Generalized pairwise comparisons for censored data: An overview

Peer-reviewed author version

DELTUVAITE-THOMAS, Vaiva; VERBEECK, Johan; BURZYKOWSKI, Tomasz; BUYSE, Marc; Tournigand, Christophe; MOLENBERGHS, Geert & THAS, Olivier
(2023) Generalized pairwise comparisons for censored data: An overview. In: Biometrical journal, 65 (2) (Art N° 2100354).

DOI: 10.1002/bimj.202100354
Handle: http://hdl.handle.net/1942/38686

Generalized Pairwise Comparisons for Censored Data. An Overview

Vaiva Deltuvaite-Thomas[1], Johan Verbeeck[2], Tomasz Burzykowski[1,2], Marc Buyse[1,2], Geert Molenberghs[2,3], and Olivier Thas[2]

[1]International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium
[2]Data Science Institute (DSI), Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Hasselt, Belgium
[3]Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), KU Leuven, Leuven, Belgium

Abstract

The method of generalized pairwise comparisons (GPC) is an extension of the well-known non-parametric Wilcoxon -Mann -Whitney test for comparing two groups of observations. Multiple generalizations of Wilcoxon-Mann-Whitney test and other GPC methods have been proposed over the years to handle censored data. These methods apply different approaches to handling loss of information due to censoring: ignoring non-informative pairwise comparisons due to censoring (Gehan, Harrell and Buyse); imputation using estimates of the survival distribution (Efron, Péron and Latta); or inverse probability of censoring weighting (IPCW, Datta and Dong). Based on the GPC statistic, a measure of treatment effect, the "net benefit", can be defined. It quantifies the difference between the probabilities that a randomly selected individual from one group is doing better than an individual from the other group. This paper aims at evaluating GPC methods for censored data, both in the context of hypothesis testing and estimation, and providing recommendations related to their choice in various situations. The methods that ignore uninformative pairs have comparable power to more complex and computationally demanding methods in situations of low censoring, and are slightly superior for high proportions ($> 40\%$) of censoring. If one is interested in estimation of the net benefit, Harrell's c index is an unbiased estimator if the proportional hazards assumption holds. Otherwise, the imputation (Efron or Peron) or IPCW (Datta, Dong) methods provide unbiased estimators in case of proportions of drop-out censoring up to 60%.

KEYWORDS: Bias; Censored outcome; Generalized pairwise comparisons; Net benefit; Statistical power;

1 Introduction

Ever since Wilcoxon laid foundations to the non-parametric testing procedures (Wilcoxon, 1945), later successfully extended by Mann and Whitney (Mann and Whitney, 1947), the idea of using pairwise comparisons has been gaining in popularity across various fields. Several statistics have built on the Mann-Whitney approach, giving rise to an entire family of Generalized Pairwise Comparisons (GPC) methods. This term, formally introduced by Buyse (Buyse, 2010), encompasses any method or test based on comparing one observation from a treatment group with an observation from the control group. Most of the GPC methods generalize the Mann-Whitney test to include multiple outcomes of interest of any type (time-to-event, continuous or categorical), possibly considered in turn based on some pre-defined order of priorities. The methods built on the prioritized comparisons include the Finkelstein-Schoenfeld method (Finkelstein and Schoenfeld, 1999), the net benefit (Buyse, 2010), the win ratio (Pocock *et al.*, 2012), the win odds (Dong *et al.*, 2020a, Brunner *et al.*, 2021), and the probabilistic index (Acion *et al.*, 2006). The non-prioritized methods include the O'Brien test (O'Brien, 1984). While the original Wilcoxon test was not using the pairwise comparisons, it can be adapted to pairwise comparisons as well (Ramchandani *et al.*, 2016, Verbeeck *et al.*, 2019).

Consider two groups of patients with X_i $(i = 1, \ldots, n_E)$ denoting the outcome for the i-th patient in the experimental group and Y_j $(j = 1, \ldots, n_C)$ denoting the same outcome for the j-th patient in the control group. We start by assigning a score to each pairwise comparison between the two groups, in a manner proposed by Gehan in his extension of the Mann-Whitney approach to censored data (Gehan, 1965):

$$U_{ij} = \begin{cases} 1, & \text{if } X_i > Y_j \\ -1, & \text{if } Y_j > X_i \\ 0, & \text{if } X_i = Y_j. \end{cases} \tag{1}$$

Based on these scores, we can construct a test statistic to test the null hypothesis of equality of outcome distributions between the treatment and control groups:

$$\widehat{\Delta} = \frac{1}{n_E n_C} \sum_{i=1}^{n_E} \sum_{j=1}^{n_C} U_{ij}. \tag{2}$$

Buyse ([Buyse, 2010]) has proposed using permutations to obtain the empirical distribution of $\widehat{\Delta}$, with two-sided p-values calculated by taking the proportion of all permutation samples with the absolute value of the test statistic greater than or equal to $\left|\widehat{\Delta}\right|$. Depending on the number of permutations, this method can become computationally demanding and time consuming.

The expected value of $\widehat{\Delta}$ is equal to:

$$\mathrm{E}(\widehat{\Delta}) = P(X_i > Y_j) - P(Y_j > X_i). \tag{3}$$

Some authors ([Buyse, 2010, Peron *et al.*, 2016]) view this quantity as a treatment effect measure called the net benefit (3). It can be seen as a difference between the probabilities that a randomly selected individual from one group is doing better than an individual from the other group.

In the remainder of the paper we will focus on the GPC test statistic defined in (2) and the related treatment effect measure, the net benefit. However, the presented results may extend to other measures (and corresponding statistics) related to the net benefit such as the probabilistic index, the win ratio, and the win odds.

Ideally, the scoring mechanism in (1) should accommodate any type of outcome. The scoring of continuous, categorical, binary, or time-to-event outcomes in the absence of censoring is straightforward. Moreover, it has been shown that the probabilistic index and net benefit are always unbiased and efficient in detecting a treatment effect in realistic clinical scenarios in a univariate setting ([Verbeeck *et al.*, 2021]). However, in many practical situations, observed values of time-to-event variables are right-censored, most often due to the limited duration of follow-up. In such scenarios there is no simple and unambiguous way of scoring pairs with one or both observations in a pair being censored, and these pairs are often considered to be uninformative.

Several extensions of the Wilcoxon-Mann-Whitney test towards censored data have been proposed. They differ in the way of handling loss of information due to the uninformative pairs. As the GPC test statistic can be viewed as a linear transformation of the Wilcoxon-Mann-Whitney test statistic, W, with

$$\Delta = 1 - 2\frac{W}{n_E n_C},$$

the same extensions can be easily applied to the GPC test as well.

The extensions that deal with censored data can be divided into three main groups. The first group uses a naïve approach where each uninformative pair receives a score of zero, as proposed by Gehan ([Gehan, 1965]), or is ignored, as in Harrell's c ([Harrell *et al.*, 1982]) and by Buyse ([Buyse, 2008]). The second group contains imputation-based methods, such as the ones proposed by Efron ([Efron, 1967]), Péron ([Peron *et al.*, 2016]), and Latta ([Latta, 1977]). These methods use the non-parametric

Kaplan-Meier estimator of the survival function to obtain an appropriate estimate of the probability one element in a pair doing better than the other. Finally, the third group includes the Inverse Probability of Censoring Weighting (IPCW) methods, such as those suggested by Datta (Datta *et al.*, 2010) and Dong (Dong *et al.*, 2020b). These methods use the Kaplan-Meier estimates of the censoring distribution to obtain the probability of censoring for each observation. The inverse of this probability is used to re-weight the observed scores.

This paper aims at comparing, in a univariate setting, the extensions of the GPC test to censored data discussed above, both in the context of hypothesis testing and estimation, and at providing recommendations. We begin with a brief review of the available extensions in Section 2, followed by a simulation study. The setting of the simulation study is described in Section 3 and the results are summarized and presented in Section 4. We follow up with a practical example of an analysis of real-life clinical trial data in Section 5. Section 6 completes the paper with a discussion and conclusions.

2 Extensions of GPC and the net benefit to censored data

2.1 Notation

Let X_i $(i = 1, \ldots, n_E)$ and Y_j $(j = 1, \ldots, n_C)$ be the event-times from treatment and control group, respectively. We assume that the times are iid within each group. We assume that observations of some of these times might be right-censored such that we observe only the time of censoring, X_i^c or Y_j^c. Therefore, for each of the observations, we observe only $X_i' = \min(X_i, X_i^c)$ or $Y_j' = \min(Y_j, Y_j^c)$ and the corresponding censoring indicators $\delta_i = \mathbf{1}(X_i < X_i^c)$ and $\epsilon_j = \mathbf{1}(Y_j < Y_j^c)$, where $\mathbf{1}(A)$ is the indicator of event A.

Furthermore, let $F(t) = P(X_i > t)$, $G(t) = P(Y_j > t)$, $H(t) = P(X_i^c > t)$, and $I(t) = P(Y_j^c > t)$ denote survival functions of X_i, Y_j, X_i^c, and Y_j^c, respectively.

Similar notation will be assumed for the joint sample of size $n_E + n_C = N$, i.e., denote the event-times in the joint sample by Z_k $(k = 1, \ldots, N)$, with the observed time $Z_k' = \min(Z_k, Z_k^c)$, where Z_k^c is the time of censoring. The corresponding censoring indicator is denoted by $\theta_k = \mathbf{1}(Z_k < Z_k^c)$. We use $J(t) = P(Z_k > t)$ to refer to the survival function of Z_k.

2.2 Naïve approaches

The first attempt at extending the scoring in formula (1) to include censored outcomes was made by Gehan ([Gehan, 1965]) as follows:

$$U_{ij}^G = \begin{cases} 1, & \text{if } X_i' > Y_j', \text{ and } \epsilon_j = 1 \\ -1, & \text{if } Y_j' > X_i', \text{ and } \delta_i = 1 \\ 0, & \text{if } X_i' = Y_j', \text{ and } \epsilon_j = \delta_i = 1 \\ 0, & \text{otherwise. The pair is uninformative.} \end{cases} \tag{4}$$

In other words, Gehan suggested assigning a score of zero to every comparison that did not lead to a clear cut decision on which of the two observations had a better outcome. The test statistic is then obtained by plugging the score (4) into formula (2):

$$\widehat{\Delta}_G = \frac{1}{n_E n_C} \sum_{i=1}^{n_E} \sum_{j=1}^{n_C} U_{ij}^G. \tag{5}$$

A fast, exact, and closed-form expression of the permutation variance of the GPC statistic under the null hypothesis $F(t) = G(t)$ is available ([Mantel, 1967, Finkelstein and Schoenfeld, 1999, Verbeeck *et al.*, 2020]). It can be used for inference based on the Gehan statistic. The formula requires scoring all possible pairs of observations, both between and within the two groups of observations. In particular, consider a joint sample, $\boldsymbol{Z} = Z_1, \ldots, Z_N$. Following (4), we define the score

$$V_{kl}^G = \begin{cases} 1, & \text{if } Z_k' > Z_l', \text{ and } \theta_l = 1 \\ -1, & \text{if } Z_l' > Z_k', \text{ and } \theta_k = 1 \\ 0, & \text{if } Z_k' = Z_l', \text{ and } \theta_k = \theta_l = 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

for $k, l = 1, \ldots, N$.

The exact permutation variance of $\widehat{\Delta}_G$ is given by

$$\text{Var}(\widehat{\Delta}_G) = \frac{1}{n_E n_C N(N-1)} \sum_{k=1}^{N} \left(\sum_{l=1}^{N} V_{kl}^G \right)^2. \tag{7}$$

Since $\widehat{\Delta}_G$ belongs to a broader class of generalized U-statistics ([Lee, 1990]), the null hypothesis,

$$H_0 : F(t) = G(t)$$

can be tested by using the asymptotic normality of $\widehat{\Delta}_G$ that follows from the theory of U-statistics. Consequently, the p-value of a two-sided test can be obtained as follows:

$$p = 2 \cdot \Phi \left(\frac{- \mid \widehat{\Delta}_G \mid}{\sqrt{\mathrm{Var}(\widehat{\Delta}_G)}} \right), \tag{8}$$

where $\Phi()$ is the standard-normal cdf and $\mathrm{Var}(\widehat{\Delta})$ is defined in (7).

It is straightforward to show that the expectation of the generalized Gehan statistic is equal to

$$\mathrm{E}(\widehat{\Delta}_G) = P(Y_j^c > X_i' > Y_j) - P(X_i^c > Y_j' > X_i). \tag{9}$$

Under the null hypothesis of equality of distributions, $H_0 : F(t) = G(t)$, we get $\mathrm{E}(\widehat{\Delta}_G) = 0$ (Gehan, 1965). However, if $F(t) \neq G(t)$, the expected value of $\mathrm{E}(\widehat{\Delta}_G)$ depends on the distributions of censoring times in the treatment groups. Thus, under the alternative hypothesis and in the presence of right-censoring, the Gehan statistic is a biased estimate of the net benefit Δ, given in (3). Note that the dependence of the expected value of the GPC statistic on the censoring distribution under Gehan scoring is well-known and it has been studied for the test based on the win-ratio statistic as well (Rauch *et al.*, 2014, Oakes, 2016, Dong *et al.*, 2020c).

The Harrell's c-index approach (Harrell *et al.*, 1982, Koziol, 2009) adopts the same scoring as defined in (4). However, the uninformative pairs are omitted when calculating the test statistic (2), i.e.,

$$\widehat{\Delta}_H = \frac{1}{n_E' n_C'} \sum_{i=1}^{n_E} \sum_{j=1}^{n_C} U_{ij}^G, \tag{10}$$

where n_C' and n_E' denote the number of uncensored observations in the control and experimental group, respectively. Buyse (Buyse, 2008) proposed the same approach when considering estimation of the net benefit, and showed its relationship with the hazard ratio under proportional hazards (PH).

The expression for the exact permutation variance, given in (7), can be adapted to the case of Harrell's c index as follows:

$$\mathrm{Var}(\widehat{\Delta}_H) = \frac{n_E n_C}{N(N-1) \sum_{k=1}^{N} \sum_{l=1}^{N} \mathbf{1}(V_{kl}^G \text{ is informative})} \sum_{k=1}^{N} \left(\sum_{l=1}^{N} V_{kl}^G \right)^2. \tag{11}$$

Construction of the formal test of the null hypothesis relies on the asymptotic normality of the U-statistic $\widehat{\Delta}_H$, as in the case of $\widehat{\Delta}_G$ (8).

The expectation of the Harrell's c index is proportional to $\mathrm{E}(\widehat{\Delta}_G)$, with the proportionality factor equal to the inverse of a probability of a pair being informative (Koziol, 2009):

$$\mathrm{E}(\widehat{\Delta}_H) = \frac{1}{P\{\mathbf{1}(U_{ij} \text{ is informative})\}} \times \mathrm{E}(\widehat{\Delta}_G) = \frac{P(X_i' , Y_j^c > Y_j) - P(Y_j', X_i^c > X_i)}{P(X_i' , Y_j^c > Y_j) + P(Y_j', X_i^c > X_i)}. \quad (12)$$

The dependence of the expectation of the Harrell's c index on the censoring distribution of the outcomes implies that, in general, $\mathrm{E}(\widehat{\Delta}_H) \neq \Delta$. However, Harrell has developed the index as a measure of separation of two survival curves under the PH assumption. Indeed, it can be shown that, under this assumption, $\mathrm{E}(\widehat{\Delta}_H) = \Delta$. Appendix 6 presents a proof of this equality in case of two exponentially distributed outcomes.

2.3 Imputation-based approaches

Efron (Efron, 1967) has criticized the dependence of the expected value of the Gehan statistic on the censoring distributions and developed an alternative and mathematically elegant method of approaching the two sample problem with right-censored data. In particular, he proposed to redefine U_{ij}^G as follows:

$$U_{ij}^E = P(X_i > Y_j \mid X_i', Y_j', \delta_i, \epsilon_j) - P(Y_j > X_i \mid X_i', Y_j', \delta_i, \epsilon_j) \quad (13)$$

If $X_i' > Y_j'$ and $\epsilon_j = 1$, or if $Y_j' > X_i'$ and $\delta_i = 1$, then U_{ij}^E takes a value of 1, or -1, as U_{ij}^G defined in (4). However, a score of 0 is no longer assigned to U_{ij}^E in case a clear comparison is impossible due to censoring. In fact, U_{ij}^E involves conditional probabilities $P(X_i > Y_j \mid X_i', Y_j', \delta_i, \epsilon_j)$ and $P(Y_j > X_i \mid X_i', Y_j', \delta_i, \epsilon_j)$, which can be estimated by using the Kaplan-Meier estimates of $\widehat{F}(t)$ and $\widehat{G}(t)$. The scores of U_{ij}^E are summarized in Table 1.

Table 1: Values of U_{ij}^E for the Efron approach to the right-censored data problem.

(δ_i, ϵ_j)	$X_i > Y_j$	$X_i = Y_j$	$X_i < Y_j$
(1,1)	1	0	-1
(0,1)	1	1	$2\frac{\widehat{F}(Y_j)}{\widehat{F}(X_i)} - 1$
(1,0)	$1 - 2\frac{\widehat{G}(X_i)}{\widehat{G}(Y_j)}$	-1	-1
(0,0)	$1 - \frac{\widehat{G}(X_i)}{\widehat{G}(Y_j)}$	0	$\frac{\widehat{F}(Y_j)}{\widehat{F}(X_i)} - 1$

The Efron-test statistic $\widehat{\Delta}_E$ is then obtained by plugging the score (13) into (2).

The exact permutation variance (Verbeeck *et al.*, 2020) estimation requires that the score between any two observations remains constant across all permutation samples. This condition does not

hold for the Efron scoring. Thus, the inference based on $\widehat{\Delta}_E$ relies on a re-sampling permutation distribution.

Efron ([Efron, 1967]) shows that $\widehat{\Delta}_E = \int_{-\infty}^{\infty} \widehat{G}(t)d\widehat{F}(t) - \int_{-\infty}^{\infty} \widehat{F}(t)d\widehat{G}(t)$ is the maximum likelihood estimate of Δ provided that the largest observation in each group is treated as uncensored.

Péron *et al.* ([Peron *et al.*, 2016]) have suggested an adaptation of Efron's scoring, differing only in that it allows the largest observation in each group to be right-censored. For details of this scoring approach we refer to Péron *et al.* ([Peron *et al.*, 2016]). Similarly to $\widehat{\Delta}_E$, the test using the statistic resulting from Péron's adaptation of Efron's scoring is based on a re-sampling permutation distribution.

The scoring systems of Efron and Péron *et al.* use separate Kaplan-Meier estimates of the survival function for the two groups of observations. Latta ([Latta, 1977]) suggested using the Kaplan-Meier estimate of the joint survival function based on the joint sample Z_k, because under the null hypothesis of no treatment effect observations are assumed to come from the same distribution.

Following Latta, the scores in Table 1 are adjusted by substituting each $\widehat{F}(X_i)$ and $\widehat{G}(Y_j)$ by $\tilde{J}(X_i)$ or $\tilde{J}(Y_j)$, with $\tilde{J}(t)$ defined as:

$$\tilde{J}(Z_k) = \begin{cases} \frac{1}{2}\{\widehat{J}(Z_k-) + \widehat{J}(Z_k+)\} & \text{if } \theta_k = 1 \\ \widehat{J}(Z_k+) & \text{if } \theta_k = 0, \end{cases} \tag{14}$$

where $\widehat{J}(t-)$ and $\widehat{J}(t+)$ correspond to the left and right limit, respectively, of the estimated survival function at time t. If the null hypothesis $F(t) = G(t)$ is true, then $\widehat{J}(t)$ is a more efficient estimator of the (common) survival function than either $\widehat{F}(X_i)$ or $\widehat{G}(Y_j)$.

The Latta-test statistic $\widehat{\Delta}_L$ is then obtained by plugging the Efron score modified by (14) into (2). As the Latta scoring uses a single survival function for both groups, the score for each pair remains stable across all permutation samples. As a result, we can use the exact permutation variance (7) to estimate $\mathrm{Var}(\widehat{\Delta}_L)$. Construction of the formal test of the null hypothesis then relies on the asymptotic normality of the U-statistic $\widehat{\Delta}_L$, as in the case of $\widehat{\Delta}_G$ (8).

It is worth noting that the Latta-test statistic is related to the Peto-Peto-Prentice modification of the log-rank test ([Latta, 1977]). Let us consider an alternative formulation of the test statistic ([Mantel, 1967]), based on the Latta scores U_{kl}^L obtained through comparisons in a joint sample Z_k. Define:

$$U_k^L = \sum_{l=1}^{N} U_{kl}^L.$$

Then, the Latta-test statistic can be obtained as:

$$\hat{\Delta}_L = \frac{1}{n_E n_C} \sum_{k=1}^{N} U_k^L \mathbf{1}(k \text{ is in treatment group}).$$

Peto and Peto, 1972 noted the linear relationship between the ranks of observations and $\tilde{J}(Z_k)$ and proposed to use statistic

$$W_P = \sum_{k=1}^{N} w_k \mathbf{1}(k \text{ is in treatment group}),$$

where

$$w_k = \begin{cases} 1 - 2\tilde{J}(Z_k) & \text{if } \theta_k = 1 \\ 1 - \tilde{J}(Z_k) & \text{if } \theta_k = 0. \end{cases} \tag{15}$$

Based on Theorem 1 in,Latta, 1977 we can show that $\hat{\Delta}_L = \frac{N}{n_E n_C} W_P$.

2.4 Methods based on Inverse Probability of Censoring Weighting

Datta (Datta *et al.*, 2010) proposed weighting informative pairs in (4) by using weights derived from the Kaplan-Meier estimates of the survival functions $H(t)$ and $I(t)$ of censored observations X_i' and Y_j', respectively. As a result, the following statistic is obtained:

$$\widehat{\Delta}_{IPCW1} = \frac{1}{n_E n_C} \sum_{i=1}^{n_E} \sum_{j=1}^{n_C} \frac{U_{ij}^G \delta_i \epsilon_j}{\widehat{H}(X_i)\widehat{I}(Y_j)}. \tag{16}$$

Datta further proves that, under independent censoring, $\mathrm{E}(\widehat{\Delta}_{IPCW1}) = \Delta$. Moreover, Stute and Wang (Stute and Wang, 1994) have shown that the Efron statistic is equal to $\widehat{\Delta}_{IPCW1}$. Note that, similarly to Efron's approach, the Datta statistic requires that the largest observation in each group is treated as uncensored.

It can be noted that the approach proposed by Datta discards much information, because only the pairs with both $\delta_i = 1$ and $\epsilon_j = 1$ are taken into consideration in $\widehat{\Delta}_{IPCW1}$. This could potentially result in a loss of efficiency if there are many censored observations in the data.

Dong (Dong *et al.*, 2020b) proposed an IPCW method, which defines two scores, K_{ij} and L_{ij}, instead of a single U_{ij}^G:

$$K_{ij} = \begin{cases} 1, & \text{if } X_i' > Y_j', \text{ and } \epsilon_j = 1 \\ 0, & \text{otherwise,} \end{cases} \tag{17}$$

and

$$L_{ij} = \begin{cases} 1, & \text{if } Y'_j > X'_i, \text{ and } \delta_i = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

The test statistic is then obtained by weighting each individual score by the inverse of the censoring probability estimated by using the Kaplan-Meier estimator of the survival functions $H(t)$ and $I(t)$:

$$\widehat{\Delta}_{IPCW2} = \frac{1}{n_E n_C} \sum_{i=1}^{n_E} \sum_{j=1}^{n_C} \left(\frac{K_{ij}}{\widehat{H}(Y'_j)\widehat{I}(Y'_j)} - \frac{L_{ij}}{\widehat{H}(X'_i)\widehat{I}(X'_i)} \right). \tag{19}$$

Construction of the formal tests of the null hypothesis based on $\widehat{\Delta}_{IPCW1}$ and $\widehat{\Delta}_{IPCW2}$ cannot rely on the use of the exact permutation variance and is based on the re-sampling permutation distribution instead.

3 Simulation study setting

We consider a setting with a single right-censored outcome observed in two groups of patients, experimental and control. We assume a sample size of $n_E = n_C = 100$.

We investigate two possible scenarios: one corresponding to situations when the PH assumption holds, and one when the assumption does not hold. The first case is simulated by using exponential distributions, while a log-normal distribution is assumed for the second case. The parameters of the simulation models are summarized in Table 2.

Table 2: Summary of scenarios and their corresponding parameters used in simulations.

Scenario	Outcome distribution	Control group parameters	Experimental group parameters
	$H_0 : F(t) = G(t)$, $\Delta = 0$		
Proportional hazards	Exp(λ)	$\lambda_C = 0.00315$	$\lambda_E = 0.00315$
Non-proportional hazards	Log-N(μ, σ^2)	$\sigma_C = 1$, $\mu_C = 4.60$	$\sigma_E = 1$, $\mu_C = 4.60$
	$H_A : F(t) \neq G(t)$, $\Delta = 0.2$		
Proportional hazards	Exp(λ)	$\lambda_C = 0.00315$	$\lambda_E = 0.00210$
Non-proportional hazards	Log-N(μ, σ^2)	$\sigma_C = 1$, $\mu_C = 4.60$	$\sigma_E = 1$, $\mu_C = 4.96$

We assume two mechanisms of censoring, because it is known (Dong *et al.*, 2020c) that the GPC statistics behave differently for them: a "drop-out" censoring that results from a uniform distribution on the interval $(0, c_E)$ for the experimental group and $(0, c_C)$ for the control, and an "administrative" censoring at time T_c due to the end of the follow-up.

Following De Backer et al. (De Backer *et al.*, 2020), we can express the proportion of drop-out censored observations, given that the censoring is uniform, as follows:

$$p_{DO,G} = \frac{\int_0^{T_c} F(t)dt}{c_G}, \tag{20}$$

with $G = E$ or C denoting the treatment group assignment.The corresponding proportion of administrative censoring is given by

$$p_{adm,G} = \frac{(c_G - T_c)}{c_G} F(T_c). \tag{21}$$

Note that the overall proportion of drop-out censored data in the sample is equal to $p_{DO} = (p_{DO,E} + p_{DO,C})/2$. Similarly, the overall proportion of administrative censoring is $p_{adm} = (p_{adm,E} + p_{adm,C})/2$.

The values of c_E, c_C, and the T_c are obtained through a numerical search, such that, when $F(t) = G(t)$, across both groups we obtain the following three scenarios:

1. a target proportion of the overall drop-out censoring alone, that is, $p_{DO} \in \{0\%, 10\%, 30\%, 50\%, 70\%, 90\%\}$, $p_{adm} = 0\%$, and $c_E = c_C$;

2. a target proportion of the overall drop-out censoring alone $p_{DO} \in \{0\%, 10\%, 30\%, 50\%, 70\%, 90\%\}$, $p_{adm} = 0\%$ and $c_E \neq c_C$. We have set c_E and c_C such that $p_{DO,E} \in \{0\%, 5\%, 25\%, 45\%, 65\%, 85\%\}$ and $p_{DO,C} \in \{0\%, 15\%, 35\%, 55\%, 75\%, 95\%\}$;

3. a constant drop out $p_{DO} = 10\%$ ($c_E = c_C$), with additional administrative censoring $p_{adm} \in \{10\%, 30\%, 50\%, 70\%, 90\%\}$.

Note that (a) and (c) imply that $H(t) = I(t)$.

For each scenario, and each combination of p_{DO} and p_{adm}, we simulate 5,000 datasets of right-censored time-to-event data in two treatment groups. Then, each of these datasets is analyzed by using all the approaches discussed in Sections 2.2–2.4. In parallel, we use the classical log-rank test for comparison of power.

4 Results

4.1 Hypothesis testing

Figure 1 presents rejection proportions computed for 5,000 simulated datasets, i.e., empirical estimates of the type-I error probability. None of the scenarios corresponding to the null hypothesis $H_0 : F(t) = G(t)$, and in the presence of equal censoring distributions $H(t) = I(t)$ (Figure 1, panels (a), (b), (e) and (f)), indicate a significant deviation of the type-I error probability from 5%. Nevertheless, in case of a combination of drop-out and administrative censoring, a slightly conservative type-I error probability control can be observed if administrative censoring proportion is high.

In case of unequal censoring distributions between the treatment groups, $H(t) \neq I(t)$ (Figure 1, panels (c) and (d)), the type-I error probability estimates differ significantly from the nominal level of 5%, especially for proportion of censoring larger than 30%. However, it is worth noting that the use of permutations consisting of re-shuffling treatment assignments across the entire data set requires the assumption that the entire vector of observations belonging to a particular patient, (X_i', δ_i) or (Y_j', ϵ_j), is equally likely to be observed both for the experimental treatment and for the control. That is, the test based on such permutations amounts to testing a hypothesis of equality of survival AND censoring distributions simultaneously. In cases with differential censoring distributions, we no longer operate under the null hypothesis, which is reflected in failure to control the type-I error probability at the nominal level.

Figure 2 presents rejection proportions computed for 5,000 simulated datasets for the scenarios with $H_a : F(t) \neq G(t)$, i.e., empirical estimates of power. If the censoring distributions are equal, $H(t) = I(t)$ (Figure 2, panels (a), (b), (e) and (f)), the power varies slightly among all of the methods, with the Latta approach leading to the highest power for the non-PH case and the PH case with large censoring. Among the other approaches, the naïve approaches (Gehan's and Harrel's) are comparable in power to the more computationally demanding ones for combined drop-out/administrative censoring, and are slightly more powerful for large proportions of drop-out censoring ($> 50\%$).

In the presence of unequal censoring distributions, $H(t) \neq I(t)$ (Figure 2, panels (b) and (d)), there is a clear difference in power of the Efron and Datta statistics as compared to the remaining tests. Note that the two statistics require that the largest observation in each group is treated as uncensored. Such artificial introduction of an additional event to the dataset, especially in the presence of high censoring, might introduce bias (see Section 4.2), which, possibly coupled with the fact that the permutation test assumes equal censoring distributions, leads to unpredictable changes in power.

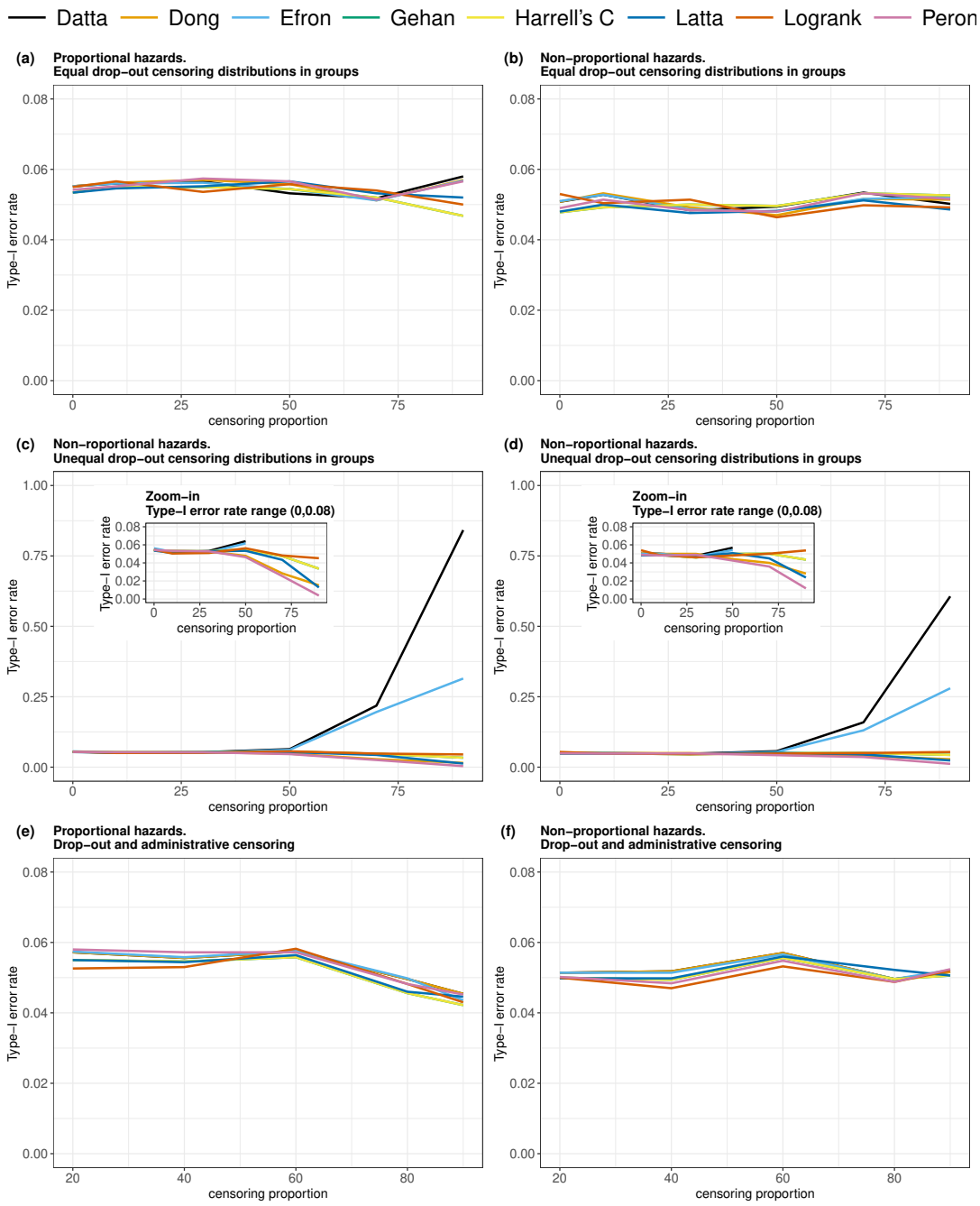

Figure 1: Empirical rejection rate over 5,000 simulated datasets. Scenarios under the null hypothesis, $H_0 : F(t) = G(t)$: for proportional hazards (panels (a), (c), and (e)) and non-proportional hazards (panels (b), (d), and (f)) scenarios.

It is of interest to compare the power of all the GPC test-statistics (with the exception of the Efron and Datta statistics in unequal censoring distributions scenarios) with the classical log-rank test. As expected, the log-rank test is more powerful in PH scenarios with low censoring, but the difference decreases for higher censoring. In non-PH scenarios, there is a clear power advantage of the GPC statistics such as, e.g., the one proposed by Latta, especially for higher

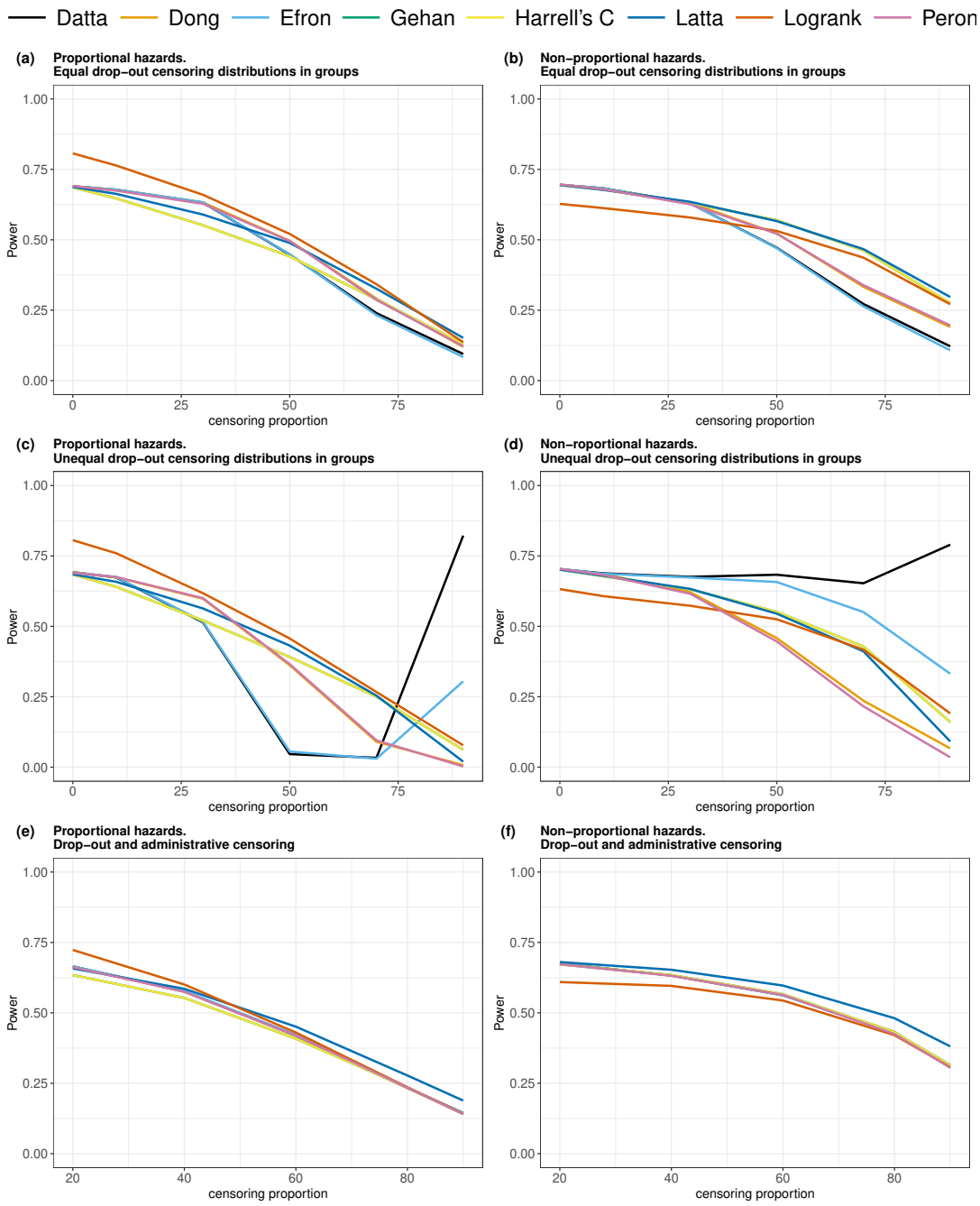

Figure 2: Empirical rejection rate over 5,000 simulated datasets. Scenarios under the alternative hypothesis, $H_a : F(t) \neq G(t)$, with true $\Delta = 0.2$: proportional hazards (panels (a), (c), and (e)) and non-proportional hazards (panels (b), (d), and (f)) scenarios.

percentages of censoring.

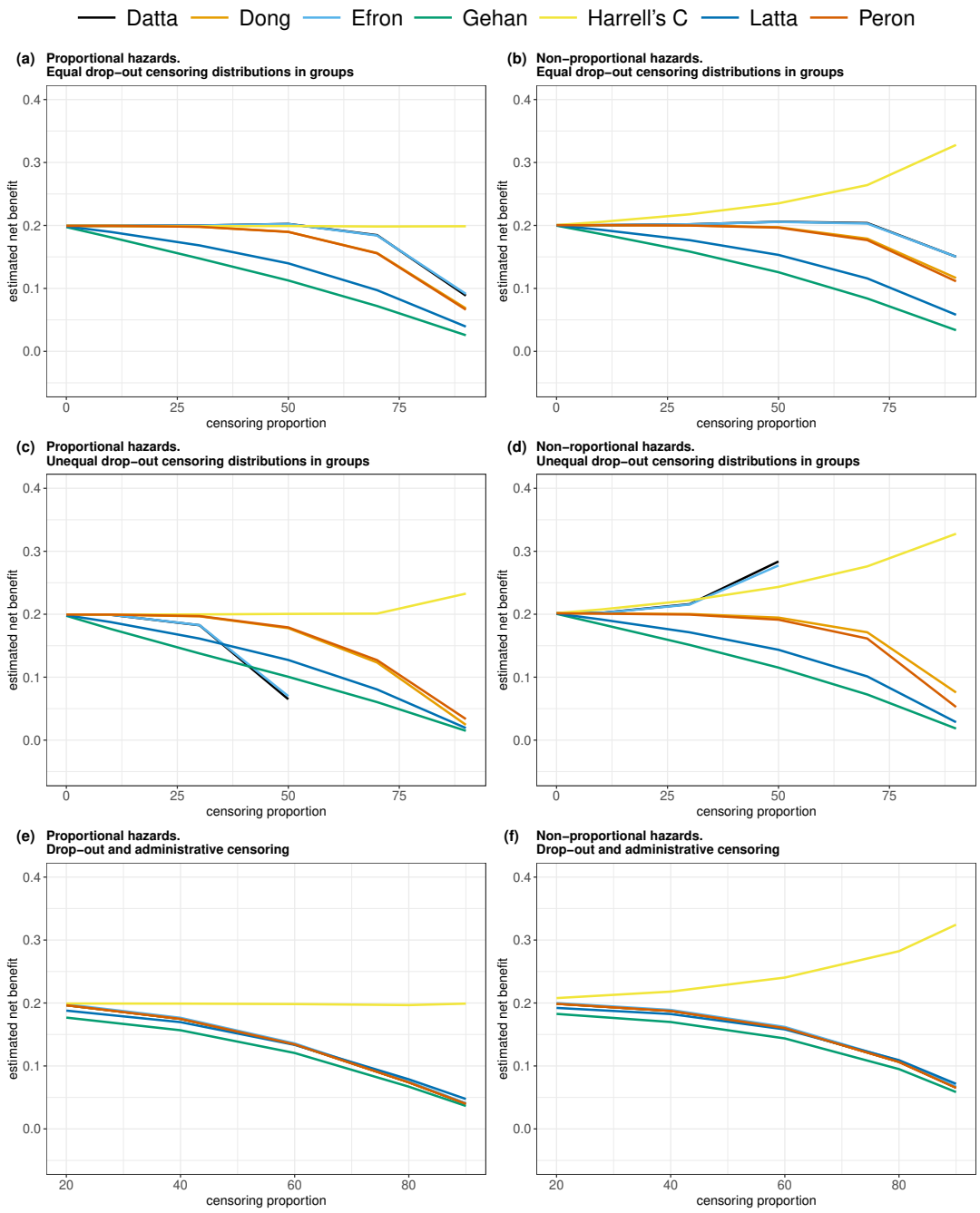

Figure 3: Estimated net benefit, mean over 5,000 simulated datasets. Scenarios under the alternative hypothesis, $H_a : F(t) \neq G(t)$, with true $\Delta = 0.2$: proportional hazards (panels (a), (c), and (e)) and non-proportional hazards (panels (b), (d), and (f)) scenarios.

4.2 Estimation of net benefit

Figure 3 presents the mean of the estimates of the net benefit computed for 5,000 datasets simulated under the alternative hypothesis, $H_a : F(t) \neq G(t)$ with $\Delta = 0.2$. The plots indicate clear differences among the different approaches.

The estimates obtained by using the two naïve approaches, i.e., Gehan's and Harrell's c index, differ considerably in terms of bias of estimation of Δ. The Gehan statistic $\widehat{\Delta}_G$ yields biased estimates in the presence of any mechanism and any proportion of censoring. In comparison, Harrell's c index, $\widehat{\Delta}_H$, is unbiased in PH scenarios (Figure 3 (a), (c) and (e)), even with a high proportion of administrative censoring. The only situation where Harrell's c shows any bias in PH scenarios is the presence of 90% censoring whenever $H(t) \neq I(t)$. It is no longer unbiased if the PH assumption does not hold (Figure 3 (b), (d) and (f)).

Whenever $H(t) = I(t)$ (Figure 3 (a), (b), (e) and (f)), the estimates obtained by using the imputation and IPCW based statistics show very similar profiles, with a notable exception of the Latta approach that yields biased estimates in all scenarios. This is not surprising, because this statistic is constructed by assuming a common survival function for both treatment groups, which does not hold under the alternative hypothesis $H_a : F(t) \neq G(t)$. The other statistics (proposed by Efron, Péron, Datta, and Dong) yield unbiased estimates for up to 50-70% drop-out censoring and start exhibiting downward bias as the censoring proportion increases further, though the Efron and Datta statistics are slightly less biased than the Péron and Dong ones in the presence of high proportions of censoring.

In scenarios where $H(t) \neq I(t)$ (Figure 3 (c) and (d)), the bias of the Péron, Latta and Dong test statistics follow similar trajectories as in scenarios with $H(t) = I(t)$. As for Efron and Datta approaches, they show a considerable bias for overall censoring of over 30%, in line with discussion in Section 4.1.

In the presence of administrative censoring, all the imputation and IPCW based statistics show strong negative bias. This observation can be justified theoretically. In particular, it can be shown that, if one observes the survival function only until some fixed time T_c, then the expected value of the statistic, irrespective of the approach, can be expressed as follows:

$$\begin{aligned}
\mathrm{E}(\widehat{\Delta}_*) &= P(X_i > Y_j) - P(Y_j > X_i) - P(T_c \leq X_i < Y_j) + P(T_c \leq X_i < Y_j) \\
&+ P(T_c \leq X_i, T_c \leq Y_j, X_i < Y_j) - P(T_c \leq Y_j, T_c \leq X_i, Y_j > X_i) = \\
&= \mathrm{E}(\widehat{\Delta}) - P(T_c \leq X_i < Y_j) + P(T_c \leq X_i < Y_j) \\
&+ P(T_c \leq X_i, T_c \leq Y_j, X_i < Y_j) - P(T_c \leq Y_j, T_c \leq X_i, Y_j > X_i). \qquad (22)
\end{aligned}$$

5 Example

We apply all the GPC approaches discussed in Sections 2.2–2.4, to right-censored data from a randomized phase III trial in advanced colorectal cancer. The trial compared two sequences of combination regimens: first-line treatment with FOLFIRI (folinic acid, fluorouracil, and

Table 3: The estimated net benefit and p-values of the test of equality of survival distributions in arms A and B, $H_0 : F(t) = G(t)$, for of the data from the phase III clinical trial in advanced colorectal cancer using different GPC statistics for censored data.

Endpoint		Naïve methods		Imputation-based methods			IPCW methods		Log-rank
		Gehan	Harrell	Efron	Péron	Latta	Datta	Dong	
1st-line	$\hat{\Delta}$	-0.049	-0.054	-0.059	-0.059	-0.053	0.059	0.059	
PFS	p-val	0.508	0.508	0.455	0.439	0.488	0.455	0.460	0.083
2nd-line	$\hat{\Delta}$	0.303	0.334	0.316	0.316	0.309	0.316	0.316	
PFS	p-val	0.001	0.001	0.001	0.003	0.001	0.001	0.001	0.066
PFS2	$\hat{\Delta}$	0.093	0.153	0.137	0.133	0.114	0.137	0.133	
	p-val	0.111	0.111	0.125	0.137	0.106	0.124	0.134	0.413

PFS: progression-free survival; PFS2: second progression-free survival.

irinotecan) followed by FOLFOX6 (folinic acid, fluorouracil, and oxaliplatin) as the second-line therapy in arm A, and the reverse sequence (FOLFOX6 in first line followed by FOLFIRI in second line) in arm B. The trial enrolled 226 patients who were randomly assigned to the two treatment arms. Six of those patients were deemed ineligible. Thus, the final analysis included the remaining 220 eligible patients, 109 in arm A and 111 in arm B.

The original analysis (Tournigand *et al.*, 2004) used the log-rank test for the comparison of the progression-free survival (PFS) of the first and the second-line treatments separately, as well as for the second progression-free survival (PFS2, defined as time from randomization to the disease progression after the second-line treatment), considered the primary endpoint in this study. The results included both an external review, and the investigator's assessment of tumor progression.

We re-analyse the investigators' data using the seven approaches discussed in Sections 2.2–2.4 considering the three outcomes of interest of the study. A summary of the results of these analyses is presented in Table 3, along with the results of the log-rank test.

None of the tests of the null hypothesis for the first-line PFS and PFS2 is statistically significant at the 5% significance level. On the other hand, all the GPC tests for the second-line PFS show a statistically significant result in favor of arm A, while the result of the log-rank test is not significant. This loss of power of the log-rank test may be expected, because there are signs that the PH assumption may not hold, as seen from the estimates of the survival functions in Figure 4.

Regarding the estimation of the net benefit, the estimates provided in Table 3 range between -0.059 and -0.049 for the first-line PFS; between 0.302 and 0.334 for the second-line PFS; and between 0.093 and 0.153 for PFS2. As the proportion of censoring was small (10% for the first-line PFS, and 7% for the second-line PFS), there is little difference between the estimates

obtained for the different statistics. It is of interest to note that the estimates yielded by the Efron, Péron, Datta, and Dong statistics are identical or very close in value.

Clearly, a small proportion of administrative censoring observed in the example data does not permit to fully appreciate the extent to which this type of censoring influences the conclusions of the GPC analysis under each of the approaches. Artificially inducing various proportions of censoring in the real-life data might be an interesting problem, beyond the focus of the present paper, that could be addressed in detail in a separate paper as a part of future research.

6 Discussion and conclusions

We performed a simulation study to evaluate the performance of several extensions of the univariate GPC test to right-censored data. The considered methods included the naïve scoring proposed by Gehan (Gehan, 1965) and Harrell (Harrell *et al.*, 1982), the imputation-based statistics of Efron (Efron, 1967), Péron *et al.* (Peron *et al.*, 2016), and Latta (Latta, 1977), and the IPCW statistics of Datta (Datta *et al.*, 2010) and Dong (Dong *et al.*, 2020b).

Before discussing general results and comparisons between the tests, it is important to single out the two tests, proposed by Efron and Datta, that are sensitive to the presence of differential censoring distributions between the treatment groups. Considering the fact that these are the two methods that require that the last observation in each group is treated as an event, it is likely that such an approach introduces information that, especially in the presence of a small number of truly observed events, distorts the estimated survival curves, and instills differences that are reflected in bias, failure to control the type-I error probability, and unpredictable

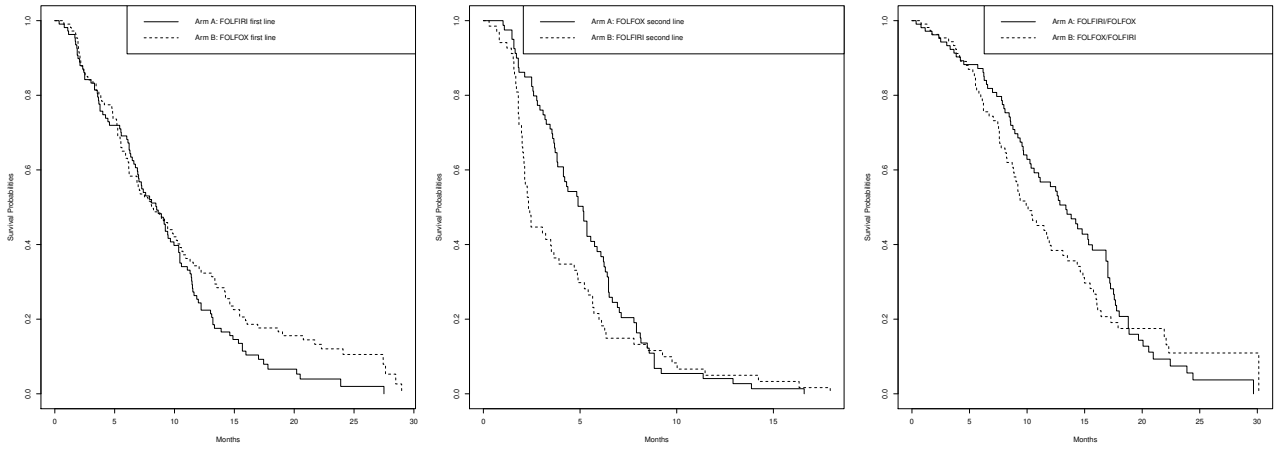

Figure 4: Kaplan-Meyer curves of progression-free survival in the first line (a) and second-line (b) therapy, and the second progression-free survival (c).

trends in power. Thus, in the discussion that follows, the references to Efron and Datta tests will exclude situations with different censoring distributions per group. In such situations, the two tests should not be used at all.

Under the null hypothesis, the nominal level of the type-I error probability was maintained provided the censoring distributions were equal between the two groups. This is reassuring, since adequate type-I error protection is essential for any statistical test. In the presence of differential censoring distributions, the type-I error probability control was not maintained, especially if censoring was larger than 30%, likely due to violation of the assumptions required for the permutation test (i.e., joint equality of survival and censoring distributions). This problem needs to be addressed separately in future research.

Under the alternative hypothesis and the PH assumption, the tests based on the GPC methods led to a marginally lower power than the log-rank test. The only exception was the presence of large proportions of administrative censoring, in which case the Latta test was the most powerful. In case of non-PH, the naïve GPC methods and the Latta test were more powerful than the log-rank test. The other tests, obtained by the methods proposed by Efron, Péron *et al.*, Datta and Dong, remained less powerful than the log-rank test in the presence of large proportions of drop-out censoring.

The estimators of the net benefit obtained by using the imputation within groups (Efron, Péron *et al.*) or IPCW (Datta, Dong) were unbiased up to 50-70% of drop-out censoring. Any bias appearing in higher proportions of censoring was most probably due to imprecise estimation of the survival functions for event-times and/or censoring caused by smaller amounts of available information. Harrell's c index remained completely unbiased in the presence of any censoring, provided that the PH assumption was fulfilled.

In the presence of administrative censoring, however, the estimators of the net benefit obtained by using the imputation within groups (Efron, Péron *et al.*) or IPCW (Datta, Dong) were negatively biased. This is due to the fact that, when the full support of the outcome distribution cannot be observed, the expected value of the estimators depends on the censoring distribution. Given that right-censoring due to insufficient follow-up is common in practice (e.g., in clinical trials), this limits the use of GPC methods for estimation purposes.

A way to circumvent the bias issue is to consider an alternative treatment-effect measure, the restricted net benefit. It is defined as the difference between the probabilities that, over the period spanning from 0 to a specific time point, a random individual from one group is doing better than an individual from the other group. This is a topic for further research.

Overall, no single method considered in this paper is unanimously superior both for hypothesis testing and estimation of the net benefit in the presence of right-censored data. If the interest

lies in estimation, then the Harrell's c index is uniformly best if the PH assumption holds. In most real-life situations, the drop-out censoring rarely exceeds 50%, thus the estimates of Δ obtained by using the Péron or Dong statistic should be equally reliable, provided there is no administrative censoring. However, in the presence of administrative censoring, all statistics will underestimate the net benefit. This caveat is important to bear in mind in the analysis of (randomized) clinical trials, in which follow-up is usually limited.

If one is interested in testing the null hypothesis of no treatment effect, i.e., $H_0 : F(t) = G(t)$, the naïve approaches proposed by Gehan and Harrell have power comparable to that of other approaches when censoring is low, and are among the most powerful when the proportion of censoring increases. Therefore, the need to use of the more complex and computationally demanding methods in the testing framework may be questioned.

This paper has evaluated the operational characteristics and point estimates of the GPC based statistics that used various extensions to account for the presence of right-censored data. We have focused on independent and non-informative censoring, frequently assumed when dealing with censored data. The comparison of various methods could be further extended to informative/dependent censoring. However, in such a case, fully non-parametric estimation of survival curves based on only observed survival information is no longer possible. Thus, one might have to resort to parametric or semi-parametric methods (Dong *et al.*, 2021) in order to correctly implement the imputation-based or IPCW approaches, sacrificing a fully non-parametric nature of the GPC procedure. The implications related to such parametric corrections of dependent censoring is a possible topic for future research.

Funding

Research partially funded by the government of Wallonia, BioWin Consortium Agreement No 7979.

Conflict of Interest

The authors have declared no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Appendix

A.1. Unbiasedeness of Harrell's c index for exponential outcomes

Let $X_i \sim \exp(\lambda_X)$, iid, $i = 1, \ldots, n_E$ and $Y_j \sim \exp(\lambda_Y)$, iid, $j = 1, \ldots, n_C$ be observations from treatment and control group respectively. Besides, we allow for some of these observations to be censored such that we observe only the time of censoring, X_i^c or Y_j^c. It is assumed that $X_i^c \sim \exp(\lambda_{X^c})$, and $Y_j^c \sim \exp(\lambda_{Y^c})$. Therefore, for each of the observations we observe only an outcome $X_i' = \min(X_i, X_i^c)$ or $Y_j' = \min(Y_j, Y_j^c)$.

The net benefit for two exponential outcomes is equal to:

$$\Delta = P(X > Y) - P(Y > X) = \frac{\lambda_Y - \lambda_X}{\lambda_Y + \lambda_X}$$

In the presence of censoring, and under the Gehan's approach, the expected net benefit value can be expressed as:

$$\mathrm{E}(\widehat{\Delta}_G) = P(XX^cY^c > Y) - P(YY^cX^c > X) = \frac{\lambda_Y - \lambda_X}{\lambda_X + \lambda_Y + \lambda_{X^c} + \lambda_{Y^c}}$$

As shown in (12), the expectation of the Harrell's c index, $\mathrm{E}(\hat{\Delta}_c)$, is equivalent to $\mathrm{E}(\hat{\Delta}_G)$, weighted by the inverse of a probability of a pair being informative.

Given that the probability for a pair being informative for two exponential outcomes with exponential censoring is:

$$P_{inf} = P(XX^cY^c > Y) + P(YY^cX^c > X) = \frac{\lambda_Y + \lambda_X}{\lambda_X + \lambda_Y + \lambda_{X^c} + \lambda_{Y^c}},$$

$\mathrm{E}(\hat{\Delta}_c)$ is therefore equal to

$$\mathrm{E}(\hat{\Delta}_c) = \frac{1}{P_{inf}} * \mathrm{E}(\hat{\Delta}_G) = \frac{\lambda_X + \lambda_Y + \lambda_{X^c} + \lambda_{Y^c}}{\lambda_Y + \lambda_X} * \frac{\lambda_Y - \lambda_X}{\lambda_X + \lambda_Y + \lambda_{X^c} + \lambda_{Y^c}} = \frac{\lambda_Y - \lambda_X}{\lambda_Y + \lambda_X} = \Delta$$

A.2. Tables of simulation study results

Table 4: Net treatment benefit estimates for the proportional hazards scenarios.

Δ	p_{cens}	**Gehan**	**Harrell's C**	**Efron**	**Peron**	**Latta**	**Datta**	**Dong**
		$\widehat{\Delta}_G$	$\widehat{\Delta}_H$	$\widehat{\Delta}_E$	$\widehat{\Delta}_P$	$\widehat{\Delta}_L$	$\widehat{\Delta}_{IPCW1}$	$\widehat{\Delta}_{IPCW2}$
Only drop-out censoring, $H(t) = I(t)$								
0	0	-0.001	-0.001	0.000	0.000	-0.001	0.000	0.000
	10	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	30	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	50	0.000	0.000	-0.001	0.000	0.000	-0.001	0.000
	70	0.000	-0.001	0.001	0.001	0.000	0.000	0.002
	90	0.000	-0.001	-0.002	-0.001	0.003	-0.006	0.000
0.2	0	0.197	0.199	0.199	0.199	0.198	0.199	0.199
	10	0.181	0.199	0.199	0.199	0.190	0.199	0.199
	30	0.148	0.199	0.200	0.198	0.168	0.200	0.198
	50	0.113	0.199	0.202	0.190	0.140	0.202	0.190
	70	0.072	0.198	0.184	0.156	0.097	0.185	0.156
	90	0.026	0.199	0.091	0.066	0.039	0.089	0.068
Only drop-out censoring, $H(t) \neq I(t)$								
0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	10	0.000	0.000	0.000	0.000	0.000	-0.001	0.000
	30	0.000	0.000	-0.008	0.000	0.000	-0.008	-0.001
	50	0.000	0.001	-0.070	0.001	0.000	-0.072	0.000
	70	0.000	0.002	-0.264	0.004	0.001	-0.307	0.000
	90	0.001	0.028	-0.217	0.006	0.002	-0.827	-0.008
0.2	0	0.197	0.199	0.199	0.199	0.198	0.199	0.199
	10	0.177	0.199	0.199	0.199	0.187	0.199	0.199
	30	0.138	0.200	0.183	0.197	0.161	0.182	0.197
	50	0.101	0.201	0.070	0.179	0.127	0.065	0.178
	70	0.060	0.201	-0.187	0.127	0.081	-0.248	0.123
	90	0.015	0.233	-0.187	0.034	0.019	-0.825	0.024
Drop-out and administrative censoring								
0	20	-0.001	-0.001	-0.001	-0.001	0.001	-0.001	-0.001
	40	-0.001	-0.001	-0.001	-0.001	0.003	-0.001	-0.001
	60	-0.001	-0.001	-0.001	-0.001	0.005	-0.001	-0.001
	80	-0.001	-0.002	-0.001	-0.001	0.007	-0.001	-0.001
	90	0.000	-0.001	0.000	0.000	0.009	0.000	0.000
0.2	20	0.177	0.199	0.198	0.197	0.188	0.197	0.197
	40	0.157	0.199	0.177	0.175	0.170	0.175	0.175
	60	0.121	0.198	0.136	0.134	0.134	0.134	0.134
	80	0.067	0.197	0.075	0.074	0.079	0.074	0.074
	90	0.037	0.199	0.041	0.040	0.048	0.040	0.040

$\widehat{\Delta}$: the mean estimated value of Δ over 5,000 simulated data sets

$p_{cens} = p_{DO} + p_{adm}$

Table 5: Empirical rejection rates for the test $H_0 : F(t) = G(t)$ for the proportional hazards scenarios.

Δ	p_{cens}	**Gehan**	**Harrell's C**	**Efron**	**Peron**	**Latta**	**Datta**	**Dong**	**Log-rank**
		p-val	p-val	p-val	p-val	p-val	p-val	p-val	p-val
Only drop-out censoring, $H(t) = I(t)$									
0	0	0.053	0.053	0.055	0.054	0.053	0.055	0.055	0.055
	10	0.055	0.055	0.056	0.055	0.055	0.056	0.056	0.057
	30	0.055	0.055	0.056	0.057	0.055	0.057	0.057	0.054
	50	0.054	0.054	0.054	0.057	0.057	0.053	0.056	0.056
	70	0.052	0.052	0.051	0.051	0.053	0.052	0.051	0.054
	90	0.047	0.047	0.057	0.057	0.052	0.058	0.057	0.05
0.2	0	0.686	0.686	0.69	0.691	0.688	0.69	0.69	0.807
	10	0.647	0.647	0.678	0.674	0.663	0.677	0.678	0.764
	30	0.552	0.552	0.633	0.628	0.59	0.632	0.633	0.66
	50	0.441	0.441	0.449	0.496	0.489	0.447	0.497	0.521
	70	0.289	0.289	0.232	0.287	0.326	0.24	0.291	0.342
	90	0.13	0.13	0.084	0.122	0.152	0.095	0.12	0.136
Only drop-out censoring, $H(t) \neq I(t)$									
0	0	0.054	0.054	0.056	0.054	0.054	0.055	0.055	0.055
	10	0.053	0.053	0.052	0.053	0.051	0.052	0.052	0.050
	30	0.053	0.053	0.052	0.053	0.053	0.053	0.053	0.051
	50	0.055	0.055	0.062	0.046	0.053	0.064	0.048	0.056
	70	0.048	0.048	0.196	0.025	0.043	0.218	0.028	0.048
	90	0.034	0.034	0.315	0.004	0.013	0.842	0.015	0.045
0.2	0	0.684	0.684	0.691	0.689	0.685	0.692	0.692	0.806
	10	0.640	0.640	0.675	0.675	0.658	0.674	0.674	0.759
	30	0.521	0.521	0.516	0.600	0.563	0.514	0.598	0.618
	50	0.391	0.391	0.055	0.366	0.432	0.046	0.362	0.457
	70	0.250	0.250	0.030	0.095	0.254	0.033	0.090	0.267
	90	0.062	0.062	0.305	0.002	0.020	0.822	0.007	0.078
Drop-out and administrative censoring									
0	20	0.055	0.055	0.057	0.058	0.055	0.057	0.057	0.053
	40	0.055	0.055	0.056	0.057	0.054	0.056	0.056	0.053
	60	0.056	0.056	0.058	0.057	0.056	0.057	0.057	0.058
	80	0.046	0.046	0.05	0.048	0.046	0.05	0.05	0.048
	90	0.042	0.042	0.044	0.045	0.045	0.045	0.045	0.043
0.2	20	0.634	0.634	0.665	0.662	0.658	0.665	0.665	0.723
	40	0.553	0.553	0.578	0.575	0.585	0.576	0.576	0.601
	60	0.408	0.408	0.422	0.419	0.451	0.419	0.419	0.43
	80	0.234	0.234	0.234	0.235	0.278	0.233	0.233	0.236
	90	0.141	0.141	0.145	0.141	0.189	0.144	0.144	0.142

p-val: the proportion of rejected H_0 over 5,000 simulations

$p_{cens} = p_{DO} + p_{adm}$

Table 6: Net treatment benefit estimates for the non-proportional hazards scenarios.

Δ	p_{cens}	**Gehan**	**Harrell's C**	**Efron**	**Peron**	**Latta**	**Datta**	**Dong**
		$\widehat{\Delta}_G$	$\widehat{\Delta}_H$	$\widehat{\Delta}_E$	$\widehat{\Delta}_P$	$\widehat{\Delta}_L$	$\widehat{\Delta}_{IPCW1}$	$\widehat{\Delta}_{IPCW2}$
Only drop-out censoring, $H(t) = I(t)$								
0	0	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	10	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	30	0.001	0.001	0.000	0.000	0.001	0.000	0.000
	50	0.001	0.001	0.000	0.001	0.001	0.000	0.001
	70	0.000	0.001	-0.002	-0.001	0.001	-0.002	-0.001
	90	0.000	0.002	-0.001	-0.001	0.004	-0.003	-0.001
0.2	0	0.200	0.201	0.200	0.200	0.200	0.200	0.200
	10	0.186	0.206	0.201	0.200	0.193	0.201	0.200
	30	0.159	0.218	0.202	0.200	0.177	0.202	0.200
	50	0.126	0.235	0.206	0.197	0.153	0.206	0.197
	70	0.084	0.264	0.203	0.177	0.116	0.204	0.179
	90	0.034	0.328	0.150	0.111	0.058	0.150	0.117
Only drop-out censoring, $H(t) \neq I(t)$								
0	0	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	10	0.000	0.000	0.001	0.000	0.000	0.001	0.001
	30	0.000	0.000	0.009	0.000	0.000	0.010	0.001
	50	0.000	0.001	0.061	0.000	0.000	0.065	0.002
	70	0.000	0.000	0.200	-0.003	0.000	0.255	0.006
	90	-0.001	-0.025	0.259	-0.009	-0.002	0.726	0.014
0.2	0	0.201	0.202	0.201	0.201	0.201	0.201	0.201
	10	0.184	0.207	0.202	0.201	0.192	0.202	0.201
	30	0.151	0.222	0.215	0.199	0.171	0.216	0.200
	50	0.115	0.243	0.277	0.191	0.143	0.284	0.195
	70	0.072	0.276	0.410	0.161	0.101	0.469	0.171
	90	0.018	0.328	0.336	0.053	0.029	0.867	0.076
Drop-out and administrative censoring								
0	20	0.000	0.001	0.000	0.000	0.002	0.000	0.000
	40	0.001	0.001	0.001	0.001	0.004	0.001	0.001
	60	0.001	0.001	0.001	0.001	0.006	0.001	0.001
	80	0.001	0.002	0.001	0.001	0.009	0.001	0.001
	90	0.000	0.000	0.000	0.000	0.009	0.000	0.000
0.2	20	0.183	0.208	0.200	0.199	0.192	0.199	0.199
	40	0.170	0.218	0.189	0.187	0.183	0.188	0.188
	60	0.144	0.240	0.162	0.160	0.158	0.160	0.160
	80	0.095	0.282	0.108	0.106	0.109	0.107	0.107
	90	0.059	0.324	0.067	0.065	0.072	0.066	0.066

$\widehat{\Delta}$: the mean estimated value of Δ over 5,000 simulated data sets

$p_{cens} = p_{DO} + p_{adm}$

Table 7: Empirical rejection rates for the test $H_0 : F(t) = G(t)$ for the non-proportional hazards scenarios.

Δ	p_{cens}	**Gehan**	**Harrell's C**	**Efron**	**Peron**	**Latta**	**Datta**	**Dong**	**Log-rank**
		p-val	p-val	p-val	p-val	p-val	p-val	p-val	p-val
Only drop-out censoring, $H(t) = I(t)$									
0	0	0.048	0.048	0.051	0.049	0.048	0.051	0.051	0.053
	10	0.049	0.049	0.053	0.051	0.050	0.053	0.053	0.050
	30	0.050	0.050	0.049	0.048	0.048	0.048	0.049	0.051
	50	0.050	0.050	0.048	0.048	0.048	0.049	0.047	0.046
	70	0.053	0.053	0.052	0.053	0.051	0.053	0.052	0.050
	90	0.053	0.053	0.052	0.052	0.049	0.050	0.051	0.049
0.2	0	0.694	0.694	0.695	0.697	0.695	0.695	0.695	0.627
	10	0.678	0.678	0.683	0.679	0.677	0.683	0.682	0.613
	30	0.631	0.631	0.630	0.626	0.635	0.630	0.631	0.579
	50	0.569	0.569	0.470	0.522	0.566	0.472	0.522	0.531
	70	0.462	0.462	0.264	0.339	0.467	0.272	0.333	0.437
	90	0.275	0.275	0.108	0.196	0.297	0.122	0.191	0.271
Only drop-out censoring, $H(t) \neq I(t)$									
0	0	0.049	0.049	0.050	0.049	0.048	0.051	0.051	0.054
	10	0.049	0.049	0.050	0.048	0.048	0.050	0.050	0.049
	30	0.046	0.046	0.047	0.049	0.047	0.048	0.050	0.046
	50	0.050	0.050	0.054	0.043	0.051	0.057	0.044	0.048
	70	0.050	0.050	0.131	0.036	0.045	0.159	0.040	0.050
	90	0.044	0.044	0.280	0.012	0.024	0.606	0.028	0.054
0.2	0	0.701	0.701	0.704	0.704	0.701	0.703	0.703	0.632
	10	0.678	0.678	0.687	0.682	0.680	0.688	0.686	0.607
	30	0.631	0.631	0.673	0.616	0.633	0.676	0.621	0.573
	50	0.551	0.551	0.657	0.446	0.545	0.683	0.458	0.525
	70	0.429	0.429	0.551	0.216	0.411	0.653	0.236	0.417
	90	0.160	0.160	0.332	0.035	0.091	0.790	0.067	0.191
Drop-out and administrative censoring									
0	20	0.050	0.050	0.051	0.050	0.050	0.051	0.051	0.050
	40	0.049	0.049	0.051	0.048	0.050	0.052	0.052	0.047
	60	0.056	0.056	0.057	0.055	0.056	0.057	0.057	0.053
	80	0.050	0.050	0.050	0.049	0.052	0.050	0.050	0.049
	90	0.051	0.051	0.052	0.052	0.051	0.052	0.052	0.052
0.2	20	0.672	0.672	0.675	0.673	0.680	0.675	0.675	0.609
	40	0.634	0.634	0.633	0.632	0.653	0.633	0.633	0.596
	60	0.566	0.566	0.562	0.564	0.597	0.563	0.563	0.544
	80	0.431	0.431	0.430	0.425	0.481	0.432	0.432	0.421
	90	0.314	0.314	0.315	0.306	0.381	0.306	0.306	0.307

p-val: the proportion of rejected H_0 over 5,000 simulations

$p_{cens} = p_{DO} + p_{adm}$

References

Acion *et al.*, 2006 Acion, A., Peterson, J., Temple, S., *et al.* (2006) Probabilistic index: an intuitive non-parametric approach to measuring the size of the treatment effects. *Stat Med* **25**, 591—602.

Anderson and Verbeeck, 2019 Anderson, W.N., Verbeeck, J. (2019) Exact Bootstrap and Permutation Distribution of Wins and Losses in a Hierarchical Trial. *https://arxiv.org/pdf/1901.10928.pdf*.

Brunner *et al.*, 2021 Brunner, E., Vandemeulebroecke, M., Mütze, T. (2021) Win odds: An adaptation of the win ratio to include ties. *Statistics in Medicine* **40**, 3367–3384.

Buyse, 2010 Buyse, M. (2010) Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine* **29**, 3245–3257.

Buyse, 2008 Buyse, M. (2008) Reformulating the hazard ratio to enhance communication with clinical investigators [Letter to the Editor]. *Clin Trials* **5**, 641–2.

Datta *et al.*, 2010 Datta, S., Bandyopadhyay, D., Satten, G. (2010) Inverse Probability of Censoring Weighted U-statistics for Right-Censored Data with an Application to Testing Hypotheses. *Scandinavian Journal of Statistics* **37**, 680–700.

De Backer *et al.*, 2020 De Backer, M., Legrand, C., Péron, J., Lambert, A., Buyse, M. (2020) *On the use of Extreme Value Tail Modeling for Generalized Pairwise Comparisons with Censored Outcomes*. Manuscript submitted for publication.

Dong *et al.*, 2020a Dong, G., Hoaglin, D.C., Qiu, J., Matsouaka, R.A., Chang, Y.-W., Wang, J., Vandemeulebroecke, M. (2020a) The Win Ratio: On Interpretation and Handling of Ties. *Statistics in Biopharmaceutical Research* **12**, 106–99.

Dong *et al.*, 2020b Dong, G., Mao, L., Huang, B., Gamalo-Siebers, M., Wang, J., You, G., Hoaglin, D.C. (2020b) The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring. *Journal of Biopharmaceutical Statistics* **30**, 882–899.

Dong *et al.*, 2020c Dong, G., Huang, B., Chang, Y.-W., Seifu, Y., Song, J., Hoaglin, D.C. (2020c) The win ratio: Impact of censoring and follow-up time and use with nonproportional hazards. *Pharmaceutical Statistics* **19**, 168–177.

Dong *et al.*, 2021 Dong, G., Huang, B., Wang, D., Verbeeck, J., Wang, J., Hoaglin, D.C. (2021) Adjusting win statistics for dependent censoring. *Pharmaceutical Statistics* **20**, 440–450.

Efron, 1967 Efron, B. (1967) The two sample problem with censored data. *Proc. 5th Berkeley Symposium on Math. Statist. and Prob.* **4**, 831–853.

Finkelstein and Schoenfeld, 1999 Finkelstein, D.M., Schoenfeld, D.A. (1999) Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine* **18**, 1341-1354.

Finkelstein and Schoenfeld, 2019 Finkelstein, D.M., Schoenfeld, D.A. (2019) Graphing the Win Ratio and its components over time. *Statistics in Medicine* **38**, 53-–61.

Gehan, 1965 Gehan, E.A. (1965) A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* **52**, 203–223.

Harrell *et al.*, 1982 Harrell, F.A., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati R.A. (1982) Evaluating the yield of medical tests. *The Journal of the American Medical Association* **247**, 2543–2546.

Harrell, 1986 Harrell, F.A., Lee, K.L., Mark, D.B. (1986) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.

Koziol, 2009 Koziol, J.A., Jia, Z. (2009) The concordance index C and the Mann-Whitney parameter Pr(X¿Y) with randomly censored data. *Biometrical Journal* **51**, 467–74.

Latta, 1977 Latta, R.B. (1977) Generalized Wilcoxon statistics for the two sample problem with censored data. *Biometrika* **63**, 633–635.

Lee, 1990 Lee, A.J. (1990) *U-Statistics: Theory and Practice*. CRC.

Mann and Whitney, 1947 Mann, H.B., Whitney, D.R. (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18**, 50–60.

Mantel, 1967 Mantel, N. (1967) Ranking Procedures for Arbitrarily Restricted Observation. *Biometrics* **23**, 65–78.

Oakes, 2016 Oakes, D. (2016) On the win-ratio statistic in clinical trials with multiple types of event. *Biometrika* **103**, 742—745.

O'Brien, 1984 O'Brien, P.C. (1984) Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

Peron *et al.*, 2016 Péron, J., Buyse, M., Ozenne, B., Roche, L., Roy, P. (2016) An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Stat Methods Med Research* **27**, 1230–1239.

Peto and Peto, 1972 Peto, R. and Peto, J. (1972) Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society* **135**, 185–207.

Pocock *et al.*, 2012 Pocock, S.J., Ariti, C.A., Collier, T.J., Wang, D. (2012) The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal* **33**, 176–182.

Ramchandani *et al.*, 2016 Ramchandani, R., Schoenfeld, D., Finkelstein, D.M. (2016) Global rank tests for multiple, possibly censored, outcomes. *Biometrics* **72**, 926-935.

Rauch *et al.*, 2014 Rauch, G., Jahn-Eimermacher, A., Brannath, W., Kieser, M. (2014) Opportunities and challenges of combined effect measures based on prioritized outcomes. *Statistics in Medicine* **33**, 1104—1120.

Stute and Wang, 1994 Stute, W., Wang, J.L. (1994) Multi-Sample U-Statistics for Censored Data. *Scandinavian Journal of Statistics* **20**, 369–374.

Tournigand *et al.*, 2004 Tournigand, C., André, T., Achille, E., Lledo, G., Flesh, M., Mery-Mignard, D., Quinaux, E., Couteau, C., Buyse, M., Ganem, G., Landi, B., Colin, P., Louvet, C., de Gramont, A. (2004) FOLFIRI Followed by FOLFOX6 or the Reverse Sequence in Advanced Colorectal Cancer: A Randomized GERCOR Study. *Journal of Clinical Oncology* **22**, 229 – 237.

Verbeeck *et al.*, 2019 Verbeeck, J., Spitzer, E., de Vries, T., van Es, G.A., Anderson, W. N., Van Mieghem, N.M., Leon, M.B., Molenberghs, G., Tijssen, J. (2019) Generalized pairwise comparison methods to analyze (non)prioritized composite endpoints. *Statistics in Medicine* **38**, 5641–5656.

Verbeeck *et al.*, 2021 Verbeeck, J., Deltuvaite-Thomas, V., Berckmoes, B., Burzykowski, T., Aerts, M., Thas, O., Buyse, M., Molenberghs, G. (2021) Unbiasedness and efficiency of nonparametric and UMVUE estimators of the probabilistic index and related statistics. *Statistical Methods in Medical Research* **30**, 747–768.

Verbeeck *et al.*, 2020 Verbeeck, J., Ozenne, B., Anderson, W.N. (2020) Evaluation of inferential methods for the net benefit and win ratio statistics. *Journal of Biopharmaceutical Statistics* **30**, 765–782.

Wilcoxon, 1945 Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**, 80–83.