informative censoring). This is done to obtain a limiting distribution which does not depend on the relationship between the event and censoring times and because such tests arise naturally from counting process theory. Small-sample and asymptotic distribution properties of three versions of this rank test statistic can be found in Koziol[30] and Schumacher.[31]

Finally, a further class of test statistics is a generalization of the two-sample median statistics for non-informative censored data, proposed by Brookmeyer and Crowley,[32] which is useful when we are interested in comparing the median survival times of the two samples rather than the difference in the hazard rate or the survival functions over time. Asymptotic distribution properties of the WKM statistics can be found in Brookmeyer and Crowley.[32]

Other two-sample procedures have been suggested in the literature. The most recent works include a midrank unification of rank tests for exact, tied and censored data proposed by Hudgens and Satten,[33] a nonparametric procedure for use when the distribution of time to non-informative censoring depends on treatment group and survival time, proposed by DiRienzo,[34] an asymptotically valid C-sample test statistic ($C \geq 2$) proposed by Heller and Venkatraman,[35] and the randomization-based log-rank test proposed by Zhang and Rosemberger.[36] Recent contributions include Wakounig et al.[37] which focuses on the nonparametric estimation of relative risk in survival tests.

1.1 Permutation tests in survival analysis

When asymptotic tests are used to compare survival functions, it is possible to find situations in which the number of failures of interest is so small that it is reasonable to question the validity of asymptotic tests. In such situations, the standard asymptotic log-rank test, for example, is frequently replaced by its corresponding log-rank permutation test – Galimberti and Valsecchi[38] and Callegaro et al.[39] This provides an exact small-sample test when the censoring patterns in the two compared populations are equal. In actual fact, when the censoring patterns are treatment dependent (i.e. informative), the observation pairs from the first population do not have the same distribution as those from the second. The failure of the asymptotic log-rank test is not only due to an inappropriate asymptotic approximation, which in turn can be replaced by an exact evaluation, but it is also due to ignoring the interdependency of the *observed risk sets*, that is the risk sets in the 2×2 contingency tables associated with the D event times, as described in Heinze et al.[40] Furthermore, since the two group sizes may be unbalanced, the asymptotic test may not be appropriate as the asymptotic distributions under the null hypothesis may be too far away from being appropriate. Some authors, among whom Kellerer and Chmelevsky,[41] Chen and Gaylor,[42] Ali,[43] Soper and Tonkonoh,[44] proposed the so-called exact procedures. These provide exact small sample tests when the censoring patterns in the samples being compared are treatment independent.

In situations like these, permutation tests may be helpful. In this section, we give a brief description of methods that require extensive computation. In the survival analysis framework, we present a review of the most common two-sample permutation tests that have been suggested in the literature. The only necessary assumption is the independence across the individuals of the pairs of primary and secondary data.

There are three different unidimensional exact conditional procedures analogous to the asymptotic log-rank test and suitable for situations with treatment dependent (informative) censoring. The first method was proposed by Heimann and Neuhaus.[45] The other two were proposed in a work by Heinze et al.[40] Callegaro et al.[39] proposed an

exact "Renyi-type" test suitable in case of treatment dependent censoring where it is assumed that, in the alternative hypothesis, treatment may also influence the censoring process, *but without considering their joint effects*. An interesting permutation contribution is given by Galimberti and Valsecchi[38] who introduced a permutation test to compare survival curves for non fixed-matched data when the number of strata increases, the stratum sizes are small, and the proportional hazard model is not satisfied.

It is worth mentioning some papers that can deal with different censoring distributions like Neuhaus[46] and Brendel et al.[47] (see also Janssen and Meyer[48]) and also with competing risks (Dobler et al.[49] and subsequent papers). All these papers assume non-informative censoring. However, the first two apply a different, but also interesting permutation technique for survival data which is asymptotically unbiased while still being finitely exact if exchangeability holds.

The joint analysis of primary and secondary time processes, as both can contain information on treatment effects, is the main proposal of our paper.

1.2 Data structure

In a two-sample problem, the whole set of observed data can be summarized by the pair of associated matrices $(\boldsymbol{X}, \boldsymbol{\delta})$

$$(\boldsymbol{X}, \boldsymbol{\delta}) = \left\{[X_{mj}(t), \delta_{mj}(t)], 0 \leq t \leq \tau,\ m = 1, \ldots, n_j,\ j = 1, 2\right\}$$

where $[X_{mj}(t), \delta_{mj}(t)]$ is the event-time profile at time t of individual m in group j)

We make the assumption that the two groups are independent and that the observations are exchangeable under the null hypothesis.

At time t for the m-th individual of the j-th group, the possible values of $(X_{mj}(t), \delta_{mj}(t))$ are

- $(., 0)$ when at time t the individual is still alive and had not left the study;
- $(x, 0),\ x \leq t$ when the individual died at time x (primary event);
- $(x, 1),\ x \leq t$ when the individual left the study at time x (censoring event).

In particular at the final time τ if the m-th individual of the j-th group is still alive and did not leave the study, the value of $(X_{mj}(t), \delta_{mj}(t))$ is $(\tau, 1)$, because this situation is equivalent to a censoring event that occurs at time τ.

Data in $(\boldsymbol{X}, \boldsymbol{\delta})$ correspond to pooling two-sample data profiles of two groups with n_1 and n_2 individuals, respectively, classified according to two levels of a treatment. That is, $(\boldsymbol{X}, \boldsymbol{\delta}) = \left\{(\boldsymbol{X}_1, \boldsymbol{\delta}_1) \uplus (\boldsymbol{X}_2, \boldsymbol{\delta}_2)\right\}$, where $\uplus$ is the pooling operator to merge together two datasets and $\{(\boldsymbol{X}_j, \boldsymbol{\delta}_j)\}$ is

$$\left\{[X_{mj}(t), \delta_{mj}(t)],\ 0 \leq t \leq \tau,\ m = 1, \ldots, n_j\right\}, \quad j = 1, 2$$

We also assume that the response variables in the two groups have unknown distributions $P_1 = P_{1\delta} \cdot P_{1X|\delta}$ and $P_2 = P_{2\delta} \cdot P_{2X|\delta}$, $(P_j \in \mathcal{P}$, where $\mathcal{P}$ is a possibly non specified nonparametric family of non-degenerate distributions) both defined on the same probability space $(\Omega, \mathcal{A})$ where $\Omega = (\mathcal{X}, \mathcal{O})$ is the sample space, $\mathcal{A}$ is a σ-algebra of events and $\mathcal{X}$ and $\mathcal{O}$ are the sample spaces for data times X and censoring indicators δ, respectively. Hence, let $\Omega = (\mathcal{X}, \mathcal{O})$ be the support of the random vector $(\boldsymbol{X}, \boldsymbol{\delta})$ and $\Omega_{/(\boldsymbol{X}, \boldsymbol{\delta})}$ the permutation sample space given $(\boldsymbol{X}, \boldsymbol{\delta})$. In this way, $(\mathcal{X}, \mathcal{O})_{/(\boldsymbol{X}, \boldsymbol{\delta})}$ is the orbit associated with the data set $(\boldsymbol{X}, \boldsymbol{\delta})$, that is the set of sufficient statistics under the null hypothesis associated with the observed data set $(\boldsymbol{X}, \boldsymbol{\delta})$, thus containing the set of all permutations $(\boldsymbol{X}^*, \boldsymbol{\delta}^*)$. Further details on the orbits can be found in Pesarin and Salmaso.[50]

In the permutation setting, let $(\boldsymbol{X}_j, \boldsymbol{\delta}_j)$ be the observed data set of n_j elements related to the jth sample $j = 1, 2$. Let us also use $\left\{(\boldsymbol{X}_j^{*b}, \boldsymbol{\delta}_j^{*b}),\ j = 1, 2\ , b = 1, \ldots, B\right\}$ to indicate a random sample of B elements from the permutation sample space $\Omega_{/(\boldsymbol{X}, \boldsymbol{\delta})}$.

2 Comparison of survival curves

As frequently occurs in survival studies, time-to-event data are characterized by incompleteness due to censoring. In particular, right-censored data occur when the unobserved and unknown time to the event of interest is more than the recorded time for which an individual was under observation. For instance, subjects in a survival study

can be lost to follow-up due to transfer to a non participating institution, or the study can finish before all the subjects have observed the event.

The focus here is the presence of complicated censoring patterns and, in particular, the type of censoring. In the right-censored survival data framework, censored data are usually assumed to originate from an underlying random process, which may or may not be related to treatment levels or to event processes. When we assume that the probability of a datum being censored does not depend on its unobserved value, then we may ignore this process and so need not specify it.

Nearly all statistical procedures for right-censored survival data are based on the assumption that censoring effects are, in a very specific sense, non-informative with respect to the distribution of survival time, i.e. unaffected by treatment levels. If the censoring distributions are equal, the censoring process does not depend on group, and observed values may be considered a random sub-sample of the complete data set. Thus, in these situations, it is appropriate to ignore the process that causes censored data when making inferences on X. Therefore, in the case of treatment independent censoring distribution, the process that causes censored data is called *ignorable* and analysis may be carried out conditionally on the actually observed data.

In contrast, when the censoring patterns are treatment dependent, the observation pairs from the first sample do not have the same distribution as those from the second sample, even when the null hypothesis on pure survival times is true. In the case of treatment dependent censoring distributions, in order to make valid inferences, the censored data process must be properly specified. Thus, the analysis of treatment dependent censoring data is much more complicated than that of treatment independent censoring data because inferences must be made by taking into consideration the data set as a whole and by specifying a proper model for the censoring pattern. In any case, the specification of a model which correctly represents the censored data process up to now seems the only way to remove the inferential bias caused by censoring.

In survival analysis, it is often of interest to test whether or not two survival time distributions are equal. We assume that observations are available on the failure times of n individuals assumed to behave independently. Focusing on the two independent sample case, researchers are often interested in comparing two therapies, two products, two processes, two treatments, etc. For the moment let us consider the case where only one response is of interest. The two samples are denoted by $(X_{11}, \ldots, X_{1n_1})$ and $(X_{21}, \ldots, X_{2n_2})$, respectively. The $X'_1 s$ constitute a random sample from the random variable X_1 with CDF F_1, and the $X'_2 s$ a random sample from X_2 with CDF F_2. The testing problem is usually formulated as testing the null hypothesis $H_0 : \{F_1(t) = F_2(t),\ \forall t \in \mathbb{R}^+\} \equiv X_1 \stackrel{d}{=} X_2$.

We now move to the situation in which both primary and censoring events can occur. Let S_j and K_j denote the marginal distribution functions of the survival and censoring times corresponding to the individuals of the jth group, respectively, and let P_j denote their joint distribution function, $j = 1, 2$. As before $[0, \tau]$ is the time interval under study. The hypotheses of interest, in the case of non-informative censoring ($K_1 = K_2 = K$), are

$$H_0 : \left\{P_1(t) = P_2(t) = P(t),\ \forall t \leq \tau\right\} = \{[S_1(t) = S_2(t)],\ \forall t \leq \tau\}$$

against

$$H_1 : \left\{P_1(t) <\neq> P_2(t),\ \text{some } t \leq \tau\right\} = \left\{S_1(t) <\neq> S_2(t),\ \text{some } t \leq \tau\right\}$$

where $<\neq>$ means either $<$, or $\neq$, or $>$. These hypotheses reflect the notion that if treatment has no effect, then two primary processes are equal, the secondary processes being equal by assumption.

In the case of treatment-dependent (informative) censoring, the hypotheses are

$$H_0 : \left\{P_1(t) = P_2(t) = P(t),\ \forall t \leq \tau\right\} = \{[S_1(t) = S_2(t)] \cap [K_1(t) = K_2(t)],\ \forall t \leq \tau\}$$

against

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

which reflects the notion that if treatment has no effect, then both primary and secondary processes are equal in two samples, whereas in the alternative, at least the primary or the secondary (or both) are not equal.

If we assume that in the null hypothesis, where treatments have exactly the same effect, all pairs in $(\boldsymbol{X}, \boldsymbol{\delta})$ of event and censoring times are jointly exchangeable with respect to subjects and groups, then such multivariate testing problems are solvable by the nonparametric combination (NPC) of dependent permutation tests, as are the

tests on main and censoring time events – Pesarin,[51] Pesarin and Salmaso.[50] It should be emphasized that the dependence structure between two processes in the alternative is too complicated to analyze, so we may only deal with it nonparametrically by NPC. NPC works within Roy's Union-Intersection principle. Thus, it is assumed that the hypotheses can be broken down into a set of sub-hypotheses, and the related partial tests are assumed to be marginally (i.e. separately, Pesarin[52]) unbiased, significant for large values and consistent.

Although some solutions presented in this chapter are exact, the most important ones are approximations because the permutation distributions of the test statistics are not exactly invariant with respect to permutations of censored data, as we shall see. However, the approximations are quite reasonably accurate in all situations, provided that the number of observed data is not too small. The approximation is due to the fact that we remove from the permutation sample space, associated with the whole data set, all those permutations where the permutation sample sizes of actually observed n-dimensional data are not large enough. In a way, similarly to the permutation analysis of missing data (see Section 7.9 of Pesarin and Salmaso[50]), we must establish a kind of restriction on the permutation space, provided that this restriction does not imply unacceptable bias on inferential conclusions.

3 Nonparametric combination-based tests

In the framework of permutation methods, it is possible to consider an analysis approach incorporating two successive stages, the first focusing on the D observed distinct event times in the pooled sample, which can be considered partial aspects of the hypothesis testing problem giving rise to a list of partial tests, and the second focusing on the combination of these partial aspects into a global aspect.

Therefore, the NPC procedure for dependent tests may be viewed as a two-phase testing procedure. In the first phase, let us suppose that $T_i : (\mathcal{X}, \mathcal{O}) \to \mathbb{R}^1$ $(i = 1, \ldots, D)$ is an appropriate univariate partial test statistic for the ith sub-hypothesis H_{0i} against H_{1i}, for which (without loss of generality) we assume that T_i is non-degenerate, marginally or separately unbiased, consistent and that large values are significant, so that they are stochastically larger in H_{1i} than in H_{0i} in both conditional and unconditional senses. In the second phase, we construct the global test statistic either simply as $T'' = \sum_{i=1}^{D} T_i$ (direct combination) or as $T'' = \psi(\lambda_1, \ldots, \lambda_i, \ldots, \lambda_D)$ by combining the permutation p-values $\lambda_i = \lambda_{T_i}$ associated with the D partial tests through a suitable combining function ψ. Hence, the combined test is a function of D dependent partial tests. In practical terms, in place of true p-values λ_i, we use their estimates $\hat{\lambda}_i$ based on B random permutations from the permutation sample space $\Omega_{/(\boldsymbol{X}, \boldsymbol{\delta})}$.

When there is a more complex data configuration (where the more interesting cases are given by testing in the presence of stratification variables, closed-testing, multi-aspect testing, etc.), the NPC may be like a multi-phase procedure characterized by several intermediate combinations, where we may, for instance, firstly combine partial tests with respect to variables within each s stratum (with $s = 1, \ldots, S$), and then combine the second-order tests with respect to strata into a single third-order combined test.

3.1 Breaking down the hypotheses

It is generally of interest to test for the global (or overall) null hypothesis that the two groups have the same underlying distribution

$$H_0^G : \left\{ (X_1, \delta_1) \stackrel{d}{=} (X_2, \delta_2) \right\}$$

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

$$H_1^G : \left\{ (X_1, \delta_1) \ <\stackrel{d}{\neq}> \ (X_2, \delta_2) \right\}$$

Let us assume that under the null hypothesis, the data $(\boldsymbol{X}, \boldsymbol{\delta})$ are jointly exchangeable with respect to the two groups. It is important to note that the pooled set of observed data $(\boldsymbol{X}, \boldsymbol{\delta})$ in the null hypothesis is a set of jointly *sufficient statistics* for the underlying observed and censoring data processes. Moreover, H_0^G implies the exchangeability of individual data vectors with respect to groups, so that the permutation multivariate testing principle is properly applicable, Pesarin.[53]

The complexity of this testing problem is such that it is very difficult to find a single overall test statistic. However, the problem may be dealt with by means of the NPC of a set of dependent permutation tests.

Hence, we consider a set of D partial tests followed by their NPC. The overall null hypothesis can be written as

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

equivalent to

$$H_0^G : \left\{ \left[\bigcap_{i=1}^{D} \left(\delta_{1i} \stackrel{d}{=} \delta_{2i} \right) \right] \cap \left[\bigcap_{i=1}^{D} \left(X_{1i} \stackrel{\mathrm{d}}{=} X_{2i} \right) | \boldsymbol{\delta} \right] \right\} = H_0^{\boldsymbol{\delta}} \cap H_0^{1X|\delta}$$

where a breakdown of H_0^G is emphasized according to the main theory of the nonparametric combination – Pesarin and Salmaso.[50] In fact a suitable way to decompose the overall null hypothesis is usually denoted by the union of the partial null hypotheses as in the Union-Intersection testing theory. The overall alternative hypothesis may be represented as

$$\begin{aligned} H_1^G &= \left\{ \bigcup_{i=1}^{D} [(X_{1i}, \delta_{1i}) \; \dot{<\neq>} \; (X_{2i}, \delta_{2i})] \right\} = \left\{ \bigcup_{i=1}^{D} H_{1i} \right\} \\ &= \left\{ \left[\bigcup_{i=1}^{D} \left(\delta_{1i} \; \dot{<\neq>} \; \delta_{2i} \right) \right] \bigcup \left[\bigcup_{i=1}^{D} \left(X_{1i} \; \dot{<\neq>} \; X_{2i} \right) | \boldsymbol{\delta} \right] \right\} = H_1^{\boldsymbol{\delta}} \bigcup H_1^{1X|\delta} \end{aligned}$$

It should, however, be highlighted that at time $t_i,\ i = 1, \ldots, D$ either one main or one censoring event occurs, so the total number of active sub-hypotheses is exactly D. Hence, the hypothesis H_0^G against H_1^G is broken down into D sub-hypotheses H_{0i} against $H_{1i},\ i = 1, \ldots, D$, in such a way that H_0^G is true if all the D null sub-hypotheses H_{0i} are jointly true and H_1^G implies that the inequality of the two distributions entails the falsity of at least one among the D null sub-hypotheses. Finally, note that the hypotheses and assumptions are such that the *permutation testing principle* applies, Pesarin.[53]

Thus, to test H_0^G against H_1^G, we consider a D-dimensional vector of real-valued test statistics $= \{_1, \ldots, _D\}$, the ith component of which is the univariate partial test for the ith sub-hypothesis H_{0i} against H_{1i}. We assume that partial tests are non-degenerate, marginally unbiased, consistent, and significant for large values. Hence, the combined test is a function of D dependent partial tests and, of course, the combination must be nonparametric, particularly with regard to the underlying dependence relation structure, which is too complex to be analyzed by means of all its unknown coefficients of dependence.

3.2 The test structure

Let us consider $t^o_{(1)} < \cdots < t^o_{(D)},\ i = 1, \ldots, D$, the ordered and distinct observed times of the event of interest. To make the notation easy, we write $t^o_{(i)}$ simply as t_i. For each subject m within the jth group ($m = 1, \ldots, n_j,\ j = 1, 2$) and for each t_i, we define $V_{mj}(t_i)$ as

$$V_{mj}(t_i) = \begin{cases} 1 & \text{if} \quad X_{mj} > t_i \\ 0 & \text{if} \quad X_{mj} \leq t_i \text{ and a primary event occurred at } X_{mj} \\ 2 & \text{if} \quad X_{mj} \leq t_i \text{ and a secondary event occurred at } X_{mj} \end{cases}$$

and the indicator of non-censored observations $O_{mj}(t_i)$ as $1 - \delta_{mj}(t_i)$, i.e.

$$O_{mj}(t_i) = \begin{cases} 0 & \text{if } V_{mj}(t_i) = 2 \\ 1 & \text{otherwise} \end{cases}$$

We also define $v_{ij} = \sum_{m=1}^{n_j} O_{mj}(t_i)$ as the number of observations that have not already been censored at time t_i in the jth group, and $v_i = \sum_{j=1}^{2} v_{ij}(t_i)$ as the number of observations that have not already been censored at time t_i in the pooled sample.

3.3 NPC test for treatment independent censoring

In this section, we consider a multidimensional permutation test in the case of treatment independent censoring (TIC-NPC). This test, proposed by Callegaro et al.,[39] is based on the assumption that censoring effects are non-informative with respect to the distribution of survival time.

In the present context, we are interested in testing the global null hypothesis

$$H_0^G = H_0^{\boldsymbol{\delta}} \cap H_0^{1X|\delta}$$

against the global alternative

$$H_1^G = H_1^{\boldsymbol{\delta}} \bigcup H_1^{1X|\delta}$$

If the censored data are treatment independent, we may proceed conditionally on the observed censoring indicator $\boldsymbol{\delta}$ and ignore $H_0^{\boldsymbol{\delta}}$, because in this context $\boldsymbol{\delta}$ does not provide any information about treatment effects. Hence, we may equivalently write the null hypothesis in the relatively simpler form

$$H_0 = H_0^{1X|\delta} : \left\{ \bigcap_{i=1}^{D} \left[\left(X_{i1} \stackrel{\mathrm{d}}{=} X_{i2} \right) | \boldsymbol{\delta} \right] \right\} = \left\{ \cap_i H_{0i}^{1X|\delta} \right\}$$

against

$$H_1 : \left\{ \bigcup_{i=1}^{D} H_{1i}^{1X|\delta} \right\}$$

The partial permutation test statistics for testing the sub-hypothesis $H_{0i}^{1X|\delta}$ against the sub-alternative $H_{1i}^{1X|\delta}$ then takes the form

$$\Gamma_i^{*1X|\delta} = \bar{S}_2^*(t_i) \sqrt{\frac{v_{i1}^*(t_i)}{v_{i2}^*(t_i)}} - \bar{S}_1^*(t_i) \sqrt{\frac{v_{i2}^*(t_i)}{v_{i1}^*(t_i)}}$$

where $\bar{S}_j(t_i) = \sum_{m=1}^{n_j} V_{mj}(t_i)\delta_{mj}(t_i)$ is the number of individuals that did not leave the study and are still alive at time t_i and the suffix * means that the statistic has been computed using a random permutation of the sample.

Note that each test statistic $\Gamma_i^{1X|\delta}$ is permutationally invariant, in mean value and variance, with respect to the sample size $v_j^* = \sum_{m=1}^{n_j} \delta_{mj}^*$, that varies according to the random attribution of units to the two groups, because units with censoring data participate in the permutation mechanism as well as all other units. Also, note that when there are no censoring values, so that $v_j^* = n_j$, $j = 1, 2$, each partial test is permutationally equivalent to the traditional two-sample permutation test for comparison of locations.

In order for the given solution to be well-defined, we must assume that v_1^* and v_2^* are jointly positive. This implies that, in general, we must consider a sort of restricted permutation strategy which consists of discarding from the analysis all points of the permutation sample space $(\mathcal{X}, \mathcal{O})_{/(\boldsymbol{X}, \boldsymbol{\delta})}$ in which even a single component of the permutation matrix $1v^*$, of actual sample sizes of valid data, is zero. Of course, this kind of restriction has no effect on inferential conclusions.

Therefore, the survival analysis may be solved by NPC of $\Gamma''(t_i) = \Gamma_{\psi}''^{1X|\delta}(t_i) = \psi_X\left(\hat{\lambda}_1^{1X|\delta}, \ldots, \hat{\lambda}_D^{1X|\delta}\right)$ where

$$\hat{\lambda}_i = \frac{\frac{1}{2} + \sum_{b=1}^{B} I\left\{\Gamma_i^{*b} \geq \Gamma_i^*\right\}}{B+1}$$

is the p-value function estimate for each $\Gamma_i^*(t_i)$ and ψ_X is a suitable combining function ψ. For details on different combining functions and related properties, we refer the reader to Pesarin and Salmaso.[50]

Note that according to Rubin,[54] we may ignore the variable δ because in this context we have assumed that it does not provide any information on treatment effects.

3.4 NPC test for treatment dependent censoring

In this section, we introduce a permutation test in the case of treatment dependent censoring (or informative censoring).

As in the previous section, we are interested in testing the global null hypothesis

$$H_0^G = H_0^{\boldsymbol{\delta}} \cap H_0^{1X|\delta}$$

against the global alternative

$$H_1^G = H_1^{\boldsymbol{\delta}} \bigcup H_1^{1X|\delta}$$

In the case of treatment dependent censoring, it is assumed that, in the alternative, treatment may also influence the censoring process. Then data $\boldsymbol{X}$ associated to the censored data process $\boldsymbol{\delta}$ must be taken into consideration. In fact, the treatment may affect both the distributions of variables X and the censoring indicator δ. Hence, in the case of a treatment-dependent censoring data model, the null hypothesis requires the joint distributional equality of the censored data processes in the two groups, giving rise to $\boldsymbol{\delta}$, and of response data $\boldsymbol{X}$ conditional on $\boldsymbol{\delta}$, i.e.

$$H_0^G : \left\{ \left[\boldsymbol{\delta}_1 \stackrel{d}{=} \boldsymbol{\delta}_2 \right] \cap \left[\left(\boldsymbol{X}_1 \stackrel{d}{=} \boldsymbol{X}_2 \right) | \boldsymbol{\delta} \right] \right\}$$

The assumed exchangeability, in the null hypothesis, of the n individual data vectors in $(\boldsymbol{X}, \boldsymbol{\delta})$, with respect to the two groups, implies that the treatment effects are null on *all* observed and unobserved variables. In other words, H_0 implies that there is no difference in distribution for the multivariate censoring indicator variables $\boldsymbol{\delta}_j$, $j = 1, 2,$ and, conditionally on $\boldsymbol{\delta}$, for actually observed data $\boldsymbol{X}$. As a consequence, it is not necessary to specify both the censored data process and the data distribution, provided that marginally unbiased permutation tests are available. In particular, it is not necessary to specify the dependence relation structure in $(\boldsymbol{X}, \boldsymbol{\delta})$ because it is nonparametrically processed by the NPC.

In this framework, the global null and alternative hypotheses, as in Section 3.3, may be broken down into the D sub-hypotheses

$$H_0^G : \left\{ \left[\bigcap_{i=1}^{D} \left(\delta_{i1} \stackrel{d}{=} \delta_{i2} \right) \right] \cap \left[\bigcap_{i=1}^{D} \left(X_{i1} \stackrel{\mathrm{d}}{=} X_{i2} \right) | \boldsymbol{\delta} \right] \right\} = \left\{ H_0^{\boldsymbol{\delta}} \cap H_0^{1X|\delta} \right\} = \left\{ \left(\bigcap_{i=1}^{D} H_{0i}^{\boldsymbol{\delta}} \right) \cap \left(\bigcap_{i=1}^{D} H_{0i}^{1X|\delta} \right) \right\}$$

against

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

where $H_{0i}^{\boldsymbol{\delta}}$ indicates the equality in distribution among the two levels of the ith marginal component of the censoring indicator (censoring process), and $H_{0i}^{1X|\delta}$ indicates the equality in distribution of the ith component of $\boldsymbol{X}$, conditional on $\boldsymbol{\delta}$.

As before $0 < t_1 < t_2 < \cdots < t_D = \tau$ are the observed distinct times to event occurrences (primary and censoring events).

For each time t_i, let r_{1i}^* and r_{2i}^* be the number of subjects *at risk* at the start of the ith period in a permutation of individual profiles

$$r_{ji}^* = \sum_{m=1}^{n_j} V_{mi}^* O_{mi}^*, \quad j = 1, 2.$$

Let d_{ji}^*, and c_{ji}^*, $j = 1, 2,$ be the number of primary and censoring events, respectively, in two groups at ith period $(t_{i-1}, t_i]$, $1 \leq i \leq D$.

Of course, $r_i = r_{1i}^* + r_{2i}^*$, $d_i = d_{1i}^* + d_{2i}^*$ and $c_i = c_{1i}^* + c_{2i}^*$ are permutation invariant quantities.

At this stage, in the spirit of Mantel–Cox, we may apply two test statistics (log-rank type) for primary and censoring events, respectively, followed by their NPC, that is

$$T_D^* = \frac{\sum_i (d_{1i}^* - \bar{d}_{1i}^*)}{\sqrt{\sum_i V_{Di}^*}}$$

and

$$T_C^* = \frac{\sum_i (c_{1i}^* - \bar{c}_{1i}^*)}{\sqrt{\sum_i V_{Ci}^*}}$$

where

$$\bar{d}_{1i}^* = d_i \frac{r_{1i}^*}{r_i} \quad \text{and} \quad \bar{c}_{1i}^* = c_i \frac{r_{1i}^*}{r_i}$$

are the permutation means and

$$V_{Di}^* = d_i \frac{r_{1i}^*}{r_i} \frac{r_{2i}^*}{r_i} \frac{r_i - d_i}{r_i - 1} \quad \text{and} \quad V_{Ci}^* = c_i \frac{r_{1i}^*}{r_i} \frac{r_{2i}^*}{r_i} \frac{r_i - c_i}{r_i - 1}$$

are the permutation variances of d_{1i}^* and c_{1i}^*, respectively.

Alternatively, in the spirit of Anderson–Darling, we may use

$$T_{D,AD}^* = \sum_i \frac{d_{1i}^* - \bar{d}_{1i}^*}{\sqrt{V_{Di}^*}}$$

and

$$T_{C,AD}^* = \sum_i \frac{c_{1i}^* - \bar{c}_{1i}^*}{\sqrt{V_{Ci}^*}}$$

followed by the NPC. Test statistics within a permutation framework are usually chosen from well-known statistics in the parametric or standard non-parametric field since such statistics are generally unbiased and consistent, as in this case.

It should be noted that T_D^*, T_C^*, $T_{D,AD}^*$ and $T_{C,AD}^*$ are nothing more than *direct combinations* of D partial tests, for primary and censoring events respectively.

We also observe that the Anderson–Darling-type statistics are the sum of standardized statistics, while the Mantel–Cox-type statistics are the standardized sum of statistics.

If using weights $w_i, 1 \leq i \leq D$, which can be different for primary and censoring events, the Mantel–Cox-type statistic T_D^* becomes

$$T_D^*(w_1, \ldots, w_D) = \frac{\sum_i w_i (d_{1i}^* - \bar{d}_{1i}^*)}{\sqrt{\sum_i V_{Di}^*}} \tag{1}$$

and the Anderson–Darling-type statistic T_D^* becomes

$$T_{D,AD}^*(w_1, \ldots, w_D) = \sum_i w_i \frac{d_{1i}^* - \bar{d}_{1i}^*}{\sqrt{V_{Di}^*}} \tag{2}$$

where

$$V^*_{Di}(w_1,\ldots,w_D) = w_i^2 d_i \frac{r^*_{1i}}{r_i}\frac{r^*_{2i}}{r_i}\frac{r_i - d_i}{r_i - 1}$$

Analogous expressions are obtained for $T^*_C(w_1,\ldots,w_D)$ and $T^*_{C,AD}(w_1,\ldots,w_D)$.

The final step of this procedure is the combination of these partial tests according to NPC Theory – Pesarin and Salmaso.[50]

4 Simulation study

We set up a simulation study to compare the results obtained using different statistical tests. We analysed the NPC tests, the statistical tests described in Wakounig et al.[37] and those made available by the Proc Lifetest of SAS[55] under the assumption of non-informative censoring.

We consider $n = n_1 + n_2$ individuals in the combined sample of two groups G_1 and G_2 where G_1 (G_2) has size n_1 (n_2). The data concerning primary events (or deaths) and censoring events are collected at each of D time points $0 < t_1 < t_2 < \cdots < t_D$ which are considered as time to event occurrences. At the i-th time point, $i = 1,\ldots,D$, for each group $G_j, j = 1, 2$, the relevant data are the number of individuals at risk $r_{i,j}$, the number of deaths $d_{i,j}$ and the number of censoring events $c_{i,j}$ in the period of time between t_{i-1} (exclusive) and t_i (inclusive), being $t_0 = 0$. This is equivalent to considering the time t_i as time to event occurrences.

The data scheme is summarized in Table 2, where $r_{j,i}$, the number of individuals of group j at risk at time t_i, is computed as the difference between $r_{j,i-1}$, the number of individuals of group j at risk at time t_{i-1}, and the number of primary and secondary events ($d_{j,i-1}$ and $c_{j,i-1}$, respectively) which occurred between t_{i-1} (exclusive) and t_i (inclusive)

$$r_{j,i} = r_{j,i-1} - (d_{j,i-1} + c_{j,i-1}), \quad j = 1, 2 \tag{3}$$

with $r_{1,0} = n_1, r_{2,0} = n_2,\ r_0 = n$ and $d_{j,0} = c_{j,0} = 0$. From equation (3), it follows that $r_{1,1} = n_1, r_{2,1} = n_2$ and $r_1 = n$.

In this study, all the tests are based on the weighted log-rank type test statistics (Mantel–Cox) $T_D(w_1,\ldots,w_D)$ and $T_C(w_1,\ldots,w_D)$

$$T^*_D(w_1,\ldots,w_D) = \sum_i w_i \frac{d^*_{1,i} - d_i \frac{r^*_{1,i}}{r_i}}{\sqrt{V_D}} \tag{4}$$

$$T^*_C(w_1,\ldots,w_D) = \sum_i w_i \frac{c^*_{1,i} - c_i \frac{r^*_{1,i}}{r_i}}{\sqrt{V_C}} \tag{5}$$

where w_j are weights which are defined as described in Table 3, $V_D = \sum_i V^*_{Di}(w_1,\ldots,w_D)$ and $V_C = \sum_i V^*_{Ci}(w_1,\ldots,w_D)$.

We use $\hat{S}(.)$ to denote the Kaplan–Meier estimator of the survival function

$$\hat{S}(t) = \prod_{t_i \le t}\left(1 - \frac{d_i}{r_i}\right)$$

Table 2. Situation at time t_i, $i = 1,\ldots,D$.

	Number of individuals		
Group	At risk at t_i	Dead	Censored
G_1	$r_{1,i}$	$d_{1,i}$	$c_{1,i}$
G_2	$r_{2,i}$	$d_{2,i}$	$c_{2,i}$
Total	r_i	d_i	c_i

Table 3. Weights for the different statistical tests.

Test Statistics	Name of test in Proc Lifetest	Name of test in Wakounig et al.[36]	Weights w_j
T_M	Logrank	Mantel	1
T_B	Wilcoxon	Breslow	r_j
T_T	Tarone–Ware		$\sqrt{r_j}$
T_P	Peto–Peto		$\tilde{S}(t_j)$
T_{P_*}	Modified Peto–Peto		$\tilde{S}(t_j)\frac{r_j}{r_j+1}$
T_H	Harrington–Fleming (p, q), $p, q \geq 0$		$(\hat{S}(t_j))^p(1-\hat{S}(t_j))^q$
T_{M_*}		Modified Mantel	$\frac{1}{\hat{G}(t_j)}$
T_R		Prentice	$\hat{S}(t_j)$
T_{R_*}		Modified Prentice	$\frac{\hat{S}(t_j)}{\hat{G}(t_j)}$

$\hat{G}(.)$ to denote the Kaplan–Meier estimator of the follow-up distribution and $\tilde{S}(.)$ to denote an estimate of the survival function given by

$$\tilde{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j + 1}\right)$$

According to the approach implemented in the Proc Lifetest of SAS,[55] the p-value for the two-sided test is computed as $2(1 - \Phi(|t_{obs}|))$ where Φ is the cumulative distribution function of the standard normal distribution and $|t_{obs}|$ is the absolute value of the observed value of $T_D(w_1, \ldots, w_D)$. The justification for using the normal distribution is that under H_0 and *non-informative censoring*, the statistic $T_D(w_1, \ldots, w_D)$ *asymptotically* follows a standard normal distribution. Working in the two-group situation, it is easy to see that Proc Lifetest computes p-values as $1 - F_{\chi_1^2}(t_{obs}^2)$, where $F_{\chi_1^2}$ is the cumulative distribution function of the Chi-square distribution with one degree of freedom. It follows that the same p-values as those computed using $|t_{obs}|$ and the standard normal distribution are obtained.

Our goal is to compare the statistical tests described in Table 3 with those that can be obtained using NPC-based tests. As a test statistic for partial tests, we consider the log-rank type Mantel statistic which is obtained using all weights w_i equal to 1. We used this statistic for both deaths (T_{MD}) and censoring events (T_{MC})

$$T^*_{MD} = T^*_D(1, \ldots, 1) = \sum_{i=1}^{D} \frac{d^*_{1,i} - d_i \frac{r^*_{1,i}}{r_i}}{\sqrt{V^*_D}}$$

$$T^*_{MC} = T^*_C(1, \ldots, 1) = \sum_{i=1}^{D} \frac{c^*_{1,i} - c_i \frac{r^*_{1,i}}{r_i}}{\sqrt{V^*_C}}$$

where V^*_D (V^*_C) is the variance of the observed number of deaths (number of censoring events) in group G_1. It is worth noting that the methodology can easily be extended to consider any of the statistics in Table 3.

We can now summarize the steps of the NPC test procedure.

- We consider B random permutations of the pooled profiles $(\boldsymbol{X}, \boldsymbol{\delta})$. For each permutation b, we compute both the test statistics $T^{(b)}_{MD}$ and $T^{(b)}_{MC}$, $b = 1, \ldots, B$. We consider $B = 1000$ permutations and omit the * to simplify the notation.
- We compute the p-values corresponding to deaths p_{MD} and censoring events p_{MC} as

$$p_{MD} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(|T^{(b)}_{MD}| \geq |T^{(0)}_{MD}|)}{B + 1}$$

$$p_{MC} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(|T_{MC}^{(b)}| \geq |T_{MC}^{(0)}|)}{B+1}$$

where $T_{MD}^{(0)}$ and $T_{MC}^{(0)}$ are the statistics corresponding to the observed data.

- We compute the p-value-like statistics as

$$\lambda_{MD}^{(b)} = \frac{\frac{1}{2} + \sum_{j \in \{1,\ldots,B\}, j \neq b} I(|T_{MD}^{(j)}| \geq |T_{MD}^{(b)}|)}{B+1} \quad b = 1, \ldots B$$
$$\lambda_{MC}^{(b)} = \frac{\frac{1}{2} + \sum_{j \in \{1,\ldots,B\}, j \neq b} I(|T_{MC}^{(j)}| \geq |T_{MC}^{(b)}|)}{B+1} \quad b = 1, \ldots B.$$

- We compute $\psi(\lambda_{MD}^{(b)}, \lambda_{MC}^{(b)})$ using the combination function Ψ. In this work, we used Tippet (ψ_T) and Fisher (ψ_F) combination functions which are defined as

$$\psi_T(x_1, x_2) = \min\{x_1, x_2\}$$
$$\psi_F(x_1, x_2) = -\log x_1 - \log x_2.$$

- We compute the p-value for both the Tippet (p_{MT}) and the Fisher (p_{MF}) combination functions

$$p_{MT} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(\min\{\lambda_{MD}^{(b)}, \lambda_{MC}^{(b)}\} \leq \min\{p_{MD}, p_{MC}\})}{B+1}$$
$$p_{MF} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(-\log \lambda_{MD}^{(b)} - \log \lambda_{MC}^{(b)} \geq -\log p_{MD} - \log p_{MC})}{B+1}$$

- As the output of the NPC test procedure, we consider both the one-response p-values p_{MD} (for deaths) and p_{MC} (for censoring events) as well as the combined p-values p_{MT} (using the Tippet combination function) and p_{MF} (using the Fisher combination function).

4.1 Different scenarios

For primary events, we set up the same scenarios as those described in Wakounig et al.,[37] see Figures 1 and 2. We consider two groups with $n_1 = n_2 = 40$ individuals and 10 equispaced times $0 < t_1 < \cdots < t_{10} = 4$ in the time interval $[0, \tau = 4]$

$$t_i = i\frac{4}{10}, \quad i = 1, \ldots, 10.$$

Survival times for group G_1 are drawn randomly from an exponential distribution with hazard function $h_1(t) = 0.5$, while survival times for group G_2 are drawn randomly from distributions with hazard function $h_2(t)$ defined as follows

(a) proportional hazards: $h_2(t) = 2h_1(t)$
(b) converging hazards: $h_2(t) = (1 + \frac{2.88}{1+5t})h_1(t)$
(c) diverging hazards: $h_2(t) = (1 + 1.86t)h_1(t)$
(d) identical hazards: $h_2(t) = h_1(t)$
(e) diverging hazards: $h_2(t) = 0.11 \exp(1.5t)$

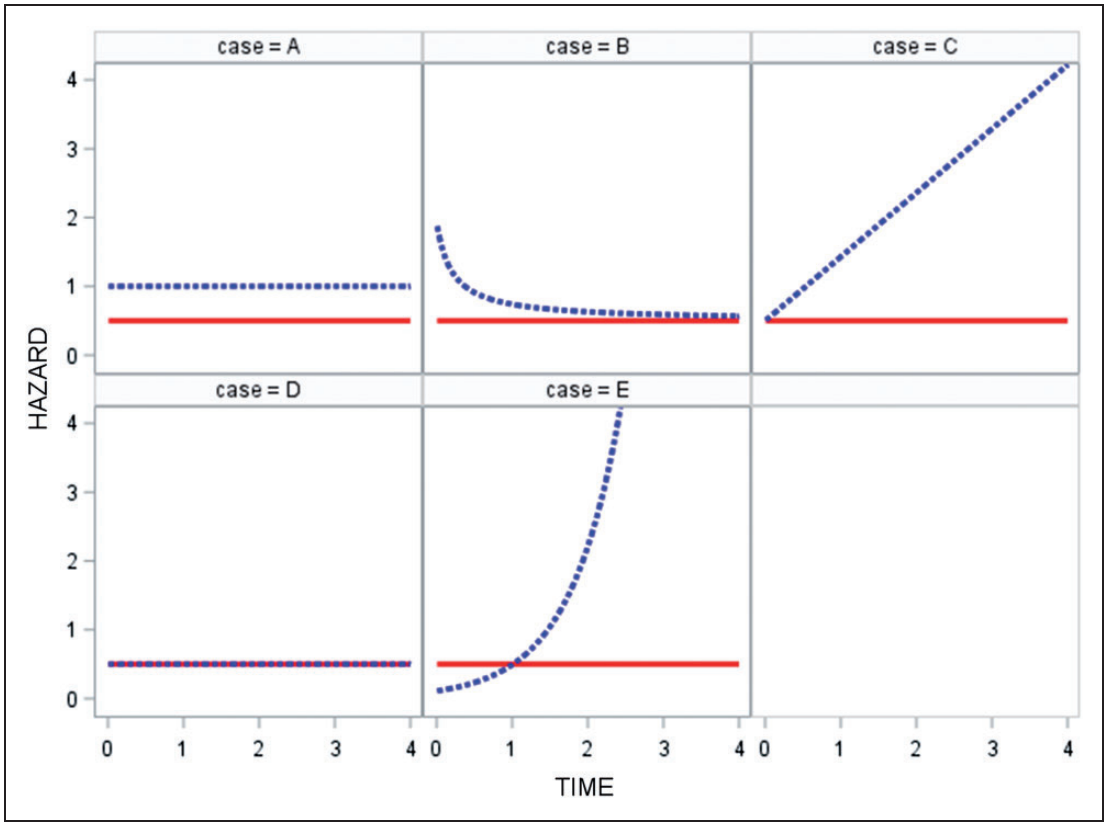

Figure 1. Hazard functions for A,B,C,D and E scenarios.

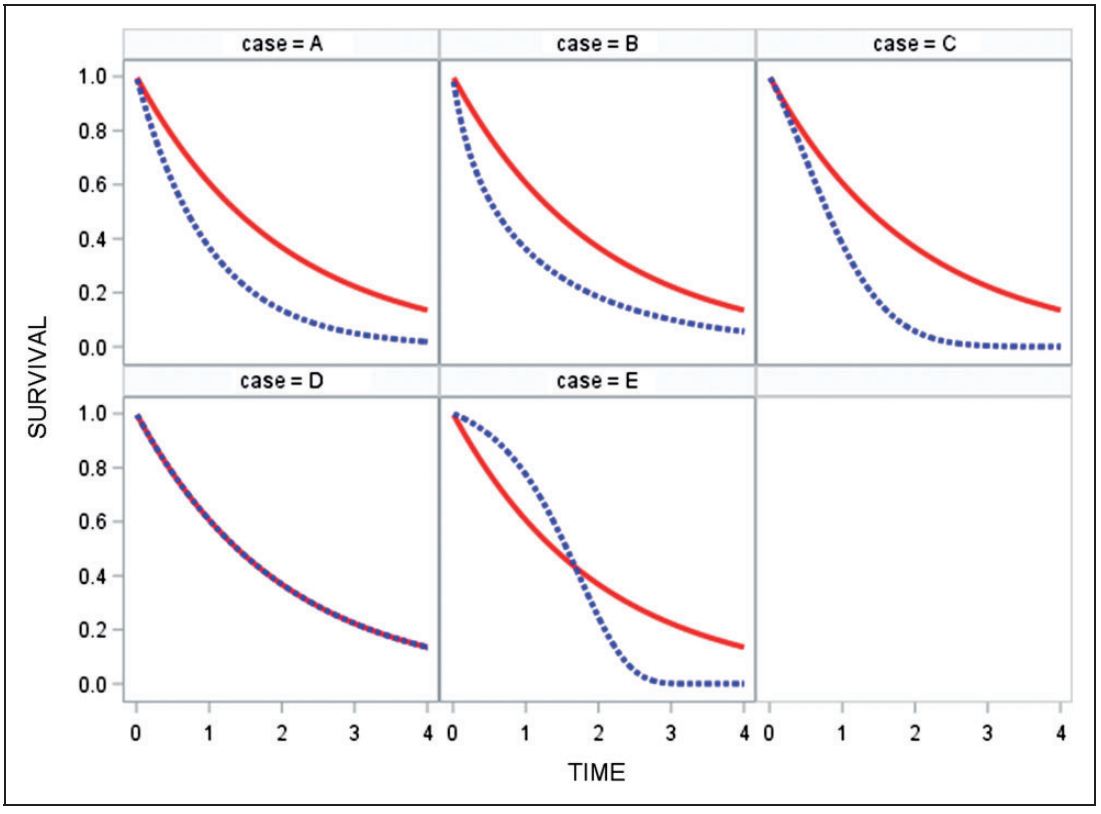

Figure 2. Survival functions for A,B,C,D and E scenarios.

Each scenario has been considered under three different conditions of censoring events:

$C_{0,0}$ no censoring events, i.e. $P(\delta_{mj} = 1) = 0 \ \forall m, j$.
$C_{10,10}$ probability of censoring events equal to 10% for both G_1 and G_2 (non-informative censoring), i.e $P(\delta_{mj} = 1) = 0.10 \ \forall m, j$.
$C_{8,2}$ probability of censoring events equal to 8% for G_1 and equal to 2% for G_2 (informative censoring), i.e. $P(\delta_{m1} = 1) = 0.08, P(\delta_{m2} = 1) = 0.02 \ \forall m$.

It follows that we considered a total of 15 scenarios $s = (s_1, s_2)$ with $s_1 \in \{A, B, C, D, E\}$ and $s_2 \in \{C_{0,0}, C_{10,10}, C_{8,2}\}$. For each scenario, we ran $N = 1000$ simulations. In Section 4.6, we consider five further scenarios that are inspired by one of the real examples that we will study in Section 5.

We observe that in the 10 scenarios $s = (s_1, s_2)$ with $s_1 \in \{A, B, C, D, E\}$ and $s_2 \in \{C_{0,0}, C_{10,10}\}$, we have no censoring or non-informative censoring. In these circumstances, the comparison between NPC and the other tests is fair. In the remaining scenarios, censoring is informative and all the tests, except NPC, should not be used. We use them mainly to show the risks that occur when a test suitable for non-informative censoring is used when censoring is informative.

The simulation data have been generated according to the procedure below, considering the generic interval $(t_{i-1}, t_i]$, $i = 1, \ldots, D$.

As the first step of the procedure, we simulate censoring events. For each individual at risk at time t_{i-1}, we consider the Bernoulli random variable $C_{mj} \sim \text{Bernoulli}\left(P(\delta_{mj} = 1)\right)$ and, using a function which generates random numbers from a specified distribution, we make a virtual experiment. If we get $C_{mj} = 1$, the individual is censored. If a censoring event does not occur (i.e. $C_{mj} = 0$), we simulate a primary event. We compute

$$\begin{aligned}\pi_{mj,i} &= P(X_{mj} \le t_i | X_{mj} > t_{i-1}, C_{mj} = 0) = \\ &= \frac{P(C_{mj} = 0)P(t_{i-1} < X_{mj} \le t_i | C_{mj} = 0)}{P(C_{mj} = 0)P(X_{mj} > t_{i-1} | C_{mj} = 0)} = \\ &= \frac{F_j(t_i) - F_j(t_{i-1})}{1 - F_j(t_i)}\end{aligned}$$

where $F_j(t) = 1 - \exp(-\int_0^t h_j(s)\mathrm{d}s)$, $j = 1, 2$. We consider the Bernoulli random variable $D_{mj} \sim \text{Bernoulli}\left(\pi_{mj,i}\right)$ and as before using a function which generates random numbers from a specified distribution, we perform a virtual experiment. If we get $D_{mj,i} = 1$, a primary event has occurred to the m-th individual of the j-th group in the time interval $(t_{i-1}, t_i]$. Otherwise the m-th individual of the j-th group will be considered to be *at risk* in the next time interval $(t_i, t_{i+1}]$.

4.2 Results

For each scenario s and for each simulation, we follow the following steps:

(1) We compute all the test statistics listed in Table 3. We use T_Q to denote one of these statistics.
(2) For each test statistic T_Q, we compute the corresponding p-value p_Q as $2(1 - \Phi(|t_{Q,obs}|))$, where Φ is the cumulative distribution function of the standard normal random variable and $|t_{Q,obs}|$ is the absolute value of the observed value of T_Q. We get p_M for Logrank, p_B for Wilcoxon, p_T for Tarone–Ware, $p_{P_\star}$ for Modified Peto–Peto, p_H for Harrington–Fleming, $p_{M_\star}$ for Modified Mantel, p_R for Prentice and $p_{R_\star}$ for Modified Prentice.
(3) Using the NPC procedure as described earlier, we get p_{MD}, p_{MC}, p_{MT} and p_{MF}.

Then, using the results of the $N = 1000$ simulations, the (α_Q^o, α) curve is built for each scenario and each test statistic T_Q computing, for each nominal $\alpha_k = \frac{k}{100}, k = 1, \ldots, 99$, the corresponding *achieved* $\alpha_{Q,k}^o$

$$\alpha_{Q,k}^o = \frac{\sum_i^N I(p_Q < \alpha_k)}{N}$$

In particular for each statistical test and for each scenario, we consider the achieved values $\alpha^o_{Q,k}$ for $k = 1, 5, 10$. It corresponds to evaluate, using the $N = 1000$ simulations, the power/size of each statistical test under the commonly used values of α_k, i.e. 1%, 5% and 10% (see Appendix I).

In the interest of saving space and of facilitating the comparison among all the tests, we define a *global* score for each test and for each scenario and we report the observed values of power/size in the tables of Section A.

In the scenarios for which the null hypothesis is not true, a test is *good* when, for a given α, the achieved alpha is *large*, which means that the *p*-values computed in the different simulations are *often* less than the nominal α; in other words, the null hypothesis is *often* rejected. The scenarios for which the null hypothesis is not true are $\{(s_1, s_2) : s_1 \in \{A, B, C, E\}, s_2 \in \{C_{0,0}, C_{10,10}, C_{8,2}\}\}$ and $(D, C_{8,2})$. We observe that for the $(D, C_{8,2})$ scenario, the null hypothesis is not true since the distributions of the censoring events are different. In these cases, we proceed as follows: For each primary-event scenario s_1 and for each nominal α_k, we compute the maximum $m^{(s_1)}_{\alpha_k}$ of all the achieved $\alpha^o_{Q,k}$ by the different methods. Then for each method Q, we compute the score $g^{(s_1)}_Q$ as the number of times over the 99 different values of the nominal α that the method Q provides a value $\alpha^o_{Q,k}$ *close to the maximum* $m^{(s_1)}_{\alpha_k}$. In practice, by *close to the maximum* we mean that the value of $\alpha^o_{Q,k}$ lies in the interval between 90% of the maximum and the maximum itself, i.e. $[0.9m^{(s_1)}_{\alpha_k}, m^{(s_1)}_{\alpha_k}]$. This interval has been defined to mitigate the effect of simulations. By definition, the scores are between 0 and 1. Finally, as a global measure of the performance of method Q, we also compute an average score g_Q as the geometric mean of the scores obtained in each single primary-event scenario.

On the other hand, in the scenarios for which the null hypothesis is true, a test is *good* when, for a given α, the achieved alpha is *close to* the nominal α. The scenarios for which the null hypothesis is true are $(D, C_{0,0})$ and $(D, C_{10,10})$. For each test Q, we compute the error statistic e_Q defined as the quadratic mean of the differences between the observed $\alpha^o_{Q,k}$ and the nominal α_k

$$e_Q = \sqrt{\frac{1}{99}\sum_{k=1}^{99}\left(\alpha^o_{Q,k} - \alpha_k\right)^2} = \sqrt{\frac{1}{99}\sum_{k=1}^{99}\left(\alpha^o_{Q,k} - \frac{k}{100}\right)^2}. \quad (6)$$

4.3 Scenarios $(s_1, C_{8,2})$; informative censoring

First we consider the primary-event scenarios A, B, C, D and E when the probability of censoring events is different. We observe that in this case, the scenario D does not correspond to the null hypothesis. As can be seen in Figure 3, NPC tests perform better than all the other competitors.

The scores are reported in Table 4. The last column shows the global score g_Q. We see that the NPC tests are the best test procedures with a global score equal to 0.95 (Tippet combination function) and 0.96 (Fisher combination function). All the other tests', global score is no higher than 0.40. We also observe that for scenarios A, B, C, D and E, the NPC scores are the highest. It is worth noting that univariate permutation tests considered as separate tests are not as good as the NPC tests in detecting the difference between the two groups (their global scores are 0.40 and 0.61 respectively against 0.95 and 0.96 of the NPC tests).

4.4 Scenario $(s_1, C_{10,10})$; non-informative censoring

We now consider the primary-event scenarios A, B, C, D and E when the probability of censoring events is equal. As we can see from the scores in Table 5, NPC tests perform very well in scenarios A, C and E (Tippet scores are always greater than 0.92). In scenario B, Tippet and Fisher NPC tests get 0.70 and 0.69, respectively, which is good but not excellent. In this case, the univariate permutation test performs very well, getting a score of 0.90. This provides a practical guideline and suggests the univariate tests could be very useful in some situations.

With respect to the $(D, C_{10,10})$ scenario, which corresponds to the null hypothesis, we observe that all the tests perform quite well, with the error statistic e_Q less than 1.4%, as we can see in Table 6.

4.5 Scenario $(s_1, C_{0,0})$; no censoring

In this scenario, there are no censoring events (formally the only censoring events will occur at the end of the study i.e. at time τ). It follows that the permutation test on the primary events and the NPC tests all give exactly the same result, i.e. $p_{MD} = p_{MT} = p_{MF}$ (p_{MC} is not computed). As Table 7 shows, the permutation tests perform very well in scenarios A, B, C and E (the total score is 0.96).

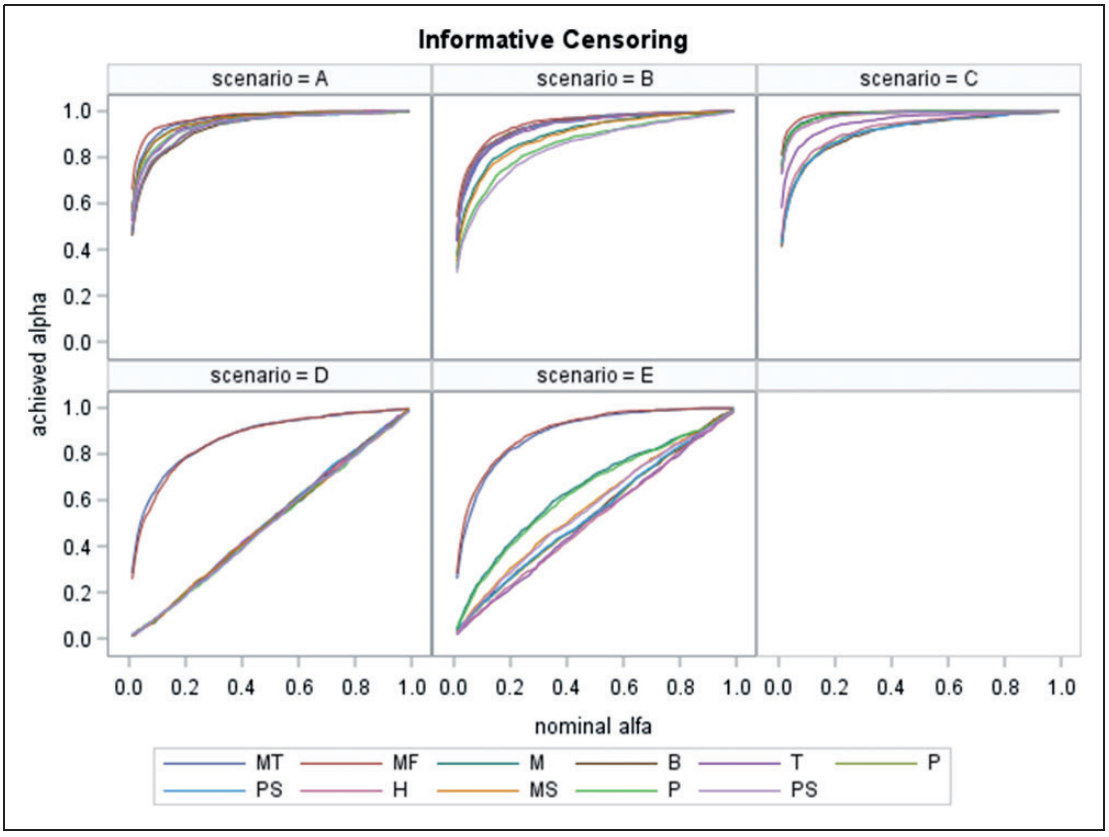

Figure 3. Achieved vs. nominal α.

Table 4. Scores for informative censoring scenarios.

Test	A	B	C	D	E	Tot
NPC Tippet	0.98	0.97	0.99	0.87	0.94	0.95
NPC Fisher	1.00	1.00	1.00	0.84	0.99	0.96
Permutation on primary events	0.96	0.81	1.00	0.10	0.14	0.40
Permutation on censoring events	0.42	0.55	0.38	1.00	1.00	0.61
Logrank/Mantel	0.98	0.83	1.00	0.10	0.13	0.40
Wilcoxon/Breslow	0.83	0.99	0.71	0.11	0.13	0.38
Tarone–Ware	0.89	0.98	0.90	0.11	0.09	0.38
Peto–Peto	0.85	0.98	0.75	0.11	0.12	0.38
Modified Peto–Peto	0.85	0.99	0.74	0.11	0.12	0.38
Harrington–Fleming	0.84	0.99	0.78	0.11	0.10	0.37
Modified Mantel	0.97	0.78	1.00	0.09	0.13	0.39
Prentice	0.92	0.63	1.00	0.09	0.14	0.37
Modified Prentice	0.90	0.59	1.00	0.10	0.11	0.36

We also observe that all the tests perform well under the null hypothesis (scenario D) with e_Q always less than 1.7% (see Table 8).

4.6 Scenario (s_1, C_{dog}); censoring similar to the 'Cancer in dogs' case

We conclude this section with the study of five further scenarios inspired by one of the real examples discussed in Section 5 (the 'Cancer in dogs' case). More precisely, we consider two groups with different sizes, i.e. $n_1 = 20$ and $n_2 = 40$ individuals. Then for each primary-event scenario $s_1 \in \{A, B, C, D, E\}$, we generate censoring events with

Table 5. Scores for non-informative censoring.

Test	A	B	C	E	Tot
NPC Tippet	0.95	0.70	0.96	0.92	0.73
NPC Fisher	0.91	0.69	0.94	0.78	0.72
Permutation on primary events	1.00	0.90	1.00	0.74	0.79
Permutation on censoring events	0.10	0.11	0.09	0.98	0.13
Logrank/Mantel	1.00	0.90	1.00	0.75	0.80
Wilcoxon/Breslow	0.94	1.00	0.65	0.86	0.88
Tarone–Ware	0.99	1.00	0.87	0.88	0.62
Peto–Peto	0.97	1.00	0.72	0.88	0.83
Modified Peto–Peto	0.97	1.00	0.71	0.87	0.86
Harrington–Fleming	0.97	1.00	0.77	0.87	0.64
Modified Mantel	1.00	0.86	1.00	0.76	0.66
Prentice	1.00	0.79	1.00	0.70	0.76
Modified Prentice	1.00	0.77	1.00	0.68	0.61

Table 6. e_Q statistics for non-informative censoring, see equation (6) for e_Q definition.

Test	e_Q
NPC Tippet	0.007
Permutation on primary e	0.009
Permutation on censoring	0.014
NPC Fisher	0.008
Logrank/Mantel	0.007
Wilcoxon/Breslow	0.014
Tarone–Ware	0.009
Peto–Peto	0.012
Modified Peto–Peto	0.012
Harrington–Fleming	0.010
Modified Mantel	0.006
Prentice	0.013
Modified Prentice	0.013

Table 7. Scores for no-censoring scenarios.

Test	A	B	C	E	Tot
Permutation on primary events	1.00	0.86	1.00	0.88	0.96
Logrank/Mantel	1.00	0.87	1.00	0.91	0.97
Wilcoxon/Breslow	0.93	1.00	0.84	0.83	0.62
Tarone–Ware	0.99	1.00	0.96	0.85	0.66
Peto–Peto	0.94	1.00	0.81	0.81	0.62
Modified Peto–Peto	0.94	1.00	0.80	0.84	0.61
Harrington–Fleming	0.93	1.00	0.84	0.83	0.62
Modified Mantel	1.00	0.83	1.00	0.91	0.73
Prentice	0.99	0.57	1.00	0.97	0.87
Modified Prentice	0.95	0.45	1.00	0.98	0.62

probabilities p_i which are different in the 10 time intervals $(t_{i-1}, t_i]$, $i = 1, \ldots, 10, t_0 = 0$ and also different for the two groups. These probabilities are approximately the same as those observed in the 'Cancer in dogs' real case and are reported in Table 9.

We obtain the scores which are detailed in Table 10.

As for scenarios $(s_1, C_{8,2})$, the NPC tests display very good behaviour.

Table 8. e_Q statistics for no censoring, see equation (6) for e_Q definition.

Test	e_Q
Permutation on primary events	0.011
Logrank/Mantel	0.010
Wilcoxon/Breslow	0.017
Tarone–Ware	0.013
Peto-Peto	0.017
Modified Peto–Peto	0.017
Harrington–Fleming	0.017
Modified Mantel	0.009
Prentice	0.012
Modified Prentice	0.011

Table 9. Probabilities of censoring similar to the 'Cancer in dogs' case.

Group	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
1	0.016	0.538	0.119	0.028	0.176	0.208	0.333	0.087	0.167	0.833
2	0.048	0.924	0.000	0.333	1.000	0.000	0.000	0.000	0.000	0.000

Table 10. Scores for censoring similar to the 'Cancer in dogs' case.

Test	A	B	C	D	E	Tot
NPC Tippet	0.99	0.99	1.00	0.97	1.00	0.99
NPC Fisher	1.00	1.00	1.00	0.94	1.00	0.99
Permutation on primary events	0.53	0.51	0.28	0.10	0.24	0.28
Permutation on censoring events	0.93	0.85	1.00	1.00	1.00	0.95
Logrank/Mantel	0.54	0.51	0.29	0.11	0.24	0.29
Wilcoxon/Breslow	0.42	0.57	0.18	0.10	0.40	0.28
Tarone–Ware	0.47	0.56	0.23	0.08	0.35	0.28
Peto–Peto	0.46	0.56	0.20	0.08	0.33	0.27
Modified Peto–Peto	0.46	0.57	0.21	0.08	0.34	0.27
Harrington–Fleming	0.47	0.56	0.20	0.08	0.32	0.27
Modified Mantel	0.51	0.52	0.24	0.09	0.30	0.28
Prentice	0.38	0.27	0.29	0.10	0.17	0.22
Modified Prentice	0.38	0.27	0.29	0.10	0.19	0.22

5 Real examples

We work on two real cases for which the data are publicly available. These examples are also studied in Wakounig et al.[37] The data sets are:

(1) 'Cancer in dogs'. Groups: 83 beagles receiving irradiation and bone marrow transplantation versus 198 control dogs. Endpoint: time till occurrence of cancer. Censoring: 90%. Source: Prentice and Marek.[56]
(2) 'Primary biliary cirrhosis'. Groups: 49 patients suffering from edema versus 263 free of edema, all of them included in a clinical trial of primary biliary cirrhosis of the liver. Endpoint: time till transplantation of the liver or death. Censoring: 60%. Source: Therneau and Grambsch.[57]

The product-limit survival estimates for the two cases are shown in Figures 4 and 5. The p-values resulting from all the tests under study are shown in Table 11.

From Table 11, we observe that for the 'Primary biliary cirrhosis' data set all the p-values, apart from the one corresponding to the permutation test on censoring events (p-value $= 0.736$), are almost zero.

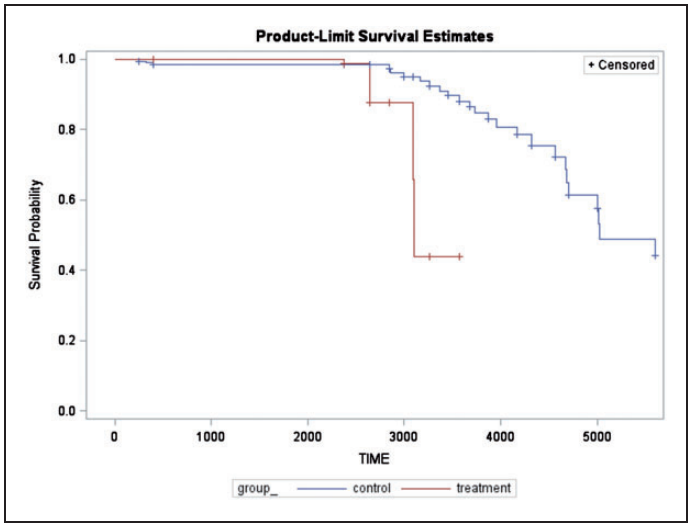

Figure 4. Product-limit survival estimates ('Cancer in dogs' data set).

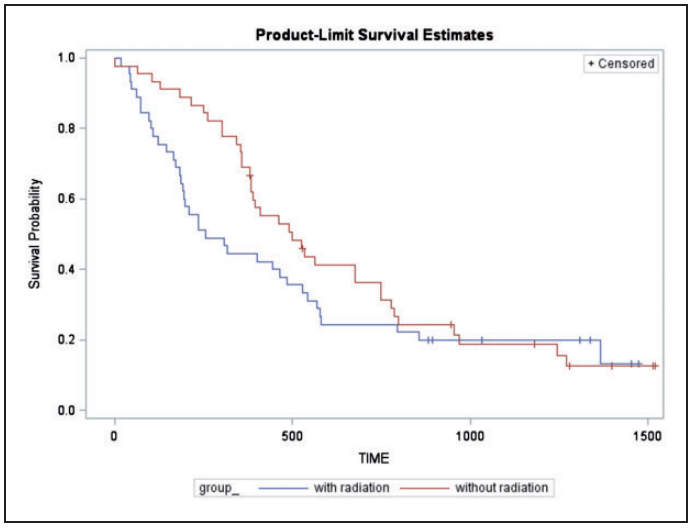

Figure 5. Product-limit survival estimates ('Primary biliary cirrhosis' data set).

Table 11. *p*-values for the real data applications.

Test	Cancer in dogs	Primary biliary cirrhosis
NPC Tippet	0.000	0.000
NPC Fisher	0.001	0.003
Logrank/Mantel	0.034	0.000
Wilcoxon/Breslow	0.526	0.000
Tarone–Ware	0.209	0.000
Peto–Peto	0.042	0.000
Modified Peto–Peto	0.043	0.000
Harrington–Fleming	0.040	0.000
Modified Mantel	0.035	0.000
Prentice	0.023	0.000
Modified Prentice	0.023	0.000

Table 12. Adjusted *p*-values for the 'Primary billiary cirrhosis' case.

Test	Unadjusted *p*-value	Adjusted Bonferroni *p*-value	Adjusted Sidak *p*-value
Permutation on primary events	0.000	0.000	0.000
Permutation on censoring events	0.736	1.000	0.930

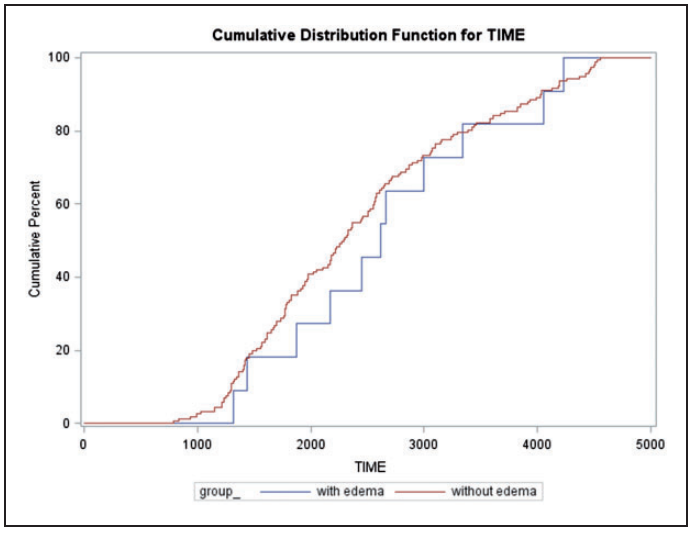

Figure 6. Empirical cumulative distribution function for censoring events.

Table 13. Adjusted *p*-values for the 'Cancer in dogs' case.

Test	Unadjusted *p*-value	Adjusted Bonferroni *p*-value	Adjusted Sidak *p*-value
Permutation on primary events	0.048	0.096	0.094
Permutation on censoring events	0.000	0.000	0.000

With the aim of further investigating the reason why the null hypothesis is rejected, we can consider the permutation tests on primary and censoring events. The original unadjusted *p*-values and the adjusted *p*-values that have been computed to take into account multiple testing issues are reported in Table 12. We observe that the censoring looks non-informative and the difference between the groups appears to be related to the different behaviour of the patients in terms of endpoint. The empirical cumulative distribution function of the censoring events is drawn in Figure 6.

For the 'Cancer in dogs' data set, all the *p*-values apart from those corresponding to the Wilcoxon/Breslow and Tarone–Ware tests are less than 5%. From Table 13, we observe that the censoring looks informative, in that it depends on the treatment effect while the difference for primary events is only weakly significant (adjusted p-values less than 10% but greater than 5%). The empirical cumulative distribution function of the censoring events is drawn in Figure 7.

In this case, it is worth noting that *p*-values of combined tests (NPC Fisher and NPC Tippet) are much lower than *p*-values of other tests because tests on censoring provide additional information related to differences between groups. This feature, i.e. the possibility to add a test on censoring, is only possible using our proposed approach.

Finally, we observe that real-life examples, where the primary event can occur due to causes unrelated to the *disease* under study, suggest an extension of our solution to competing risks, see Dobler et al.[49] and subsequent papers.

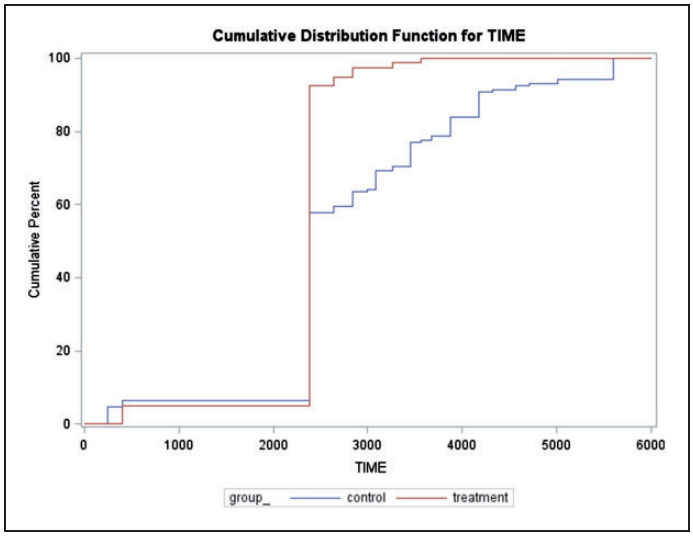

Figure 7. Empirical cumulative distribution function for censoring events.

6 Conclusion

Our simulation study covered a wide range of interesting scenarios for both primary and censoring events (15 different scenarios in total). The scenarios for primary events (A, B, C, D and E) are exactly the same as those studied by Wakounig et al.[37] We analysed the NPC tests, the weighted log-rank tests described in Wakounig et al.[37] and those made available by the Proc Lifetest of SAS.[55]

In relation to informative censoring, NPC tests showed very good behaviour. Their power was more than two times that of the other competitors for all the primary event scenarios.

NPC tests also performed very well in the presence of non-informative censoring or no censoring. In only one situation (primary event scenario B and non-informative censoring) did the combined test's performance decrease. In such a situation, only a test for primary events should be adopted.

In summary, we believe these noticeable results establish our NPC testing method as the standard for the analysis of survival processes. Indeed, it generally behaves at least as well as the best traditional tests when censoring is assumed to be non-informative, and where traditional tests are specialized. Moreover, its behaviour generally increases in power when both primary and secondary aspects are present. In this framework, using multiple testing techniques, NPC makes it possible to test which aspect is significant, if any, while controlling the family-wise error, thus providing a more comprehensive answer than traditional tests.

It is worth noting that all presented NPC tests can easily be extended to multivariate problems by combining multiple endpoints with the same fashion as for primary and censoring events. We are also hopeful that it is possible to extend them to competing risk models.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics* 2000; **42**: 12–25.
2. Aalen O. Nonparametric inference for a family of counting processes. *Ann Stat* 1978; **6**: 701–726.
3. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemo Rep Part 1* 1966; **50**: 163–170.
4. Peto R and Peto J. Asymptotically efficient rank invariant test procedures. *J R Stat Soc Ser A* 1972; **135**: 185–207.

5. Cox D. Regression models and life tables (with discussion). *J R Stat Soc B* 1972; **34**: 187–220.
6. Savage IR. Contributions to the theory of rank order statistics-the two-sample case. *Ann Math Stat* 1956; **27**: 590–615.
7. Gilbert J. *Random censorship*. Unpublished PhD Thesis, University of Chicago, Chicago, Illinois, 1962.
8. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 1965; **52**: 203–223.
9. Breslow N. A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika* 1970; **57**: 579–594.
10. Prentice RL, Kalbfleisch JD, Peterson AV Jr, et al. The analysis of failure times in the presence of competing risks. *Biometrics* 1978; **34**: 541–554.
11. Tarone RE and Ware J. On distribution-free tests for equality of survival distributions. *Biometrika* 1977; **64**: 156–160.
12. Harrington DP and Fleming TR. A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**(3): 553–566.
13. Gaugler T, Kim D and Liao S. Comparing two survival time distributions: an investigation of several weight functions for the weighted logrank statistic. *Commun Stat Simul Comput* 2007; **36**: 423–435.
14. Jones MP and Crowley J. A general class of nonparametric tests for survival analysis. *Biometrics* 1989; **45**: 157–170.
15. Tarone RE. Tests for trend in life table analysis. *Biometrika* 1975; **62**: 679–690.
16. Jonckheere AR. A distribution-free k-sample test against ordered alternatives. *Biometrika* 1954; **41**: 133–145.
17. Brown Jr BW, Hollander M and Korwar RM. *Nonparametric tests of independence for censored data with application to heart transplant studies*. Technical Report, DTIC Document, 1973.
18. Gill RD. Censoring and stochastic integrals. *Stat Neerland* 1980; **34**: 124–124.
19. O'Brien PC. A nonparametric test for association with censored datapages. *Biometrics*. JSTOR, 1978, pp.243–250.
20. Fleming TR and Harrington DP. *Counting processes and survival analysis*. Vol. 169, Hoboken, New Jersey, USA: John Wiley & Sons, 2011.
21. Breslow NE, Edler L and Berger J. A two-sample censored-data rank test for acceleration. *Biometrics* 1984; **40**: 1049–1062.
22. Fleming TR, Harrington DP and O'sullivan M. Supremum versions of the log-rank and generalized wilcoxon statistics. *J Am Stat Assoc* 1987; **82**: 312–320.
23. Lee SH, Lee EJ and Omolo BO. Using integrated weighted survival difference for the two-sample censored data problem. *Comput Stat Data Anal* 2008; **52**: 4410–4416.
24. Kosorok MR and Lin CY. The versatility of function-indexed weighted log-rank statistics. *J Am Stat Assoc* 1999; **94**: 320–332.
25. Andersen PK, Borgan O, Gill RD, et al. *Statistical models based on counting processes*. [Springer Series in Statistics] New York: Springer, 2012.
26. Andersen PK, Borgan O, Gill R, et al. Linear nonparametric tests for comparison of counting processes, with applications to censored survival data, correspondent paper. *Int Stat Rev* 1982; **50**: 219–244.
27. Gaugler T, Kim D and Liao S. Comparing two survival time distributions: an investigation of several weight functions for the weighted logrank statistic. *Commun Stat Simul Comput* 2007; **36**: 423–435.
28. Pepe MS and Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* 1989; **45**: 497–507.
29. Pepe MS and Fleming TR. Weighted Kaplan-Meier statistics: large sample and optimality considerations. *J R Stat Soc Ser B* 1991; **53**: 341–352.
30. Koziol J. A two sample Cramér–von Mises test for randomly censored data. *Biomet J* 1978; **20**: 603–608.
31. Schumacher M. Two-sample tests of Cramér–von Mises-and Kolmogorov–Smirnov-type for randomly censored data. *Int Stat Rev* 1984; **52**: 263–281.
32. Brookmeyer R and Crowley J. A confidence interval for the median survival time. *Biometrics* 1982; **38**: 29–41.
33. Hudgens M and Satten G. Midrank unification of rank tests for exact, tied, and censored data. *J Nonpara Stat* 2002; **14**: 569–581.
34. DiRienzo A. Nonparametric comparison of two survival-time distributions in the presence of dependent censoring. *Biometrics* 2003; **59**: 497–504.
35. Heller G and Venkatraman E. A nonparametric test to compare survival distributions with covariate adjustment. *J R Stat Soc Ser B* 2004; **66**: 719–733.
36. Zhang Y and Rosenberger WF. On asymptotic normality of the randomization-based logrank test. *Nonpara Stat* 2005; **17**: 833–839.
37. Wakounig S, Heinze G and Schemper M. Non-parametric estimation of relative risk in survival and associated tests. *Stat Meth Med Res* 2015; **24**: 856–870.
38. Galimberti S and Valsecchi MG. Multivariate permutation test to compare survival curves for matched data. *BMC Med Res Meth* 2013; **13**: 1.
39. Callegaro A, Pesarin F and Salmaso L. Test di permutazione per il confronto di curve di sopravvivenza. *Stat Appl* 2003; **15**: 241–261.
40. Heinze G, Gnant M and Schemper M. Exact log-rank tests for unequal follow-up. *Biometrics* 2003; **59**: 1151–1157.
41. Kellerer AM and Chmelevsky D. Small-sample properties of censored-data rank tests. *Biometrics* 1983; 675–682.

42. Chen JJ and Gaylor DW. The upper percentiles of the distribution of the logrank statistic for small numbers of tumors. *Commun Stat Simul Comput* 1986; **15**: 991–1002.
43. Ali M. Exact versus asymptotic tests of trend of tumor prevalence in tumorigenicity experiments: a comparison of p-values for small frequency of tumors. *Drug Inform J* 1990; **24**: 727–737.
44. Soper K and Tonkonoh N. The discrete distribution used for the log-rank test can be inaccurate. *Biomet J* 1993; **35**: 291–298.
45. Heimann G and Neuhaus G. Permutational distribution of the log-rank statistic under random censorship with applications to carcinogenicity assays. *Biometrics* 1998; **54**: 168–184.
46. Neuhaus G. Conditional rank tests for the two-sample problem under random censorship. *Ann Stat* 1993; **21**: 1760–1779.
47. Brendel M, Janssen A, Mayer CD, et al. Weighted logrank permutation tests for randomly right censored life science data. *Scand J Stat* 2014; **41**: 742–761.
48. Janssen A and Mayer CD. Conditional studentized survival tests for randomly censored models. *Scand J Stat* 2001; **28**: 283–293.
49. Dobler D, Pauly M, et al. Bootstrapping Aalen-Johansen processes for competing risks: handicaps, solutions, and limitations. *Elect J Stat* 2014; **8**: 2779–2803.
50. Pesarin F and Salmaso L. *Permutation tests for complex data: theory, applications and software*. Chichester, West Sussex, United Kingdom: John Wiley & Sons, 2010.
51. Pesarin F. *Multivariate permutation tests: with applications in biostatistics*. Vol. 240, Chichester, West Sussex, United Kingdom: Wiley, 2001.
52. Pesarin F. *Permutation tests: multivariate*. Wiley StatsRef: Statistics Reference Online, 2016.
53. Pesarin F. Some elementary theory of permutation tests. *Commun Stat Theor Meth* 2015; **44**: 4880–4892.
54. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–592.
55. SAS Institute Inc. *SAS/STAT(R) 9.2 user's guide* 2004; 2nd ed. .
56. Prentice RL and Marek P. A qualitative discrepancy between censored data rank tests. *Biometrics* 1979; **35**: 861–867.
57. Therneau TM and Grambsch PM. *Modeling survival data: extending the Cox model*. New York, USA: Springer Science & Business Media, 2000.

Appendix 1. Power and size of the statistical tests

In this section, for each of scenario and each statistical test, we report the achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ corresponding to the commonly used nominal $\alpha_k, k = 1, 5, 10$ values, i.e. $\alpha_1 = 1\%$, $\alpha_5 = 5\%$ and $\alpha_{10} = 10\%$.

Appendix 1.1. Scenarios $(s_1, C_{8,2})$, informative censoring.

Table 14. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for $(A, C_{8,2})$.

Nominal α	1%	5%	10%
NPC Tippet	0.567	0.811	0.899
NPC Fisher	0.661	0.870	0.927
Permutation on primary events	0.553	0.788	0.874
Permutation on censoring events	0.249	0.461	0.576
Logrank/Mantel	0.567	0.792	0.877
Wilcoxon/Breslow	0.460	0.691	0.793
Tarone–Ware	0.523	0.746	0.829
Peto–Peto	0.475	0.708	0.806
Modified Peto–Peto	0.475	0.703	0.806
Harrington–Fleming	0.476	0.703	0.802
Modified Mantel	0.565	0.790	0.877
Prentice	0.541	0.764	0.845
Modified Prentice	0.536	0.749	0.834

Table 15. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for $(B, C_{8,2})$.

Nominal α	1%	5%	10%
NPC Tippet	0.435	0.678	0.804
NPC Fisher	0.543	0.753	0.845
Permutation on primary events	0.357	0.594	0.716
Permutation on censoring events	0.295	0.500	0.604
Logrank/Mantel	0.373	0.605	0.722
Wilcoxon/Breslow	0.463	0.729	0.829
Tarone–Ware	0.438	0.703	0.804
Peto–Peto	0.456	0.723	0.823
Modified Peto–Peto	0.458	0.724	0.823
Harrington–Fleming	0.459	0.722	0.826
Modified Mantel	0.347	0.583	0.706
Prentice	0.319	0.518	0.628
Modified Prentice	0.300	0.492	0.607

Table 16. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for $(C, C_{8,2})$.

Nominal α	1%	5%	10%
NPC Tippet	0.726	0.916	0.958
NPC Fisher	0.807	0.940	0.976
Permutation on primary events	0.745	0.911	0.953
Permutation on censoring events	0.223	0.435	0.573
Logrank/Mantel	0.762	0.915	0.952
Wilcoxon/Breslow	0.412	0.657	0.767
Tarone–Ware	0.580	0.797	0.876
Peto–Peto	0.437	0.669	0.775
Modified Peto–Peto	0.429	0.664	0.770
Harrington–Fleming	0.450	0.698	0.792
Modified Mantel	0.739	0.898	0.937
Prentice	0.752	0.914	0.950
Modified Prentice	0.727	0.894	0.936

Table 17. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for $(D, C_{8,2})$.

Nominal α	1%	5%	10%
NPC Tippet	0.284	0.542	0.659
NPC Fisher	0.257	0.506	0.631
Permutation on primary events	0.011	0.045	0.080
Permutation on censoring events	0.365	0.628	0.754
Logrank/Mantel	0.012	0.047	0.084
Wilcoxon/Breslow	0.010	0.045	0.091
Tarone–Ware	0.012	0.046	0.087
Peto–Peto	0.010	0.048	0.089
Modified Peto–Peto	0.009	0.046	0.089
Harrington–Fleming	0.011	0.050	0.088
Modified Mantel	0.011	0.047	0.089
Prentice	0.015	0.049	0.091
Modified Prentice	0.016	0.047	0.092

Table 18. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for $(E, C_{8,2})$.

Nominal α	1%	5%	10%
NPC Tippet	0.259	0.508	0.666
NPC Fisher	0.281	0.567	0.691
Permutation on primary events	0.038	0.145	0.266
Permutation on censoring events	0.314	0.589	0.716
Logrank/Mantel	0.041	0.154	0.264
Wilcoxon/Breslow	0.033	0.085	0.151
Tarone–Ware	0.016	0.066	0.126
Peto–Peto	0.034	0.083	0.149
Modified Peto–Peto	0.035	0.085	0.157
Harrington–Fleming	0.021	0.072	0.119
Modified Mantel	0.019	0.096	0.167
Prentice	0.039	0.143	0.251
Modified Prentice	0.018	0.091	0.162

Appendix 1.2. Scenarios $(s_1, C_{10,10})$, non-informative censoring.

Table 19. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for $(A, C_{10,10})$.

Nominal α	1%	5%	10%
NPC Tippet	0.412	0.676	0.758
NPC Fisher	0.363	0.636	0.752
Permutation on primary events	0.529	0.748	0.824
Permutation on censoring events	0.008	0.049	0.101
Logrank/Mantel	0.545	0.760	0.834
Wilcoxon/Breslow	0.414	0.667	0.776
Tarone–Ware	0.485	0.730	0.811
Peto–Peto	0.456	0.696	0.798
Modified Peto–Peto	0.451	0.697	0.796
Harrington–Fleming	0.451	0.700	0.800
Modified Mantel	0.539	0.749	0.830
Prentice	0.538	0.756	0.833
Modified Prentice	0.531	0.742	0.827

Table 20. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for $(B, C_{10,10})$.

Nominal α	1%	5%	10%
NPC Tippet	0.269	0.510	0.636
NPC Fisher	0.255	0.495	0.630
Permutation on primary events	0.343	0.619	0.730
Permutation on censoring events	0.011	0.053	0.113
Logrank/Mantel	0.372	0.626	0.735
Wilcoxon/Breslow	0.458	0.712	0.819
Tarone–Ware	0.447	0.693	0.793
Peto–Peto	0.457	0.704	0.805
Modified Peto–Peto	0.454	0.705	0.805
Harrington–Fleming	0.457	0.703	0.814
Modified Mantel	0.354	0.605	0.723
Prentice	0.345	0.601	0.703
Modified Prentice	0.327	0.584	0.691

Table 21. Achieved $\alpha^{o}_{Q,k}, k = 1, 5, 10$ for $(C, C_{10,10})$.

Nominal α	1%	5%	10%
NPC Tippet	0.588	0.809	0.888
NPC Fisher	0.499	0.774	0.861
Permutation on primary events	0.676	0.881	0.933
Permutation on censoring events	0.009	0.040	0.098
Logrank/Mantel	0.698	0.882	0.936
Wilcoxon/Breslow	0.342	0.577	0.693
Tarone–Ware	0.496	0.731	0.824
Peto–Peto	0.392	0.613	0.727
Modified Peto–Peto	0.380	0.605	0.722
Harrington–Fleming	0.414	0.660	0.753
Modified Mantel	0.651	0.855	0.910
Prentice	0.696	0.877	0.939
Modified Prentice	0.650	0.851	0.909

Table 22. Achieved $\alpha^{o}_{Q,k}, k = 1, 5, 10$ for $(D, C_{10,10})$.

Nominal α	1%	5%	10%
NPC Tippet	0.013	0.053	0.096
NPC Fisher	0.008	0.046	0.095
Permutation on primary events	0.011	0.038	0.089
Permutation on censoring events	0.012	0.058	0.116
Logrank/Mantel	0.011	0.042	0.095
Wilcoxon/Breslow	0.009	0.044	0.100
Tarone–Ware	0.010	0.038	0.095
Peto–Peto	0.012	0.036	0.097
Modified Peto–Peto	0.012	0.035	0.097
Harrington–Fleming	0.010	0.037	0.103
Modified Mantel	0.009	0.046	0.095
Prentice	0.010	0.049	0.093
Modified Prentice	0.009	0.046	0.094

Table 23. Achieved $\alpha^{o}_{Q,k}, k = 1, 5, 10$ for $(E, C_{10,10})$.

Nominal α	1%	5%	10%
NPC Tippet	0.013	0.076	0.140
NPC Fisher	0.016	0.070	0.136
Permutation on primary events	0.026	0.079	0.150
Permutation on censoring events	0.005	0.060	0.106
Logrank/Mantel	0.025	0.086	0.151
Wilcoxon/Breslow	0.044	0.139	0.220
Tarone–Ware	0.018	0.067	0.122
Peto–Peto	0.035	0.122	0.186
Modified Peto–Peto	0.037	0.127	0.193
Harrington–Fleming	0.025	0.086	0.145
Modified Mantel	0.016	0.062	0.111
Prentice	0.026	0.086	0.146
Modified Prentice	0.016	0.062	0.110

Appendix 1.3. Scenarios $(s_1, C_{0,0})$, no censoring.

Table 24. Achieved $\alpha^{o}_{Q,k}, k = 1, 5, 10$ for $(A, C_{0,0})$.

Nominal α	1%	5%	10%
Permutation on primary events	0.607	0.843	0.900
Logrank/Mantel	0.638	0.852	0.903
Wilcoxon/Breslow	0.498	0.748	0.847
Tarone–Ware	0.572	0.811	0.888
Peto–Peto	0.511	0.749	0.844
Modified Peto–Peto	0.507	0.745	0.840
Harrington–Fleming	0.498	0.748	0.847
Modified Mantel	0.637	0.845	0.908
Prentice	0.572	0.782	0.851
Modified Prentice	0.560	0.762	0.840

Table 25. Achieved $\alpha^{o}_{Q,k}, k = 1, 5, 10$ for $(B, C_{0,0})$.

Nominal α	1%	5%	10%
Permutation on primary events	0.346	0.584	0.704
Logrank/Mantel	0.365	0.595	0.716
Wilcoxon/Breslow	0.495	0.723	0.810
Tarone–Ware	0.457	0.688	0.798
Peto–Peto	0.489	0.713	0.803
Modified Peto–Peto	0.489	0.715	0.805
Harrington–Fleming	0.495	0.723	0.810
Modified Mantel	0.348	0.576	0.700
Prentice	0.256	0.450	0.582
Modified Prentice	0.233	0.430	0.550

Table 26. Achieved $\alpha^{o}_{Q,k}, k = 1, 5, 10$ for $(C, C_{0,0})$.

Nominal α	1%	5%	10%
Permutation on primary events	0.859	0.963	0.989
Logrank/Mantel	0.873	0.967	0.990
Wilcoxon/Breslow	0.536	0.767	0.841
Tarone–Ware	0.712	0.884	0.937
Peto–Peto	0.497	0.728	0.816
Modified Peto–Peto	0.488	0.722	0.810
Harrington–Fleming	0.536	0.767	0.841
Modified Mantel	0.840	0.959	0.977
Prentice	0.872	0.967	0.988
Modified Prentice	0.838	0.957	0.976

Table 27. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for $(D, C_{0,0})$.

Nominal α	1%	5%	10%
Permutation on primary events	0.004	0.046	0.077
Logrank/Mantel	0.006	0.042	0.079
Wilcoxon/Breslow	0.011	0.044	0.089
Tarone–Ware	0.010	0.046	0.083
Peto–Peto	0.012	0.043	0.089
Modified Peto–Peto	0.012	0.044	0.088
Harrington–Fleming	0.011	0.044	0.089
Modified Mantel	0.008	0.044	0.083
Prentice	0.007	0.046	0.103
Modified Prentice	0.011	0.046	0.098

Table 28. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for $(E, C_{0,0})$.

Nominal α	1%	5%	10%
Permutation on primary events	0.125	0.335	0.449
Logrank/Mantel	0.139	0.342	0.458
Wilcoxon/Breslow	0.016	0.069	0.135
Tarone–Ware	0.027	0.110	0.188
Peto–Peto	0.014	0.092	0.158
Modified Peto–Peto	0.014	0.098	0.157
Harrington–Fleming	0.016	0.069	0.135
Modified Mantel	0.055	0.183	0.296
Prentice	0.139	0.342	0.458
Modified Prentice	0.055	0.183	0.296

Appendix 1.4. Scenarios (s_1, C_{dog}), censoring similar to the 'Cancer in dogs' case

Table 29. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for (A, C_{dog}).

Nominal α	1%	5%	10%
NPC Tippet	0.563	0.782	0.861
NPC Fisher	0.643	0.823	0.896
Permutation on primary events	0.200	0.406	0.539
Permutation on censoring events	0.535	0.732	0.809
Logrank/Mantel	0.195	0.420	0.547
Wilcoxon/Breslow	0.195	0.396	0.510
Tarone–Ware	0.198	0.410	0.531
Peto–Peto	0.200	0.408	0.525
Modified Peto–Peto	0.200	0.412	0.524
Harrington–Fleming	0.199	0.413	0.527
Modified Mantel	0.200	0.416	0.546
Prentice	0.160	0.372	0.494
Modified Prentice	0.154	0.363	0.491

Table 30. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for (B, C_{dog}).

Nominal α	1%	5%	10%
NPC Tippet	0.629	0.834	0.902
NPC Fisher	0.708	0.863	0.919
Permutation on primary events	0.285	0.502	0.601
Permutation on censoring events	0.571	0.748	0.810
Logrank/Mantel	0.286	0.503	0.604
Wilcoxon/Breslow	0.297	0.520	0.637
Tarone–Ware	0.294	0.519	0.635
Peto–Peto	0.295	0.519	0.640
Modified Peto–Peto	0.295	0.519	0.639
Harrington–Fleming	0.301	0.516	0.638
Modified Mantel	0.284	0.505	0.603
Prentice	0.169	0.347	0.457
Modified Prentice	0.169	0.331	0.428

Table 31. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for (C, C_{dog}).

Nominal α	1%	5%	10%
NPC Tippet	0.499	0.709	0.802
NPC Fisher	0.545	0.756	0.833
Permutation on primary events	0.118	0.292	0.411
Permutation on censoring events	0.507	0.698	0.794
Logrank/Mantel	0.118	0.293	0.412
Wilcoxon/Breslow	0.064	0.208	0.300
Tarone–Ware	0.082	0.231	0.336
Peto–Peto	0.088	0.227	0.338
Modified Peto–Peto	0.083	0.222	0.331
Harrington–Fleming	0.091	0.240	0.349
Modified Mantel	0.104	0.275	0.390
Prentice	0.149	0.318	0.440
Modified Prentice	0.140	0.303	0.435

Table 32. Achieved $\alpha^o_{Q,k}, k = 1, 5, 10$ for (D, C_{dog}).

Nominal α	1%	5%	10%
NPC Tippet	0.563	0.759	0.840
NPC Fisher	0.506	0.728	0.826
Permutation on primary events	0.011	0.055	0.100
Permutation on censoring events	0.627	0.820	0.896
Logrank/Mantel	0.009	0.057	0.098
Wilcoxon/Breslow	0.009	0.056	0.119
Tarone–Ware	0.011	0.052	0.115
Peto–Peto	0.010	0.052	0.111
Modified Peto–Peto	0.010	0.053	0.114
Harrington–Fleming	0.010	0.054	0.113
Modified Mantel	0.011	0.054	0.107
Prentice	0.007	0.044	0.103
Modified Prentice	0.005	0.043	0.104

Table 33. Achieved $\alpha^{o}_{Q,k}, k = 1, 5, 10$ for (E, C_{dog}).

Nominal α	1%	5%	10%
NPC Tippet	0.621	0.810	0.884
NPC Fisher	0.646	0.842	0.900
Permutation on primary events	0.064	0.208	0.317
Permutation on censoring events	0.669	0.855	0.913
Logrank/Mantel	0.064	0.210	0.321
Wilcoxon/Breslow	0.092	0.284	0.428
Tarone–Ware	0.082	0.263	0.394
Peto–Peto	0.076	0.258	0.387
Modified Peto–Peto	0.079	0.261	0.392
Harrington–Fleming	0.075	0.248	0.377
Modified Mantel	0.069	0.228	0.348
Prentice	0.043	0.165	0.284
Modified Prentice	0.045	0.179	0.303