UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Evaluating the estimation performance of Generalized Pairwise Comparisons methods for censored survival data

Auteure : Ly Binh An Tran

Promoteurs : Prof. Catherine Legrand et Dr. Mickaël De Backer

Lectrice : Prof. Ingrid Van Keilegom

Année académique 2019-2020

UCL
Université
catholique
de Louvain

Institut de Statistiques, Biostatistiques et Sciences Actuarielles

Evaluating the estimation performance of Generalized Pairwise Comparisons methods for censored survival data

Promotrice:
Prof. Catherine LEGRAND

Co-promoteur:
Dr. Mickaël DE BACKER

Lectrice:
Prof. Ingrid VAN KEILEGOM

Mémoire présenté en vue de l'obtention du
Master en statistique, orientation biostatistique par:

Ly Binh An TRAN

Louvain-La-Neuve
Juin 2020

Acknowledgments

First and foremost, I would like to express my gratitude to my promoter, Prof. Catherine Legrand, for giving me the opportunity to do such an interesting project, and more importantly, for her precious instructions during the research work. Without her kind guidance, this thesis would have taken far longer to complete.

I would especially like to thank Dr. Mickaël De Backer, my co-promotor, for his constructive comments and very careful correction on each of my documents. He enlightened me about statistical methodologies and simulation tool, and was always approachable, affable for all my questions.

I am also grateful to Prof. Ingrid Van Keilegom who served on my Master's committee. I greatly appreciate the valuable time she has taken to review this work.

Contents

Introduction and research objectives

Clinical trials are powerful experimental studies for assessing the effect of medical interventions on health-related outcomes. A proper clinical trial not only requires a good design but also an appropriate statistical analysis. To avoid misleading conclusion and ensure the credibility of the trial, data analysis must be carried out with careful consideration regarding the research purposes and the nature of treatment outcomes. When it comes to a time-to-event outcome, the comparison of treatment effects between two experimental arms mostly relies on the hazard ratio which is often estimated through a Cox proportional hazard model [1]. However, it has been recently argued that this traditional measure is not always suitable and sufficient to report the results of clinical trials, especially when the condition of proportional hazards is not met [2, 3]. The proposals for new measures as alternatives or complements to hazard ratio are thus being very welcomed.

Moreover, results analyzed based on one single endpoint may sometimes fail to convey the proper clinical value of a treatment. For instance, a drug claimed to be efficacious may no longer be recommended if one takes into account its safety profile [4].Therefore, to make use of all clinically important events, there is a growing interest in novel methods which are able to incorporate multiple outcomes in the analysis. Such strategies would provide investigators with different perspectives to better inform clinical practice and decision making.

Among the recently proposed measures of treatment effect, net benefit estimated using Generalized Pairwise Comparisons (GPC) procedure seems to be highly attractive. This measure, defined as the probability for a random patient in the treatment group to have a better outcome than a random patient in the control group minus the opposite probability, not only analyzes time-to-event data but can also be applied to any type of outcomes. More importantly, GPC can easily extend to several prioritized outcomes and include in the analysis a threshold of clinical significance which is the smallest change in treatment outcome so that the treatment can be considered meaningful in practice. This method was first proposed by Buyse [5] (called standard GPC herein) and has demonstrated its usefulness in a database of pancreatic cancer trials [4, 6]. Nevertheless, the technique suffers some drawbacks when dealing with censored data and thus needs to be improved. Currently, four GPC approaches have been suggested to overcome the challenges left by the standard technique: (1) Peron original method (PO method); (2) Peron method with

correction (PC method), (3) method based on the Extreme Value Theory (EVT method); and (4) Restricted net benefit method (RNB method). Considering that each approach has its own limitation and may not applicable to all situations, this Master's thesis has been conducted to compare these new estimators (mainly between PO, PC and EVT) and to identify which method would be the most appropriate in certain cases. To achieve this goal, we will perform an extensive simulation study to examine the performance of net benefit estimators in various scenarios. It should be noted that even though GPC can handle several outcomes simultaneously, within this work, we will restrict ourselves to a single time-to-event variable with emphasis on the impact of censored data on the bias of estimates. We will also consider the variance of estimators as an aspect to evaluate their quality.

In order to respond to the research question above, this thesis is structured in five chapters. In the first chapter, we give a brief introduction about clinical trials and survival analysis, focusing towards the measures of treatment effect for time-to-event outcome. Chapter 2 describes the standard GPC and highlights its shortcomings which need to be addressed by the new GPC procedures. Chapter 3 summarizes the methodology and the reported results of four new GPCs. By discussing the limitations of these approaches, we can clearly point out several scenarios that would be interesting to investigate in the simulation study. The implementation and results of the simulation will be presented in Chapter 4, followed by the general conclusions and discussion in Chapter 5.

Chapter 1

Clinical trials and survival analysis

1.1 Overview of clinical trials

According to the definition by US National Institutes of Health's clinical trials [7], clinical trials are the research studies which are conducted in human volunteers (healthy participants or patients with a specific health issue) and are intended to test new medical products such as drugs or devices, new medical procedures, or new diagnostic tests. These new medical approaches may be compared to a standard one that is already available, or to a placebo that contains no active ingredients, or to no intervention. To ensure the complete safety of volunteers and protection of human rights, clinical trials are highly monitored and regulated through the standard ethical guidelines, codes of conduct and national laws. Before an experimental treatment can be considered for testing in humans, it must undergo intense laboratory testing and animal research. Once the safety and potential effectiveness of the treatment have passed preclinical studies and been approved for human research, clinical trials will be carefully designed and typically proceed through four phases, each phase has a different purpose [8, 9]:

- **Phase I**: Experimental treatment is tested for the first time in human to evaluate safety, to identify side effects, as well as to learn about optimal dosage range. At a particular tested dose, there are typically three enrolled participants. If no unacceptable toxicity is observed in any participant, one can move to the next higher dose. If two or three participants experience the toxicity, the study is terminated at this dose. If the toxicity occurs in one participant, the study is repeated at the same dose, usually with another three participants. If no toxicity is further seen in the additional individuals, the dose is escalated. Otherwise, the study is terminated, the current dose, or perhaps the previous dose, is declared to be the maximum tolerated dose. Generally, 5-8 doses will be tested and fewer than 30 participants are included in phase I of trial.

- **Phase II**: Experiment is conducted in a larger group of people (from 50 to 200) to determine the effectiveness of treatment and to further study its safety. Phase II is

commonly designed into two steps. The first step (phase IIA) is to check whether the treatment has no or little therapeutic effect. If the estimated effect does not exceed the minimal specified activity, the treatment will not be further considered. Otherwise, more participants would be added to better estimate the response rate (phase IIB). The number of additional participants would depend on the desired precision of estimation.

- **Phase III**: Experimental treatment is given to a large group of people (from several hundred to several thousand) and is compared to a standard treatment, or a best supportive care, or a placebo. The purpose is to confirm effectiveness, to monitor adverse effects and to collect information for the safe use of new treatment.

- **Phase IV**: Studies are conducted after the new treatment has been marketed to track its safety in the general population and to get more information about benefits and optimal use.

Typically, it takes around 10 to 15 years for a new treatment to complete clinical trials and receive market approval. In clinical trials, safety and efficacy of treatment or intervention are determined by measuring several outcomes in participants. Outcome variables can be discrete (e.g. cure or no cure, infection or no infection...), continuous which are accessed on a numeric scale (e.g. blood pressure, quality of life scores...) or time-to-event. The latter possesses unique characteristics since here, one is not only interested in whether or not an event occurred, but also when that event occurred. This kind of information is not always fully available (i.e. presence of censoring), which creates some special challenges when analyzing time-to-event data. As time-to-event is of interest in this work, more description about this type of outcome will be presented in the next section.

1.2 Survival analysis

1.2.1 Time-to-event outcome

Time-to-event outcome is the variable of interest in the statistical procedures that are generally called survival analysis. Time-to-event data arise when the focus of research is time until an event occurs [10, 11]. They are commonly referred as survival data since death is often the event of interest, notably in oncology and cardiology. But they can be any designated experience that may happen to an individual such as tumor remission, myocardial infarction, discharge from hospital... In most clinical trials, investigators may not be able to obtain the fully observed data as some subjects may not experience the event during the follow-up time. The data recorded for these subjects are called "censored", their survival time is only known to occur within a certain period of time. If the event occurs beyond the study, then the data is right-censored. If the event has already occurred before the study, the data is left-censored. And data are interval-censored if the event is known to fall in an interval but not exactly when. Right-censored data are the most frequently

observed in clinical trials; and within this thesis, from now on, by mentioning "censored data", we imply the right censoring.

With right-censored data, one knows that the patient stays alive at least as long as the period C that he/she has been followed (i.e. censored time), but if the relevant outcome occurs after that, one can not obtain the complete survival time T. T is only fully determined if the patient experienced the event of interest during his/her follow-up time. Therefore, to express time-to-event data, one would need a pair of random variables (T', θ) for each individual:

- $T' = \min(T, C)$: length of time during which the event has not been observed.
- $\theta = \mathbb{I}(T < C)$: indicator of whether the end of the above time period corresponds to an event occurrence or a censorship.

The censored issue may emerge due to the three following reasons: (1) patient has not yet experienced the event by the close of the study; (2) patient is lost during the follow-up; (3) patient withdraws from the study because of different event. In the first situation, the study is terminated at a prespecified time point before all the patients experience the event. These censored times, called Type I censoring, may be equal among the participants if they all enter the study on common date. If patients are enrolled in staggered manner, the administrative censored times will be different from individual to individual. Figure 1.1 illustrates the observed and censored times for four individuals. Survival time for patient 2 is completely observed; whereas, only the lower bound of the survival time is known for the three remaining patients. Unlike patient 1 who is lost before the end of the study, patient 3 and 4 are censored because of the study termination. Also, the type I censored time of patient 3 is longer than that of patient 4 as the former entered the study before the latter.

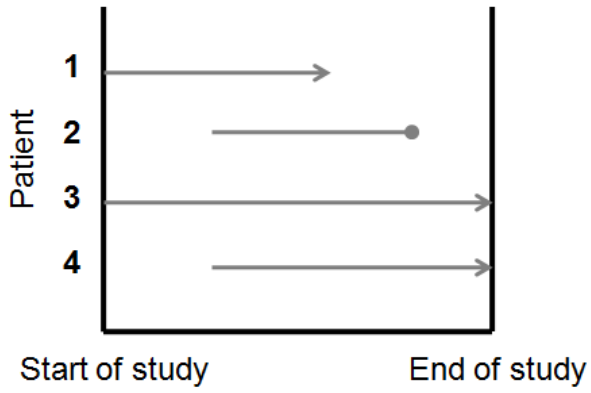

Figure 1.1: Illustration of observed and censored data of a study where patients are enrolled in staggered manner. Gray dot represents the fully observed event. Arrow represents the censored observation.

In most analytical approaches, censored time C is commonly assumed to be independent of survival time T as the relation between them are not identifiable in general. This condition is satisfied if censoring occurs due to administration or other reasons that are totally unrelated to the event of interest. In the cases where patients withdraw from the study because they have suffered side effects or because their health condition has improved, the independence of T and C may be in doubt. The assumption of independent censoring is also applied to the GPC methods presented in this work.

1.2.2 Key functions

Let T_i for $i = 1, ..., n$ be iid samples of the non-negative continuous random variable $T \sim F$. T represents the time until the event of interest. To characterize the behavior of T, the three functions below are mostly considered [10, 11]:

- **Survival function** $S(t)$: Defined as the probability that a random individual will survive beyond time t:

  $$S(t) = \mathbb{P}(T > t).$$

  S(t) is a continuous and strictly decreasing function. It is equal to 1 at $t = 0$ and tends to 0 as the time approaches infinity.

- **Hazard function** $\lambda(t)$: Defined as:

  $$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{1.1}$$

  Hazard function measures the "instantaneous risk" of occurrence of an event right after time t given that the individual has survived up to time t. It is a positive function which is not necessarily increasing or decreasing, and has no upper bound. Furthermore, one has:

  $$\begin{aligned}\mathbb{P}(t \leq T < t + \Delta t | T \geq t) &= \frac{\mathbb{P}(t \leq T < t + \Delta t, T \geq t)}{\mathbb{P}(T \geq t)} \\ &= \frac{\mathbb{P}(t \leq T < t + \Delta t)}{S(t)} \\ &= \frac{F(t + \Delta t) - F(t)}{S(t)}.\end{aligned}$$

  So,

  $$\begin{aligned}\lambda(t) &= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\ &= \frac{1}{S(t)} \frac{\mathrm{d}}{\mathrm{d}t} F(t) \\ &= -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t).\end{aligned} \tag{1.2}$$

- **Cumulative hazard function** $\Lambda(t)$: Defined as the integral of the hazard function from time 0 to time t:

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s. \tag{1.3}$$

  $\Lambda(t)$ describes the "total exposure risk" up to time t for a survivor. It is an increasing function taking the values from 0 to infinity. Moreover, from the formula (1.2) and (1.3), one can get: $\Lambda(t) = -\log S(t)$.

Clearly, there is a relationship between the survival function, hazard function and cumulative hazard function. If one knows one of these functions, one can derive the corresponding and vice versa (Table 1.1)

Table 1.1: Relationship between survival function, hazard and cumulative hazard

	$S(t)$	$\lambda(t)$	$\Lambda(t)$
$S(t)$		$-\frac{\mathrm{d}}{\mathrm{d}t}\log S(t)$	$-\log S(t)$
$\lambda(t)$	$\exp[-\int_0^t \lambda(s)\mathrm{d}s]$		$\int_0^t \lambda(s)\mathrm{d}s$
$\Lambda(t)$	$\exp[-\Lambda(t)]$	$\frac{\mathrm{d}}{\mathrm{d}t}\Lambda(t)$	

1.2.3 Kaplan-Meier estimation

Kaplan-Meier (KM) is the most common method for non-parametrically estimating the survival function. The formula to calculate KM estimator $\widehat{S}(t)$ is given as below [12]:

$$\widehat{S}(t) = \prod_{t_{(i)} \leq t} \frac{r_{(i)} - d_{(i)}}{r_{(i)}}, \tag{1.4}$$

where $t_{(i)}$ is the duration of the study up to the i^{th} ordered event time, $d_{(i)}$ is the number of events occurring at time $t_{(i)}$, and $r_{(i)}$ is the number of individuals at risk just prior to $t_{(i)}$. Note that censored observations are not counted as "at risk". The idea behind KM procedure is that the estimation of $S(t)$ can be divided into a series of intervals based on the ordered event times, and the probability of surviving at the end of a time interval depends conditionally on the probability of surviving at the start of this time interval.

That is:

$$\begin{aligned} S(t_{(i)}) &= \mathbb{P}(T > t_{(i)}) \\ &= \mathbb{P}(T > t_{(i)},\, T > t_{(i-1)}) \\ &= \mathbb{P}(T > t_{(i-1)})\mathbb{P}(T > t_{(i)} | T > t_{(i-1)}) \\ &= S(t_{(i-1)})\mathbb{P}(T > t_{(i)} | T \geq t_{(i)}). \end{aligned}$$

The conditional probability $\mathbb{P}(T > t_{(i)} | T \geq t_{(i)})$ can be estimated as the fraction of individuals who are at risk at time $t_{(i)}$ but do not die at this time:

$$\widehat{\mathbb{P}}(T > t_{(i)} | T \geq t_{(i)}) = \frac{r_{(i)} - d_{(i)}}{r_{(i)}}.$$

Using the recursive expansion, one can then construct the estimate $\widehat{S}(t)$ of the survival function as displayed in the formula (1.4). Intuitively, there are two important remarks which can be deduced from this formula. First, KM estimator is a step function with the jumps corresponding to the moment when one moves from an observed event time to another. The size of the jumps not only depends on the number of events $d_{(i)}$ but also on the censoring pattern as the censored individuals prior to time $t_{(i)}$ are excluded from the risk set $r_{(i)}$. Second, $\widehat{S}(t)$ is well determined up to the largest event time. If the largest observation corresponds to an occurrence of event, $\widehat{S}(t)$ will attain 0 at this time point. If the largest observation is censored, it is impossible to estimate $S(t)$ beyond this because one does not know until when the last survivor would experience the event.

1.2.4 Hazard ratio and proportionality assumption

In order to express the treatment effect for time-to-event outcome, hazard ratio (HR) is the most commonly used and may be nearly universal measure when comparing the survival of two groups [1]. HR is defined as the hazard for one individual divided by the hazard for a different individual, and can be estimated using the Cox proportional hazards model that is written as follows [13]:

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) \exp(\beta^T \mathbf{Z})$$

where $\lambda_0(t)$ is the baseline hazard function, β is the parameter vector and $\mathbf{Z}$ is the covariate vector. Cox model can be estimated using a semi-parametric approach; concretely, it is still possible to estimate the exponential term (i.e. parametric component) without specifying the baseline hazard (i.e. non-parametric component). It means that one does not need to assume about the survival time distribution to make inference about HR. This appealing feature is one of the reasons why Cox model is very popular in medical research. Nevertheless, the use and interpretation of Cox model require a vital condition, called proportional hazards (PH) assumption. To understand this, let us consider two individuals with covariate $\mathbf{Z_1}$ and $\mathbf{Z_2}$, we then have:

$$\text{HR} = \frac{\lambda(t|\mathbf{Z_1})}{\lambda(t|\mathbf{Z_2})} = \frac{\lambda_0(t)\exp(\beta^T\mathbf{Z_1})}{\lambda_0(t)\exp(\beta^T\mathbf{Z_2})} = \frac{\exp(\beta^T\mathbf{Z_1})}{\exp(\beta^T\mathbf{Z_2})}$$

As can be seen in the above equation, the baseline hazard $\lambda_0(t)$ is canceled out in the final expression. Thus, HR does not involve time t. It implies that hazard functions for any two individuals are proportional at all time points; or equivalently, HR is constant over time even though hazards themselves may change continuously. Given the importance of PH assumption, numerous methods, both graphical and test-based, have been suggested to verify its validity [11]. This will be not discussed here as it goes beyond the scope of the thesis.

Under the PH assumption, HR can be interpreted in a similar way as a relative risk, it indicates how many times more (or less) a patient would likely experience the event if he/she receives the experimental treatment rather than the control one. If the PH assumption does not hold, the explanation for treatment difference using HR would be more complicated or less straightforward. On the other hand, even when the PH is plausible, the relative difference between two survival curves reflected by HR is still not a completely satisfactory measure due to the lack of absolute component. Indeed, considering the mathematical formula to calculate hazard (equation (1.1)), the absolute values of hazard themselves may be very difficult to clinically interpret. This fact would limit the translation of hazard difference into an interesting and transparent treatment benefit.

1.2.5 Other measures of treatment effect

Given the undesirable features of HR cited in the previous section, there is a recent encouragement to accommodate other measures as alternatives or complements to the conventional HR [3]. Several proposed measures for evaluating the treatment effect on time-to-event data can be listed as below; some of them are the absolute measures which compute the actual difference between groups and the others are the relative ones which describe the results based on the ratio:

- **Difference in survival probabilities at a single or a few points in time** [14]: A simplest way for quantifying the group difference is comparing the survival probabilities at a given time point t. Such measure can be easily obtained from KM estimators but the treatment effect is clearly influenced by the choice of t. Two survival curves could be judged significantly different at time t but not at other points. This issue would become more problematic when the hazard ratio is not proportional. Therefore, this kind of comparison is not recommended unless a few clinically relevant points can be justified as prior to the analysis.

- **Difference in median survival times** [15]: Another way to assess treatment effect is comparing the time that corresponds to the 50% survival probability. The median survival time is easily computed using KM estimation and intuitive for clinicians, but

it is only estimable when the study is long enough so that one half of the participants had the event. Of course, one can choose a lower percentile if the survival does not drops below 50%. However, this measure is based on a single point and thus unable to take the entire survival curve into account; except for the special case where the survival time follows an exponential distribution. In the latter situation, the median survival time is proportional to the inverse of hazard; hence, the ratio of the median survival times can be used to estimate the hazard ratio.

- **Difference between restricted mean survival times** [2, 16]: Mean survival time μ, which is the expected time that patients are still alive from the start of study, can be estimated non-parametrically as the area under the survival curve $S(t)$; however, the computation is only feasible if the survival curve reaches 0. If it is not the case, one must restrict the analysis up to a chosen time point ϵ, called the restricted mean survival time μ_ϵ:

$$\mu_\epsilon = E[\min(T, \epsilon)] = \int_0^\epsilon S(s)\mathrm{d}s.$$

  The difference in restricted mean survival times between the treatment and the control group is then the area between the two survival curves. This measure provides the mean gain or the mean loss in the survival time of the treatment group as compared to the control within a window time ϵ. Such interpretation is still valid whether hazards are proportional or not; however, it depends on the choice of the restricted time ϵ which may be quite arbitrary and tricky, especially when two survival curves cross each other. To make use of all the available information, ϵ can be set to the minimum time of the largest observed event in the comparison groups; or it can be prespecified at the stage of study design based on the clinical relevance. Furthermore, the interpretation may be criticized if the survival probabilities at ϵ are too far from 0.

- **Ratio of restricted mean survival times** [17]: One can also evaluate the treatment effect in two groups by computing the ratio of their restricted mean survival times; i.e. the ratio of the area under the survival curves up to a given time point ϵ. This measure shares the same advantages and disadvantages as the difference in restricted mean survival times.

- **Difference between unrestricted mean survival times** [3]: An alternative approach to estimate the mean survival time when the estimated survival curve does not reach 0 is to make an assumption about the distribution of survival function. This strategy can take the entire survival curve into account and does not require the PH assumption either. The key issue here is to select a suitable parametric model and to check its fit to the data. If the data are too immature (i.e. the survival probability is far from 0 at the largest follow-up time), it would be difficult to choose the appropriate distribution and the estimation for model parameters would be less precise.

- **Net benefit**: This measure is estimated by GPC procedure, and is defined as the probability for a random patient in the treatment group to have a better outcome than a random patient in the control group minus the opposite probability [5]. All the relevant methods analyzing net benefit, which is the main focus of the work in this thesis, will be further discussed in the next chapters.

- **Win ratio**: This measure was first proposed by Pocock *et al.* to analyze multiple outcomes in cardiovascular diseases by accounting for clinical priorities [18]. It summarizes the ratio of number of patients who fared better versus worse on the experimental group. Hence, the idea of Win ratio is very similar to net benefit but it displays the treatment effect as a relative difference. Recently, Oakes *et al.* revealed that Win ratio estimate could be affected by the follow-up time [19] and how to tackle this issue is still under research [20].

Among the absolute measures, net benefit appears to be highly appealing in the sense that it is able to combine several outcomes of different nature. Indeed, when evaluating a new drug, not only efficacy but safety is also the main concern driving the decision of adopting a treatment. However, these two outcomes are often reported independently, which makes it hard to access benefit and risk on a comparable scale. Hence, if it is possible to make a direct link of all the important events, clinicians would have sounder information to rationally make decision in practice.

Chapter 2

Standard Generalized Pairwise Comparisons

2.1 Overview

In order to compare two groups of observations, many parametric and non-parametric tests have been constructed for a single variable or for several variables in multivariate analysis [10, 21]. However, none of these available tests are designed to tackle the situation where the comparison is performed based on several variables of different types or a variable measuring different occasions. Therefore, Buyse has extended the idea behind the Wilcoxon-Mann-Whitney test to carry out the pairwise comparisons in a more general setting [5].

Briefly, pairwise comparisons require to compare all possible pairs of individuals, one taken from the treatment group (group Trt) and the other taken from the control group (group Ctr). Let assume the outcomes in group Trt $(X_1, \ldots, X_n)$ be iid copies of X, the outcomes in group Ctr $(Y_1, \ldots, Y_m)$ be iid copies of Y, and X and Y being independent. Then, one will have totally $n \times m$ pairs to compare. Pair is classified as "favorable" or "Win" if the outcome X_i of the i^{th} individual $(i = 1, ..., n)$ in group Trt is better than the outcome Y_j of the j^{th} individual $(j = 1, ..., m)$ in group Ctr, "unfavorable" or "Loss" if the outcome of the individual in group Trt is worse than that of the individual in group Ctr, "neutral" if there is no difference between the outcomes of two individuals, or "uninformative" if it is impossible to determine which one has better outcome (e.g. due to missing data or censored observation). The net benefit, denoted Δ, which is defined as the probability for a random patient in the treatment group to have a better outcome than a random patient in the control group minus the opposite probability, is then given by:

$$\Delta = \mathbb{P}(X > Y + \tau) - \mathbb{P}(Y > X + \tau),$$

where τ is the threshold of clinical relevance. This threshold is crucial to ensure that a treatment would work in practice. For instance, a drug may statistically significantly

reduce the heart rate by average of 2 or 3 beats per minute but it would not be clinically significant as the effect is not large enough to improve the life quality of patients. The rule to define "better outcome" depends on the type of variables and is presented as follows:

- **Binary variable**: Table 2.1 displays the possible situations when comparing a binary outcome with $X_i = 1$ (or $Y_j = 1$) indicating success, and $X_i = 0$ (or $Y_j = 0$) indicating failure.

- **Continuous variable**: Assuming that larger values of X (and Y) are preferable to their smaller values when evaluating the treatment effect, and the effect would be considered meaningful only if the difference between the values of these two variables exceeds the clinically relevant threshold τ. The status of a pair can be decided as shown in Table 2.2.

- **Time-to-event variable**: As mentioned in Section 1.2.1, the true survival times X and Y are not observed for all individuals in the case of right censoring. The observable times are the minimum between the event times and the censored times (denoted X_i^c and Y_j^c). That is, the observed samples are:

$$\begin{cases} (X_i', \theta_i) & \text{where } X_i' = \min(X_i, X_i^c) \text{ and } \theta_i = \mathbb{I}(X_i < X_i^c). \\ (Y_j', \eta_j) & \text{where } Y_j' = \min(Y_j, Y_j^c) \text{ and } \eta_j = \mathbb{I}(Y_j < Y_j^c). \end{cases}$$

  Table 2.3 shows all possible situations when comparing time-to-event outcomes with τ as the threshold of clinical relevance. As compared to other types of variables, one has a new situation, "uninformative", besides the favorable, unfavorable and neutral categories. These "uninformative" pairs are the critical part of GPCs for time-to-event data and will be developed more in the following sections.

Table 2.1: Generalized pairwise comparisons for binary variable

Pairwise comparison	*Pair is*
$X_i = 1$, $Y_j = 0$	Favorable
$X_i = 1$, $Y_j = 1$	Neutral
$X_i = 0$, $Y_j = 0$	Neutral
$X_i = 0$, $Y_j = 1$	Unfavorable

Table 2.2: Generalized pairwise comparisons for continuous variable

Pairwise comparison	*Pair is*
$X_i - Y_j > \tau$	Favorable
$\lvert X_i - Y_j \rvert < \tau$	Neutral
$X_i - Y_j < -\tau$	Unfavorable

Table 2.3: Generalized pairwise comparisons for time-to-event variable

(θ_i, η_j)	$X'_i - Y'_j > \tau$	$X'_i - Y'_j < -\tau$	$\lvert X'_i - Y'_j \rvert < \tau$
(1, 1)	Favorable	Unfavorable	Neutral
(0, 1)	Favorable	Uninformative	Uninformative
(1, 0)	Uninformative	Uninformative	Unfavorable
(0, 0)	Uninformative	Uninformative	Uninformative

Importantly, the idea of net benefit can be easily extend to several outcomes, providing that the outcomes are prioritized and handled in descending order of priority. The highest priority is assigned to the variable considered as the most clinically relevant. Outcome of lower priority is taken into account only when a pair is uninformative or neutral for outcomes of higher priority (Table 2.4).

Table 2.4: Status of a pair defined by GPC when considering two prioritized outcomes

Outcome with higher priority	*Outcome with lower priority*	*Pair is*
Favorable	Ignored	Favorable
Unfavorable	Ignored	Unfavorable
Uninformative/neutral	Favorable	Favorable
Uninformative/neutral	Unfavorable	Unfavorable
Uninformative/neutral	Uninformative/neutral	Uninformative/neutral

Furthermore, prioritized outcomes can be variables of different nature. It means that time-to-event can be combined with binary variables such as toxicities, or with continuous variables such as quality of life scores. On the other hand, several outcomes can be defined as successive thresholds of a single treatment measure. For instance, in ophthalmology, visual acuity is evaluated based on the number of letters correctly read on a standardized visual acuity chart. Prioritized outcomes in GPC could be then the difference in the number of letters before and after treatment, which may varies from 30 letters (corresponding to high improvement in vision) to less than 5 letters (corresponding to minor improvement) [5]. It is also possible to apply GPC for repeated observations. In longitudinal study, for example, a later measure between the groups is usually more relevant than an earlier one as it represents a sustained treatment effect over time. The later difference would thus be assigned to higher priority in GPC.

2.2 Estimation and testing

For the l^{th} outcome measure ($l = 1, ..., L$) numbered from 1 (highest priority level) to L (lowest priority level), there are two indicators associated with a couple of observation (i, j): the scoring indicator s_{ijl} defined as:

$$s_{ijl} = \begin{cases} 1 & \text{if the } l^{th} \text{ outcome is favorable ,} \\ -1 & \text{if the } l^{th} \text{ outcome is unfavorable ,} \\ 0 & \text{if the } l^{th} \text{ outcome is neutral or uninformative ,} \end{cases}$$

and the weight w_{ijl}. This weight is naturally initialized to 1 for each pair and for the highest priority outcome. With $l > 1$, w_{ijl} is set to 0 if the pair is classified as favorable or unfavorable for the higher priority outcome, and to 1 if the pair is uninformative or neutral for the higher priority. The net benefit Δ can be then estimated as the net difference between the proportion of pairs in favor of treatment and not in favor of treatment:

$$\widehat{\Delta} = \frac{\sum_{l=1}^{L} \sum_{i=1}^{n} \sum_{j=1}^{m} s_{ijl}.w_{ijl}}{n.m}. \tag{2.1}$$

A randomization test can be used to test the null hypothesis H_0: $\Delta = 0$. As a remark, randomization test is considered as a permutation test which is based on random assignment (not random sampling) [22]. To perform the randomization test, a statistic that is relevant to the purpose of research is first defined and computed for the experimental data. Next, the data are rearranged in a manner consistent with the random assignment which has been used to conduct the actual experiment (i.e. keeping all experimental design features, except the groups). For instance, if n subjects were randomly assigned to each of two experimental groups, there would be $(2n)!/n!n!$ different ways to re-allocate these subjects to the two groups. The test statistic is calculated for each of these different assignments, and so a distribution of possible statistics could be generated. Finally, the statistical significance of the observed statistic can be determined by examining its position within the above distribution. Therefore, the idea behind the randomization test is that if the null hypothesis of no difference between the groups is true, the test statistic should be mostly uninfluenced by the different assignments. This test does not require any assumption regarding the density or mass probability functions and is thus distribution-free. However, it is only valid under the condition that subjects are randomly allocated to the experimental groups or data are exchangeable between groups. The latter condition is to ensure that under the null hypothesis, the joint distribution of observations is invariant for all the permuted samples [23]. In other words, it is only the observations that matter and not the order in which they are considered. Intuitively, in clinical research, participants in the experimental arms are exchangeable if the difference in outcome is attributed to the assigned treatments and not to the difference in characteristics of groups. For instance, the exchangeability assumption may be doubtful if age of participants, which is probably a factor influencing treatment response, is not comparable among the groups.

Returning to the GPC testing procedure, a confidence interval for Δ and the significance probability (p-value) can be calculated based on the empirical distribution of $\widehat{\Delta}$ under H_0. Let Δ_{obs} and Δ_k ($k = 1, ..., K$ with K being the number of random assignments) be the net benefit calculated for the observed and permuted samples, respectively. $\Delta_{\alpha/2}$ is the value that leaves at most $\alpha/2(\%)$ values of Δ_k to its left, and $\Delta_{1-\alpha/2}$ be the

value that leaves at most $\alpha/2(\%)$ values of Δ_k to its right. The $(1-\alpha)$ confidence interval under H_0 for Δ is given by $[\Delta_{obs} + \Delta_{\alpha/2}; \Delta_{obs} + \Delta_{1-\alpha/2}]$. Let k_1 be the number of Δ_k for which $\Delta_k \geq \Delta_{obs}$, and k_2 be the number of Δ_k for which $|\Delta_k| \geq |\Delta_{obs}|$. The p-value is calculated as k_1/K for a one-sided test, and k_2/K for a two-sided test. Other approaches such as bootstrap resampling or the asymptotic distribution can also be used to make the inference about Δ [24].

Additionally, it has been demonstrated that GPC for a single variable is analogous to the standard non-parametric tests for traditional measures of treatment effect [5]. If the variable of interest is binary, net benefit Δ is equal to the absolute risk difference and a link can be established between GPC and Fisher's exact test. If the variable is continuous, GPC is equivalent to the Wilcoxon rank-sum test in the special case where $\tau = 0$, and Δ is related to the effect size through the standard normal cumulative distribution function. For a time-to-event variable, GPC is equivalent to Gehan's generalized Wilcoxon rank-sum test when $\tau = 0$ [25], and Δ can be expressed as a function of the hazard ratio:

$$\Delta = \frac{1 - \mathrm{HR}}{1 + \mathrm{HR}}.$$

2.3 Advantages and limitations

Net benefit by GPC is an extremely general measure of treatment effect as the procedure can handle variables of any type as long as one has a unique way to define "a better outcome" for the considered variables. The measure is also quite easy to interpret. Intuitively, net benefit is positive if the treatment is better than the control, and negative if the treatment is worse than the control. It would be equal to 100% if the treatment is more effective for all patients. Another key advantage of the GPC developed by Buyse is the possibility of simultaneously analyzing several prioritized outcomes [5]. This feature is valuable when one needs to express beneficial and harmful effects of a treatment on a common scale. A clinically relevant threshold can also be incorporated for each outcome, allowing researchers to put in evidence any difference between groups as they wish.

In the context of survival analysis, GPC can extend easily to various time-to-event endpoints. For instance, in advanced cancer, time to disease progression is another clinically important event besides time to death. A traditional way to make use of all these events of interest is to define a single composite endpoint [26]. However, this approach is based on the time course; i.e. it considers only the first occurred event and ignores the later one regardless of the higher clinical importance of the latter. Whereas, GPC is more flexible, it can use either time to death or time to disease progression depending on the available data of the higher priority outcome. If clinicians are more interested in the survival time of patients following cancer therapy, this outcome is considered as the first priority and is taken to compute the net benefit. In the case where the patient stays alive until the end of the study (i.e. absence of the first prioritized outcome), the second outcome which is the

time to disease progression would then be employed. Moreover, in clinical statistics, time-to-event outcomes are usually analyzed under the PH assumption and treatment effect is reported as hazard ratio. Net benefit has been shown to link to HR in the case of a single time-to-event variable and is still interpretable whether the PH assumption is satisfied or not. This is highly advantageous for the situation where HR changes with folllow-up time; for instance, early effect captured in trials with cytotoxic chemotherapies [27] or late survival differences typically observed with modern immunotherapies [28]. GPC does not require the model hypothesis either as it is completely non-parametric.

However, the main limitation of the GPC procedure developed by Buyse is that the scoring rule cannot extract all the useful information of censored times, leading to a loss of power in many cases. The larger the amount of censoring is, the more pairs are classified as uninformative. And since the score of these pairs is set to 0, the numerator in the formula (2.1) will get closer to 0, whereas the denominator (i.e. the total number of pairs) is unchanged. Consequently, the estimated net benefit will become more and more biased towards no effect treatment. This bias can be corrected using the proportion of informative pairs of the entire trial (denoted f) as follows [29, 30]:

$$\widehat{\Delta}_{corr} = \frac{\widehat{\Delta}}{f}.$$

Nevertheless, it is barely applicable for the case consisting of one outcome. Such a simple correction would not be able to tackle the possible dependence between the higher and the lower priority endpoints at the pair level. As analyzing multiple endpoints is one of the main interests of GPCs, this correction will not be further mentioned in this document. Instead we will discuss the Peron method with correction which proceeds the rectification at the pair level (see Section 3.2).

Chapter 3

GPC methods for censored data

3.1 Peron original method

As mentioned in the previous chapter, the standard GPC is not optimal to analyze time-to-event data with censored observations as the censored times are completely ignored in this procedure. To make use of the information of censored times, Peron *et al.* suggested replacing the "definite" status of Win, Loss, or Neutral by a corresponding probability for each pair to belong to one of these three categories [30]. As such, the pairwise score for time-to-event outcome of a pair of observations (i, j) would be computed as:

$$s_{ij} = \mathbb{P}(X_i > Y_j + \tau | X_i', Y_j', \theta_i, \eta_j, S_X, S_Y) - \mathbb{P}(Y_j > X_i + \tau | X_i', Y_j', \theta_i, \eta_j, S_X, S_Y), \tag{3.1}$$

where X' and Y' are the observable times which are the minimum between the event times and the censoring times, $S_X(t) = \mathbb{P}(X > t)$ and $S_Y(t) = \mathbb{P}(Y > t)$ are the survival function of group Trt and Ctr. $\mathbb{P}(X_i > Y_j + \tau | X_i', Y_j', \theta_i, \eta_j, S_X, S_Y)$, $\mathbb{P}(Y_j > X_i + \tau | X_i', Y_j', \theta_i, \eta_j, S_X, S_Y)$, and $\mathbb{P}(|X_i - Y_j| \leq \tau | X_i', Y_j', \theta_i, \eta_j, S_X, S_Y)$ denote the probability for a pair to be favorable, unfavorable and neutral, respectively. For instance, if the observed times (x_i', y_j') of two individuals of a pair are both censored and $|x_i' - y_j'| < \tau$, the probability of this pair to be favorable is given by:

$$\begin{aligned}
\mathbb{P}(X_i > Y_j + \tau | X_i' = x_i', Y_j' = y_j', \theta_i = 0, \eta_j = 0, S_X, S_Y) &= \mathbb{P}(X_i > Y_j + \tau | X_i > x_i', Y_j > y_j', S_X, S_Y) \\
&= \frac{\mathbb{P}(X_i > Y_j + \tau, X_i > x_i', Y_j > y_j' | S_X, S_Y)}{\mathbb{P}(X_i > x_i' | S_X)\mathbb{P}(Y_j > y_j' | S_Y)} \\
&= \frac{\mathbb{P}(X_i > Y_j + \tau | S_X)\mathbb{P}(Y_j > y_j' | S_Y)}{S_X(x_i')S_Y(y_j')} \\
&= \frac{1}{S_X(x_i')S_Y(y_j')} \int_{y_j'}^{\infty} \mathbb{P}(X_i > t + \tau | S_X) \mathrm{d}F_Y(t) \\
&= \frac{-1}{S_X(x_i')S_Y(y_j')} \int_{y_j'}^{\infty} S_X(t + \tau) \mathrm{d}S_Y(t).
\end{aligned} \tag{3.2}$$

Similarly, the probability of this pair to be unfavorable can be computed as:

$$\mathbb{P}(Y_j > X_i + \tau | X_i', Y_j', \theta_i = 0, \eta_j = 0, S_X, S_Y) = \frac{-1}{S_X(x_i')S_Y(y_j')} \int_{x_j'}^{\infty} S_Y(t+\tau) \mathrm{d}S_X(t), \quad (3.3)$$

and then the pairwise score is calculated as the formula (3.1). Assuming that the censoring distribution is independent of the time-to-event distribution, the probabilities to be a Win, Loss and Neutral can be estimated based on $\widehat{S}_X(t)$ and $\widehat{S}_Y(t)$, the KM estimates of S_X, S_Y. The $\widehat{s}_{ij}$ for each combination $\{X_i', Y_j', \theta_i, \eta_j\}$ have been fully derived and presented in the paper of Peron *et al* [30]. Obviously, this score would be equivalent to that of Buyse procedure in the case where pairs can be decidedly classified as "Favorable", "Unfavorable" or "Neutral" as described in Table 2.3; i.e. the Peron's score for those pairs would be 1, -1 or 0, respectively. Complex scores are only needed to compute the pairs classified as "Uninformative" in Table 2.3. Hence, by using censored times and survival curves, Peron *et al.* were able to improve the quantity of informative scores as compared to that of the standard procedure. It is worth noting that this new scoring rule leads to all "completely" informative pairs only when the entire survival curve can be estimated. If the largest observation is censored, as discussed in Section 1.2.3, the survival curve can barely be estimated up to a finite time t_M corresponding to the largest event time. Whereas, for some pairs like the above example, one has to integrate the estimated survival functions and their derivatives (i.e. their jumps) until infinity. When there is no more jumps after the last event time, the probabilities computed from this integral are underestimated since $-\int_0^{\infty} \mathrm{d}\widehat{S}(t) \neq 1$. As a consequence, the probability of being Win, Loss, Neutral calculated for some pairs may not sum up to 1; the remaining probability corresponding to the non-estimated part will be considered "uninformative" (Figure3.1).

The procedure can be generalized to include L outcomes by using a more sophisticated weight. This weight is now defined as the probability for a given pair being neutral ($\widehat{v}_{ijl}$) or uninformative ($\widehat{u}_{ijl}$) based on all the outcomes of higher priority:

$$\widehat{w}_{ijl} = \begin{cases} 1 & \text{if } l = 1. \\ \prod_1^{l-1}(\widehat{v}_{ijl} + \widehat{u}_{ijl}) & \text{otherwise.} \end{cases}$$

As compared to the standard method, the score s_{ijl} no longer takes the "definite" values $\{-1, 0, 1\}$ but can be any value in [-1, 1]. Also, the weight $\widehat{w}_{ijl}$ varies between 0 and 1 instead of being 0 or 1. The formula (2.1) can then be rewritten as:

$$\widehat{\Delta} = \frac{\sum_{l=1}^{L} \sum_{i=1}^{n} \sum_{j=1}^{m} \widehat{s}_{ijl}.\widehat{w}_{ijl}}{n.m}. \quad (3.4)$$

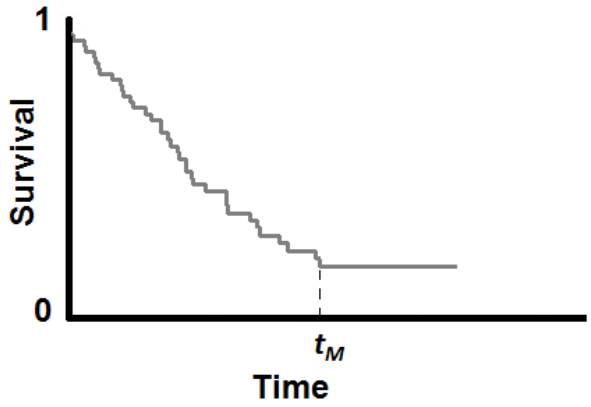

Figure 3.1: Illustration of Kaplan-Meier survival curve for a dataset of which the last observation is censored. There is no longer jump after the largest event time t_M, the remaining part that cannot be estimated is considered as uninformative.

3.2 Peron method with correction

Based on Table 2.3, it can be deduced that the number of uninformative pairs would increase when the censoring proportion becomes larger. With GPC procedure described above (i.e. PO method), if the survival function is not fully estimated, the score of these uninformative pairs involving the non-estimated part will be ignored when computing the pairwise score $\widehat{s}_{ijl}$. As a result, if the censoring proportion is too important, very few pairs would remain "completely" informative (i.e. containing the uninformative score $\widehat{u}_{ijl} > 0$), and thus the ultimately estimated net benefit would be shrunk towards 0 which is the value of no treatment effect. Peron *et al.* have tried to deal with the bias by converting the uninformative proportion into informative one [31]. To do that, they make the strong assumption that uninformative pairs would behave on average like informative pairs. As such, they can compute the contribution of the uninformative score of each pair to the pairwise score of that pair based on the average contribution of the informative scores of all pairs. They introduce p_{win}, p_{loss} and $p_{neutral}$ as the probabilities of pairs classified as Win, Loss and Neutral in the entire study. Their corresponding estimates, denoted $\widehat{p}_{win}$, $\widehat{p}_{loss}$ and $\widehat{p}_{neutral}$, are computed by averaging the empirical probability of being favorable, unfavorable or neutral over all the pairs, respectively. The corrected score is then estimated as follows:

$$\widehat{s}^{c}_{ijl} = \widehat{s}_{ijl} + \frac{\widehat{p}_{win} - \widehat{p}_{loss}}{\widehat{p}_{win} + \widehat{p}_{loss} + \widehat{p}_{neutral}} \widehat{u}_{ijl} \ . \tag{3.5}$$

When considering several outcomes, the corrected net benefit can be easily calculated using the formula (3.4), providing that the neutral score (v_{ijl}) and the weight (w_{ijl}) are corrected accordingly, as detailed below:

$$\widehat{v}^c_{ijl} = \widehat{v}_{ijl} + \frac{\widehat{p}_{neutral}}{\widehat{p}_{win} + \widehat{p}_{loss} + \widehat{p}_{neutral}} \widehat{u}_{ijl} \ ,$$

$$\widehat{w}^c_{ijl} = \begin{cases} 1 & \text{if } l = 1. \\ \prod_1^{l-1} \widehat{v}^c_{ijl} & \text{otherwise.} \end{cases}$$

3.3 Extreme Value Theory method

Peron's scoring rule relies on KM estimation and is thus only appropriate if it is possible to estimate the survival functions at all time points. If the last observation is censored, the tail of survival distributions can not be consistently estimated using KM method. Furthermore, the performance of PO procedure may get poorer in the situations when one needs to take a deeper dive into the tail. For instance, as can be intuitively observed from the formulas (3.2) and (3.3), when the threshold of clinical relevance τ is larger, one has to go further in the tail to be able to detect the difference. Also, when the treatment effect is more important, i.e. the survival times of the treatment group is much longer as compared to those of the control, one has to evaluate S_Y at the observed times of group Trt and therefore needs to explore more the tail of S_Y. To address this challenge, De Backer proposed a method using the result of Extreme Value Theory (EVT) built on the approach of peaks-over-threshold [32]. The idea here is to improve the overall estimation of the survival distribution by fitting a parametric model to only the tail while keeping using the KM estimator for the remaining. A full description of this approach can be found in the report of De Backer [32].

To briefly remind some main results of EVT related to the context of GPC, let us consider now only one group; e.g. group Trt of which the outcome variable $X \sim F_X(t)$. As its name suggests, EVT involves the studies about the behavior of large observations where the sample information may be very limited. The Pickands-Balkema-de Haan theorem states that for a wide class of underlying distributions $F_X(t)$, the conditional distribution of the extreme values, which exceed some sufficiently high threshold z, can be approximated by the Generalized Pareto Distribution (GPD) [33, 34]. That is for $z \to \infty$:

$$\mathbb{P}(X - z > t | X > z) \to H(t) := \left[1 + \frac{\xi t}{\sigma_z}\right]_+^{-1/\xi} ,$$

where $[t]_+ = \max(t, 0)$, $\sigma_z > 0$ and ξ are the scale and shape parameters for the GPD $H(t)$, respectively. $H(t)$ is generalized in the sense that it subsumes most of the usual parametric models for modeling the tail. It belongs to the attraction domain of Weibull distribution if $\xi < 0$, of Fréchet distribution if $\xi > 0$ or of Gumbel distribution if $\xi = 0$ [35]. Of course, there is a sufficient regularity condition to guarantee the tail convergence, called the extended regular variation [36]. As the theorem is barely applicable to the extreme values (i.e. the tail), survival curve will be estimated into two parts: estimate the first part

using a nonparametric technique until a threshold z (to be distinguished with the clinical threshold τ previously mentioned), and then approximate parametrically the tail using the result of EVT. Since for $t > 0$ and $z > 0$,

$$\mathbb{P}(X \geq t + z) = \mathbb{P}(X \geq t + z | X > z)\mathbb{P}(X \geq z),$$

a semiparametric estimator for $S_X(t) = \mathbb{P}(X > t)$ can be constructed as:

$$\widehat{S}_{EVT}(t) = \begin{cases} \widehat{S}_X(t) & \text{if } t \leq z \ , \\ \widehat{S}_X(z)\left[1 + \dfrac{\widehat{\xi}(t-z)}{\widehat{\sigma}_z}\right]_+^{-1/\widehat{\xi}} & \text{if } t > z \ , \end{cases}$$

where $\widehat{S}_X(t)$ is the nonparametric Kaplan-Meier estimator, $\widehat{\xi}$ and $\widehat{\sigma}_z$ are estimators of ξ and σ_z. The choice of z depends on the trade-off between variance and bias. If z is too high, the variance of the parameter estimates will be large due to the limited number of the excesses. If z is too low, the EVT asymptotic approximation will be less reasonable to employ since the available information used to estimate the GPD would be too far away from the tail and not adequate enough to precisely capture the tail behavior; consequently, a bias may be induced. For the present application, a z corresponding to 80% quantile of the observed event times seems to be acceptable; i.e. one will devote 20% of the uncensored data to estimating the parameters of GPD. This has been justified by the simulation results of Alvarez-Iglesias [37] as well as those of De Backer [32]. Once z is selected, $\widehat{\xi}$ and $\widehat{\sigma}_z$ can be obtained by maximizing the likelihood. The censored likelihood function is:

$$L(X_i^z, \theta_i; \xi, \sigma_z) = \prod_{i=1}^{k} \left(\frac{1}{\sigma_z}\left[1 + \frac{\xi X_i^z}{\sigma_z}\right]_+^{-\frac{1}{\xi}-1}\right)^{\theta_i} \left(\left[1 + \frac{\xi X_i^z}{\sigma_z}\right]_+^{-\frac{1}{\xi}}\right)^{1-\theta_i},$$

where X_i^z are the excesses above the threshold z (i.e. $X_i^z = X_i' - z$ with $X_i' > z$) and k is the number of the excesses. Maximizing this function is equivalent to minimizing the negative log-likelihood:

$$(\widehat{\xi}, \widehat{\sigma_z}) = \arg\min_{(\xi,\sigma_z)} \left\{ \log(\sigma_z) \sum_{i=1}^{k} \theta_i + \sum_{i=1}^{k} \left(\frac{1}{\xi} + \theta_i\right) \log\left[1 + \frac{\xi X_i^z}{\sigma_z}\right] \right\}.$$

In addition, one should keep in mind that the regularity assumption of EVT implies that the decreasing survival function would attain 0 at a certain time point. Hence, GPC using EVT tool is only valid under the hypothesis that all patients would experience the event of interest if the follow-up period is long enough, i.e. in the absence of cure fraction. Extending GPC to the latter case, however, is beyond the scope of this work.

3.4 Restricted net benefit method

Restricted net benefit method (RNB) proposed by Kicinski *et al.* solves the problem of censored observations with a different perspective [38]. Considering the fact that survival

time beyond the end of the trial may depend on many unknown factors, the authors proposed to use only the available data instead of extrapolating the missing information, and hence do not need to make any "unverifiable" assumption as the other methods do. The motivation here is very similar to that of restricted mean survival time previously mentioned in Section 1.2.5. That is in the case where the information collected from the trial is insufficient for a reliable inference, one may prefer using an alternative estimator which is robust and provides a straightforward clinical interpretation. Concretely, Kicinski *et al.* restrict the analysis of net benefit to the period of time $[0, \epsilon]$ covered by the study. Here, ϵ is the time prespecified at the design stage depending on the clinical relevance or the feasibility of conducting the trial. The formula (3.1) would thus be redefined as:

$$\begin{aligned} s_{ij} &= \mathbb{P}(\min(X_i, \epsilon) > \min(Y_j, \epsilon) + \tau | X_i', Y_j', \theta_i, \eta_j, S_X, S_Y) \\ &\quad - \mathbb{P}(\min(Y_j, \epsilon) > \min(X_i, \epsilon) + \tau | X_i', Y_j', \theta_i, \eta_j, S_X, S_Y) \\ &= \mathbb{P}(\min(X_i, \epsilon) > Y_j + \tau | X_i', X_j', \theta_i, \eta_j, S_X, S_Y) - \mathbb{P}(\min(Y_j, \epsilon) > X_i + \tau | X_i', Y_j', \theta_i, \eta_j, S_X, S_Y) \end{aligned} \tag{3.6}$$

To compute the two probabilities in the above equation, one has to distinguish the cases where there is no censoring, where only one observation of the pair is censored and when both observations are censored [38]. For instance, for a pair with only the observation from the treatment group censored, the first probability in the equation (3.6) is given by:

$$\begin{aligned} &\mathbb{P}(\min(X_i, \epsilon) > Y_j + \tau | X_i > x_i', Y_j = y_j', S_X, S_Y) \\ &= \mathbb{P}(X_i > Y_j + \tau, Y_j + \tau < \epsilon | X_i > x_i', Y_j = y_j', S_X, S_Y) \\ &= \frac{\mathbb{P}(X_i > Y_j + \tau, Y_j + \tau < \epsilon, X_i > x_i' | Y_j = y_j', S_X, S_Y)}{\mathbb{P}(X_i > x_i' | Y_j = y_j', S_X, S_Y)} \\ &= \frac{\mathbb{P}(y_j' + \tau < \epsilon)\mathbb{P}(X_i > \max(y_j' + \tau, x_i') | S_X)}{\mathbb{P}(X_i > x_i' | S_X)} \\ &= \begin{cases} 0 & \text{if } y_j' + \tau \geq \epsilon. \\ \frac{S_X(\max(y_j' + \tau, x_i'))}{S_X(x_i')} & \text{if } y_j' + \tau < \epsilon. \end{cases} \end{aligned}$$

The second probability in the equation (3.6) for this pair is given by:

$$\begin{aligned} &\mathbb{P}(\min(Y_j, \epsilon) > X_i + \tau | X_i > x_i', Y_j = y_j', S_X, S_Y) \\ &= \frac{\mathbb{P}(Y_j > X_i + \tau, X_i + \tau < \epsilon, X_i > x_i' | Y_j = y_j', S_X, S_Y)}{\mathbb{P}(X_i > x_i' | Y_j = y_j', S_X, S_Y)} \\ &= \frac{\mathbb{P}(x_i' < X_i < \min(y_j', \epsilon) - \tau | S_X)}{\mathbb{P}(X_i > x_i' | S_X)} \\ &= \begin{cases} 0 & \text{if } x_i' \geq \min(y_j', \epsilon) - \tau. \\ \frac{S_X(x_i') - S_X(\min(y_j', \epsilon) - \tau)}{S_X(x_i')} & \text{if } x_i' < \min(y_j', \epsilon) - \tau. \end{cases} \end{aligned}$$

To obtain the estimator of pairwise score s_{ij}, the survival distributions are replaced by the KM estimators. Then, the estimator of restricted net benefit Δ_R can be computed using the formula (3.4). As the survival curves are always estimable up to the time point ϵ, one has no longer the uninformative pairs and thus obtains un unbiased estimate. However, the interpretation for Δ_R should be modified since it now explains the difference between groups within a specified period of study, and not the entire profile of treatment effect as Δ does.

3.5 Reported results and limitations

The currently reported results and limitations of the four GPC methods for censored data are summarized in Table 3.1. The three "correction" methods (PC, EVT and RNB) are all based on the main framework of the PO procedure where pairwise scores are estimated based on censored times and survival curves. The main limitations of PO method are the underestimation in the presence of censoring and the inflation of the number of uninformative pairs when the censoring proportion increases, leading to a bias towards 0. Peron *et al.* corrected their own method by converting the uninformative score into the informative counterpart. As this correction is carried out at the pair level, it can be applied for the GPC analysis with several outcomes; however, PC approach is only valid under the condition that uninformative pairs are exchangeable with informative pairs. In another work, De Backer tackled the bias of PO method by using the results of EVT to extrapolate the tail of the survival curves and thus improve the overall estimation. Unlike PC and EVT, RNB method of Kicinski *et al.* discards the information unobtainable from the trial, and only computes the relevant probabilities until when the survival functions remain estimable. As such, their net benefit estimation is unbiased but the interpretation of results may be less persuasive as it depends on the arbitrary restriction time (i.e. the estimate only reflects the treatment effect within this selected period). Different strategies of these approaches are graphically illustrated in Figure 3.2.

Heretofore, PO method has been assessed together with PC method in the work of Peron [31] and directly compared with EVT method by De Backer [32]. Peron *et al.* have investigated PC estimator in many scenarios where clinical threshold is null or non null, where hazards are proportional or not proportional, and where there is one time-to-event outcome or one time-to-event outcome combined with one binary outcome. In addition, they varied the censoring proportion p_c, which is the censoring due to the withdrawal of patients from the study or due to the lost during follow-up, from 0% to 90%. In most of the scenarios, it has been found that PC procedure is barely slightly better than PO method when up to 70% of data are censored. At 90% of censoring, the estimation with correction (PC) is much better than those without correction (PO) but still very biased. In the work of De Backer, the authors introduced the type I censoring proportion (denoted p_c^M) in their simulation settings besides the classical one p_c. Of note, since type I censored observations occur at the end of the study due to the early termination, they would be

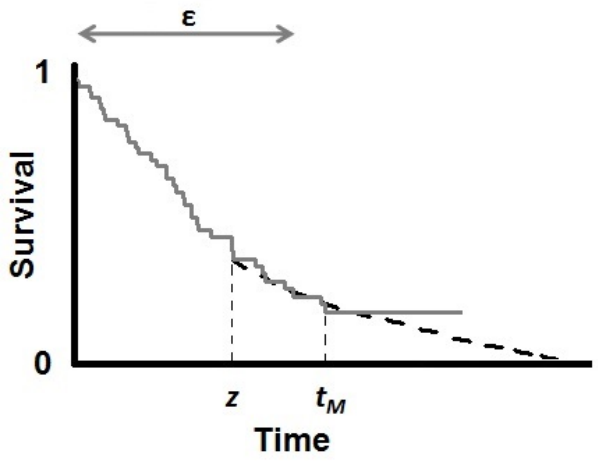

Figure 3.2: Graphical illustration of the strategies of GPC procedures. The solid line represents the KM estimator of survival function. The dashed line represents the tail estimated by EVT method (i.e. the survival part beyond the threshold z). PO method can only estimate the pairwise score up to a limited time t_M, the survival part beyond t_M is considered uninformative. PC method converts this uninformative score into informative one. The prespecified period ϵ corresponds to the restriction time analyzed with RNB method.

expected to cause more perturbation to the tail estimation than the classical censoring. De Backer also varied these two censoring proportions (p_c from 0% to 45% and p_c^M from 0% to 20%) and showed that the PO estimation is very sensitive in the presence of type I censoring; whereas the estimation by EVT still well resists if the p_c^M remains low and p_c is not too high. Furthermore, EVT estimator has been demonstrated to outperform in the cases where one has to explore further into the tail (e.g. when threshold of clinical relevance τ grows or when net benefit Δ is high) as anticipated (see Section 3.3).

Considering that the previous studies have already investigated the classical censoring at very high rates, in this work, we will not focus on the estimation performance when increasing p_c. We will keep the p_c as low as 10% or 20%, which is the drop-out rate generally anticipated in clinical trials [39]. On the other hand, we will aim at exploring the scenarios where PC and EVT method may fail. Several potential scenarios can be rationalized as follows:

- As can be seen in the formula (3.5), the correction of PC procedure involves the attribution of a weight to the uninformative score and this weight depends on the estimated proportion of Win, Loss and Neutral pairs. The calculation of these latter quantities would become more delicate if one needs to evaluate the survival curves far in the tail, for example when τ is large. Therefore, we would suggest varying

the clinically relevant difference τ in the first simulation scenario to see if the PC correction is still effective.

- PC method may be more useful when the Type I censoring is high but the classical censoring remains low. In this case, EVT tool would certainly have more difficulties to extrapolate the tail due to the lack of extreme observation; whereas, the assumption about the same behavior of informative and uninformative pairs could be plausible in some situations. Hence, in the next potential scenario, we will increase the proportion of type I censoring much higher than what has been tested in the previous studies.

- It has been well known that the PC estimation would be problematic if its inherent assumption is violated, for example when the hazards are not proportional. Indeed, this assumption implies that the average of treatment effect obtained during the follow-up can carry over to beyond that. Whereas, under non PH, the magnitude of treatment benefit at later time points would be different from that earlier. Hence, if there are too many censored patients and they actually experience the event of interest after the study period, PC estimator will not be able to correctly infer the treatment benefit. This issue may become more pronounced if the change of hazards is more considerable.

- Since EVT method has not been investigated yet in non-PH settings, it would definitely be interesting to examine the quality of EVT estimator in such cases. Particularly, we will distinguish the non-PH scenarios with and without the plausibility of EVT regularity assumption.

Table 3.1: Summary of main results and limitations of GPC methods

Method	*Simulation scenarios*	*Main results*	*Limitations*
1. PO	**A.** One time-to-event outcome and proportional hazards **B.** One time-to-event outcome and late treatment effect (HR: 1 → 0.3) **C.** One time-to-event outcome and early treatment effect (HR: 0.5 →1) **D.** One time-to-event outcome and one binary outcome, τ=1 * In all scenarios: X ~ Exp(λ_X), Y ~ Exp($\lambda_Y = 1$) and $p_C \in$ {0%, 15%, 30%, 50%, 70%)	- Higher power than the standard method - Estimator starts to be biased when $p_C > 30\%$ but always better than that of the standard procedure	- Unable to estimate consistently survival function when the last observation is censored. - Inflation of number of uninformative pairs when p_C increases.
2. PC	**A.** One time-to-event outcome and proportional hazards, τ=0 **B.** One time-to-event outcome and proportional hazards, τ=50 (40 percentile of the time-to-event distribution in group C) **C.** One time-to-event outcome and early treatment effect (HR: 0.5 →1), τ=0 **D.** One time-to-event outcome and late treatment effect (HR: 1 → 0.3), τ=0 **E:** One time-to-event outcome and one binary outcome, τ=50 (40 percentile of the time-to-event distribution in group C) * In all scenarios: X ~ Exp(λ_X), Y ~ Exp($\lambda_Y = 1$) and $p_C \in$ {0%, 20%, 50%, 70%, 90%)	- PO and PC method have the same power and type I error. - Performance of the estimator may depend on case study. - In general, PC method is only slightly better than PO method when $p_C > 70\%$	- Strong assumption about the similar behavior of informative and uninformative pairs.

Table 3.1: (cont.)

3. EVT	**A.** Varying censoring proportion, (fixing Δ=1/3, τ=0) $p_C \in \{0\%, 15\%, 30\%, 45\%\}$ and $p_c^M \in \{0\%, 10\%, 20\%\}$ **B.** Varying threshold τ (with $(p_C, p_c^M) \in \{(30\%, 10\%),(15\%, 20\%)\}$) $\tau \in \{0, 1, 1.5\}$ **C.** Varying net benefit Δ (with $(p_C, p_c^M) \in \{(30\%, 10\%),(15\%, 20\%)\}$ and τ=0) $\Delta \in \{0, 1/3, 1/2\}$ **D.** The same settings above but $X \sim$ Log-normal(μ, σ^2) * In all scenarios, $X \sim \text{Exp}(\lambda_X)$ (except Scenario 3D), $Y \sim \text{Exp}(\lambda_Y = 1)$ and proportional hazards	- PO method is still good enough until (p_C= 30%, p_c^M= 10%) or (p_C= 15%, p_c^M= 20%). Above these, EVT method is better with lower bias. - EVT method is better than PO method when τ or Δ increases. - Similar conclusion when varying time-to-event distribution.	- Problem of choosing the threshold z when the proportion of censoring is too high.
4. RNB	No data	No data	- Less persuasive interpretation as it depends on the restriction time, especially when survival curves are far from 0.

Chapter 4

Simulation study

4.1 Implementation

In this chapter, we conduct a simulation study including several scenarios to evaluate the performance of different GPC procedures. For each scenario, data will be generated for the two groups (treatment Trt and control Ctr) with sample size $n = m = 500$ patients. True survival times are assumed to follow an exponential distribution, with a scale parameter $\lambda_Y = 1$ for group Ctr in all scenarios; the hazard λ_X of group Trt will vary depending on the investigated scenario. In both groups, censored times X^c and Y^c are assumed to be uniformly distributed on the interval $(0, M)$; type I censoring due to the termination of study at t_M is also added. M and t_M (with $M > t_M$) are selected to reach the target value of p_c (classical censoring) and p_c^M (type I censoring) in group Trt by numerically solving these two equations [32]:

$$p_c = \mathbb{P}(C < X, C < t_M) = \frac{1}{M}\int_0^{t_M} S_X(t)\mathrm{d}t,$$

$$p_c^M = \mathbb{P}(t_M < X < C) + \mathbb{P}(t_M < C < X) = \frac{(M - t_M)}{M} S_X(t_M).$$

The first equation is due to the fact that censored times are uniformly distributed and the other equation follows from the fact that a Type I censored observation is fully observed at a time greater than t_M, or is censored at a time greater than t_M. Of note, since the same distribution is used to generate censored times in the two groups, and M and t_M are chosen focusing only on the treatment group, it implies that if the net benefit $\Delta > 0$, we will have a larger censoring proportion in the treatment group than in the control group, and this difference will be more important when the treatment effect increases. According to the definition of net benefit Δ, the true value of this quantity can be computed as follows:

$$
\begin{aligned}
\Delta &= \mathbb{P}(X > Y + \tau) - \mathbb{P}(Y > X + \tau) \\
&= -\int_0^\infty S_X(t+\tau)\mathrm{d}S_Y(t) + \int_0^\infty S_Y(t+\tau)\mathrm{d}S_X(t) \\
&= \int_0^\infty S_X(t+\tau)f_Y(t)\mathrm{d}t - \int_0^\infty S_Y(t+\tau)f_X(t)\mathrm{d}t \\
&= \int_0^\infty S_X(t+\tau)S_Y(t)\lambda_Y(t)\mathrm{d}t - \int_0^\infty S_Y(t+\tau)S_X(t)\lambda_X(t)\mathrm{d}t \\
&= \int_0^\infty \exp\Big[-\int_0^\infty \lambda_X(t+\tau)\mathrm{d}t\Big]\exp\Big[-\int_0^\infty \lambda_Y(t)\mathrm{d}t\Big]\lambda_Y(t)\mathrm{d}t \\
&\quad -\int_0^\infty \exp\Big[-\int_0^\infty \lambda_Y(t+\tau)\mathrm{d}t\Big]\exp\Big[-\int_0^\infty \lambda_X(t)\mathrm{d}t\Big]\lambda_X(t)\mathrm{d}t, \qquad (4.1)
\end{aligned}
$$

where f_X and f_Y is the density function of X and Y, respectively.

The data generation process is executed as follows:

- Draw a random sample of $n+m$ censored times: $(x_1^c, ..., x_n^c, y_1^c \ldots, y_m^c) \sim Unif(0, M)$ and randomly allocate the censored times to group Trt and Ctr.
- Draw two random samples of true survival times:

  - Group Trt: In the PH cases, λ_X is constant and $(x_1, x_2, \ldots, x_n)$ can be simulated directly from $Exp(\lambda_X)$. In the non PH cases, λ_X is a function of t; the survival times will be simulated from the hazard function using the principle described in the paper of Bender *et al.* [40]:

  $$X = \Lambda^{-1}[\log(U)] \text{ with } U \sim Unif(0, 1),$$

  where $\Lambda = \int_0^t \lambda_X(s)\mathrm{d}s$ is the cumulative hazard function.

  - Group Ctr: $(y_1, y_2, \ldots, y_m) \sim Exp(\lambda_Y)$ with $\lambda_Y = 1$ in all scenarios.
- Obtain the observed values: $x_i' = \min(x_i, x_i^c, t_M)$ for group Trt $(i = 1, 2, ..., n)$, and $y_j' = \min(y_j, y_j^c, t_M)$ for group Ctr $(j = 1, 2, ..., m)$.
- Obtain the censoring indicators: $\theta_i = \mathbb{I}(x_i < x_i^c, x_i < t_M)$ for group Trt, and $\eta_j = \mathbb{I}(y_j < y_j^c, x_j < t_M)$ for group Ctr.

For each generated dataset, the following net benefit estimators will be computed:

- **Peron original method (PO)**: Using the R package BuyseTest [41].
- **Peron method with correction (PC)**: Using the R package BuyseTest with the option for applying the correction based on the average score of informative pairs [41].

- **Extreme Value Theory method (EVT)**: The entire estimation procedure has been re-coded by De Backer as it is not straightforward to use the package BuyseTest when considering the semiparametric estimator; notably, one has to distinguish many situations depending on the thresholds z_T and z_C above which the survival distribution starts to be parametrically estimated for the treatment and the control group [32]. Here we will fix these thresholds at 80% quantile of the uncensored data in all scenarios.

These above steps will be repeated $nsim \in \{100; 500\}$ times; the mean of the estimated values and their empirical variance will be then calculated. The bias and mean squared error (MSE) of these estimators can be assessed based on the true net benefit Δ obtained by numerically solving the formula (4.1):

$$\text{Bias}(\widehat{\Delta}) = \frac{\sum_{k=1}^{nsim} \widehat{\Delta}_k}{nsim} - \Delta,$$

$$\text{MSE}(\widehat{\Delta}) = \frac{\sum_{k=1}^{nsim} (\widehat{\Delta}_k - \Delta)^2}{nsim}.$$

4.2 Investigated scenarios

As discussed in Section 3.5, we will examine the situations which can potentially help us unravel the usefulness of each GPC estimator. In the first two data generating process (DGP), we consider the scenarios where the hazards of group Trt and Ctr are proportional.

- **DGP 1**: $\lambda_X = 0.5$, i.e. HR = 0.5 and constant over time. $p_c = 20\%$ and $p_c^M = 20\%$. The threshold of clinical relevance will be varied: $\tau \in \{0, 1, 1.5\}$ to investigate the impact of this factor on the estimated net benefit. We may expect that the performance of PO and PC estimators would deteriorate when τ gets larger as we must go further to the tail of survival curves to detect the difference between two groups. The HR and representative survivals curves in this scenario are illustrated in Figure 4.1.

- **DGP 2**: $\lambda_X = 0.5$, i.e. HR = 0.5 and constant over time. $p_c = 10\%$ and $\tau = 0$. The type I censoring proportion will be varied: $p_c^M \in \{40\%, 50\%, 60\%\}$ (Figure 4.2) so that we can investigate the impact of p_c^M on the estimation performance when the underlying assumption of all GPC procedures are satisfied. PO estimator may be expected to outperform EVT as the latter would possess too little fully observed data to extrapolate the tail.

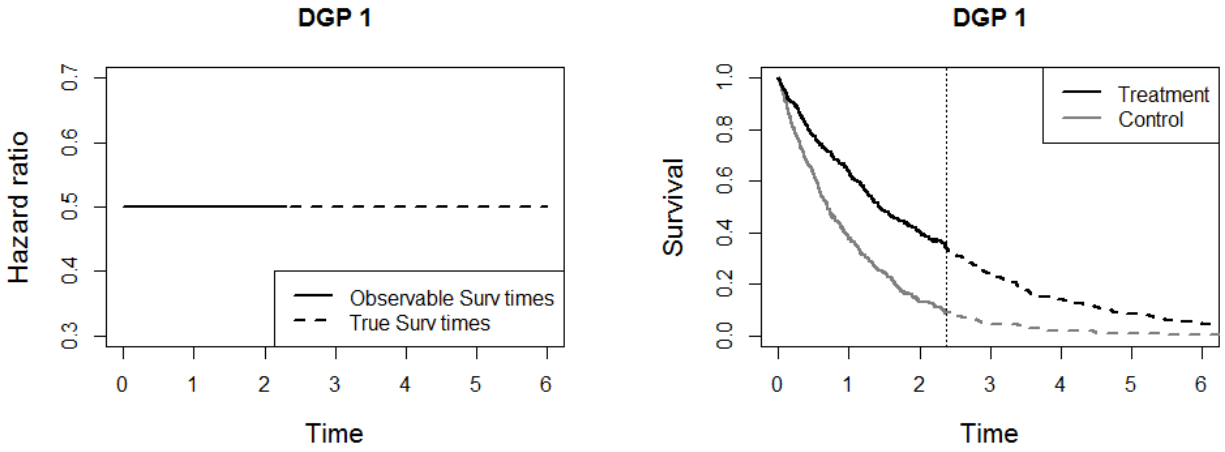

Figure 4.1: Hazard ratio (left panel) and estimated survival curves of one representative dataset (right panel) in DGP 1 ($p_c = 20\%$ and $p_c^M = 20\%$). The solid and dashed lines represent the observable and true survival times, respectively. The dotted vertical line represents the time t_M at which the study is terminated.

In the remaining settings, HR is no longer constant but increases (early treatment effect) or decreases (late treatment effect) over time. The hazard of group Trt in the next two DGPs (DGP 3 and DGP 4) is let vary but in a regular manner so that the underlying assumption of EVT is still plausible. The latter estimator would thus be expected to be the most effective; whereas, PC estimator may be in trouble as its assumption is violated in the non-PH cases.

- **DGP 3** (early treatment effect): $\lambda_X(t) = 0.1 + \alpha t$ with $p_c = 20\%$, $p_c^M = 20\%$, $\tau = 0$ and $\alpha \in \{0.1, 0.15, 0.2\}$. α is let vary to examine how the magnitude of changing hazard would impact the estimation quality. We may expect that the faster the hazard changes, the less effective the correction of PC method would be, due to its inherent assumption. As we fix the censoring rates, changing the slope also implies changing the study termination time t_M (Figure 4.3).
- **DGP 4** (late treatment effect): $\lambda_X(t) = \max(0.8 - \alpha t, 0.1)$ with $p_c = 20\%$, $p_c^M = 20\%$, $\tau = 0$ and $\alpha \in \{0.1, 0.15, 0.2\}$. Similarly to DGP 3, α is let vary with the expect that the faster the hazard changes, the less effective the correction of PC method would be. As we fix the censoring rates, changing the slope also implies changing the study termination time t_M (Figure 4.4).

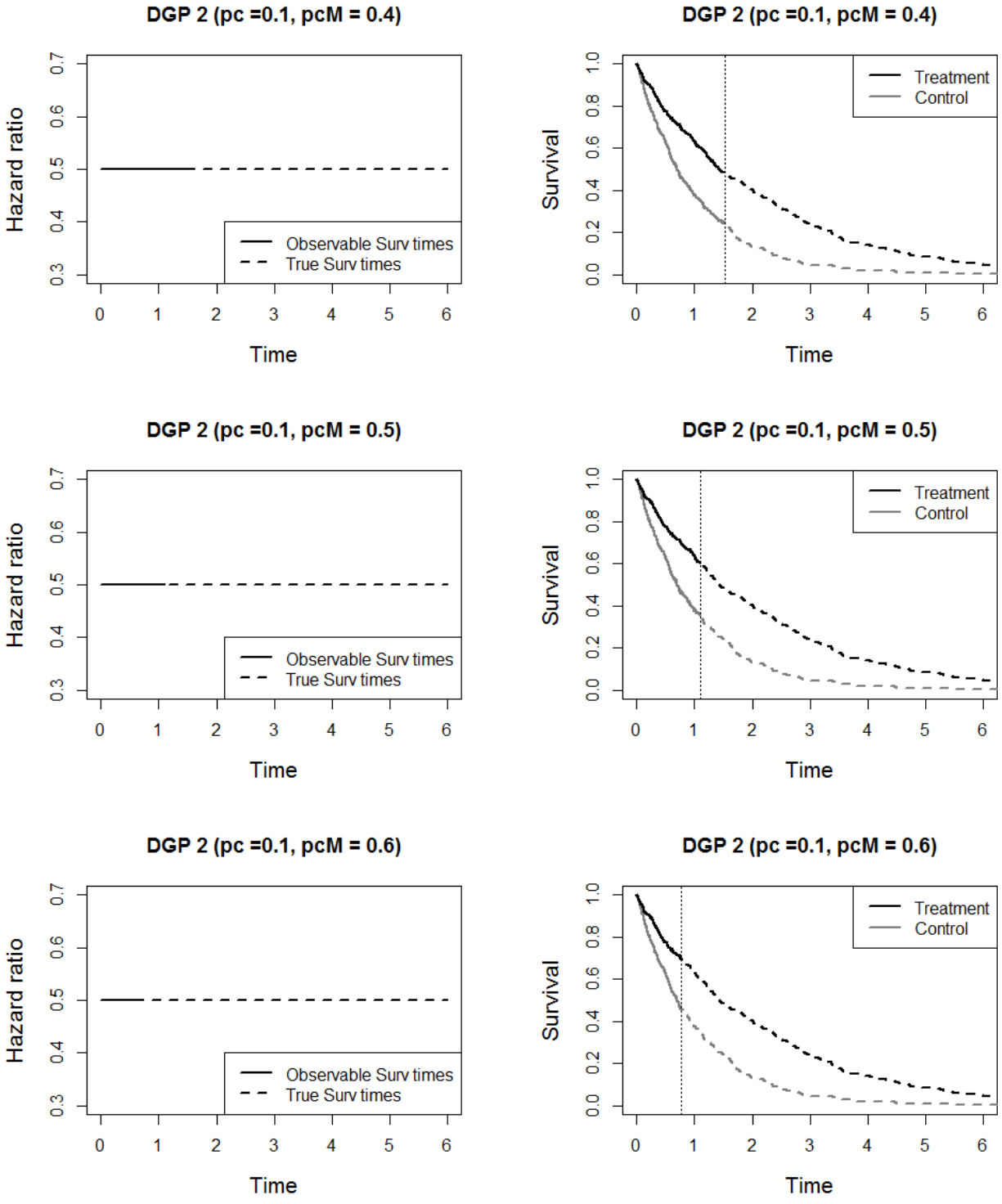

Figure 4.2: Hazard ratio (left panel) and estimated survival curves of one representative dataset (right panel) in DGP 2 when varying the proportion of type I censoring (p_c = 10% and $\tau = 0$).The solid and dashed lines represent the observable and true survival times, respectively. The dotted vertical line represents the time t_M at which the study is terminated.

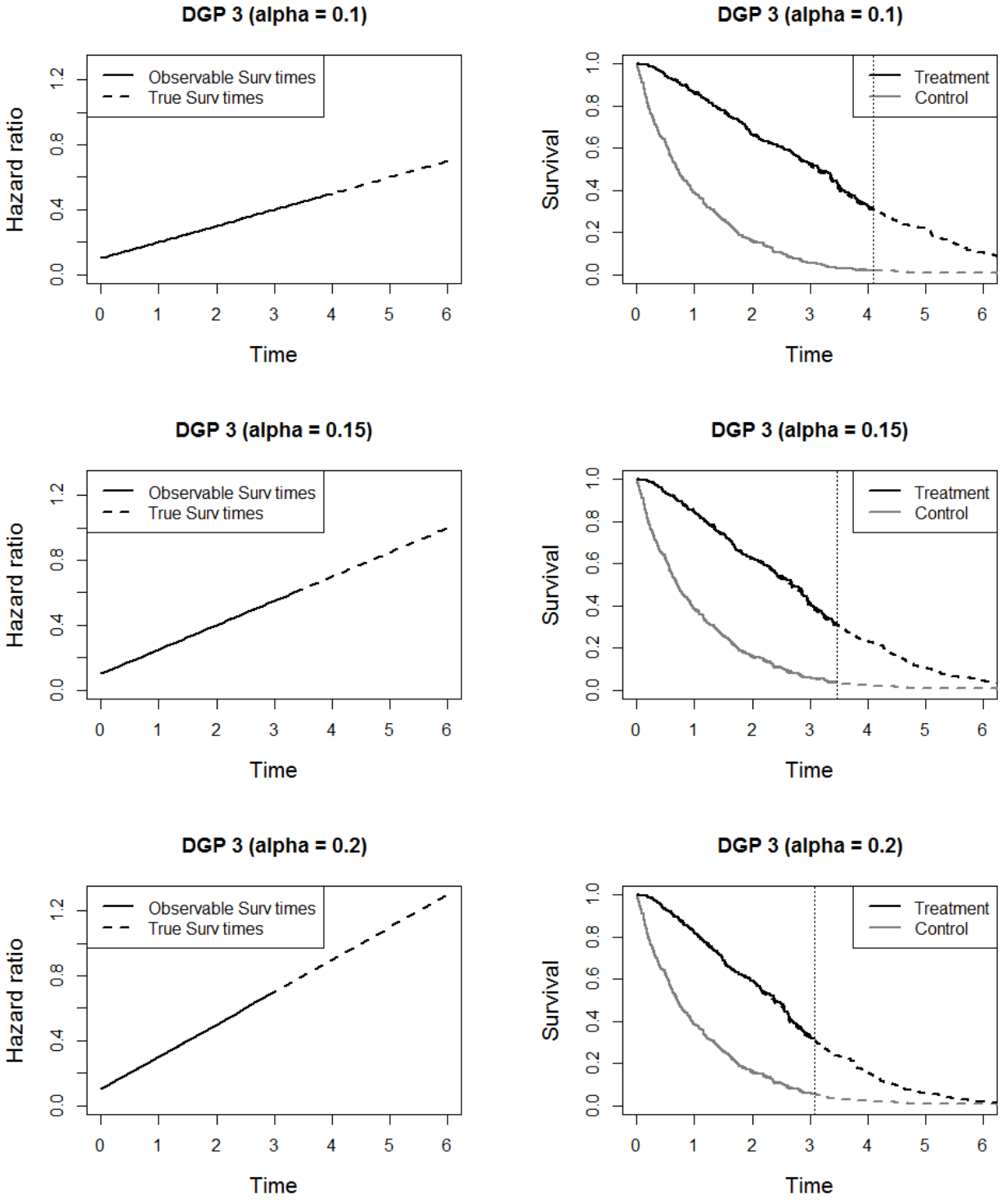

Figure 4.3: Hazard ratio (left panel) and estimated survival curves of one representative dataset (right panel) in DGP 3 ($p_c = 20\%$, $p_c^M = 20\%$ and $\tau = 0$). The solid and dashed lines represent the observable and true survival times, respectively. The dotted vertical line represents the time t_M at which the study is terminated.

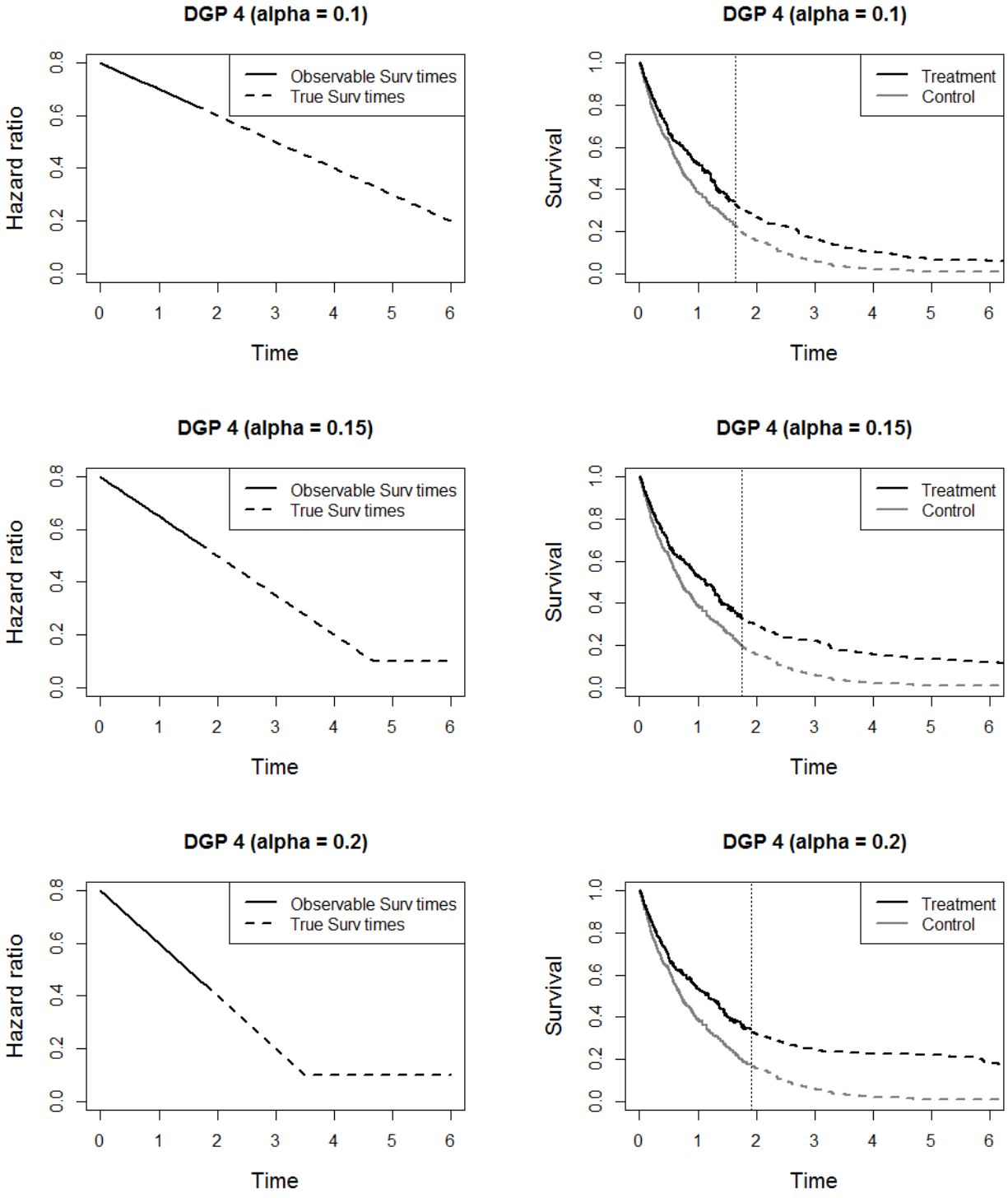

Figure 4.4: Hazard ratio (left panel) and estimated survival curves of one representative dataset (right panel) in DGP 4 ($p_c = 20\%$, $p_c^M = 20\%$ and $\tau = 0$). The solid and dashed lines represent the observable and true survival times, respectively. The dotted vertical line represents the time t_M at which the study is terminated.

In the two following scenarios (DGP 5 and DGP 6), the hazard of group Trt is changed in such a way that the survival function S_X would vary irregularly. EVT estimator may fail in these cases as the regularity assumption is no longer realistic. To generate such situation, we will simulate the survival times from different distributions. More specifically, one part of the survival function will follow an exponential distribution characterized by a constant hazard; and the other part will correspond to a Weibull distribution with a linear hazard. The hazard λ_X will be selected based on the shape of the survival curve; (i.e. graphical inspection).

- **DGP 5** (early treatment effect):

$$\lambda_X(t) = \begin{cases} \alpha & \text{if } t \leq 3 \\ \alpha + 2(t-3) & \text{if } t > 3 \end{cases}$$

  with $p_c = 20\%$, $p_c^M = 20\%$, $\tau = 0$ and $\alpha \in \{0.3, 0.35, 0.4\}$. Time t_M at which the study is terminated will be varied when changing the value α (Figure 4.5).

- **DGP 6** (late treatment effect):

$$\lambda_X(t) = \begin{cases} \alpha & \text{if } t \leq 1.5 \\ \max(\alpha - 0.5(t-1.5), 0.1) & \text{if } t > 1.5 \end{cases}$$

  with $p_c = 20\%$, $p_c^M = 20\%$, $\tau = 0$ and $\alpha \in \{0.5, 0.55, 0.6\}$. Time t_M at which the study will be terminated would be varied when changing the value α (Figure 4.6).

The two last DGPs (DGP 7 and DGP 8) are added with two purposes: first, to observe the behavior of PO estimator in more non-PH cases, especially when HR increases over time; second, to assess the behavior of EVT estimator when the EVT assumption seems more plausible than in the case of DGP 5 and DGP 6.

- **DGP 7** (early treatment effect): $\lambda_X(t) = \min(0.2 + \alpha t, 1)$ with $p_c = 20\%$, $p_c^M = 20\%$, $\tau = 0$ and $\alpha \in \{0.2, 0.4, 0.8\}$ (Figure 4.7).

- **DGP 8** (late treatment effect): $\lambda_X(t) = \max(1 - \alpha t, 0.2)$ with $p_c = 20\%$, $p_c^M = 20\%$, $\tau = 0$ and $\alpha \in \{0.2, 0.43, 0.55\}$ (Figure 4.8).

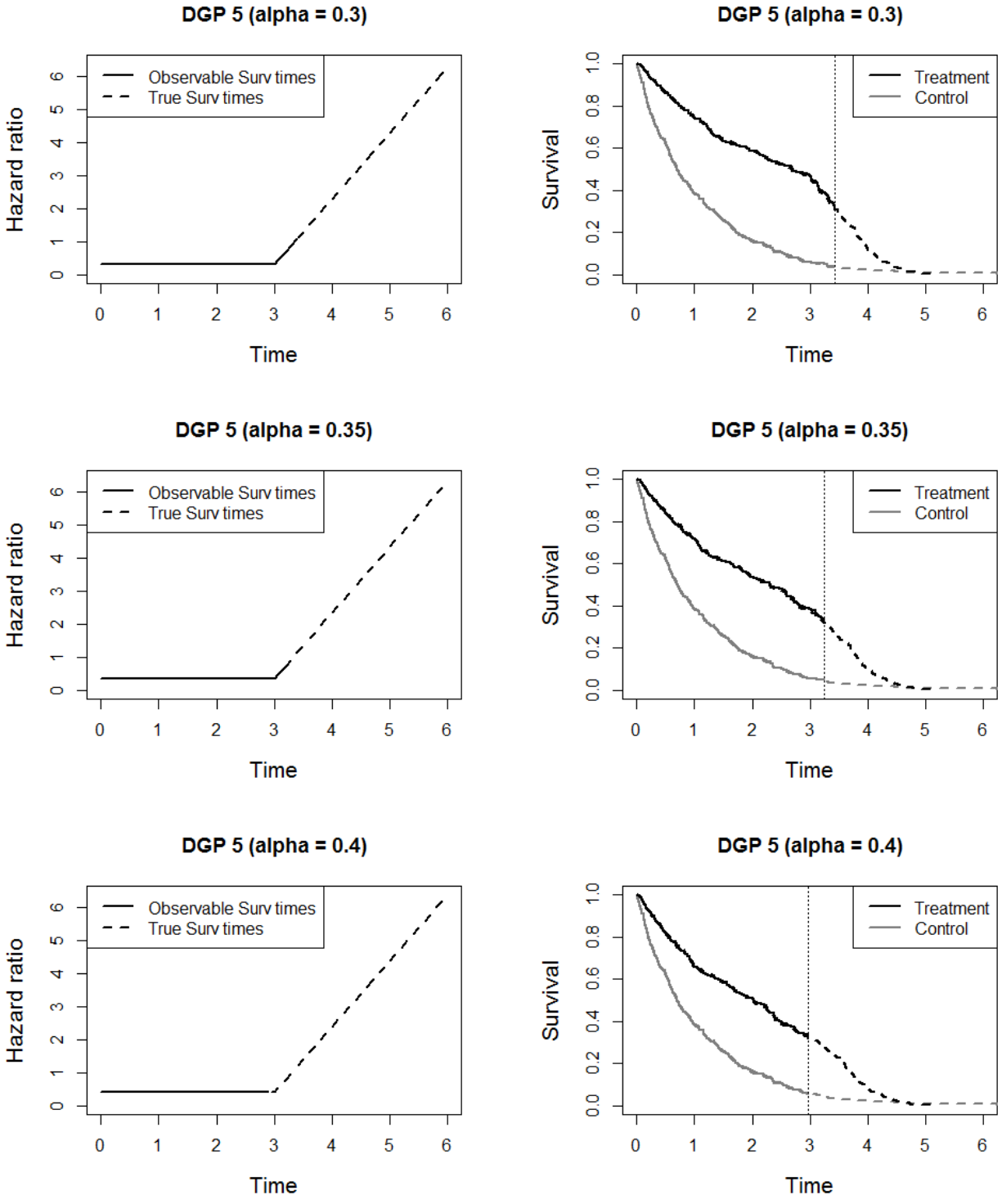

Figure 4.5: Hazard ratio (left panel) and estimated survival curves of one representative dataset (right panel) in DGP 5 ($p_c = 20\%$, $p_c^M = 20\%$ and $\tau = 0$). The solid and dashed lines represent the observable and true survival times, respectively. The dotted vertical line represents the time t_M at which the study is terminated.

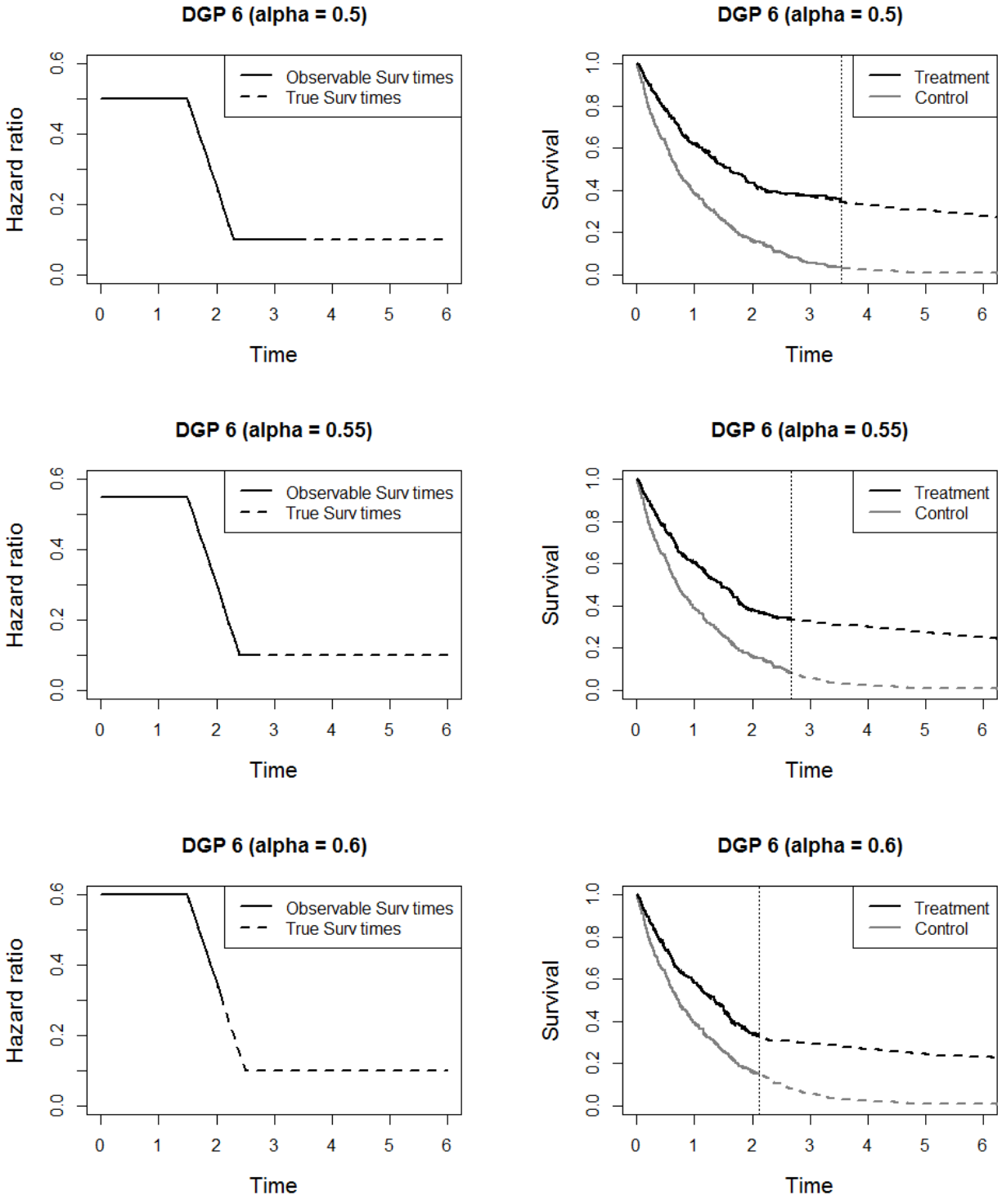

Figure 4.6: Hazard ratio (left panel) and estimated survival curves of one representative dataset (right panel) in DGP 6 ($p_c = 20\%$, $p_c^M = 20\%$ and $\tau = 0$). The solid and dashed lines represent the observable and true survival times, respectively. The dotted vertical line represents the time t_M at which the study is terminated.

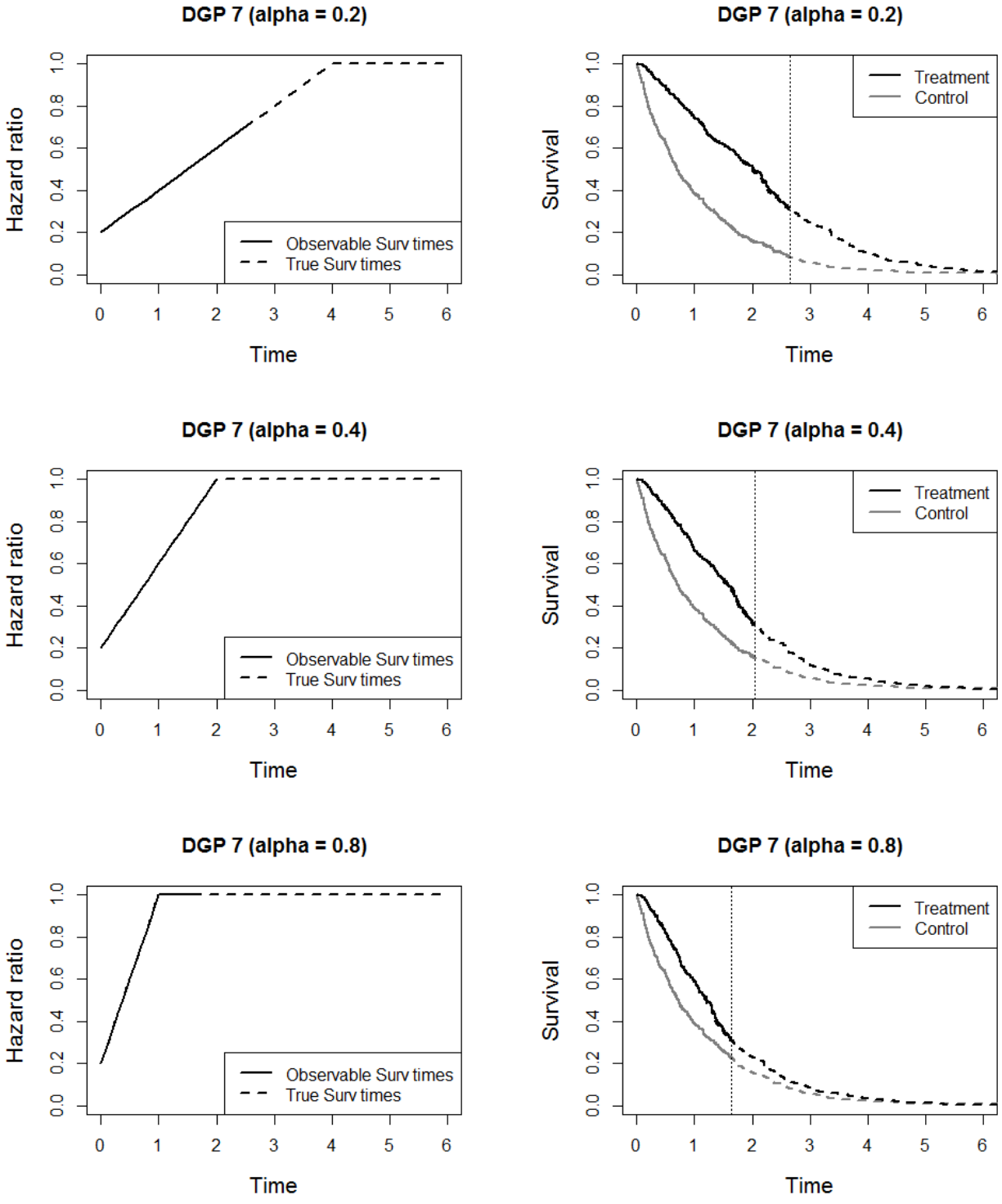

Figure 4.7: Hazard ratio (left panel) and estimated survival curves of one representative dataset (right panel) in DGP 7 ($p_c = 20\%$, $p_c^M = 20\%$ and $\tau = 0$). The solid and dashed lines represent the observable and true survival times, respectively. The dotted vertical line represents the time t_M at which the study is terminated.

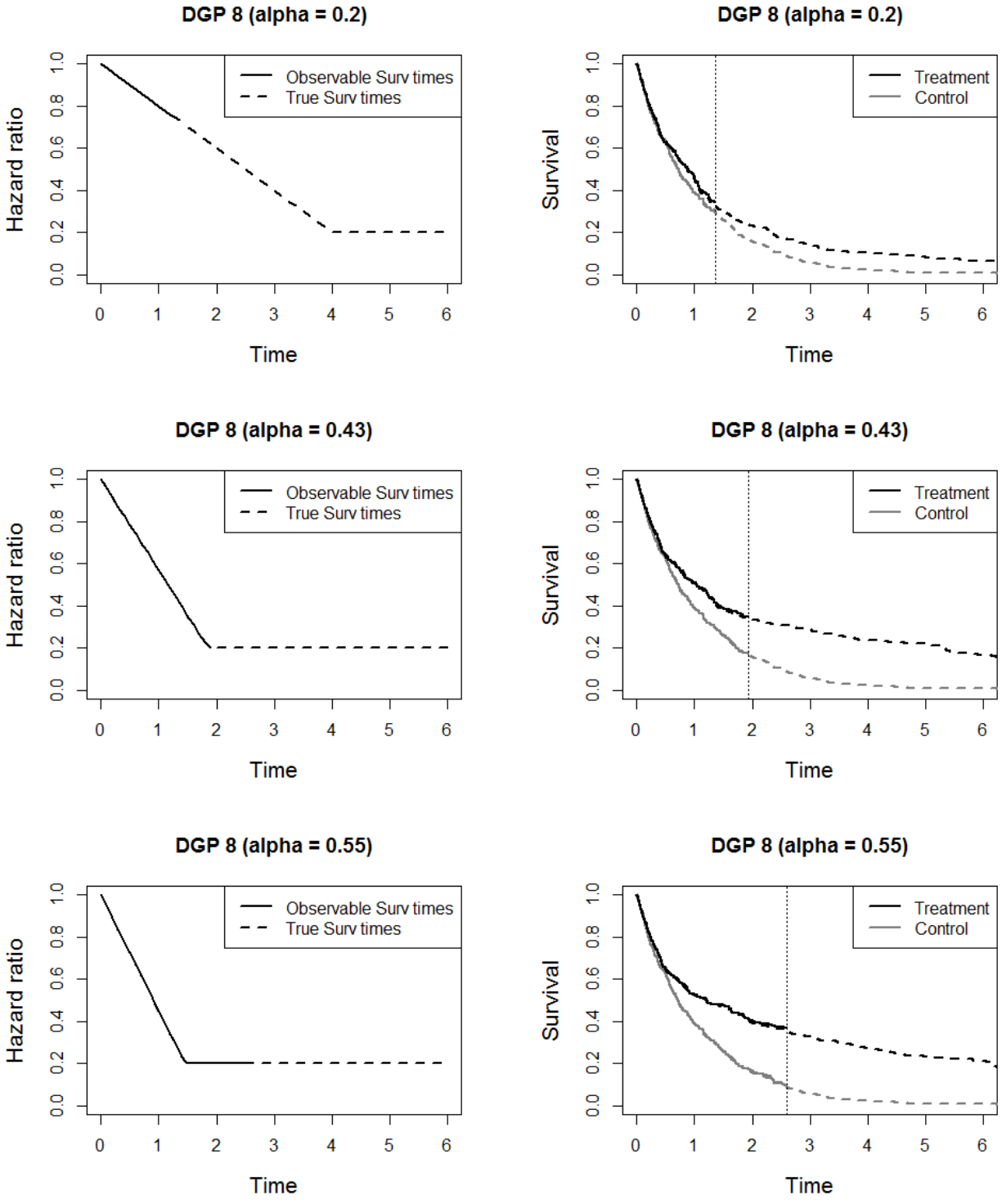

Figure 4.8: Hazard ratio (left panel) and estimated survival curves of one representative dataset (right panel) in DGP 8 ($p_c = 20\%$, $p_c^M = 20\%$ and $\tau = 0$). The solid and dashed lines represent the observable and true survival times, respectively. The dotted vertical line represents the time t_M at which the study is terminated.

4.3 Results

The true net benefits along with the mean, the variance, the bias and the MSE of PO, PC and EVT estimators in all simulation scenarios are calculated and summarized in the Appendices. Except for the EVT estimator in DGP 2, the variances of all estimators are fairly similar; hence, the difference in MSE among the considered estimators mostly come from their bias. For the following, we will focus mainly on the bias to evaluate the estimation quality of net benefit.

Figure 4.9 illustrates the results of DGP 1 for different number of simulations when increasing the clinically significant difference τ between two groups. Not surprisingly, the EVT estimation is the most effective in this case. The scoring rule of PO method, which is based on the KM estimation, is negatively impacted by censoring, and this impact becomes more remarkable when we are forced to explore further into the tail as τ grows. This is well in accordance with the previous results reported by De Backer [32]. The correction used in PC procedure can improve the net benefit estimation of PO method but it is still biased when a $\tau \neq 0$ is applied. This can be explained on the one hand by the shrinking of the correction weight along with the increase of the neutral probability; and on the other hand by the fact that all relevant terms to calculate the weight in the formula (3.5) would be less accurate since the estimation of Win and Loss probabilities will be more complicated with large τ. It can be also noted that the result is almost unchanged when we increase the number of simulations.

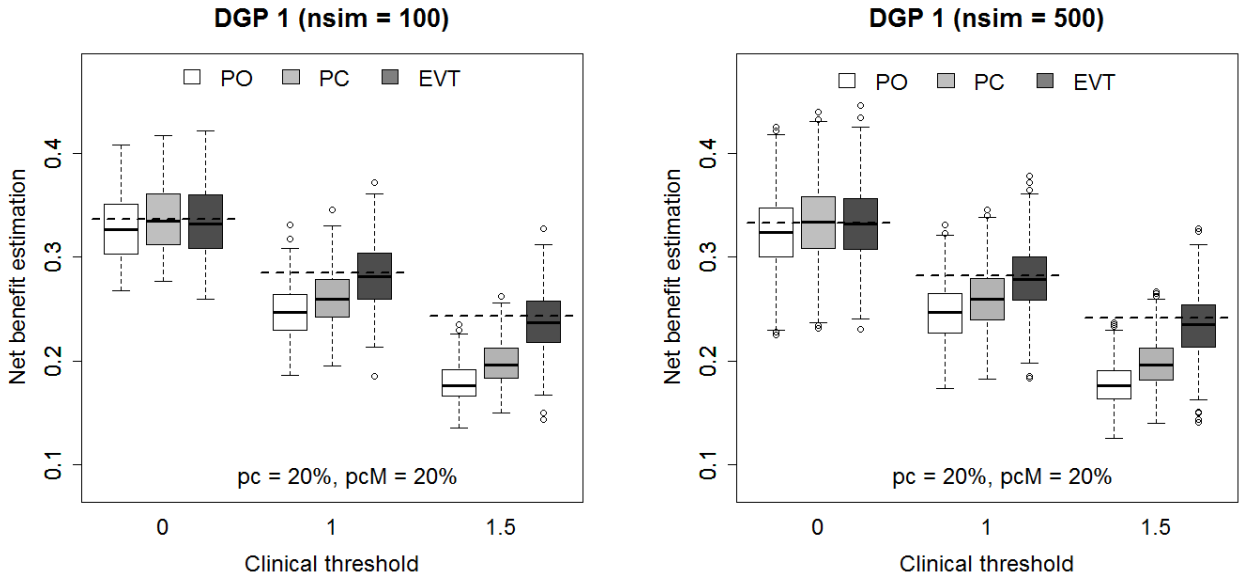

Figure 4.9: Net benefit estimation obtained from 100 and 500 datasets generated in DGP 1. The dashed line represents the true net benefit.

In DGP 2 scenario (Figure 4.10), we set a low proportion of classical censoring ($p_c = 10\%$) and clinical threshold $\tau = 0$ so that we can evaluate the impact of only type I censoring. As expected, the quality of PO estimation dramatically deteriorates at high p_c^M. The EVT estimator somewhat resists at $p_c^M = 40\%$; however, it can be noticed that its variance starts to be larger than that of the PO and PC estimators. This is due to the fact that we need 20% of the uncensored data for the parametric estimation of the tail. When the overall censoring rate grows, we will have less and less fully observed values for this estimation procedure. At $p_c^M = 50\%$, not only the variance but also the bias of EVT noticeably worsen since the asymptotic results of EVT is no longer appropriate to apply. In fact, when the type I censoring proportion becomes more and more important (i.e. the study is terminated very early), the 80% quantile of the observed event times gets smaller and smaller, and thus not "extreme" enough to provide the adequate tail fitting. At $p_c^M = 60\%$, it is unable to estimate the net benefit in many generated datasets using EVT tool as there is too little excesses for reliably justifying the model. Whereas, the PC estimator persists well regardless of high type I censoring rate. It should be emphasized that the hazards are proportional in this scenario and hence, the underlying assumption of PC method is totally plausible here. Again, the results obtained with 100 and 500 repetitions are very similar, we will thus run only 100 simulations for the remaining scenarios. On the other hand, considering the deterioration of PC along with the value τ in DGP 1, we re-run DGP 2 but set τ to 0.5 instead of 0 as previously. Figure 4.11 clearly indicated that PC is no longer effective with $\tau \neq 0$, it is even much more biased than EVT.

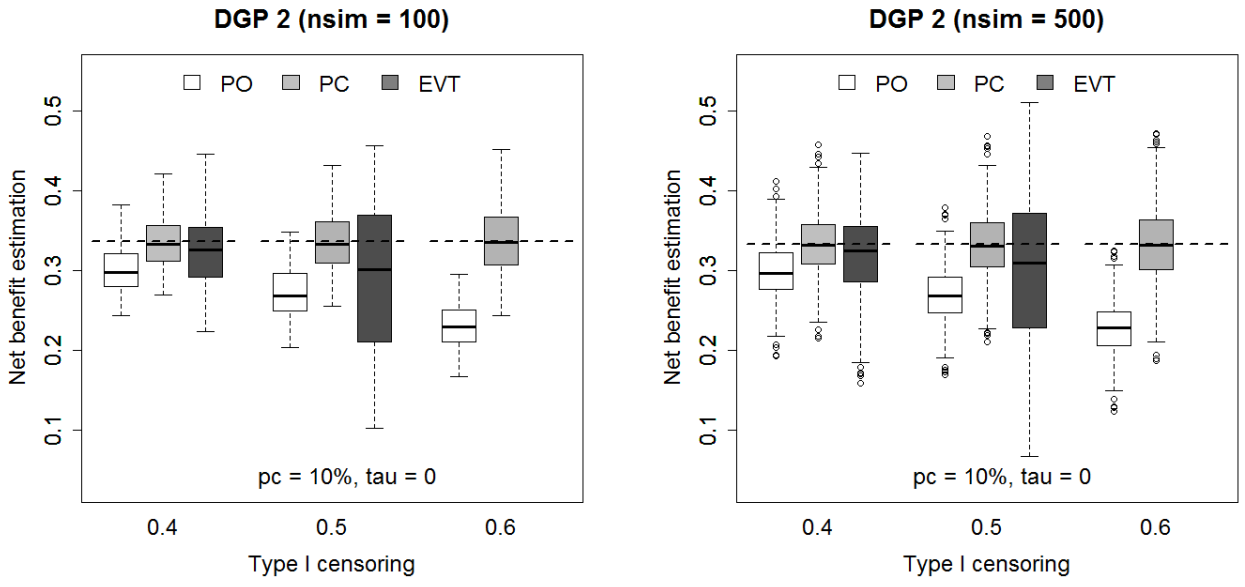

Figure 4.10: Net benefit estimation obtained from 100 and 500 datasets generated in DGP 2. The dashed line represents the true net benefit.

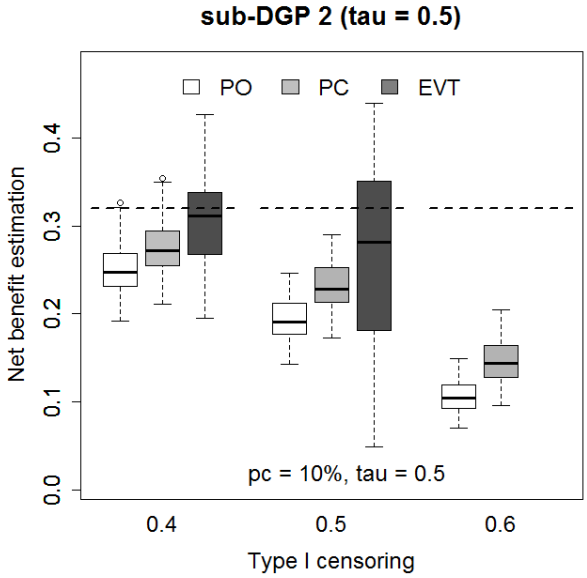

Figure 4.11: Net benefit estimation obtained from 100 dataset generated in the sub-scenario DGP 2 ($\tau = 0.5$ instead of $\tau = 0.5$ as in the main DGP 2). The dashed line represents the true net benefit.

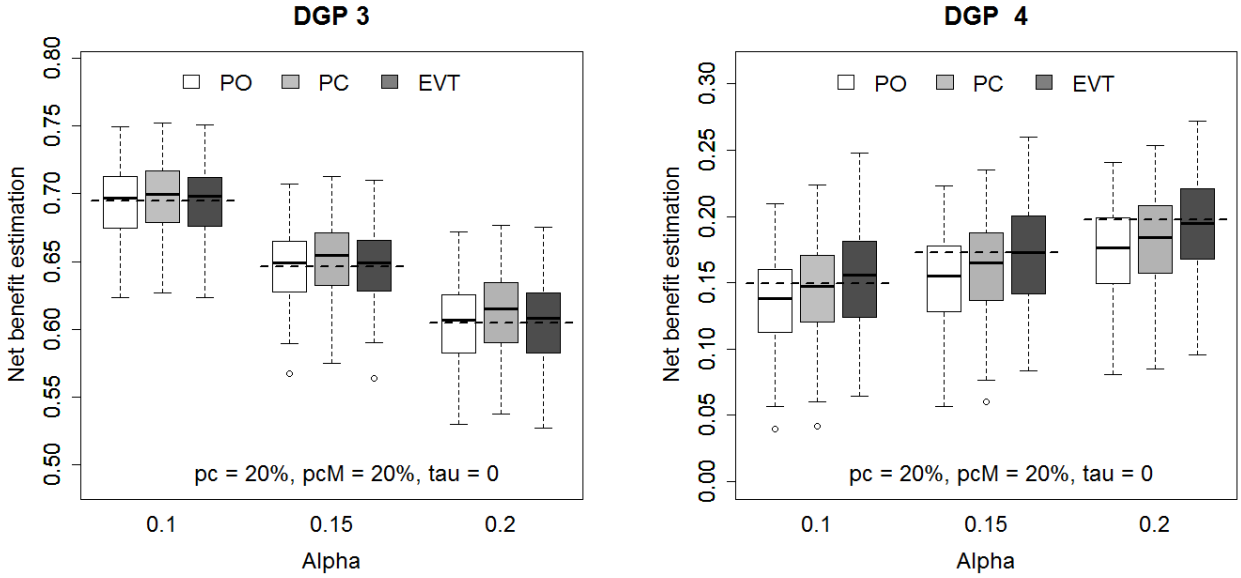

Figure 4.12: Net benefit estimation obtained from 100 datasets generated in DGP 3 (left panel) and DGP 4 (right panel). The dashed line represents the true net benefit.

To evaluate the estimation quality of GPC methods in the context of non proportional hazards, in DGP 3 and DGP 4, we set $p_c = 20\%$, $p_c^M = 20\%$ and $\tau = 0$, which is the setting where PC and EVT procedures perform quite equivalently under the PH condition (Figure

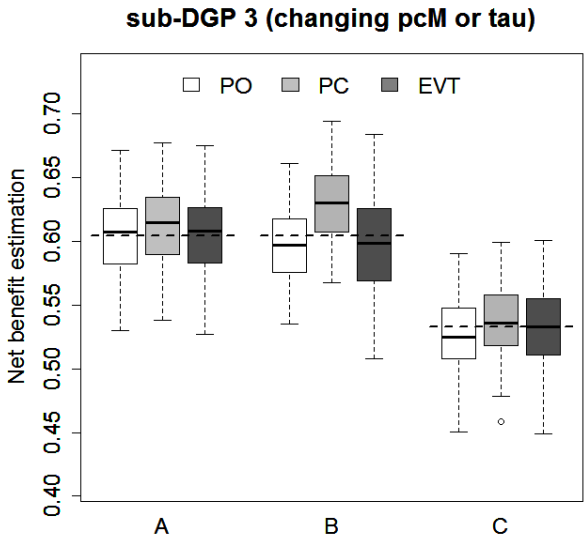

Figure 4.13: Net benefit estimation obtained from 100 datasets generated in the sub-scenarios of DGP 3 ($\alpha = 0.2$): (A) $p_c = 20\%$, $p_c^M = 20\%$ and $\tau = 0$; (B) $p_c = 30\%$, $p_c^M = 30\%$ and $\tau = 0$; (C): $p_c = 20\%$, $p_c^M = 20\%$ and $\tau = 0.8$. The dashed line represents the true net benefit.

4.9). As can be seen in Figure 4.12, EVT tool provides the best net benefit estimate in all the cases; of note, the regularity assumption is plausible in these two scenarios. On contrary, due to the violation of the assumption about the same behavior of informative and uninformative pairs, the net benefit of PC method is consistently overestimated when HR increases overtime and underestimated when HR decreases overtime. This can be explained by the fact that PC procedure implicitly imposes the average of net benefit obtained during the follow-up on that of the later period; whereas, in reality, the treatment effect in long term would be smaller in case of early treatment effect (DGP 3) and better in the case of late treatment effect (DGP 4). The bias becomes more pronounced when we increase the slope α of the hazard function since the quicker the HR changes, the less reasonable the underlying assumption of PC would be. For PO estimation, the result in DGP 4 is totally as expected; i.e. the estimate is shrunk towards 0 due to the presence of censored data. In the case of early treatment effect (DGP 3), surprisingly, the net benefit estimation by PO procedure is almost unbiased. As demonstrated in DGP 1 and DGP 2, the performance of PO estimator is very poor when the clinical threshold is not null or when the censoring rates are high. Hence, we carry out another two sub-scenarios for DGP 3 to verify if the PO estimation can still work well under the tough situations (Figure 4.14). As compared to DGP 3A (i.e. DGP 3 with $\alpha = 0.2$), we apply in DGP 3B higher censoring proportions ($p_c = 30\%$ and $p_c^M = 30\%$) but still keep $\tau = 0$. In DGP 3C, the censoring proportions are unchanged (i.e. $p_c = 20\%$ and $p_c^M = 20\%$) but τ is now set to 0.8. We found that with higher censoring rate, PC estimator clearly gets worse; whereas, PO and EVT estimators are only slightly biased but PO is better than EVT regarding the variance. However, EVT

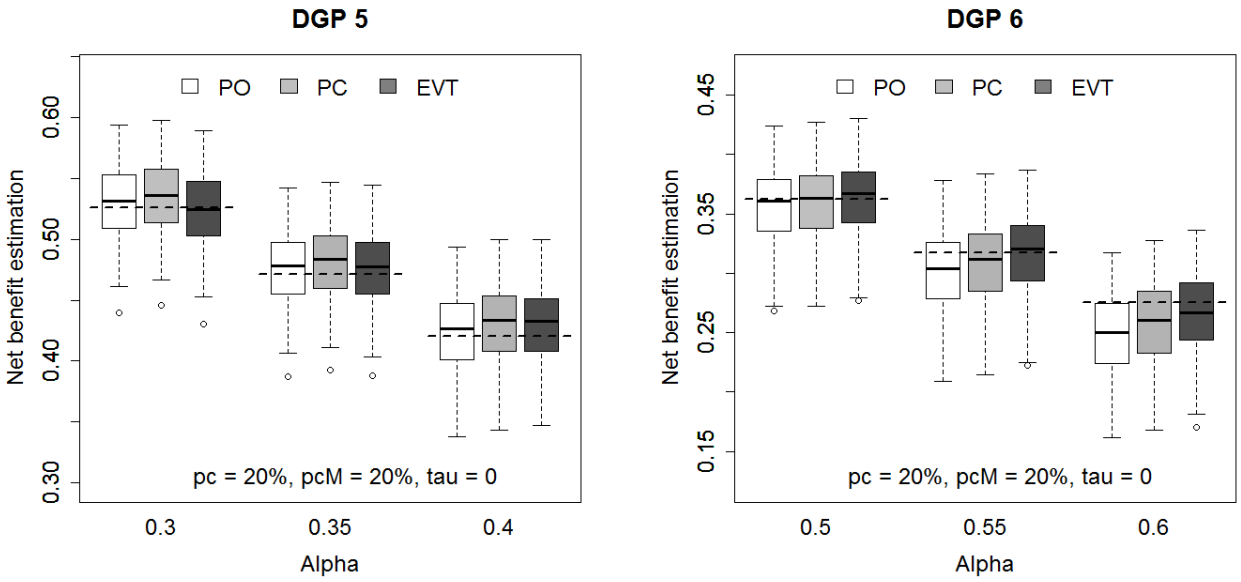

Figure 4.14: Net benefit estimation obtained from 100 datasets generated in DGP 5 (left panel) and DGP 6 (right panel). The dashed line represents the true net benefit.

resists better than PO when a $\tau \neq 0$ is applied, PO is also less effective as compared to itself when $\tau = 0$. Therefore, the performance of PO when treatment effect manifests at early stage seems to be more sensitive to the clinical threshold than to the censoring proportions.

In the next settings (DGP 5 and DGP 6), we focus on the behavior of EVT estimator when its inherent assumption is breached (Figure 4.14). By changing the value α, we move the time t_M and thus change the available information about the tail of the survival function. EVT method relies on some largest observed events to fit the model. Therefore, if the regularity of the unobservable tail distribution does not reach far enough back into the observable region, EVT estimator will be not precise. For instance, in DGP 5 with $\alpha = 0.3$, the extreme values obtained from the study somewhat reflects the behavior of the unobservable survival times and thus the net benefit estimated by EVT is still good. However, when t_M decreases ($\alpha = 0.35$ and $\alpha = 0.4$), EVT estimator starts to be overestimated since the survival part beyond t_M goes down much faster than what can be captured by EVT. Similarly, in DGP 6 with $\alpha = 0.6$, EVT estimator is underestimated as the "true" tail drops slower than what are observable before t_M. The behavior of PO and PC estimators are quite similar as previously shown in DGP 3 and DGP 4. PO performs very well in the case of early treatment effect, it is even better than EVT when the regularity assumption is severely violated. PC is generally the least reliable estimator in non-PH cases. Moreover, it seems that EVT estimator resists quite well even though its assumption is not met. Indeed, EVT is still the most effective in the case of late treatment effect.

The two final scenarios DGP 7 and DGP 8 are added to corroborate the results observed

in the previous DGPs. As can be seen in Figure 4.15, PO estimator consistently shows good property when it comes to the early treatment effect. More interestingly, we are able to observe again the robustness of EVT. Indeed, EVT method provides the unbiased estimate of net benefit in all of the cases and completely outperforms PO and PC procedure, no matter whether the treatment expresses its therapeutic effect early or lately. It should be noted that the regularity assumption is not plausible either in these two settings as the we still use the changing hazard functions; however, the violation is not as severe as in DGP 5 and DGP 6 (see the survival curves in Figure 4.5 to Figure 4.8).

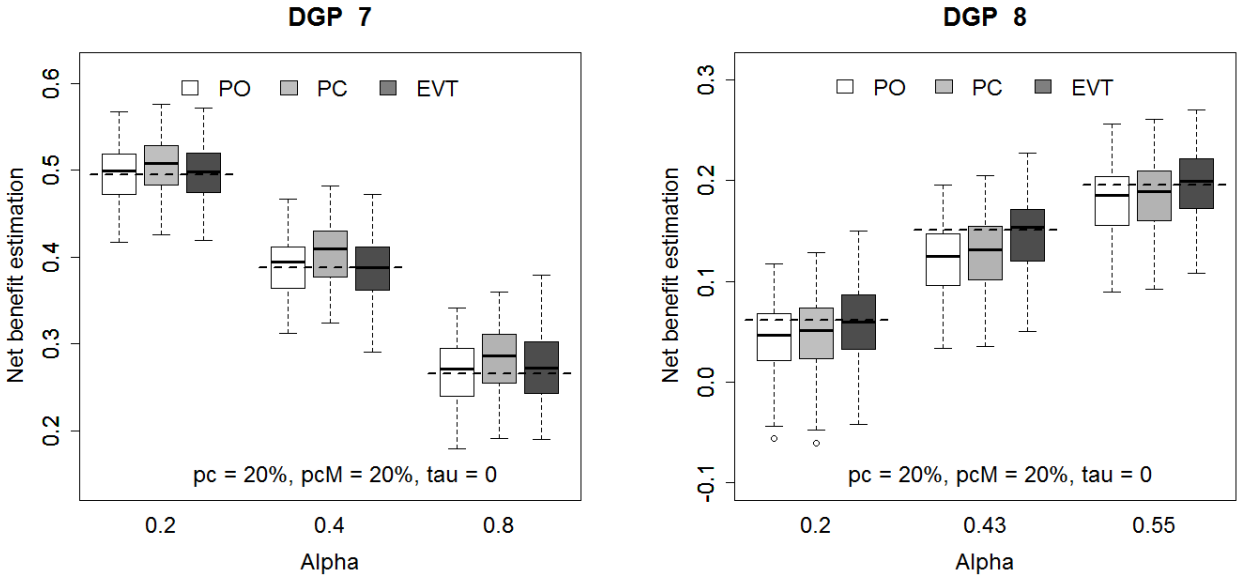

Figure 4.15: Net benefit estimation obtained from 100 datasets generated in DGP 7 (left panel) and DGP 8 (right panel). The dashed line represents the true net benefit.

Finally, as mentioned in Section 3.3, one of the challenges of EVT method is that one must have a sufficient number of fully observed data to reliably estimate the parameters of the tail distribution. Therefore, its performance may no longer be decent if one possesses smaller sample size. To verify this, we re-run DPG 1 (PH case) and DGP 3 (non PH case) with a smaller number of patients in each group; i.e. $n = 100$ instead of $n = 500$. Figure 4.16 reported the results obtained with different sample sizes. The variance of all estimators dramatically increases as we have much fewer pairs to compare. In term of bias, the tendency for the performance of PO and PC as well as EVT are not much influenced by changing sample size. This implies that the EVT estimation for tail distribution depends on the quality of available information rather than the amount of information. For instance, herein we have 100 patients in group Trt and $p_c = 20\%$, $p_c^M = 20\%$; i.e. there are on average 60 fully observed data and thus only 12 observations are used to model the tail; and yet the EVT estimator works pretty well. Whereas, in DGP 2 (Figure 4.2), we have 500 patients in group Trt and $p_c = 10\%$, $p_c^M = 50\%$; i.e. there are 200 fully observed data

and thus about 40 observations are used to approximate the tail but the EVT estimate is noticeably biased. It means that what is more crucial for EVT procedure is how "extreme" the observations are and not how many observations are devoted to the tail modeling. Of course, the estimation will be more precise if one has larger sample size; however, as observed herein, no estimator can do better than other regarding the variance criterion.

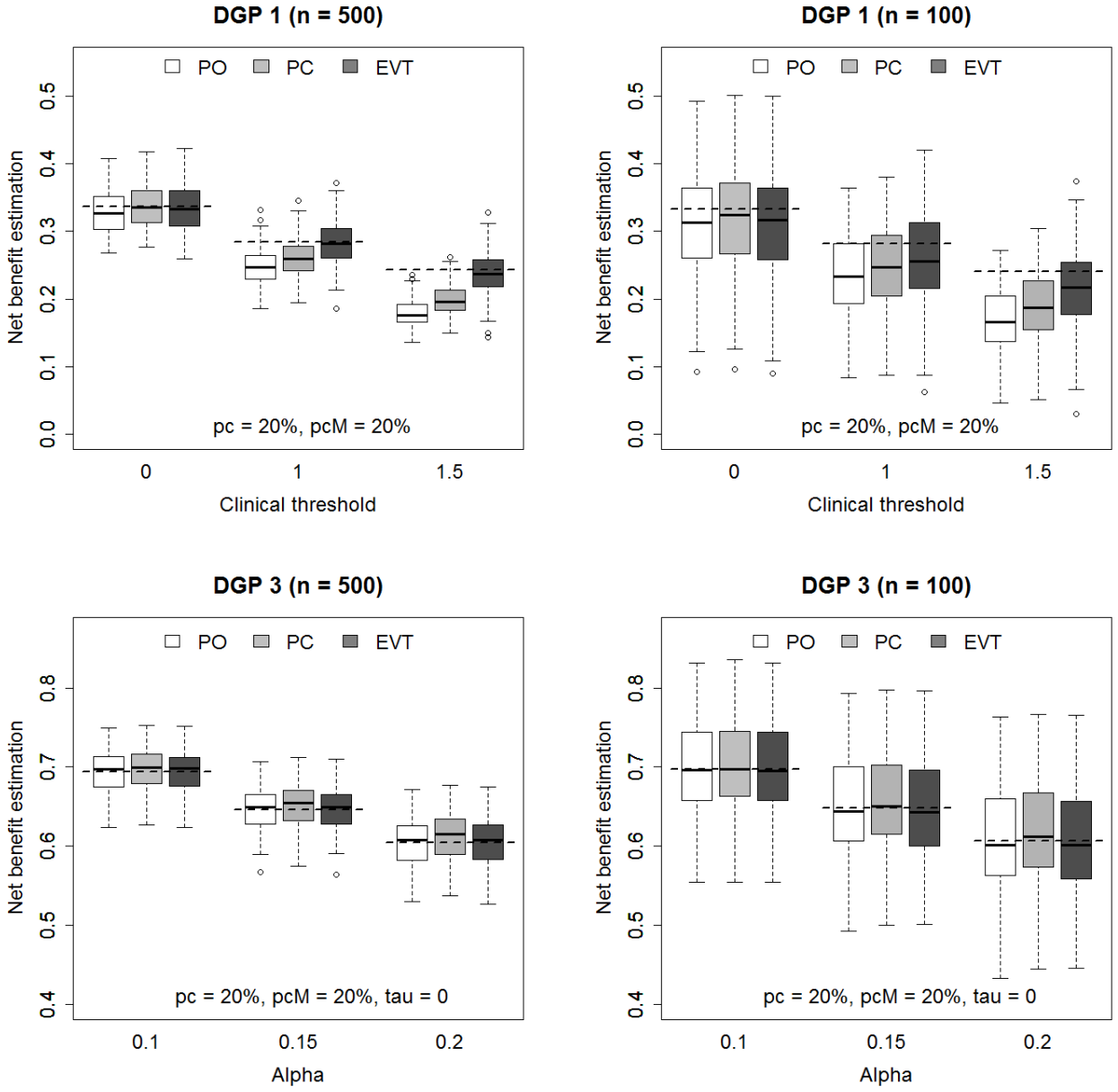

Figure 4.16: GPC estimation for 100 datasets of different sample sizes ($n = 500$ left panel versus $n = 100$ right panel) generated in DGP 1 (top panel) and DGP 3 (bottom panel). The dashed line represents the true net benefit.

Chapter 5

Conclusions and Discussion

With the aim at evaluating the estimation performance of GPC methods for censored data, in this work, we have conducted eight main simulation settings and compared the bias and the variance of net benefit estimated by PO, PC and EVT procedure.

In general, we found that the estimation procedure using EVT tool performs well in most of the cases whether the hazard ratio is constant or not. More importantly, EVT estimator appeared to be quite robust, it is only biased when the regularity assumption is severely breached. And even in such case, the bias of EVT is still less important than that of PO (except for early treatment effect) and PC. In practice, we may inspect the survival curves, if their behavior seems not too much "weird", using EVT will be the least "risky" option as compared to other GPC methods. A careful discussion with clinicians would also be necessary to determine if the survival in long term would probably differs from the observable survival times. Furthermore, the biggest issue of this method is that one must have a sufficient number of "extreme" data to reliably approximate the tail distribution. Therefore, EVT would no longer be decent in the situations where the censoring proportion is too important.

Unlike EVT method, the usefulness of PC procedure is quite limited. It somewhat improves the PO estimation but only under the condition that the assumption about the similar behavior of informative and uninformative pairs is plausible. This is "unverifiable"; however, at least we know that if the hazards are not proportional, PC method will fail and should be excluded from our options right at the beginning. The only scenario where PC outperforms EVT is when the hazard ratio is constant over time and the type I censoring proportion is very high. Nevertheless, it is no longer the case if a $\tau \neq 0$ is applied.

The result obtained for PO estimation is the most unexpected as this method resists almost perfectly in the case of early treatment effect regardless of high censoring rates. It is even better than EVT when the assumption of the latter is not satisfied. However, the quality of PO estimator would drop if a non null clinical threshold is introduced.

In summary, the first crucial step to do before analyzing net benefit would be to verify if the effect of treatment versus control is constant or changing over time. If this is changing, are survival curves likely to evolve weirdly or would they probably behave differently after the study termination? Next, based on the results obtained herein, some recommendations for the choice of GPC methods can be given as follows:

- **Constant treatment effect:**

  - Null clinical threshold: PC could be the method of choice even with very high censoring rate.

  - Non null clinical threshold: EVT could be the most appropriate, provided that the censoring proportion is not too high.

- **Early treatment effect:**

  - Null clinical threshold: PO could be the most effective method.

  - Non null clinical threshold: EVT could be the least risky option if the censoring proportion is not too high and if the regularity assumption is "unlikely" severely violated.

- **Late treatment effect:** EVT could be the optimal tool if the censoring proportion is not too high and if the regularity assumption is "unlikely" severely violated.

In the extreme case where PO, PC and EVT methods are put into question, one should switch to the restricted net benefit. The estimator obtained from this approach does not represent the comparison of the entire survival experience, but no better solution can be achieved if one possesses so poor information. It should be emphasized that the term "high" or "very high" censoring used here is a relative definition; i.e. no decided cut-off values. Within this simulation study, a total censoring proportion of 40% can be considered as "high" and a proportion greater than 50% may be considered as "very high". Nevertheless, as observed through simulations, estimation performance of GPC methods not only depends on the censoring proportions but also on the clinically relevant threshold and notably the intrinsic effect of the treatment. With a certain proportion of censoring, it may be harmful for estimating the net benefit in some situations but may be harmless for other cases. The type I censoring has also more impact on the estimation quality than the classical censoring [32]. Therefore, the choice of GPC method should be done in a flexible manner based on the actual data and taking into account all the relevant factors.

Finally, as EVT appears to be very promising method to deal with censored data, studies to improve the EVT estimation may be of primary interest in the future. It can be done by considering the alternative approaches to gain the accuracy of tail approximation. Besides the maximum likelihood being employed in the present study, several methods for EVT built on peak-over-threshold have been suggested to estimate the parameters of Generalized Pareto Distribution such as PWM (probability weighted moment) [42],

EPM (elemental percentile method) [43] or PML (penalized maximum likelihood) which combines the flexibility of ML estimator and the robustness of PWM estimator [44]. These estimators, however, need to be adapted when applied to censored data. EVT estimation in the context of censorship has not been extensively studied yet, very few papers can be found in the literature up to date [45, 46, 47, 48]. Moreover, it is important to point out that we always used herein the same threshold z which is the 80% quantile of the fully observed events; whereas, the choice of z should be tailored according to the data since the optimal value may vary depending on various factors such as the tail distribution itself, the sample size, the amount of censoring... The automatic approaches for selecting EVT threshold have recently emerged but their performance in the presence of censoring has not been known [49]. Therefore, if one wishes to implement EVT tool for net benefit analysis in practice, how to determine z would also be a crucial task to be addressed in the upcoming research.

Appendices

Table A1: Simulation results of DGP 1 (nsim = 100 and n = m = 500)

$\boldsymbol{\tau}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0	0.3333	PO	0.3277	0.0011	-0.0056	0.0011
		PC	0.3371	0.0011	0.0038	0.0011
		EVT	0.3362	0.0012	0.0029	0.0012
1	0.2817	PO	0.2487	0.0007	-0.0330	0.0018
		PC	0.2617	0.0008	-0.0200	0.0012
		EVT	0.2813	0.0010	-0.0004	0.0010
1.5	0.2405	PO	0.1791	0.0004	-0.0614	0.0041
		PC	0.1995	0.0004	-0.0410	0.0021
		EVT	0.2363	0.0010	-0.0042	0.0010

Table A2: Simulation results of DGP 1 (nsim = 500 and n = m = 500)

$\boldsymbol{\tau}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0	0.3333	PO	0.3244	0.0013	-0.0090	0.0014
		PC	0.3338	0.0014	0.0004	0.0014
		EVT	0.3329	0.0014	-0.0004	0.0014
1	0.2817	PO	0.2466	0.0008	-0.0351	0.0020
		PC	0.2596	0.0009	-0.0221	0.0014
		EVT	0.2789	0.0010	-0.0028	0.0010
1.5	0.2405	PO	0.1771	0.0005	-0.0643	0.0045
		PC	0.1975	0.0006	-0.0431	0.0024
		EVT	0.2339	0.0010	-0.0067	0.0010

Table A3: Simulation results of DGP 1 (nsim = 100 and n = m = 100)

$\boldsymbol{\tau}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0	0.3333	PO	0.3096	0.0067	-0.0237	0.0072
		PC	0.3181	0.0069	-0.0152	0.0070
		EVT	0.3138	0.0068	-0.0195	0.0072
1	0.2817	PO	0.2331	0.0040	-0.0487	0.0064
		PC	0.2463	0.0044	-0.0355	0.0056
		EVT	0.2609	0.0049	-0.0209	0.0052
1.5	0.2405	PO	0.1670	0.0021	-0.0735	0.0075
		PC	0.1868	0.0026	-0.0537	0.0055
		EVT	0.2151	0.0042	-0.0254	0.0048

Table A4: Simulation results of DGP 2 (nsim = 100 and n = m = 500)

$\boldsymbol{p_c^M}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.4	0.3333	PO	0.3025	0.0010	-0.0309	0.0019
		PC	0.3364	0.0012	0.0031	0.0012
		EVT	0.3253	0.0021	-0.0081	0.0022
0.5	0.3333	PO	0.2723	0.0010	-0.0611	0.0047
		PC	0.3359	0.0015	0.0026	0.0015
		EVT	0.2880	0.0098	-0.0453	0.0118
0.6	0.3333	PO	0.2310	0.0009	-0.1023	0.0114
		PC	0.3376	0.0019	0.0043	0.0019
		EVT	N/A	N/A	N/A	N/A

Table A5: Simulation results of DGP 2 (nsim = 500 and n = m = 500)

p_c^M	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.4	0.3333	PO	0.2991	0.0013	-0.0343	0.0024
		PC	0.3329	0.0015	-0.0004	0.0015
		EVT	0.3201	0.0028	-0.0133	0.0030
0.5	0.3333	PO	0.2692	0.0012	-0.0642	0.0053
		PC	0.3324	0.0018	-0.0009	0.0018
		EVT	0.2973	0.0097	-0.0360	0.0109
0.6	0.3333	PO	0.2275	0.0011	-0.1059	0.0123
		PC	0.3326	0.0023	-0.0007	0.0023
		EVT	N/A	N/A	N/A	N/A

Table A6: Simulation results of sub-DGP 2 with $\tau = 0.5$ (nsim = 100 and n = m = 500)

p_c^M	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.4	0.3170	PO	0.2508	0.0008	-0.0662	0.0052
		PC	0.2742	0.0009	-0.0428	0.0027
		EVT	0.3066	0.0023	-0.0107	0.0024
0.5	0.3170	PO	0.1932	0.0006	-0.1239	0.0159
		PC	0.2300	0.0008	-0.0871	0.0084
		EVT	0.2641	0.0108	-0.0529	0.0135
0.6	0.3170	PO	0.1065	0.0004	-0.2106	0.0447
		PC	0.1467	0.0007	-0.1703	0.0296
		EVT	N/A	N/A	N/A	N/A

Table A7: Simulation results of DGP 3 (nsim = 100 and n = m = 500)

$\boldsymbol{\alpha}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.1	0.6951	PO	0.6928	0.000	-0.0022	0.0006
		PC	0.6963	0.0006	0.0013	0.0006
		EVT	0.6938	0.0006	-0.0013	0.0006
0.15	0.6472	PO	0.6450	0.0007	-0.0023	0.0007
		PC	0.6509	0.0007	0.0037	0.0007
		EVT	0.6460	0.0007	-0.0012	0.0007
0.2	0.6052	PO	0.6037	0.0008	-0.0015	0.0008
		PC	0.6118	0.0008	0.0066	0.0008
		EVT	0.6047	0.0008	-0.0005	0.0008

Table A8: Simulation results of sub-DGP 3 with $\alpha = 0.2$ (nsim = 100 and n = m = 500)

Sub-scenario	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
A	0.6052	PO	0.6037	0.0008	-0.0015	0.0008
		PC	0.6118	0.0008	0.0066	0.0008
		EVT	0.6047	0.0008	0.0005	0.0008
B	0.6052	PO	0.5940	0.0008	-0.0112	0.0009
		PC	0.6281	0.0008	0.0229	0.0013
		EVT	0.5960	0.0015	-0.0092	0.0016
C	0.5330	PO	0.5256	0.0007	-0.0075	0.0007
		PC	0.5362	0.0007	0.0031	0.0007
		EVT	0.5328	0.0007	0.0003	0.0007

Table A9: Simulation results of DGP 3 (nsim = 100 and n = m = 100)

$\boldsymbol{\alpha}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.1	0.6951	PO	0.6948	0.0034	-0.0003	0.0033
		PC	0.6982	0.0033	0.0031	0.0033
		EVT	0.6946	0.0034	-0.0005	0.0033
0.15	0.6472	PO	0.6463	0.0040	-0.0010	0.0039
		PC	0.6518	0.0039	0.0046	0.0038
		EVT	0.6457	0.0039	-0.0015	0.0039
0.2	0.6052	PO	0.6042	0.0046	-0.0010	0.0046
		PC	0.6122	0.0046	0.0070	0.0046
		EVT	0.6032	0.0045	-0.0020	0.0045

Table A10: Simulation results of DGP 4 (nsim = 100 and n = m = 500)

$\boldsymbol{\alpha}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.1	0.1492	PO	0.1359	0.0013	-0.0133	0.0014
		PC	0.1446	0.0014	-0.0046	0.0014
		EVT	0.1521	0.0016	0.0028	0.0016
0.15	0.1720	PO	0.1533	0.0013	-0.0187	0.0016
		PC	0.1620	0.0014	-0.0100	0.0015
		EVT	0.1707	0.0016	-0.0013	0.0015
0.2	0.1973	PO	0.1733	0.0013	-0.0240	0.0018
		PC	0.1818	0.0014	-0.0155	0.0016
		EVT	0.1920	0.0014	-0.0053	0.0014

Table A11: Simulation results of DGP 5 (nsim = 100 and n = m = 500)

$\boldsymbol{\alpha}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.3	0.5269	PO	0.5296	0.0009	0.0026	0.0009
		PC	0.5348	0.0009	0.0079	0.0010
		EVT	0.5236	0.0010	-0.0033	0.0010
0.35	0.4722	PO	0.4751	0.0010	0.0029	0.0010
		PC	0.4808	0.0010	0.0086	0.0010
		EVT	0.4746	0.0010	0.0024	0.0010
0.4	0.4211	PO	0.4231	0.0011	0.0019	0.0010
		PC	0.4297	0.0011	0.0086	0.0011
		EVT	0.4294	0.0011	0.0083	0.0011

Table A12: Simulation results of DGP 6 (nsim = 100 and n = m = 500)

$\boldsymbol{\alpha}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.5	0.3624	PO	0.3564	0.0011	-0.0060	0.0011
		PC	0.3599	0.0011	-0.0025	0.0011
		EVT	0.3639	0.0011	0.0015	0.0011
0.55	0.3175	PO	0.3017	0.0012	-0.0158	0.0014
		PC	0.3085	0.0012	-0.0090	0.0013
		EVT	0.3160	0.0012	-0.0016	0.0012
0.6	0.2750	PO	0.2485	0.0012	-0.0265	0.0019
		PC	0.2583	0.0013	-0.0167	0.0016
		EVT	0.2657	0.0013	-0.0093	0.0013

Table A13: Simulation results of DGP 7 (nsim = 100 and n = m = 500)

$\boldsymbol{\alpha}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.2	0.4956	PO	0.4941	0.0009	-0.0015	0.0009
		PC	0.5043	0.0010	0.0088	0.0010
		EVT	0.4970	0.0010	0.0014	0.0010
0.4	0.3886	PO	0.3893	0.0011	0.0007	0.0011
		PC	0.4047	0.0011	0.0162	0.0014
		EVT	0.3861	0.0012	-0.0025	0.0012
0.8	0.2659	PO	0.2682	0.0012	0.0022	0.0012
		PC	0.2843	0.0013	0.0184	0.0016
		EVT	0.2721	0.0016	0.0062	0.0016

Table A14: Simulation results of DGP 8 (nsim = 100 and n = m = 500)

$\boldsymbol{\alpha}$	Δ	**GPC method**	**Mean**($\widehat{\Delta}$)	**Var**($\widehat{\Delta}$)	**Bias**($\widehat{\Delta}$)	**MSE**($\widehat{\Delta}$)
0.2	0.0601	PO	0.0436	0.0012	-0.0165	0.0015
		PC	0.0474	0.0015	-0.0127	0.0016
		EVT	0.0591	0.0017	-0.0010	0.0017
0.43	0.1507	PO	0.1224	0.0013	-0.0283	0.0020
		PC	0.1285	0.0014	-0.0222	0.0019
		EVT	0.1475	0.0014	-0.0031	0.0014
0.55	0.1954	PO	0.1812	0.0013	-0.0142	0.0015
		PC	0.1859	0.0013	-0.0096	0.0014
		EVT	0.1984	0.0013	0.0030	0.0013

Bibliography

[1] Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. Journal of clinical epidemiology 63: e1-e37 (2010).

[2] Royston P and Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Medical Research Methodology 13: 152 (2013).

[3] Saad ED, Zalcberg JR, Péron J, et al. Understanding and communicating measures of treatment effect on survival: can we do better? Journal of the National Cancer Institute 110: 232–240 (2017).

[4] Péron J, Roy P, Ding K, et al. Assessing the benefit-risk of new treatments using generalised pairwise comparisons: the case of erlotinib in pancreatic cancer. British Journal of Cancer 112: 971–976 (2015).

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

[6] Péron J, Roy P, Conroy T, et al. An assessment of the benefit-risk balance of FOLFIRINOX in metastatic pancreatic adenocarcinoma. Oncotarget 7: 82953–82960 (2016).

[7] https://clinicaltrials.gov

[8] Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials (3rd Ed). Springer Science & Business Media, New York (1998).

[9] Robert A. Statistiques des essais cliniques. Syllabus (2008).

[10] Klein JP and Moeschberger ML. Survival analysis - Techniques for censored and truncated data (2nd Ed). Springer Statistics for Biology and Health, New York (1997).

[11] Kleinbaum DG and Klein M. Survival analysis - A Self-Learning Text (3rd Ed). Springer Statistics for Biology and Health, New York (1996).

[12] Kaplan EL and Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association 53: 457-481 (1958).

[13] Cox DR. Regression models and life tables. Journal of the Royal Statistical Society, B 34: 187-220 (1972).

[14] Seruga B, Pond GR, Hertz PC, et al. Comparison of absolute benefits of anticancer therapies determined by snapshot and area methods. Annals of Oncol 23: 2977–2982 (2012).

[15] Schnipper LE, Davidson NE, Wollins DS, et al. American Society of Clinical Oncology Statement: A Conceptual Framework to Assess the Value of Cancer Treatment Options. Journal of Clinical Oncology 33: 2563–2577 (2015).

[16] Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. Journal of Clinical Oncology 32: 2380–2385 (2014).

[17] Trinquart L, Jacot J, Conner SC and Porcher R. Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. Journal of Clinical Oncology 34: 1813–1819 (2016).

[18] Pocock SJ, Ariti CA, Collier TJ and Wang D. 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 (2012).

[19] Oakes D. On the win-ratio statistic in clinical trials with multiple types of event. Biometrika103: 742-745 (2016).

[20] Finkelstein DM and Schoenfeld DA. Graphing the Win Ratio and its components over time. Statistics in Medecine 38: 53-61 (2019).

[21] Wassertheil-Smoller S and Kim MY. Statistical analysis of clinical trials. Seminar in Nuclear Medecine 40: 357-363 (2010).

[22] Edgington ES and Onghena P. Randomization Tests (4st Ed). Chapman and Hall/CRC, New York (2007).

[23] Good P. Permutation tests. A practical guide to resampling methods for testing hypothesis. Springer Science & Business Media, New York (1994).

[24] Verbeeck J, Ozenne B and Anderson WN. Evaluation of inferential methods for the net benefit and win ratio statistics. Journal of Biopharmaceutical Statistics: 1–18 (2020).

[25] Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika 52: 203–224 (1965).

[26] Freemantle N, Calvert M, Wood J, Eastaugh J and Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? Journal of the American Medical Association 289: 2554–2559 (2003).

[27] Johnston S, Pippen J, Pivot X, et al. Lapatinib combined with letrozole versus letrozole and placebo as first-line therapy for postmenopausal hormone receptor-positive metastatic breast cancer. Journal of Clinical Oncology 27: 5538–5546 (2009).

[28] Maio M, Grob JJ, Aamdal S, et al. Five-year survival rates for treatment-naive patients with advanced melanoma who received ipilimumab plus dacarbazine in a phase III trial. Journal of Clinical Oncology 33: 1191–1196 (2015).

[29] Buyse M. Reformulating the hazard ratio to enhance communication with clinical investigators. Clinical Trials 5: 641–642 (2008).

[30] Peron J, Buyse M, Ozenne B, Roche L and Roy P. An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. Statistical Methods in Medical Research 27: 1230–1239 (2018).

[31] Peron J, Idlhaj J, Maucort-Boulch D, et al. Unbiased estimate of the net survival benefit in the presence of censored observations. Statistics in Medicine (under revision).

[32] De Backer M. The use of Extreme Value Theory for GPC estimation with censoring. Report (2019).

[33] Pickands J. Statistical Inference Using Extreme Order Statistics. The Annals of Statistics 3: 119–131 (1975).

[34] Balkema AA and Haan L de. Residual Life Time at Great Age. The Annals of Probability 2: 792–804 (1974).

[35] Embrechts P, Klüppelberg C and Mikosch T. Modelling Extremal Events for Insurance and Finance. Springer-Verlag, Berlin (1997).

[36] Haan L de and Ferreira A. Extreme Value Theory: An introduction. Springer Series in Operations Research and Financial Engineering, New York (2006).

[37] Alvarez-Iglesias A, Newell J, Scarrott C and Hinde J. Summarising censored survival data using the mean residual life function. Statistics in Medicine 34: 1965–1976 (2015).

[38] Kicinski M. Restricted generalized pairwise comparisons. Report (2018).

[39] Wittes J. Sample size calculations for randomized controlled trials. Epidemiology Review 24: 39–53 (2002).

[40] Bender R, Augustin T and Blettner M. Generating survival time to simulate Cox proportional hazards models. Statistics in Medicine 24:1713-1723 (2005).

[41] https://github.com/bozenne/BuyseTest

[42] Hosking JRM and Wallis JR. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics 29: 339-349 (1987).

[43] Castillo E and Hadi A. Fitting the generalized Pareto distribution to data. Journal of the American Statistical Association 92: 1609-1620 (1997).

[44] Coles SG and Dixon MJ. Likelihood-based inference for extreme value models. Extremes 2: 5-23 (1999).

[45] Einmahl J, Fils-Villetard A and Guillou A. Statistics of extremes under random censoring. Bernoulli 14: 207-227 (2008).

[46] Gomes MI and Neves MM. Estimation of the extreme value index for randomly censored data. Biometrical Letters 48: 1-22 (2011)

[47] Worms J and Worms R. New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. Extremes 17: 337-358 (2014).

[48] Beirlant J, Worms J and Worms R. Estimation of the extreme value index in a censorship framework: Asymptotic and finite sample behavior. Journal of Statistical Planning and Inference 202: 31–56 (2019).

[49] Scarrott C and MacDonald A. A review of extreme value threshold estimation and uncertainty quantification. Revstat Statistical Journal 10: 33–60 (2012).

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**