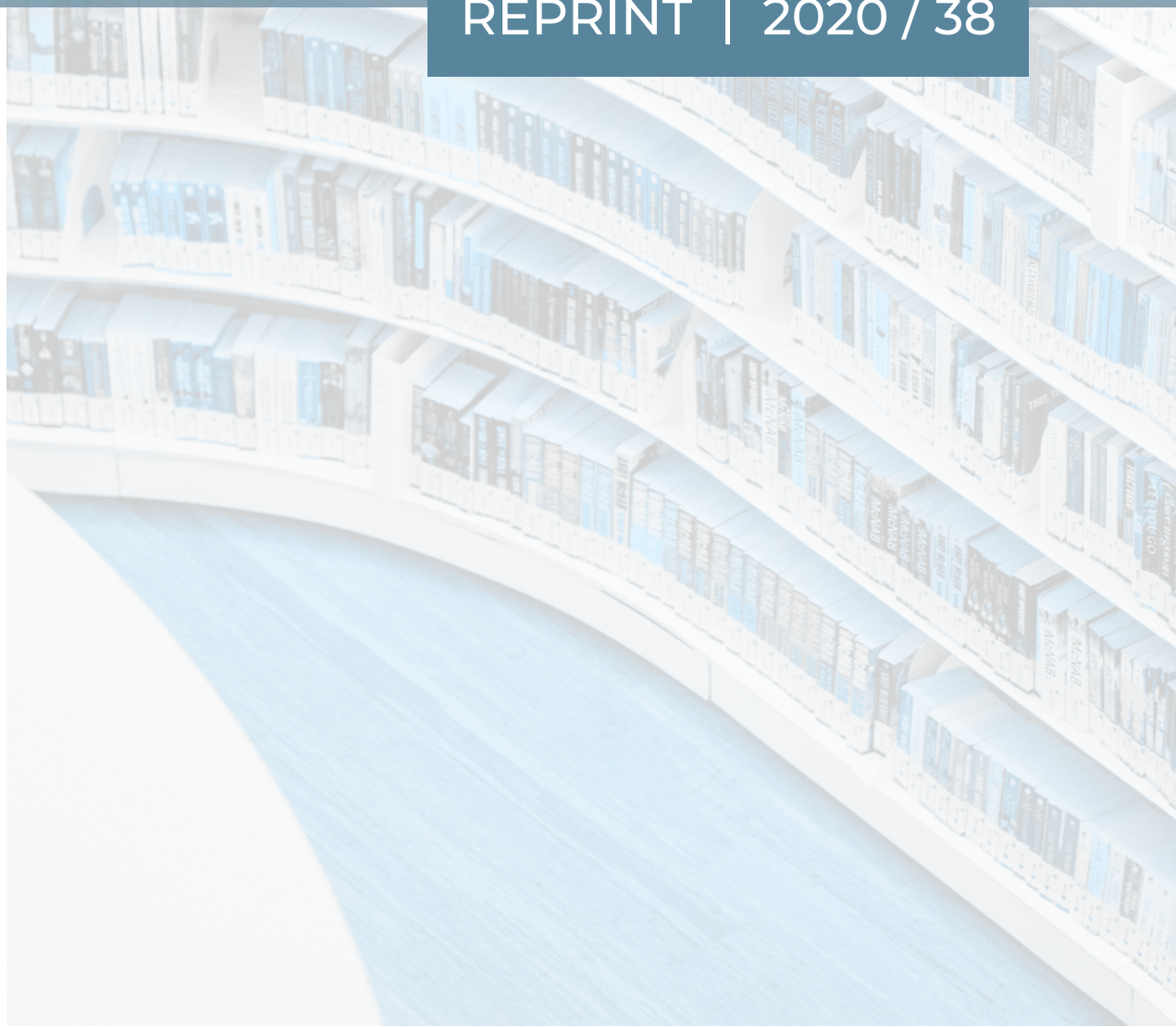
A NEW MEASURE OF TREATMENT EFFECT IN CLINICAL TRIALS INVOLVING COMPETING RISKS BASED ON GENERALIZED PAIRWISE COMPARISONS

Cantagallo, E., De Backer, M. , Kicinski, M., Ozenne, B., Collette, L., Legrand, C., Buyse, M. and J. Péron

REPRINT | 2020 / 38



UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

Received: 25 November 2019 | Revised: 31 July 2020 | Accepted: 4 August 2020

DOI: 10.1002/bimj.201900354

Biometrical Journal

RESEARCH PAPER

A new measure of treatment effect in clinical trials involving competing risks based on generalized pairwise comparisons

Eva Cantagallo[1] | **Mickaël De Backer**[2,*] | **Michal Kicinski**[1,*] | **Brice Ozenne**[3,4] | **Laurence Collette**[1] | **Catherine Legrand**[2] | **Marc Buyse**[5,6] | **Julien Péron**[7,8,9,*]

[1] Statistics Department, European Organisation for Research and Treatment of Cancer (EORTC), Brussels, Belgium

[2] Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), LIDAM, UCLouvain, Louvain-la-Neuve, Belgium

[3] Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet, Copenhagen, Denmark

[4] Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark

[5] International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium

[6] Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Hasselt University, Diepenbeek, Belgium

[7] CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Université Lyon 1, Villeurbanne, France

[8] Service de Biostatistique et Bioinformatique, Hospices Civils de Lyon, Lyon, France

[9] Oncology Department, Hospices Civils de Lyon, Pierre Bénite, France

Correspondence
Eva Cantagallo, Statistics Department, European Organisation for Research and Treatment of Cancer (EORTC), Avenue Emmanuel Mounier 83/11, 1200 Brussels, Belgium.
Email: eva.cantagallo@eortc.org

*Contributed equally to this work.

Funding information
Innoviris; EORTC Cancer Research Fund; Government of Wallonia (Biowin Consortium agreement no. 7979), Grant/Award Number: 7979

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

Abstract

In survival analysis with competing risks, the treatment effect is typically expressed using cause-specific or subdistribution hazard ratios, both relying on proportional hazards assumptions. This paper proposes a nonparametric approach to analyze competing risks data based on generalized pairwise comparisons (GPC). GPC estimate the net benefit, defined as the probability that a patient from the treatment group has a better outcome than a patient from the control group minus the probability of the opposite situation, by comparing all pairs of patients taking one patient from each group. GPC allow using clinically relevant thresholds and simultaneously analyzing multiple prioritized endpoints. We show that under proportional subdistribution hazards, the net benefit for competing risks settings can be expressed as a decreasing function of the subdistribution hazard ratio, taking a value 0 when the latter equals 1. We propose four net benefit estimators dealing differently with censoring. Among them, the Péron estimator uses the Aalen–Johansen estimator of the cumulative incidence functions to classify the pairs for which the patient with the best outcome could not be determined due to censoring. We use simulations to study the bias of these estimators and the size and power of the tests based on the net benefit. The Péron estimator was approximately unbiased when the sample size was large and the censoring distribution's support sufficiently wide. With one endpoint, our approach showed a comparable power to a proportional subdistribution hazards model even under

proportional subdistribution hazards. An application of the methodology in oncology is provided.

KEYWORDS
clinical trial, competing risks, generalized pairwise comparisons, multicriteria analysis, survival analysis

1 | INTRODUCTION

In randomized oncology trials, primary endpoints are often time-to-event endpoints, such as overall survival. The treatment effect is commonly reported through the hazard ratio (HR) estimated from a Cox proportional hazards model (Cox, 1972). Other existing measures include the difference between survival probabilities at certain time point(s), the difference/ratio of restricted mean survival times (e.g., Saad et al., 2018), and the odds ratio from a proportional odds survival model (e.g., Martinez, Sinha, Wang, Lipsitz, & Chappell, 2017).

When several events are of interest, composite endpoints such as progression-free or relapse-free survival have become popular, in particular as they increase the statistical power as a result of an increased event rate compared to considering one event only (McCoy, 2018). However, this approach uses only the time to the earliest event, which may not be the one of primary interest to clinicians and patients (Pocock, Ariti, Collier, & Wang, 2012).

A novel approach called generalized pairwise comparisons (GPC) was recently proposed to describe the treatment effect in two-group randomized clinical trials (Buyse, 2010). This method estimates a measure of treatment effect called the net benefit and allows the simultaneous analysis of multiple endpoints, prioritized according to their clinical relevance. Moreover, a clinically relevant threshold can be defined for each endpoint, allowing to focus on clinically relevant differences in the outcomes between the two groups. In particular, the method can be used when several events of different clinical relevance are of interest, such as progression of disease and death (Buyse, 2010).

Competing risks frequently occur in clinical trials. For example, the event of interest of an oncology trial can be death due to cancer, while patients can die from other causes as well, with the occurrence of this competing event precluding the occurrence of the event of interest (Gooley, Leisenring, Crowley, & Storer, 1999). Common statistical methods for competing risks analysis in a clinical trial context comprise the Cox-type proportional cause-specific and subdistribution hazards models (Kalbfleisch & Prentice, 1980; Fine & Gray, 1999). The two approaches allow the estimation and testing of the treatment effect through two different quantities: the cause-specific hazard ratio for the former and the subdistribution hazard ratio (sHR) for the latter. Both models rely on several assumptions, including that of proportionality of the corresponding hazards, which are often violated. Moreover, these measures of treatment effect are rather difficult to interpret. A third approach is the Cox-type proportional marginal hazards model, which requires the assumption that the joint distribution of the competing risks is identifiable (e.g., Emura, Shih, Ha, & Wilke, 2020). Other available methods for the analysis of competing risks data include the pseudo-observation method (Andersen, Klein, & Rosthøj, 2003) and direct binomial regression (Scheike, Zhang, & Gerds, 2008), to name a few. However, the proportional cause-specific and subdistribution hazards models remain the most frequently used approaches in practice in clinical trials contexts, possibly due to their simplicity.

In this paper, we propose to use GPC in a competing risks setting. We define the net benefit in the context of competing risks and study the relationship with the sHR. We propose several estimators of the net benefit and investigate their performance using simulations. Besides the possibility to create formal decision rules taking into account several prioritized endpoints and thresholds of clinically relevant difference, our approach is nonparametric and does not rely on the proportional hazards assumption.

The paper is organized as follows. In Section 2, the classical GPC approach for time-to-event endpoints is briefly summarized. Section 3 is dedicated to the extension of GPC to competing risks, covering in particular the definition of the net benefit, its relationship with the sHR, and its estimation. In Section 4, the results of an extensive simulation study comparing various proposed net benefit estimators is presented. Section 5 illustrates the proposed methodology through its application to data from an oncology trial. Finally, the last section discusses the results and provides some concluding remarks.

2 | GENERALIZED PAIRWISE COMPARISONS

GPC estimate the net benefit, defined as the difference between the probability that a patient from the treatment group has a better outcome than a patient from the control group and the probability of the opposite situation (Buyse, 2010). For a time-to-event endpoint, denoted by Y^T and Y^C in the treatment and control group, respectively, this corresponds to:

$$\Delta_\tau^* = \mathrm{P}(Y^T > Y^C + \tau) - \mathrm{P}(Y^C > Y^T + \tau),$$

with τ a threshold of clinically relevant difference, representing the minimal difference considered to be relevant. The net benefit ranges from −1 to 1, and is equal to −1 when the treatment is uniformly worse than the control, to 0 when both treatments have similar effects, and to 1 if the treatment is uniformly better than the control (Buyse, 2010). This parameter is estimated by the pairwise comparison of every patient from one group with every patient from the other group. All possible pairs composed of one patient from the treatment group and one patient from the control group are formed, and a pairwise score is computed for each pair. The net benefit estimator is equal to the average pairwise score over all pairs. A pair is considered as a win when the comparison is in favor of the treatment (i.e., the treatment patient has a better outcome than the control patient), as a loss when it is in favor of the control, and as neutral when both patients of the pair have similar outcomes. Wins are assigned a score of 1, losses a score of −1, and neutral pairs a null score.

In the presence of censoring, one might not be able to determine the patient with the best outcome. When such pairs are assigned a null score, this method is equivalent to the Gehan's modification of the Wilcoxon test (Gehan, 1965). This however introduces a bias in the net benefit estimate (Buyse, 2008; Efron, 1967; Péron, Buyse, Ozenne, Roche, & Roy, 2018). An extension of the GPC approach handles censoring using estimates of the survival function to compute the contribution of a pair to the net benefit (Péron et al., 2018).

When multiple endpoints are of interest, they can be simultaneously analyzed using the GPC approach. The outcomes need to be prioritized by investigators, or potentially by patients, in decreasing order of clinical relevance. Pairs are first compared based on the highest priority endpoint. Pairs classified as neutral or pairs that could not be classified due to the lack of data are then analyzed based on the second priority endpoint. This continues until all pairs are actually classified as a win or a loss or until all endpoints of interest have been used in the analysis. An overall net benefit is then calculated as the sum of the contribution of each endpoint.

A more extensive description of the GPC approach for time-to-event endpoints and this extension can be found elsewhere (Buyse, 2010; Péron et al., 2018). In the next section, the adaptation of the GPC approach to competing risks settings is presented in detail.

3 | GENERALIZED PAIRWISE COMPARISONS FOR COMPETING RISKS

Throughout this paper, a study with two groups of patients, a treatment (T) and a control (C) group, and two competing events, that is, an event of interest and a competing event, is considered. These two events are assumed complementary, meaning that they are mutually exclusive and that one of them will eventually occur. Let $\check{X}^g$ be the event time of a patient in group g ($g \in \{\mathrm{T}, \mathrm{C}\}$) and E^g be the event type, with $E^g = 1$ for the event of interest and $E^g = 2$ for the competing event. Similar to Gray (1988) and Fine and Gray (1999), we define the times to the event of interest and to the competing event, respectively, in group g as:

$$X_1^g = \check{X}^g \times \mathrm{I}(E^g = 1) + \infty \times \mathrm{I}(E^g = 2),$$
$$X_2^g = \check{X}^g \times \mathrm{I}(E^g = 2) + \infty \times \mathrm{I}(E^g = 1),$$

where $\mathrm{I}(\cdot)$ is the indicator function. A patient experiencing the competing event therefore has an infinite time to the event of interest and vice versa.

The next subsection is dedicated to the definition of the net benefit in such a context and to its relationship with the sHR of the event of interest. Then, four different estimators of the net benefit are proposed.

3.1 | Net benefit definition

We define the net benefit in a competing risks setting as the probability that an individual from the treatment group has a longer time to the event of interest, by at least τ units, than an individual from the control group minus the probability of the opposite situation:

$$\Delta_\tau = \mathrm{P}\left(X_1^T > X_1^C + \tau\right) - \mathrm{P}\left(X_1^C > X_1^T + \tau\right),$$

using the convention that $\mathrm{P}(\infty > \infty) = 0$. Note that a favorable outcome is observed when either the event of interest occurs both for a patient from the treatment group and the control group with a shorter time for the control group patient or the competing event occurs for a patient from the treatment group and the event of interest for a patient from the control group.

Without competing risks, the net benefit is directly related to the HR when the hazards are proportional. More precisely, in the case of a single time-to-event outcome with a null clinical threshold τ, the net benefit and the HR are linked by the following relationship: $\Delta_0^* = (1 - \mathrm{HR})/(1 + \mathrm{HR})$ (Buyse, 2010; Moser & McCann, 2008). It turns out that in a competing risks setting, the net benefit can be expressed as a decreasing function of the sHR of the event of interest when the subdistribution hazards are proportional. The sHR of the event of interest is defined as $\mathrm{sHR} = h_1^T(t)/h_1^C(t)$, with $h_1^g(t)$ the subdistribution hazard function of the event of interest in group g. Let $F_k^g(t) = \mathrm{P}(\tilde{X}^g < t, E^g = k)$ be the cumulative incidence function (CIF) of event type k ($k \in \{1, 2\}$) in group g. In the case of a single outcome with a null clinically relevant threshold, the relationship between the net benefit and the sHR is (see Appendix A.1):

$$\Delta_0 = \frac{1 - \mathrm{sHR}}{1 + \mathrm{sHR}}\left(1 - \lim_{t\to\infty} S_1^T(t)S_1^C(t)\right),$$

with $S_k^g(t) = 1 - F_k^g(t)$, that is, the probability for a patient of group g not to fail from event type k by time t. This relationship indicates a consistent interpretation of both measures regarding the treatment effect. Indeed, a sHR smaller than 1 corresponds to a positive net benefit, both suggesting a beneficial effect of the treatment compared to the control for the event of interest. When the sHR is equal to 1, the net benefit is null, both corresponding to the absence of treatment effect. Finally, a sHR larger than 1 corresponds to a negative net benefit, indicating a harmful effect of the treatment as compared to control. This formula also shows that the range of Δ_0 values depends on the incidence of the competing event.

3.2 | Estimation

Suppose the two groups of patients are independent. Let $(\tilde{X}_i^T, E_i^T)$, $i = 1, \ldots, n$, and $(\tilde{X}_j^C, E_j^C)$, $j = 1, \ldots, m$, be independent and identically distributed samples from $(\tilde{X}^T, E^T)$ and $(\tilde{X}^C, E^C)$, respectively. Let also (X_{1i}^T, X_{2i}^T) and (X_{1j}^C, X_{2j}^C) be independent and identically distributed samples from (X_1^T, X_2^T) and (X_1^C, X_2^C), respectively. We define H_i^T and H_j^C as the censoring time, assumed independent from $(\tilde{X}_i^T, E_i^T)$ and $(\tilde{X}_j^C, E_j^C)$. Under right censoring, we observe $\bar{X}_i^T = \min(\tilde{X}_i^T, H_i^T)$ and $\gamma_i^T = E_i^T \mathrm{I}(\tilde{X}_i^T \leq H_i^T)$ in the treatment group, and $\bar{X}_j^C = \min(\tilde{X}_j^C, H_j^C)$ and $\gamma_j^C = E_j^C \mathrm{I}(\tilde{X}_j^C \leq H_j^C)$ in the control group.

Let s_{ij} be the score of pair (i, j), also called pairwise score, and $\hat{s}_{ij}$ its estimator. The net benefit is estimated as:

$$\hat{\Delta}_\tau = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{s}_{ij}.$$

We describe below various estimators of the net benefit in a competing risks setting, differing in the way of handling censoring.

3.2.1 | Gehan scoring rule

To develop the estimator, we followed the same logic as used by Buyse (2010) in a GPC analysis of a time-to-event variable without competing risks. The pairwise score is equal to 1 for pairs that are decidedly wins (favorable pairs), to −1 for pairs that are decidedly losses (unfavorable pairs) and to 0 for pairs decidedly neutral in terms of the time until the event of interest (Table 1). Note that in the competing risks setting, consistent with the definition of Δ_τ, pairs with a patient from the treatment group having the competing event and a patient from the control group the event of interest are classified as

TABLE 1 Gehan scoring rule in a competing risks setting

$\left(\gamma_i^T, \gamma_j^C\right)$	Event times	Classification	$\hat{s}_{ij}$
(1,1)	$X_{1i}^T - X_{1j}^C > \tau$	Win	1
	$X_{1i}^T - X_{1j}^C < -\tau$	Loss	−1
	$\lvert X_{1i}^T - X_{1j}^C \rvert \leq \tau$	Neutral	0
(1,0)	$X_{1i}^T - \tilde{X}_{1j}^C > \tau$	Indecisive	0
	$X_{1i}^T - \tilde{X}_{1j}^C < -\tau$	Loss	−1
	$\lvert X_{1i}^T - \tilde{X}_{1j}^C \rvert \leq \tau$	Indecisive	0
(0,1)	$\tilde{X}_{1i}^T - X_{1j}^C > \tau$	Win	1
	$\tilde{X}_{1i}^T - X_{1j}^C < -\tau$	Indecisive	0
	$\lvert \tilde{X}_{1i}^T - X_{1j}^C \rvert \leq \tau$	Indecisive	0
(0,0)	-	Indecisive	0
(2,1)	-	Win	1
(1,2)	-	Loss	−1
(2,2)	-	Neutral	0
(2,0)	-	Indecisive	0
(0,2)	-	Indecisive	0

favorable irrespective of the time of the event. Pairs for which both patients experienced the competing event are classified as neutral. Pairs for which the patient with the largest time to the event of interest cannot be determined due to censoring are considered indecisive, as opposed to decisive pairs, with an associated score of 0. Because the patients from decisive and indecisive pairs may have different outcomes, the exclusion of the latter may cause a bias in the analysis, similarly to the case without competing risks (Efron, 1967).

3.2.2 | Péron scoring rule

The second estimator is based on the idea of Efron (1967) and Péron et al. (2018) to analyze a time-to-event variable using GPC in the absence of competing risks. In brief, a score of 1, −1, or 0 is assigned to pairs for which it is observed which patient had a longer time to event, similarly to the Gehan scoring rule. When a pair cannot be classified with certainty due to censoring, the pairwise score is based on the conditional probabilities that a pair is a win and a loss, estimated using the Kaplan–Meier estimator (Kaplan & Meier, 1958).

Here, we propose an estimator of the net benefit defined in the presence of competing risks based on a similar idea. The pairwise score s_{ij} is defined as the difference in the probabilities for a pair to be a win and a loss, conditionally on the observed times and the event indicators:

$$\begin{aligned} s_{ij} = &\mathrm{P}\left(X_{1i}^T > X_{1j}^C + \tau | \tilde{X}_i^T, \tilde{X}_j^C, \gamma_i^T, \gamma_j^C\right) \\ &-\mathrm{P}\left(X_{1j}^C > X_{1i}^T + \tau | \tilde{X}_i^T, \tilde{X}_j^C, \gamma_i^T, \gamma_j^C\right). \end{aligned} \tag{1}$$

The two probabilities in (1) are abbreviated by $p_{ij,win}$ and $p_{ij,loss}$, respectively, in the remainder of this paper. The pairwise score is estimated by $\hat{s}_{ij} = \hat{p}_{ij,win} - \hat{p}_{ij,loss}$, with $\hat{p}_{ij,win}$ and $\hat{p}_{ij,loss}$ being estimators of the conditional probabilities $p_{ij,win}$ and $p_{ij,loss}$, respectively.

The probability for a pair to be neutral, $p_{ij,neu}$, is defined as:

$$\begin{aligned} p_{ij,neu} = &\mathrm{P}\left((\lvert X_{1i}^T - X_{1j}^C \rvert \leq \tau) \cap (E_i^T = 1) \cap (E_j^C = 1) | \tilde{X}_i^T, \tilde{X}_j^C, \gamma_i^T, \gamma_j^C\right) \\ &+ \mathrm{P}\left((E_i^T = 2) \cap (E_j^C = 2) | \tilde{X}_i^T, \tilde{X}_j^C, \gamma_i^T, \gamma_j^C\right), \end{aligned}$$

so that $p_{ij,win} + p_{ij,loss} + p_{ij,neu} = 1$. In the remainder of this paper, $\hat{p}_{ij,neu}$ denotes an estimator of $p_{ij,neu}$. With the Péron scoring rule, pairs are classified into the win, loss, and neutral categories with a weight equal to the estimated conditional probability of belonging to that category. The pairwise score s_{ij} can be expressed as a function of the CIFs, and $\hat{s}_{ij}$ is

obtained by replacing the CIFs in s_{ij} with their Aalen–Johansen (AJ) estimator (Aalen & Johansen, 1978). The derivation and estimation of the conditional probabilities $p_{ij,win}$, $p_{ij,loss}$, and $p_{ij,neu}$ for different values of $\tilde{X}_i^T, \tilde{X}_j^C$, γ_i^T, and γ_j^C is available in the Supporting Information. As an illustration, consider a pair of observed event times $(\tilde{x}_i^T, \tilde{x}_j^C)$ with event indicators $(\gamma_i^T, \gamma_j^C) = (1,0)$ such that $\tilde{x}_i^T > \tilde{x}_j^C + \tau$. Since the two events are mutually exclusive, the probability that this pair is a win is equal to:

$$\begin{aligned}
&\mathrm{P}\left(X_{1i}^T > X_{1j}^C + \tau | X_{1i}^T = \tilde{x}_i^T, X_{1j}^C > \tilde{x}_j^C, X_{2j}^C > \tilde{x}_j^C\right) \\
&= \frac{\mathrm{P}\left((\tilde{x}_i^T > X_{1j}^C + \tau) \cap (X_{1j}^C > \tilde{x}_j^C) \cap (X_{2j}^C > \tilde{x}_j^C)\right)}{\mathrm{P}\left(\check{X}_j^C > \tilde{x}_j^C\right)} \\
&= \frac{\mathrm{P}\left(\tilde{x}_j^C < X_{1j}^C < \tilde{x}_i^T - \tau\right)}{1 - \mathrm{P}\left(\check{X}_j^C < \tilde{x}_j^C\right)} \\
&= \frac{F_1^C(\tilde{x}_i^T - \tau) - F_1^C(\tilde{x}_j^C)}{1 - F_1^C(\tilde{x}_j^C) - F_2^C(\tilde{x}_j^C)},
\end{aligned} \tag{2}$$

and is estimated by $(\hat{F}_1^C(\tilde{x}_i^T - \tau) - \hat{F}_1^C(\tilde{x}_j^C))/(1 - \hat{F}_1^C(\tilde{x}_j^C) - \hat{F}_2^C(\tilde{x}_j^C))$, with $\hat{F}_k^g(t)$ denoting the AJ estimator of the CIF of event type k in group g.

The proposed estimator is only defined when $\hat{F}_1^T(t)$, $\hat{F}_2^T(t)$, $\hat{F}_1^C(t)$, and $\hat{F}_2^C(t)$ are defined for all $t > 0$. This is not the case when the observation with the longest follow-up time is censored in one or both patient group(s). To handle this issue, we use the available information to compute lower bounds of the probabilities of interest. Let t_{max}^g be the time up to which the AJ estimator of the CIFs in group g is available. For terms that cannot be estimated, the last available value of the corresponding CIF is used instead, resulting in an estimate of a lower bound of the probabilities of interest. Consider, for example, a situation where the last observation in the control group is censored before $\tilde{x}_i^T - \tau$. The estimator of $F_1^C(\tilde{x}_i^T - \tau)$ in (2) is not defined in this case. A lower bound of this term can however be estimated by $\hat{F}_1^C(t_{max}^C) \leq \hat{F}_1^C(\tilde{x}_i^T - \tau)$, which is used instead of $\hat{F}_1^C(\tilde{x}_i^T - \tau)$ in the formula. The same procedure is used for all terms that cannot be computed. The sum of the estimated lower bound probabilities of a pair to be a win, a loss, and neutral may be smaller than 1, and the remaining weight is assigned to the probability of being indecisive.

3.2.3 | Pairwise correction

To further deal with the issue of pairs with an indecisive component, a pairwise correction can be used. This correction is applied in combination with either of the two scoring rules presented above with the aim of redistributing the indecisive component of pairs to their decisive component. Let $\hat{w}_{win}$, $\hat{w}_{loss}$, and $\hat{w}_{neu}$ be the estimated proportion of wins, losses, and neutral pairs, respectively:

$$\hat{w}_z = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{w}_{ij,z},$$

with $z \in \{win, loss, neu\}$. For the Gehan scoring rule, $\hat{w}_{ij,win}$, $\hat{w}_{ij,loss}$, and $\hat{w}_{ij,neu}$ are indicator functions taking a value 1 when a pair is decidedly classified as a win, a loss, and a neutral pair, respectively, and 0 otherwise. For the Péron scoring rule, $\hat{w}_{ij,win} = \hat{p}_{ij,win}$, $\hat{w}_{ij,loss} = \hat{p}_{ij,loss}$, and $\hat{w}_{ij,neu} = \hat{p}_{ij,neu}$. Assuming that indecisive pairs behave on average like decisive pairs, the pairwise correction redistributes the weight of the former to the latter based on the observed proportions of wins, losses, and neutral pairs, and computes corrected weights as $\hat{w}_{ij,z}^{cr} = \hat{w}_{ij,z} + \hat{w}_z(1 - \hat{w}_{ij,win} - \hat{w}_{ij,loss} - \hat{w}_{ij,neu})/(\hat{w}_{win} + \hat{w}_{loss} + \hat{w}_{neu})$. The corrected estimator is given by:

$$\hat{\Delta}_\tau^{cr} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{s}_{ij}^{cr},$$

with $\hat{s}_{ij}^{cr} = \hat{w}_{ij,win}^{cr} - \hat{w}_{ij,loss}^{cr}$.

With a single outcome, the pairwise correction corresponds to the multiplication of the uncorrected net benefit by the inverse of the estimated proportion of decisive pairs:

$$\begin{aligned}\hat{\Delta}_\tau^{cr} &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\hat{w}_{ij,win}^{cr} - \hat{w}_{ij,loss}^{cr}\right)\\ &= \hat{w}_{win} - \hat{w}_{loss} + (1 - \hat{w}_{win} - \hat{w}_{loss} - \hat{w}_{neu})\frac{\hat{w}_{win} - \hat{w}_{loss}}{\hat{w}_{win} + \hat{w}_{loss} + \hat{w}_{neu}}\\ &= \frac{\hat{w}_{win} - \hat{w}_{loss}}{\hat{w}_{win} + \hat{w}_{loss} + \hat{w}_{neu}}\\ &= \frac{\hat{\Delta}_\tau}{\hat{w}_{win} + \hat{w}_{loss} + \hat{w}_{neu}}.\end{aligned}$$

All together, four estimators of the net benefit can be considered: based on the Gehan scoring rule and without the pairwise correction (uncorrected Gehan scoring rule), based on the Gehan scoring rule and with the pairwise correction (corrected Gehan scoring rule), based on the Péron scoring rule and without the pairwise correction (uncorrected Péron scoring rule), and based on the Péron scoring rule and with the pairwise correction (corrected Péron scoring rule).

3.3 | Hypothesis testing

Similarly as for other GPC procedures (Buyse, 2010; Ozenne & Péron, 2019), bootstrapping or a permutation test can be used for hypothesis testing, and bootstrapping for confidence interval (CI) estimation. However, permutation tests rely on an assumption of the same censoring in the two groups. This is due to the fact that, when randomly reassigning the treatment labels among observations, we assume that the joint distribution of $(\tilde{X}_i^T, \gamma_i^T)$ is the same as that of $(\tilde{X}_j^C, \gamma_j^C)$ under the null hypothesis, which in turn amounts to requiring the same censoring distribution under the null hypothesis.

4 | SIMULATION STUDY

Simulations were performed to investigate the bias and variance of the proposed estimators, the coverage probabilities of the CIs, and the size and power of the tests using the net benefit. The GPC approach is implemented in the `R` package `BuyseTest` (Ozenne & Péron, 2019), which can be freely downloaded from `GitHub`. Source code to reproduce the results of the simulations is available as Supporting Information.

Four scenarios were considered:

(i) a scenario under the null hypothesis with $\Delta_0 = 0$ with proportional subdistribution hazards (PH) (or equivalently, sHR = 1),
(ii) a scenario under the alternative hypothesis with PH and sHR = 0.5,
(iii) a scenario under the alternative hypothesis with an early treatment effect (i.e., sHR increases toward 1 over time), and
(iv) a scenario under the alternative hypothesis with a delayed treatment effect (i.e., sHR decreases toward 0 over time).

The data were simulated from improper Gompertz distributions (Jeong & Fine, 2006; Latouche & Porcher, 2007). In brief, the cumulative distribution function of the Gompertz distribution can be written as:

$$F(t;\beta,\nu) = 1 - \exp\left(\frac{\beta}{\nu}(1 - \exp(\nu t))\right).$$

By choosing $\nu < 0$ and $|\beta| < \infty$, one gets an improper distribution with $\lim_{t\to\infty} F(t;\beta,\nu) = 1 - \exp(\frac{\beta}{\nu}) < 1$. The associated subdistribution hazard function is $h(t) = \beta\exp(\nu t)$. Assuming a Gompertz distribution for the times to the event of interest and to the competing event in both groups, with parameters (β_k^g, ν_k^g) for event type k in group g, one has sHR =

TABLE 2 Parameters of the Gompertz distribution used for data simulation

	Group T				Group C			
Scenario	β_1^T	ν_1^T	β_2^T	ν_2^T	β_1^C	ν_1^C	β_2^C	ν_2^C
(i) H_0	0.2396	−0.30	0.1794	−0.30	0.2396	−0.30	0.1794	−0.30
(ii) PH	0.1198	−0.30	0.3333	−0.30	0.2396	−0.30	0.1794	−0.30
(iii) Early effect	0.0958	−0.2315	0.3245	−0.30	0.2396	−0.30	0.1794	−0.30
(iv) Delayed effect	0.0639	−0.1725	0.1173	−0.10	0.0799	−0.10	0.0598	−0.10

$(\beta_1^T \exp(\nu_1^T t))/(\beta_1^C \exp(\nu_1^C t))$. Data under H_0 can then be simulated with $\beta_1^T = \beta_1^C$ and $\nu_1^T = \nu_1^C$, and under H_1 with PH by setting $\nu_1^T = \nu_1^C$. The four scenarios considered in this paper were simulated with the parameters displayed in Table 2. The parameters were chosen to reach a proportion of events of interest of 0.55 in the control group in all scenarios and such that the maximum value of the CIFs of both events sum up to 1 in each group. In scenarios (iii) and (iv), as the PH assumption is not fulfilled, the population parameter depends on time. The parameters of the Gompertz distributions for these two scenarios were chosen so that the sHR as estimated by a proportional subdistribution hazards model in the absence of censoring equals 0.5 (Fine & Gray, 1999). The resulting proportion of events of interest in the treatment group is 0.550, 0.329, 0.339, and 0.309 for scenarios (i) to (iv), respectively. The CIFs for the four scenarios are displayed on Figure A.2.1 in Appendix A.2. The inverse transform method was used to sample the time to events from the CIFs (Devroye, 1986). The true value of the net benefit was 0, 0.233, 0.246, and 0.208 in scenarios (i) to (iv), respectively. Two censoring distributions were considered in each scenario: exponential distribution that is characterized by infinite support, and uniform distribution which prohibits the observation of the tails of $S_1^T(t)$ and $S_1^C(t)$. The parameters of the censoring distributions were selected to reach three different overall proportions of censored observations: 0%, 30%, and 60%. The same censoring distribution was used in both groups. Within each scenario and for each censoring distribution, the sample size was increased across the censoring rate to keep a constant number of events of interest. Three different numbers of events of interest were considered: 20, 66, and 5500. The second number of events corresponds to a power of 80% to detect a sHR of 0.5 with a maximal type I error probability of 5% (Latouche, Porcher, & Chevret, 2004). The settings with 66 events were used to study the power, the type I error probability, and the coverage of CIs. Although the third number of events of interest (i.e., 5500) has little practical implications, it was considered to investigate the properties of our approach with a large sample size. In each of the 72 settings, 1000 data sets were simulated.

The net benefit was estimated for each data set with a null clinical threshold using the uncorrected and corrected Gehan scoring rule and the uncorrected and corrected Péron scoring rule. The bootstrap percentile method was used for construction of CIs and to test the hypothesis $H_0 : \Delta_0 = 0$ (with 1000 resamples). We compared the performance of our approach to a proportional subdistribution hazards model (Fine & Gray, 1999) using the Wald test for hypothesis testing.

4.1 | Bias and variance

Figures 1 to 4 show the net benefit distribution (mean ± standard deviation) computed via each scoring rule, for different amounts of censoring for scenarios (i) to (iv), respectively. Under the null hypothesis (scenario (i), Figure 1), the estimators yielded approximately unbiased net benefit estimates for all sample sizes, censoring distributions, and amounts of censoring. In the presence of censoring, the uncorrected Péron estimator had as expected a larger variance than the uncorrected Gehan estimator, since the former scoring rule attempts at classifying more pairs than the latter. Also, the corrected estimators had larger variances than the uncorrected ones, consistently with the formula of the pairwise correction to compute the corrected weights (see Section 3.2.3).

Under proportional hazards (scenario (ii), Figure 2), the estimators showed approximately no bias in the absence of censoring. Note that the four estimators are equivalent in the absence of censoring and that they consequently yield the same net benefit estimates. With censoring, the various estimators were characterized by a different size of bias. Expectedly, as the censoring rate increased, the uncorrected Gehan estimator became increasingly biased toward 0, due to the increasing proportion of indecisive pairs. In contrast, the corrected Gehan estimator overestimated the net benefit and, as expected, its variance was much larger. Regarding the uncorrected Péron scoring rule, the estimator was increasingly biased toward 0 across the censoring rates, but less biased than the uncorrected Gehan estimator. The performance of the uncorrected Péron estimator depended on the censoring distribution. With an exponential censoring distribution, the estimator was approximately unbiased with a large sample size, even under heavy censoring. However, with extensive

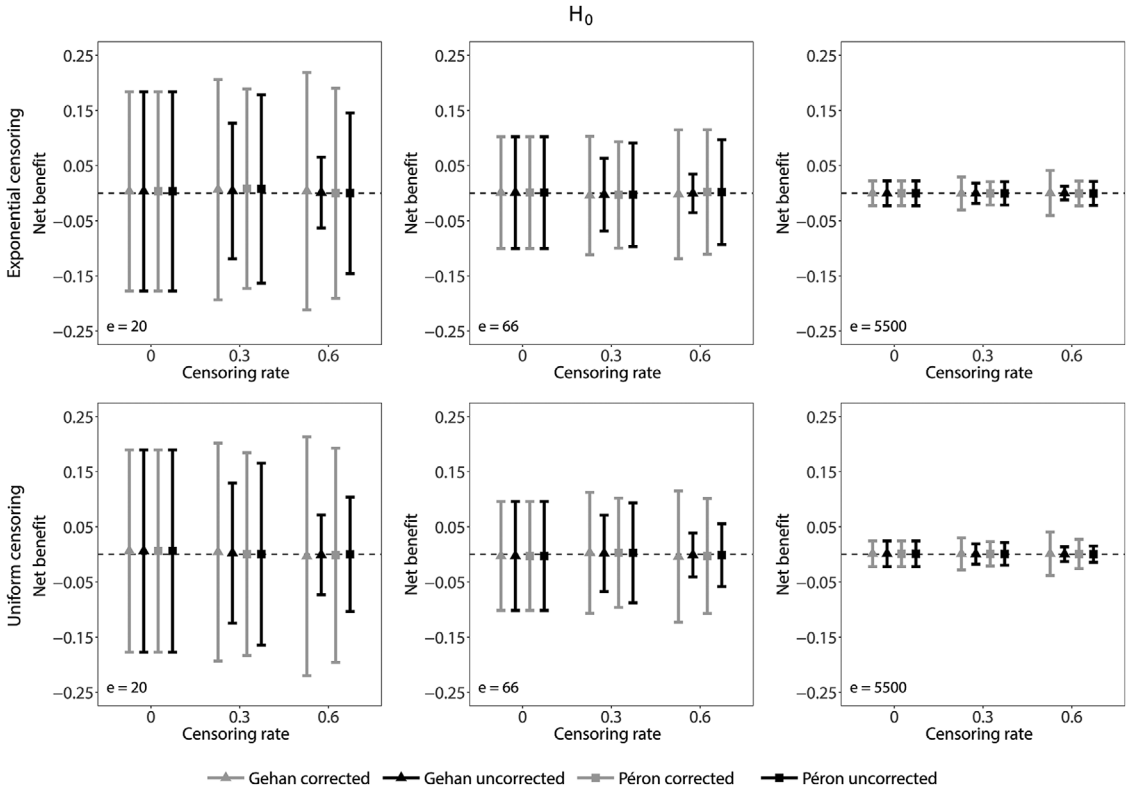


FIGURE 1 Net benefit distribution (mean ± standard deviation) computed via each (un)corrected scoring rule with respect to censoring amount in scenario (i). The dashed line represents the true net benefit and e denotes the number of events of interest

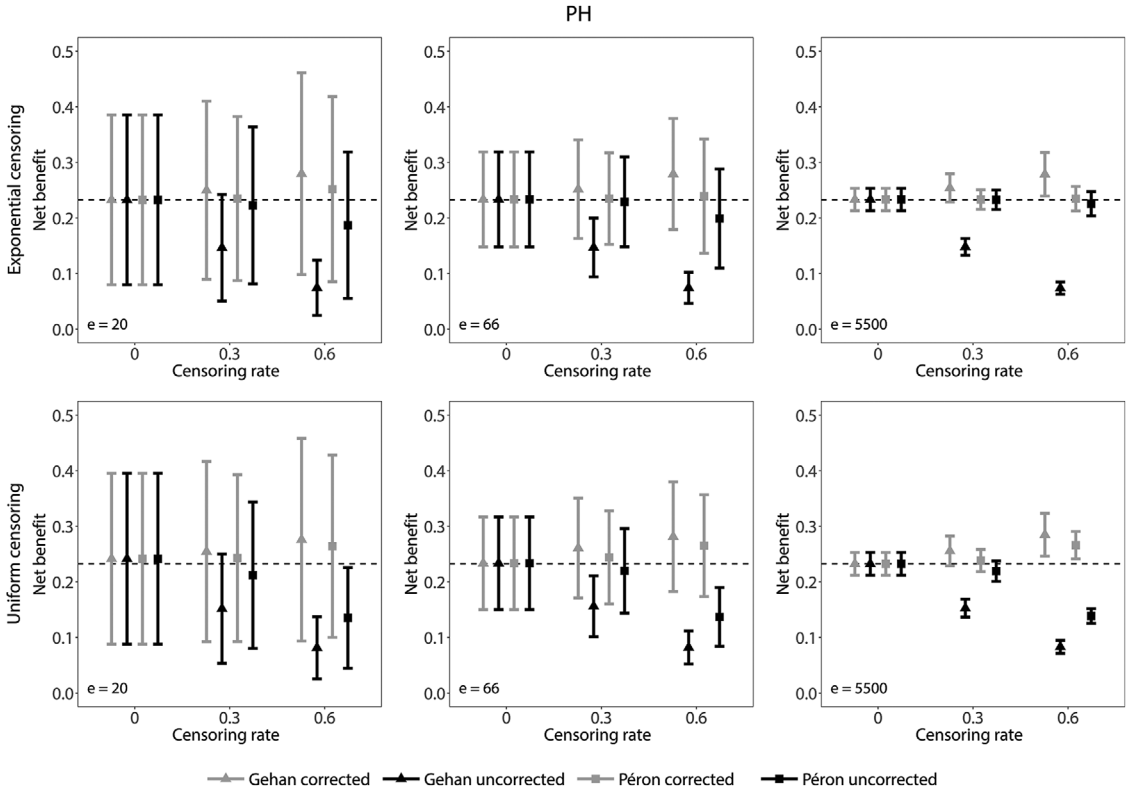


FIGURE 2 Net benefit distribution (mean ± standard deviation) computed via each (un)corrected scoring rule with respect to censoring amount in scenario (ii). The dashed line represents the true net benefit and e denotes the number of events of interest

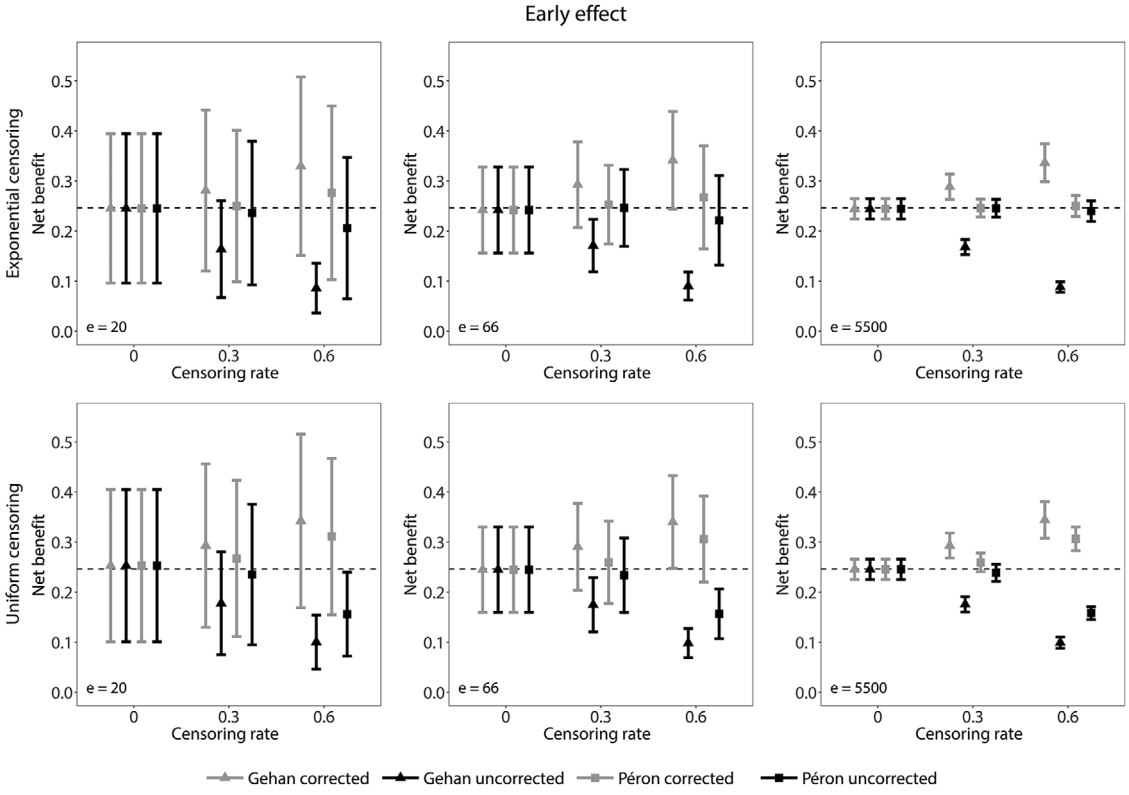


FIGURE 3 Net benefit distribution (mean ± standard deviation) computed via each (un)corrected scoring rule with respect to censoring amount in scenario (iii). The dashed line represents the true net benefit and e denotes the number of events of interest

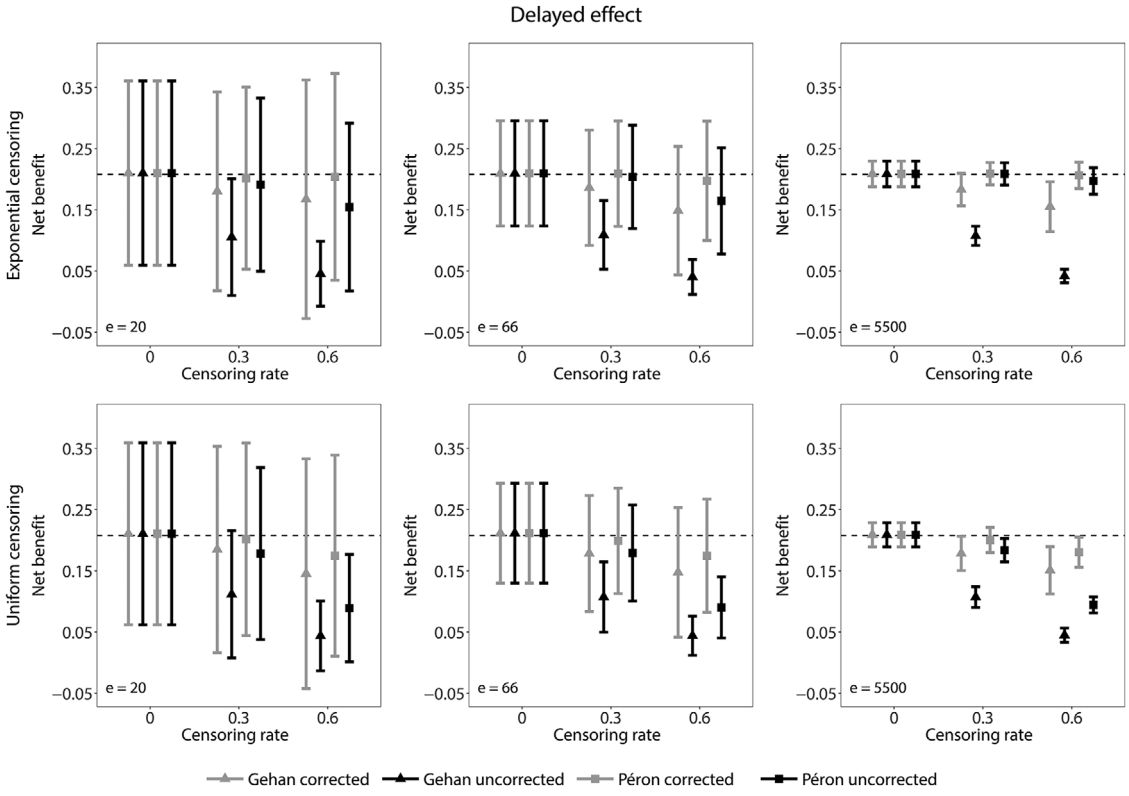


FIGURE 4 Net benefit distribution (mean ± standard deviation) computed via each (un)corrected scoring rule with respect to censoring amount in scenario (iv). The dashed line represents the true net benefit and e denotes the number of events of interest

uniform censoring, the estimator remained biased with a large sample size. This result is related to the support of the censoring distribution. With exponential censoring, the infinite support allows for an accurate estimation of the whole CIFs when increasing the sample size. In contrast, the support of the uniform distribution is limited and individuals with times to events larger than the upper bound of the support of the censoring distribution are censored. In that case, the tails of $S_1^T(t)$ and $S_1^C(t)$ cannot be estimated, even with a large sample size. Nevertheless, the size of the bias of the uncorrected Péron estimator was negligible for small censoring irrespective of the sample size and the censoring distribution. The corrected Péron estimator showed a smaller bias compared to the uncorrected estimator and, again, its variance was larger, especially under extensive censoring.

Also in the case of an early treatment effect (scenario (iii), Figure 3), the estimators were approximately unbiased in the absence of censoring. The influence of the censoring distribution, the proportion of censoring, and the sample size were similar as in scenario (ii). Finally, for the case of a delayed treatment effect (scenario (iv), Figure 4), the results were again similar. However, in contrast to scenarios (ii) and (iii), the corrected Gehan and Péron methods underestimated and not overestimated the net benefit. Nonetheless, the estimator using the Péron scoring rule was again almost unbiased with exponential censoring in the case of a large sample size and in any case less biased than the ones using the Gehan scoring rule.

4.2 | Power and coverage probability

Table 3 displays the coverage probability of the CIs and type I error probabilities in scenario (i), and the coverage probability and power in scenarios (ii) to (iv), respectively, for the settings with a sample size corresponding to 66 events of interest. Under the null hypothesis (scenario (i)), the coverage probability of the CIs was close to 0.95 for all estimators, and similar across the censoring rates for both censoring distributions. The type I error probability was close to 0.05 in most cases.

Under proportional hazards (scenario (ii)), the coverage probability was close to 0.95 for all estimators when the censoring rate was 30% besides the uncorrected Gehan scoring rule, for which the coverage probability dropped significantly at a censoring rate of 30% and reached 0 at a censoring rate of 60%. This is due to the increasing underestimation toward 0 of the net benefit by this scoring rule as the censoring amount increases. The uncorrected Péron estimator performed better under exponential censoring than uniform censoring. This is consistent with the bias results, as the uncorrected Péron estimator showed more bias toward 0 in the case of uniform censoring compared to exponential censoring for large censoring rates. The pairwise correction increased the coverage probability of the estimators, in particular under heavy censoring. Regarding power, it was close to 0.80 for all estimators besides the Péron scoring rule for which it was lower in the exponential censoring case and higher for uniform censoring at a censoring rate of 60%.

In the case of an early treatment effect (scenario (iii)), the CIs based on the uncorrected Gehan scoring rule still exhibited the highest drop in the coverage probability. However, the corrected Gehan scoring rule yielded coverage probabilities substantially lower than 0.95 as well. The coverage probability of the CIs based on the uncorrected Péron scoring rule was approximately correct for a censoring rate of 30%. However, it dropped dramatically under extensive uniform censoring. The pairwise correction improved the coverage probability of the estimator using the Péron scoring rule under heavy censoring. Regarding power, the estimators based on all scoring rules showed a similar power to the test based on the proportional subdistribution hazards model for most settings.

Finally, in the case of a delayed treatment effect (scenario (iv)), the coverage probability of the uncorrected Gehan scoring rule decreased again drastically with the increasing censoring rate. As expected, the coverage of the CIs for the sHR from the proportional subdistribution hazards model was much lower than 0.95 as well, as the proportional subdistribution hazards assumption was violated. The coverage probability for the corrected Péron scoring rule was close to 0.95. The power of the tests based on the Gehan scoring rule was the lowest when censoring was present for both censoring distributions. The tests using the Péron scoring rule showed a higher power compared to the proportional subdistribution hazards model in the presence of censoring.

5 | APPLICATION TO ONCOLOGY DATA

We illustrate our approach using the data from the EORTC 22993-08993 trial on small-cell lung cancer (Slotman, Faivre-Finn, & Kramer, et al., 2007). As brain metastases are common in small-cell lung cancer, the multicenter, phase 3, randomized trial compared prophylactic cranial irradiation (PCI, experimental group) versus no further treatment (control

TABLE 3 Coverage probability and type I error probability/power in simulation study ("PH model" refers to the proportional subdistribution hazards model)

			Coverage probability			Type I error/Power		
			Censoring rate					
Scenario	**Censoring dist.**	**Estimator**	**0**	**0.3**	**0.6**	**0**	**0.3**	**0.6**
(i) H_0	Exponential	Gehan corrected	0.952	0.944	0.955	0.048	0.056	0.045
		Gehan uncorrected	0.951	0.95	0.954	0.049	0.05	0.049
		Péron corrected	0.952	0.953	0.94	0.048	0.047	0.06
		Péron uncorrected	0.948	0.955	0.935	0.052	0.046	0.065
		PH model	0.95	0.956	0.958	0.05	0.044	0.042
	Uniform	Gehan corrected	0.948	0.943	0.94	0.052	0.057	0.06
		Gehan uncorrected	0.952	0.943	0.944	0.048	0.057	0.056
		Péron corrected	0.948	0.94	0.951	0.052	0.06	0.05
		Péron uncorrected	0.95	0.94	0.953	0.051	0.06	0.047
		PH model	0.949	0.94	0.94	0.052	0.061	0.061
(ii) PH	Exponential	Gehan corrected	0.939	0.941	0.915	0.787	0.789	0.774
		Gehan uncorrected	0.945	0.66	0	0.784	0.788	0.772
		Péron corrected	0.948	0.951	0.927	0.791	0.809	0.718
		Péron uncorrected	0.94	0.951	0.892	0.787	0.803	0.717
		PH model	0.956	0.953	0.94	0.798	0.792	0.794
	Uniform	Gehan corrected	0.948	0.929	0.912	0.779	0.812	0.785
		Gehan uncorrected	0.949	0.697	0	0.783	0.812	0.782
		Péron corrected	0.952	0.939	0.911	0.79	0.831	0.84
		Péron uncorrected	0.953	0.944	0.447	0.783	0.825	0.839
		PH model	0.953	0.949	0.954	0.787	0.817	0.789
(iii) Early effect	Exponential	Gehan corrected	0.932	0.903	0.807	0.806	0.914	0.92
		Gehan uncorrected	0.933	0.702	0.001	0.81	0.914	0.92
		Péron corrected	0.931	0.952	0.927	0.812	0.883	0.807
		Péron uncorrected	0.94	0.953	0.904	0.807	0.885	0.812
		PH model	0.945	0.938	0.909	0.786	0.905	0.921
	Uniform	Gehan corrected	0.945	0.907	0.821	0.812	0.893	0.932
		Gehan uncorrected	0.941	0.735	0.003	0.811	0.891	0.928
		Péron corrected	0.948	0.941	0.89	0.807	0.884	0.926
		Péron uncorrected	0.948	0.944	0.48	0.818	0.887	0.924
		PH model	0.948	0.946	0.938	0.788	0.895	0.919
(iv) Delayed effect	Exponential	Gehan corrected	0.939	0.938	0.914	0.717	0.518	0.299
		Gehan uncorrected	0.939	0.546	0	0.715	0.527	0.289
		Péron corrected	0.941	0.938	0.952	0.714	0.697	0.545
		Péron uncorrected	0.947	0.941	0.883	0.711	0.694	0.548
		PH model	0.955	0.907	0.763	0.784	0.605	0.349
	Uniform	Gehan corrected	0.949	0.931	0.91	0.712	0.492	0.277
		Gehan uncorrected	0.953	0.574	0.003	0.711	0.492	0.279
		Péron corrected	0.95	0.947	0.93	0.705	0.648	0.482
		Péron uncorrected	0.95	0.927	0.318	0.715	0.648	0.477
		PH model	0.955	0.884	0.736	0.784	0.555	0.295

TABLE 4 GPC analysis of time to BM. CI stands for confidence interval and p is a two-sided p-value. A null clinically relevant threshold is used. Percentile bootstrap was used with 1000 resamples

Scoring rule	Correction	Win (%)	Loss (%)	Neutral (%)	Indecisive (%)	$\hat{\Delta}$	95% CI	p
Gehan	No	37.09	10.57	31.71	20.63	0.2652	[0.165, 0.369]	<0.0001
Gehan	Yes	46.73	13.32	39.95	0	0.3341	[0.212, 0.451]	<0.0001
Péron	No	41.91	15.71	41.22	1.15	0.2621	[0.148, 0.386]	<0.0001

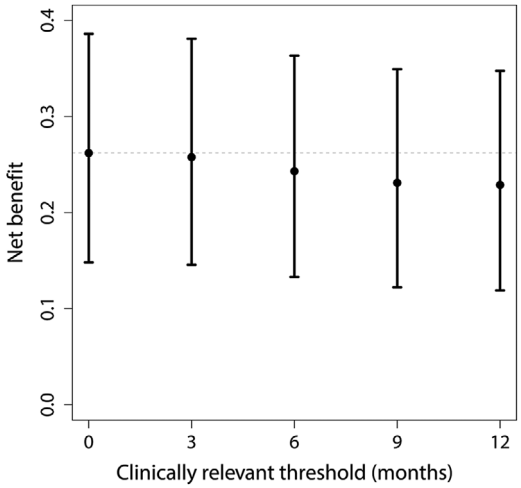


FIGURE 5 Net benefit estimates for different values of clinically relevant threshold for time to BM. The uncorrected Péron scoring rule was used. The dot is the point estimate and the bars represent 95% confidence intervals

group) after response to chemotherapy. In total, 286 patients were enrolled (143 in each group). The primary endpoint was the time to symptomatic brain metastases (BM), and death without evidence of BM was defined as a competing risk. Secondary endpoints included overall and disease-free survival. At the time of analysis, 24 of the 143 patients in the experimental group (16.8%) and 59 of the 143 in the control group (41.3%) had developed BM. The cumulative incidence of BM at 6 and 12 months was 4.4% and 14.6%, respectively, in the irradiation group, and 32% and 40.4%, respectively, in the control group. The cumulative incidences were significantly different between the two groups (Gray's test: $p<0.001$). The trial concluded a reduction of the incidence of BM with PCI as compared to the control (Slotman et al., 2007).

Time to BM was analyzed by the Gehan and Péron scoring rules with a null clinically relevant threshold with and without the pairwise correction (see Table 4). The uncorrected Gehan analysis yielded a positive net benefit of 0.265 (95% CI: [0.165, 0.369]), with a proportion of indecisive pairs around 20%. This value of the net benefit indicates that the difference between the probability to have a larger time to BM with PCI and the probability to have a larger time to BM with the control was equal to 0.265. The application of the pairwise correction increased the estimated net benefit to 0.334 (95% CI: [0.212, 0.451]). This corrected net benefit corresponds to the ratio of the uncorrected net benefit and the proportion of decisive pairs (see Section 3.2.3). Using the Péron scoring rule yielded a net benefit of 0.262 (95% CI: [0.148, 0.386]), and a proportion of indecisive pairs as low as 1.15%. With the amount of censored observations for that endpoint being low (14%), the CIFs could be almost fully estimated (the maximum value of the estimated CIFs summed up to 0.987 and 1 in the control and PCI groups, respectively). The uncorrected Péron scoring rule allowed therefore to reduce significantly the amount of indecisive pairs compared to the uncorrected Gehan scoring rule. These GPC analyses show that the net benefit was significantly different from 0 and suggest a conclusion consistent with that of the trial, that is, that PCI has a beneficial impact on the time to BM in small-cell lung cancer patients as compared to the control.

As a comparison, the sHR from a proportional subdistribution hazards model (Fine & Gray, 1999) was equal to 0.335 (95% CI: [0.212, 0.530]; $p<0.0001$). The proportional subdistribution hazards assumption was however not satisfied (Wald test for the linear change of log(sHR) over time: $p = 0.0007$), making the interpretation of the sHR difficult.

In the GPC framework, sensitivity analyses can be performed to study the impact of the clinical threshold. Figure 5 displays the net benefit estimated using the uncorrected Péron scoring rule with respect to the clinically relevant threshold for time to BM. Up to thresholds as high as 12 months, little differences were observed in the net benefit estimates. This

sensitivity analysis provides the additional information that PCI has a beneficial impact on time to BM as compared to the control even when only differences in the outcome larger than 3, 6, 9, or 12 months are considered as clinically relevant.

Finally, in order to assess the treatment effect on the competing event, we performed a log-rank test to compare the cause-specific hazards of the competing event between the groups. There was no significant difference in the cause-specific hazards of death without BM ($p = 0.305$).

6 | DISCUSSION

In this paper, we extended the GPC approach for time-to-event endpoints to a competing risks setting. The treatment effect was described by the net benefit, defined as the difference between the probability that a patient from one group has a larger time to the event of interest than a patient from the other group and the probability of the opposite situation. In terms of true parameters, we showed that, under proportional subdistribution hazards, the net benefit with a null threshold of clinical relevance can be expressed as a decreasing function of the sHR.

We proposed four estimators of the net benefit based on pairwise comparisons. In the presence of censoring, the estimator assigning score 0 to pairs that could not be classified due to censoring (Gehan scoring rule) showed a large bias and a poor coverage of CIs. The estimator using the CIFs estimates to compute the pairwise score for these pairs (Péron scoring rule) was approximately unbiased with sufficient sample size when the support of the censoring distribution was wide as compared to the support of the time-to-events distributions. In this case, the estimates of the CIFs were available for most of the relevant time interval. In contrast, when the tails of the distributions could not be estimated, the estimator using the Péron scoring rule was biased. The size of the bias was reduced by the application of the pairwise correction. A positive bias was however observed in most of the settings, which may be considered worse than a negative bias when it comes to treatment effect estimation. Also, the variance was expectedly increased upon application of the correction.

We compared the power and size of the tests using the net benefit to a proportional subdistribution hazards model (Fine & Gray, 1999). The tests based on all proposed estimators approximately controlled the type I error probability at the nominal level. In general, the power was comparable to that of the Wald test in the proportional subdistribution hazards model, even under the assumption of proportional subdistribution hazards. One drawback of the latter model linked to this assumption is the fact that, when it is not fulfilled, the true population parameter estimated by the model depends on time. The net benefit does not suffer from this limitation and has a meaningful interpretation even in the case of nonproportional subdistribution hazards. Moreover, it showed with the Péron scoring rule a substantially higher power than the proportional subdistribution hazards model in the presence of censoring when the treatment effect was delayed.

In light of the results of this paper, the estimator based on the Péron scoring rule is preferable to the one using the Gehan scoring rule as it is characterized by a smaller bias and a better coverage of CIs. However, this approach requires that the CIFs can be estimated almost entirely in order to show good performance. When the censoring amount was small and/or when the support of the censoring distribution was sufficiently wide as compared to the support of the time-to-events distributions, the approach showed desirable properties. However, it yielded biased estimates when the CIFs could not be estimated due to the small support of the censoring distribution, which may happen frequently in a clinical trial. A potential alternative to overcome this limitation could be to use a parametric estimate of the tails of the CIFs when reasonable assumptions can be made about them. Another possibility would be to develop a restricted GPC approach, similar to the principle of restricted mean survival times. The analysis would be limited to a period of time for which a consistent estimator of the CIFs is available.

One should keep in mind that the net benefit, similarly to the sHR from a proportional subdistribution hazards model, is affected by the effect of the treatment on the competing event. There is a risk to wrongly conclude that a treatment is beneficial regarding the event of interest when it actually only increases the incidence of the competing event. To avoid this situation, the cause-specific hazards of the competing event can be compared between the two groups. Another possibility could be to estimate the net benefit for the competing event.

Unlike the proportional subdistribution hazards model, our approach does not allow for covariate adjustment. Stratified analysis can be performed to deal with confounding of categorical covariates with a limited number of categories. However, the proposed approach may not be suitable for observational studies, which typically require to adjust for several confounding factors. Its use is therefore more appropriate in randomized clinical trials, in which prognostic factors are equally balanced across treatment groups thanks to randomization.

Nevertheless, as compared to standard competing risks analysis methods, the use of the net benefit to describe treatment differences has a number of advantages. First, it is a fully nonparametric approach that does not rely on the proportional hazards assumption. Second, the population parameter can be interpreted from a clinical perspective. Third, the framework provides a natural way to incorporate thresholds of clinical relevance in the analysis, bringing additional information. Last but not least, treatment effects can be described and formally compared in terms of several prioritized endpoints.

ACKNOWLEDGMENTS

The research leading to these results received funding from the Brussels-Capital Region (Innoviris), and from the Government of Wallonia (Biowin Consortium agreement no. 7979). Eva Cantagallo's fellowship at EORTC was supported by a grant from the EORTC Cancer Research Fund. The authors wish to thank the EORTC Radiation Oncology Group and Lung Cancer Group for sharing their data.

CONFLICT OF INTEREST

Marc Buyse is an employee and stockholder of IDDI. The other authors have declared no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Eva Cantagallo https://orcid.org/0000-0003-0232-6605
Brice Ozenne https://orcid.org/0000-0001-9694-2956

REFERENCES

Aalen, O. O., & Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, *5*, 141–150.

Andersen, P. K., Klein, J. P., & Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, *90*, 15–27.

Buyse, M. (2008). Reformulating the hazard ratio to enhance communication with clinical investigators (Letter). *Clinical Trials*, *5*, 641–642.

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

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, *34*, 187–220.

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer.

Efron, B. (1967). The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *4*, 831–853.

Emura, T., Shih, J-H., Ha, I. D., & Wilke, R. A. (2020). Comparison of the marginal hazard model and the sub-distribution hazard model for competing risks under an assumed copula. *Statistical Methods in Medical Research*, *29*, 2307–2327.

Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, *94*, 496–509.

Gehan, E. A. (1965). A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika*, *52*, 650–653.

Gooley, T. A., Leisenring, W., Crowley, J., & Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. *Statistics in Medicine*, *18*, 695–706.

Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, *16*, 1141–1154.

Jeong, J-H., & Fine, J. (2006). Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society*, *55*, 187–200.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

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

Latouche, A., & Porcher, R. (2007). Sample size calculations in the presence of competing risks. *Statistics in Medicine*, *26*, 5370–5380.

Latouche, A., Porcher, R., & Chevret, S. (2004). Sample size formula for proportional hazards modelling of competing risks. *Statistics in Medicine*, *23*, 3263–3274.

Martinez, E. E., Sinha, D., Wang, W., Lipsitz, S. R., & Chappell, R. J. (2017). Tests for equivalence of two survival functions: Alternative to the tests under proportional hazards. *Statistical Methods in Medical Research*, *26*, 75–87.

McCoy, C. E. (2018). Understanding the use of composite endpoints in clinical trials. *Western Journal of Emergency Medicine*, *19*, 631–634.
Moser, B. K., & McCann, M. H. (2008). Reformulating the hazard ratio to enhance communication with clinical investigators. *Clinical Trials*, *5*, 248–252.
Ozenne, B., & Péron, J. (2019). BuyseTest: Implementation of the generalized pairwise comparisons. R package version 1.7.4.
Péron, J., Buyse, M., Ozenne, B., Roche, L., & Roy, P. (2018). An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Statistical Methods in Medical Research*, *27*, 1230–1239.
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.
Saad, E. D., Zalcberg, J. R., Péron, J., Coart, E., Burzykowski, T., & Buyse, M. (2018). Understanding and communicating measures of treatment effect on survival: Can we do better? *Journal of the National Cancer Institute*, *110*, 232–240.
Scheike, T. H., Zhang, M-J., & Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, *95*, 205–220.
Slotman, B., Faivre-Finn, C., & Kramer, G. et al. (2007). Prophylactic cranial irradiation in extensive small-cell lung cancer. *New England Journal of Medicine*, *357*, 664–672.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Cantagallo E, De Backer M, Kicinski M, Ozenne B, Collette L, Legrand C, Buyse M, Péron J. A new measure of treatment effect in clinical trials involving competing risks based on generalized pairwise comparisons. *Biometrical Journal*. 2020;1–17. https://doi.org/10.1002/bimj.201900354

APPENDIX

A.1 | Relationship between net benefit and subdistribution hazard ratio

We write here

$$\begin{aligned}\Delta_0 &= \mathrm{P}(X_1^T > X_1^C) - \mathrm{P}(X_1^C > X_1^T)\\ &= (A_1) - (A_2).\end{aligned}$$

Starting with the term (A_1), we have that

$$\begin{aligned}(A_1) &= \mathrm{P}(X_1^T > X_1^C, E^C = 1)\\ &= \int_0^\infty S_1^T(t)\,\mathrm{d}F_1^C(t).\end{aligned}$$

Now, if we assume that the subdistribution hazard of the event of interest in both groups are proportional, we have for some finite constant $K > 0$,

$$\frac{h_1^T(t)}{h_1^C(t)} = K, \quad \forall\, t > 0,$$

where $h_1^g(t)$ is the subdistribution hazard function corresponding to X_1^g. We may rewrite then in this context, $\forall\, t > 0$,

$$\begin{aligned}S_1^T(t) &= \exp\left(-KH_1^C(t)\right),\\ \mathrm{d}F_1^C(t) &= h_1^C(t)\exp\left(-H_1^C(t)\right)\mathrm{d}t,\end{aligned}$$

with $H_1^C(t) = \int_0^t h_1^C(s)\,\mathrm{d}s$. We then obtain

$$(A_1) = \int_0^\infty h_1^C(t)\exp\left(-(K+1)H_1^C(t)\right)\mathrm{d}t$$
$$= \frac{1}{K+1}\Big(1 - \lim_{t\to\infty} S_1^T(t)S_1^C(t)\Big).$$

A similar development for (A_2) leads to

$$(A_2) = \frac{K}{K+1}\Big(1 - \lim_{t\to\infty} S_1^T(t)S_1^C(t)\Big),$$

which leads to the desired result.

A.2 | True cumulative incidence functions in simulated scenarios

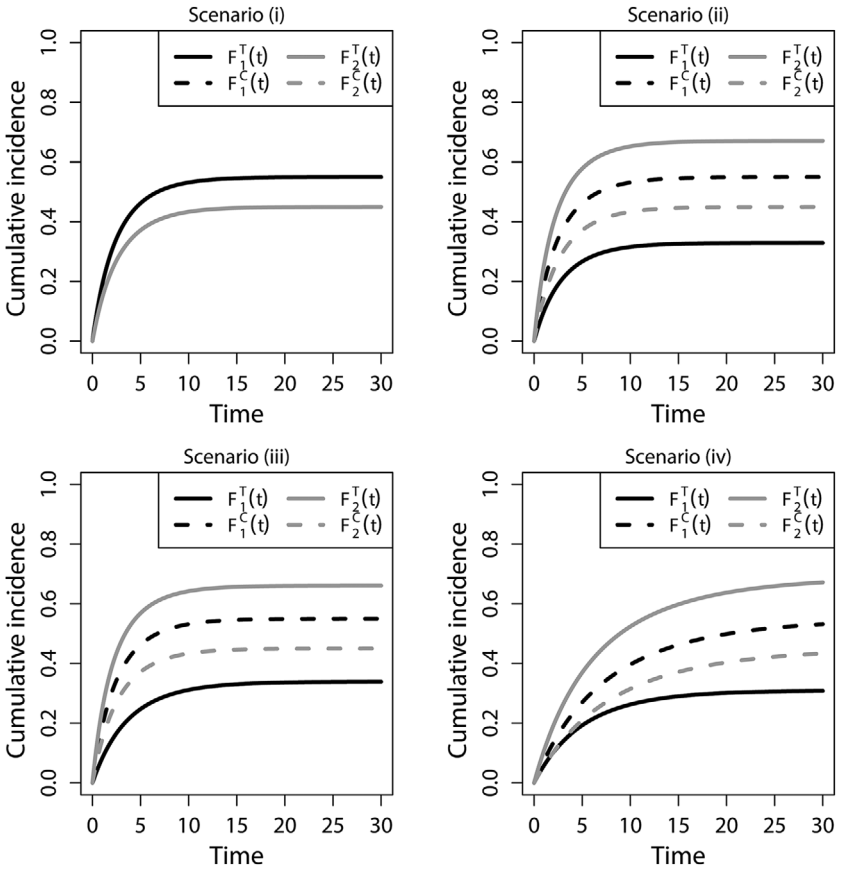


FIGURE A.2.1 True cumulative incidence functions in each scenario. In scenario (i), $F_1^T(t)$ and $F_1^C(t)$ are overlapping, as well as $F_2^T(t)$ and $F_2^C(t)$