Master's Thesis

Resampling-based Inference for Restricted Mean Survival Times

Author:
David Jesse

Supervisors:
Prof. Dr. Tim Friede
Dr. Cynthia Huber

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science (M.Sc.) in Applied Statistics

University Medical Center Göttingen
Department of Medical Statistics

May 2024

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ECOG	Eastern Cooperative Oncology Group
GEE	Generalized Estimating Equations
GLM	Generalized Linear Model
GWDG	Gesellschaft für wissenschaftliche Datenverarbeitung Goettingen
HC	Heteroscedasticity-consistent
HR	Hazard Ratio
i.i.d.	independent and identically distributed
IJ	Infinitesimal Jackknife
IPCW	Inverse Probability of Censoring Weighting
NPH	Non-proportional Hazards
PD-L1	Programmed Death Ligand 1
PH	Proportional Hazards
RMST	Restricted Mean Survival Time
RNG	Random Number Generator

1 Introduction

In clinical trials, often the outcome of interest is the time from some well-defined time origin until a certain type of event occurs, e.g. the time from randomization to a treatment until death. Such type of data is called *time-to-event* or *survival* data and has the special property that, typically, the event times of some subjects are only incompletely observed. For instance, the study might end before an individual has actually experienced the event of interest. This phenomenon is called *censoring* and requires tailored statistical methods that are able to deal with it. This is one reason why survival analysis has evolved into an own strand of research within statistics (Collett 2015).

Many of the most commonly used statistical methods for effect estimation and hypothesis testing for survival data have a foundation on the *proportional hazards* (PH) assumption and the *hazard ratio* (HR) as the effect size of interest. For instance, the Cox model (Cox 1972) directly makes this assumption and outputs an estimate of the HR. The log-rank test (Peto and Peto 1972), on the other hand, does not explicitly assume PH but loses power when this assumption is violated, i.e. in scenarios with *non-proportional hazards* (NPH). Such situations occur frequently in modern clinical trials with time-to-event endpoints (Dormuth et al. 2022). For instance, in oncology trials the comparison of treatments with different effect mechanisms often results in delayed treatment effect patterns where the survival curves separate or cross each other at some later point in time (Bardo et al. 2024). Yet, the violation of the PH assumption might also be more subtle. The consequences of such a violation are that statistical testing procedures can suffer from a loss of power and that the interpretation of the effect estimate – the hazard ratio – becomes ambiguous. Therefore, it would be desirable to have valid statistical methods at hand that make fewer assumptions and yield interpretable results.

Statistical methods for analyzing time-to-event data under non-proportional hazards can be categorized into two classes (Bardo et al. 2024): The first class focuses on hypothesis testing and mainly aims at robustifying the otherwise commonly used log-rank test against different kinds of NPH alternatives (Royston and B. Parmar 2020). In contrast, other approaches pay more attention to the selection of an appropriate population-level summary measure, which remains interpretable under both, proportional and non-proportional hazards. The choice of a corresponding statistical model or testing procedure then follows in a subsequent step based on the selected summary measure (Quartagno et al. 2023). One of such summary measures, which has gained a lot of popularity in recent years, is the *restricted mean survival time* (RMST). Loosely, the RMST is defined as the area under the (estimated) survival curve from time 0 up to some (pre-)specified time point t^*, which corresponds to the expected survival time within this observation window (Royston and Parmar 2011). An effect measure can

then be constructed by considering different two-sample contrasts such as the difference or the ratio of the RMSTs (Uno et al. 2014).

For the RMST, there exists an analytical point estimator based on the Kaplan-Meier estimator of the survival function as well as different estimators of the associated standard error. From these estimators, Wald-type tests and confidence intervals can be constructed (Hasegawa et al. 2020). However, these methods rely on asymptotic theory and have been shown to suffer from an inflated type I error rate when dealing with small to moderate sample sizes (Horiguchi and Uno 2020). For this reason, Horiguchi and Uno (2020) have developed a permutation test, which in their simulation study outperformed the standard asymptotic test with respect to controlling the type I error. Nonetheless, this method still has some shortcomings as pointed out by Ditzhaus et al. (2023). First, they note that classical permutation tests rely on the exchangeability assumption, which in the context of survival analysis translates to equal distribution functions of both, survival and censoring times, across the groups being compared. This property of the permutation test has not been challenged in the simulation study by Horiguchi and Uno (2020) as for the assessment of the type I error rate both distribution functions have been considered to be equal. Second, as mentioned by Horiguchi and Uno (2020) themselves, their permutation test cannot be used for constructing analogous confidence intervals. Therefore, Ditzhaus et al. (2023) have taken up on this idea and created a studentized permutation test for RMST contrasts, which should solve both aforementioned issues. In a simulation study, the authors demonstrate that their proposed method succeeds in doing so.

Another approach for estimating and testing RMST contrasts is based on so-called *pseudo-observations* (Andersen and Pohar Perme 2010). The main motivation of this approach is to circumvent the need to deal with censored observations and apply standard regression techniques to answer a given research question. This makes it possible to directly evaluate covariate effects on arbitrary survival quantities without the need to convert an adjusted hazard regression model to the scale of that quantity and thereby estimating the effects indirectly. It does this by calculating pseudo-observations of the survival quantity of interest for each subject in a first step and estimating a generalized linear model with these pseudo-observations as the response variable in a second step. For conducting statistical inference based on this model it has been suggested to use a sandwich-type estimator for the covariance matrix of the regression coefficients as the pseudo-observations cannot be considered to be independent and identically distributed in finite-sample settings (Andersen 2003). The usage of pseudo-observations in survival analysis has previously been proposed, in particular in the context of restricted mean survival times (Royston and Parmar 2011; Andersen et al. 2017; Ambrogi et al. 2022). However, to the best of our knowledge, it has not been investigated with a special emphasis on its performance under small to moderate sample sizes. Considering that

the usage of sandwich-type estimators has been shown to have good small sample properties in other contexts (e.g. Long and Ervin 2000) we hypothesize that this could also be the case for estimating RMST contrasts using pseudo-observations. Because of that, we propose the application of pseudo-observations methods for estimating and testing RMST differences in scenarios with small to moderate sample sizes and want to investigate their operating characteristics in this thesis. If such approaches do have a satisfactory performance and can compete with the studentized permutation method there are some potential advantages to them inherited from the flexibility of generalized linear models. For instance, if we wanted to evaluate the effect of a continuous covariate instead of a categorical one this could easily be achieved using pseudo-observations but would not be possible in the same way using the studentized permutation method by Ditzhaus et al. (2023). Even if a potential continuous covariate was not of central interest itself but should only be adjusted for in the analysis it would not be immediately clear how to proceed. Although these aspects will not be the central focus of this thesis they serve as one motivation for investigating these approaches for the problem of handling small sample sizes in the first place.

In summary, we pursue two goals in this thesis: First, we want to replicate the results from the simulation study by Ditzhaus et al. (2023) regarding the asymptotic and their proposed studentized permutation test for two-sample RMST differences in the sense that we can draw the same conclusions from it as they did. Second, we want to amplify this benchmark by proposing two further approaches based on pseudo-observations. The first of these two approaches employs an asymptotic test for conducting inference based on pseudo-observation regression models using a sandwich-type covariance matrix estimator. With the second approach we try to refine this method by replacing the asymptotic test with a nonparametric bootstrap test.

With these goals in mind, the remainder of this thesis is structured as follows: In Section 2 we first introduce basic concepts and notations used in survival analysis. Furthermore, we briefly elucidate the proportional hazards assumption and the hazard ratio as corresponding effect measure as well as statistical methods employing these concepts. For situations in which this assumption is not met we introduce the difference in restricted mean survival times as an alternative estimand for the hazard ratio. Next, we present different methods for conducting statistical inference based on the RMST in Section 3. Here, we start by explicitly formulating the statistical testing problem and then continue by explaining all methods under investigation in this thesis in detail. These methods include the asymptotic test, the studentized permutation test as well as the two aforementioned methods based on pseudo-observations. Following this, we present the design and results of our simulation study benchmarking these methods in Section 4. To complement the simulation study with practical examples we apply these methods on some real-world data sets in Section 5. This also has the purpose of

appraising the potential advantages of using pseudo-observation approaches. Finally, we conclude our main findings and discuss some potential ideas for further research in Section 6.

2 Survival Analysis under Non-proportional Hazards

In this section, we present some fundamental concepts and quantities used in survival analysis and in doing so introduce the basic notation used throughout this thesis, for which we orientate ourselves towards Collett (2015, Chapter 1) and Bardo et al. (2024). Moreover, a special emphasis is placed on the prevalence of the proportional hazards assumption in survival analysis. Eventually, we introduce the restricted mean survival time as a model-free summary measure that can be used for analyzing survival data under non-proportional hazards.

2.1 Notation and Basic Concepts

Let $T \in \mathbb{R}^+$ denote a random variable for the time to event and $C \in \mathbb{R}^+$ a second random variable for the censoring time. In real-world applications, T and C are never observed simultaneously. Instead, for an individual i we only observe $t_i = \min(T_i, C_i)$. Furthermore, we record if the event has been observed for an individual or is censored using the event indicator $\delta_i = \mathbf{1}\{T_i \leq C_i\}$. Thus, ignoring potential covariates, the time-to-event outcome of a single observation consists of the tuple (t_i, δ_i). This thesis focuses on two-sample settings in clinical trials. For this reason, we introduce the treatment indicator variable $Z \in \{0, 1\}$ where $Z = 1$ denotes the experimental and $Z = 0$ the control treatment group. Alternatively, the group associations may also be indicated using the subscript j with the same attributions as for Z. Lastly, $\mathbf{x}_i \in \mathbb{R}^p$ is used for denoting a vector of p covariates associated with an individual i. Besides the treatment assignment, this vector might record other characteristics of interest such as the age of a patient. Correspondingly, $\mathbf{X} \in \mathbb{R}^{n \times p}$ denotes the design matrix for a whole sample of size n.[1]

The random variable T can be characterized by its *distribution function*

$$F(t) = \mathrm{P}(T \leq t) = \int_0^t f(u)\, du\,, \tag{1}$$

with f being the *density function* of that random variable. The distribution function has the interpretation as the probability of the random survival time T being less than

[1]The length (number of columns) of the vector (design matrix) may be increased by one in the context of regression modelling for the inclusion of an intercept term.

or equal to some value t and is therefore also referred to as the *cumulative incidence function* in the context of survival analysis. Often, however, the interpretation of the *survival function* appears to be more intuitive, which is just one minus the distribution function, i.e.

$$S(t) = 1 - F(t) = \mathrm{P}(T > t)\,. \tag{2}$$

Hence, it describes the probability of surviving beyond some time point t. Another important function used in survival analysis is the *hazard function*. Its formal definition is

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}\,, \tag{3}$$

where Δt denotes an increment of time. Since we are considering the limit $\Delta t \to 0$ the hazard function can be understood as the instantaneous failure rate at some time point t and is therefore also called the *hazard rate*. From the hazard rate, the *cumulative hazard function* can be derived:

$$H(t) = \int_0^t h(u)\,du \tag{4}$$

The cumulative hazard function can be interpreted as the "cumulative risk of an event by time t" (Collett 2015, Chapter 1). Among practitioners, the interpretation of the latter two quantities is often unclear and can lead to confusion. Their usefulness, however, lies in their interrelation with $f(t)$, $F(t)$ and $S(t)$, respectively. Using standard results from probability theory, the following results can be obtained, among others:

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ S(t) &= \exp(-H(t)) \end{aligned}$$

Accordingly, if one of the aforementioned functions is given all of the remaining ones can be derived from it. Many of the methods employed in survival analysis are actually built upon the hazard function. Its relevance shall be further underlined in the following subsection.

2.2 Non-proportional Hazards

We now consider a setting with two independent populations, e.g. one coming from a control ($j = 0$) and the other coming from an experimental ($j = 1$) treatment group in a clinical trial:

$$T_j \sim S_j \quad (j = 0, 1)$$

As pointed out in the previous subsection, the distributional assumptions could also be expressed in terms of the hazard functions. Given these, we can define the relative

hazard between the experimental and the control group, more commonly known as the *hazard ratio*:

$$HR(t) = \frac{h_1(t)}{h_0(t)} \quad \text{for } h_0(t) > 0 \tag{5}$$

Using this general definition, it is clear that the hazard ratio is a function of time. However, many of the standard statistical methods used in survival analysis are based on the assumption that the hazards are proportional over time, implying that the hazard ratio is independent of it:

$$HR(t) = \frac{h_1(t)}{h_0(t)} = \theta \quad \forall\, t \tag{6}$$

Here, $\theta \in (0, \infty)$ is some constant highlighting this assumption, which is known as the *proportional hazards* (PH) assumption. This assumption may or may not hold, which can be illustrated using the Weibull distribution as a theoretical example. The hazard function of the Weibull distribution is (Kalbfleisch and Prentice 2002, Chapter 2.2)

$$h(t) = \lambda\alpha(\lambda t)^{\alpha-1}\,, \tag{7}$$

with *shape* parameter α and *scale* parameter λ. For the hazard ratio, we then have:

$$\begin{aligned}\frac{h_1(t)}{h_0(t)} &= \frac{\lambda_1\alpha_1(\lambda_1 t)^{\alpha_1-1}}{\lambda_0\alpha_0(\lambda_0 t)^{\alpha_0-1}} \\ &= \frac{\lambda_1^{\alpha_1}\alpha_1}{\lambda_0^{\alpha_0}\alpha_0} \cdot \frac{t^{\alpha_1-1}}{t^{\alpha_0-1}}\end{aligned}$$

Here, we can see that the Weibull distribution exhibits proportional hazards if the shape parameter is the same for both populations, i.e. $\alpha_0 = \alpha_1 = \alpha$. In this case, the hazard ratio is given by $(\lambda_1/\lambda_0)^\alpha$. If this is not the case, however, the hazards are non-proportional. Figure 1 illustrates those two situations, depicting the survival and hazard functions as well as the hazard ratio of the two groups with differently parametrized Weibull distributions.

We can also use real-world data sets to contrast situations with hazards that are rather proportional or non-proportional over time. Such examples are given in Figure 2 where the estimated survival and hazard functions are shown. The former are estimated using the nonparametric Kaplan-Meier estimator of the survival function, while the latter are derived from fitting flexible parametric models (Royston and Parmar 2002) with $(3, 2)$ degrees of freedom. Moreover, the time-varying hazard ratio derived from these estimates of the hazard functions is portrayed in the bottom row of the plot. In addition to that, the dashed line represents the estimate of the time-independent hazard ratio obtained from the Cox model (Cox 1972).

In the left column, we use the "diabetic" data set from the `{survival}` R package

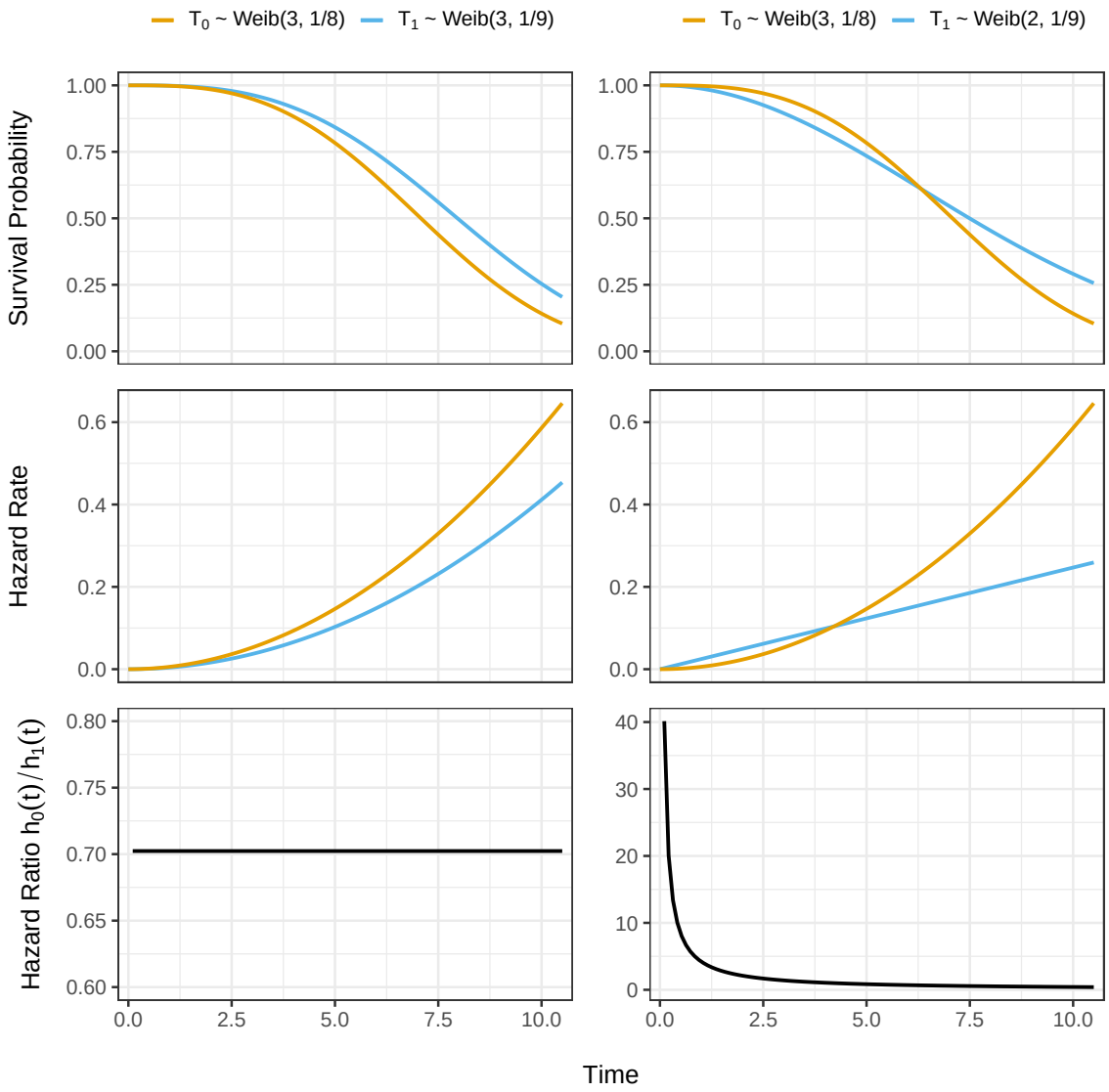

Figure 1 Weibull survival models of two populations with different parameter values. In the left column the shape parameters of the two populations are equal, resulting in a scenario with proportional hazards. In the right column the shape parameters differ, implying non-proportional hazards.

(Therneau 2023), which contains results from a trial of laser coagulation for the treatment of diabetic retinopathy. We can see that the survival curves do not cross a single time and diverge fairly evenly across time. Accordingly, the estimated hazard functions are proportional to the greatest extent. Although the hazard ratio fluctuates somewhat in the initial phase of the study, it remains within a relatively small range overall, not deviating much from the constant hazard ratio obtained from the Cox model.

On the other hand, the data set presented in the right column is reconstructed from a phase-III randomized controlled trial by Robert et al. (2015). Here, the population under investigation consisted of patients with advanced melanoma. The treatments analyzed and compared with each other were dacarbazine and nivolumab. In comparison to the "diabetic" data set, we can see that the estimated survival functions cross each other in the initial phase of the study multiple times. After around 2.5 months we can observe an effect of the nivolumab therapy leading to a separation of the survival curves. From the profile of the estimated hazard functions, we can also conclude that there is less support for assuming proportional hazards. While the hazard function of the dacarbazine group monotonically increases over time, we can see that, for the nivolumab group, it increases in the first 2.5 months but decreases afterward. Also, the curve of the time-dependent hazard ratio exhibits greater variability.

Among the statistical models that employ the proportional hazards assumption, perhaps the best-known and most widely used one is the previously mentioned semiparametric *Cox proportional hazards model* (Cox 1972):

$$h(t \mid Z) = \exp(\beta Z) h_0(t) \tag{8}$$

The Cox model is a regression model for covariate effects on the hazard scale. Here, we employ our previous notation, $h_0(t)$ being the hazard function of the control group and Z being the treatment indicator with $Z = 1$ indicating an allocation to the experimental treatment group. In this context, $\exp(\beta)$ can be interpreted as the hazard ratio as it has been defined in (6). Nonetheless, the Cox model (8) can be extended to the more general case of multiple covariates $\mathbf{x}$ of different types (categorical, continuous) and corresponding regression coefficients $\boldsymbol{\beta}$. The interpretation of $h_0(t)$ would then also change to a more generic baseline hazard not solely attributed to the control treatment group. The main reason for the popularity of the Cox model is its semiparametric nature: While an estimate of the hazard ratio ($\exp(\beta)$) is obtained, the baseline hazard $h_0(t)$ does not need to be specified or even estimated at all. This sets it apart from many parametric alternatives, such as accelerated failure time models, for which a strict assumption about the conditional distribution of T needs to be made. Since deciding on a plausible parametric distribution for time-to-event data is a difficult task in many practical applications, especially when this decision needs to be made a priori, the Cox

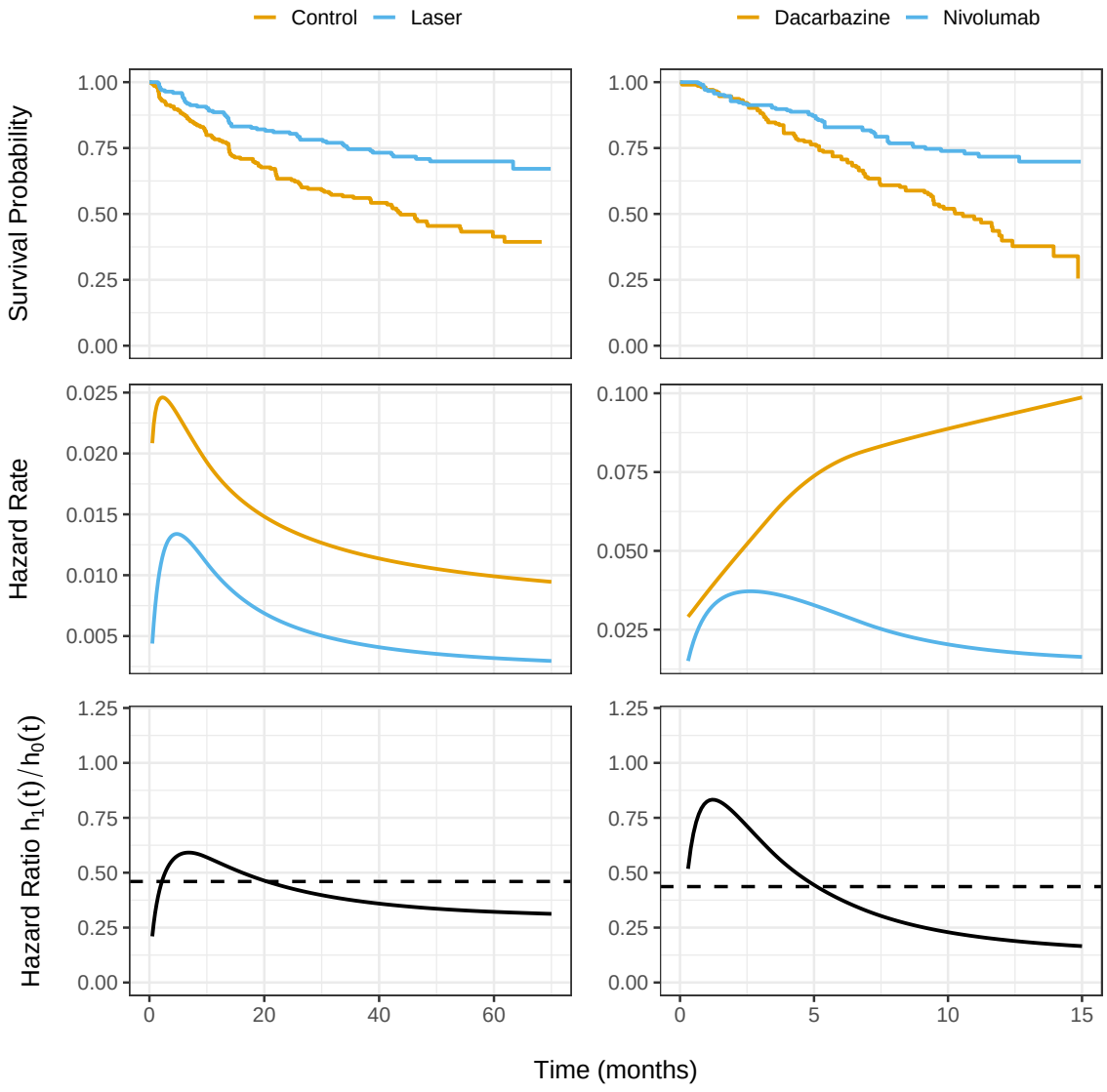

Figure 2 Estimated survival functions for the R "diabetic" data set (left) and data reconstructed from Robert et al. (2015) (right). For the former the proportional hazards assumption can be considered plasuible, whereas the latter provides less support for making this assumption.

model and the hazard ratio have become the default choice for effect estimation in these settings. Nonetheless, the strong assumption of proportional hazards persists, which is why some authors have become more skeptical about its usage for effect estimation (Royston and Parmar 2011).

Another routine method used in survival analysis is the log-rank test (Peto and Peto 1972). It is a non-parametric test for the hypotheses

$$H_0 : S_1(t) = S_0(t) \text{ for all } t \quad \text{vs.} \quad H_1 : S_1(t) \neq S_0(t) \text{ for at least one } t\,. \tag{9}$$

While the log-rank test for the testing problem (9) is valid under both, proportional and non-proportional hazards, it loses power in the case of NPH (Rufibach 2019). However, the bigger issue can be considered the fact that the underlying estimand of the log-rank test is the hazard ratio of the Cox model (Rufibach 2019). Thus, while the log-rank test may still be applied, the interpretation of the associated effect estimate becomes ambiguous.

2.3 Restricted Mean Survival Time

We now address the restricted mean survival time (RMST) and its contrasts as alternative effect measures for the hazard ratio as it is defined in (6). Mathematically, the RMST is the expectation of the transformation $f(T \,|\, t^*) = \min(T,\, t^*)$ of the random variable T where t^* is the restriction time that needs to be specified by the researcher. The RMST can be interpreted as the expected survival time from the time origin 0 up to the restriction time t^* (Royston and Parmar 2011). Using exemplary numbers, Royston and Parmar (2011) provide a clinical interpretation of the RMST: "[Y]our life expectancy with X treatment and Z disease over the next 18 months is 9 months". Moreover, considering a two-sample comparison based on the RMST difference, they further extend this interpretation (Royston and Parmar 2011): "[T]reatment A increases your life expectancy during the next 18 months by 2 months, compared with treatment B." The RMST can be calculated by integrating the survival function from time 0 to t^* (Royston and Parmar 2011), i.e.

$$\mu(t^*) = \mathbb{E}[\min(T,\, t^*)] = \int_0^{t^*} S(u)\, du\,. \tag{10}$$

The reason for considering a restricted version of the mean is right-censoring of the survival data. For the nonparametric Kaplan-Meier estimator of the survival function to reach 0 and therefore be able to estimate an unrestricted mean, the largest event time needs to be uncensored, which is rarely the case. On the other hand, employing parametric models in survival analysis is often a difficult task. Even if we had a model

with a decent in-sample fit, we cannot be sure whether the extrapolation of that model to the tail of the survival distribution is adequate (Royston and Parmar 2011).

Using the definition (10) of the RMST, actual effect measures can be obtained by considering ratios $\mu_1(t^*)/\mu_0(t^*)$ or differences $\mu_1(t^*) - \mu_0(t^*)$ of the RMSTs, e.g. between two treatment groups (Uno et al. 2014). The main advantage of using the RMST and group contrasts thereof in comparison to the prevalent hazard ratio is that it is a model-free effect measure and has a clear clinical interpretation, though this might be subject to discussion (Freidlin and Korn 2019). To further enhance the understanding of the RMST, Figure 3 provides a visual demonstration of it based on survival functions of the exponential distribution. The left panel depicts the situation for a single sample while the right panel illustrates the two-sample RMST difference as an effect measure.

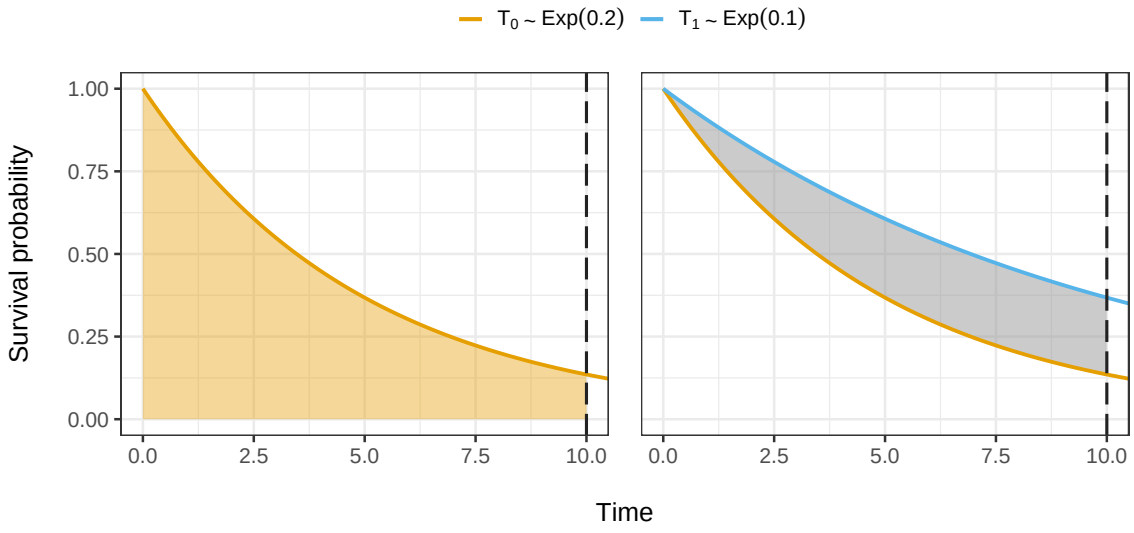

Figure 3 Illustration of the RMST: The colored area under the survival curve (left panel) depicts the one-sample RMST. The grayed area between the two survival curves (right panel) illustrates the RMST difference between two populations.

Now, the question remains how to estimate $\mu(t^*)$ for an observed data set. Broadly, a point estimator for $\mu(t^*)$ can be obtained by integrating an estimate of the survival function:

$$\hat{\mu}(t^*) = \int_0^{t^*} \hat{S}(u)\, du$$

In principle, any estimator $\hat{S}$ can be applied but most commonly the non-parametric Kaplan-Meier estimator (Kaplan and Meier 1958) is used:

$$\hat{S}(t) = \begin{cases} 1 & \text{for } t < t_1 \\ \prod_{t_j \leq t} \left(1 - \frac{d_j}{Y_j}\right) & \text{for } t \geq t_1 \end{cases}$$

Here, t_j $(j = 1, 2, \ldots, m)$ denotes the ordered and distinct event times observed in the survival data. Moreover, Y_j denotes the number of individuals at risk just prior

to t_j and d_j the number of observed events at t_j. The reasons for the usage of the Kaplan-Meier estimator for estimating the RMST are twofold. On the one hand, using a non-parametric estimator avoids making any possibly false assumptions about the shape of the survival curve. On the other hand, using this estimator, closed-form solutions for both, the point and the variance estimator, are available and there is no need to rely on computational methods like numerical integration or resampling. Specifically, the point estimator utilizing the Kaplan-Meier estimator of the survival function can be written as (Hasegawa et al. 2020)

$$\hat{\mu}(t^*) = \sum_{j=0}^{D} (t_{j+1} - t_j)\hat{S}(t_j)\,, \tag{11}$$

where, like before, $t_1 < t_2 < \ldots < t_D$ denote the ordered and distinct event times observed in the survival data and $t_{D+1} = t^*$. An estimator for the variance of the point estimator (11) based on Greenwood's formula is given by (Hasegawa et al. 2020)

$$\widehat{\mathrm{Var}}[\hat{\mu}(t^*)] = \sum_{j=1}^{D} \left[\sum_{i=j}^{D} (t_{i+1} - t_i)\hat{S}(t_i)\right]^2 \frac{d_j}{Y_j(Y_j - d_j)}\,. \tag{12}$$

Alternatively, the variance can be estimated by

$$\widehat{\mathrm{Var}}[\hat{\mu}(t^*)] = \sum_{j=1}^{D} \left[\sum_{i=j}^{D} (t_{i+1} - t_i)\hat{S}(t_i)\right]^2 \frac{d_j}{Y_j^2}\,, \tag{13}$$

which is derived by plugging in the Nelson-Aalen estimator of the cumulative hazard rate into the theoretical variance formula (Ditzhaus et al. 2023). Both variance estimators are valid but since most authors employ the estimator (12) we use it here per default as well.

3 Statistical Inference for Restricted Mean Survival Times

In this section, we present the different statistical methods under investigation for the following testing problem: As indicated in Section 2, we are in the setting of a randomized controlled trial with a time-to-event endpoint and want to compare an experimental treatment ($j = 1$) to a reference or placebo treatment ($j = 0$). The survival and censoring times are assumed to be independent between all subjects and identically distributed within each group, formally

$$T_{ji} \sim S_j, \quad C_{ji} \sim G_j \quad (j = 0, 1;\, i = 1, \ldots, n_j).$$

Here, S_j and G_j denote the survival functions of the survival and censoring times of group j, respectively, and $i \in \{1, \ldots, n_j\}$ indicates the i-th individual within that group. Furthermore, it is to note that T_{ji} and C_{ji} are assumed to be mutually independent, i.e. the survival time does not impact the censoring process or vice versa. As each subject has been randomized to either one of the groups, we do not need to worry about potential confounding variables.

Since we are not willing to assume that the proportional hazards assumption holds and want to have an interpretable estimand available, we construct the statistical analysis based on the difference in restricted mean survival times. For a given restriction time t^*, this leads to the specification of the following pair of hypotheses:

$$H_0 : \mu_1(t^*) - \mu_0(t^*) = 0 \quad \text{vs.} \quad H_1 : \mu_1(t^*) - \mu_0(t^*) \neq 0 \tag{14}$$

In the following subsections, we present the different methods studied in this thesis for conducting such a test.

3.1 Asymptotic Test

In this subsection, we present the asymptotic test for the testing problem (14) following Hasegawa et al. (2020). A Wald-type test statistic can be constructed using the point and variance estimators (11) and (12), respectively:

$$Z = \frac{\hat{\mu}_1(t^*) - \hat{\mu}_0(t^*)}{\sqrt{\widehat{\mathrm{Var}}[\hat{\mu}_1(t^*)] + \widehat{\mathrm{Var}}[\hat{\mu}_0(t^*)]}} \tag{15}$$

Using the martingale central limit theorem (Kalbfleisch and Prentice 2002, Chapter 5.5) in combination with the functional δ-method, it can be shown that, asymptotically, this test statistics follows a standard normal distribution under the null hypothesis (Zhao et al. 2016), i.e.

$$Z \overset{\text{a, } H_0}{\sim} \mathcal{N}(0, 1)\,. \tag{16}$$

Using this property, we can construct a statistical test for a given significance level $\alpha \in (0, 1)$ that is asymptotically valid:

$$\varphi^{(\text{asy})} = \mathbf{1}\{|Z| > z_{1-\alpha/2}\} \tag{17}$$

Here, $z_{1-\alpha/2}$ denotes the $1 - \alpha/2$-quantile of the standard normal distribution. In line with the statistical test (17), a symmetric $1 - \alpha$ confidence interval for the RMST difference can be constructed:

$$CI^{(\text{asy})} = \left[\hat{\mu}_1(t^*) - \hat{\mu}_0(t^*) \mp z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\hat{\mu}_1(t^*)] + \widehat{\mathrm{Var}}[\hat{\mu}_0(t^*)]}\right] \tag{18}$$

3.2 Studentized Permutation Test

Next, we present the studentized permutation test introduced by Ditzhaus et al. (2023) as an alternative to the asymptotic method presented before. In principle, this approach is not exceedingly different compared to the asymptotic one. In fact, it relies on the same type of non-parametric estimators of the RMST and its variance. We should note, however, that Ditzhaus et al. (2023) derive and use the variance estimator (13) instead of (12). What distinguishes their method from the asymptotic test is the fact that it does not make a fixed assumption about the distribution of the test statistic (15) under the null hypothesis but aims at estimating it nonparametrically from the observed data. Adopting the notation by Ditzhaus et al. (2023), we again compute the Wald-type test statistic (15) based on the original data $(\mathbf{t}, \boldsymbol{\delta}) \equiv \{(t_{ji}, \delta_{ji}) : j = 0, 1; i = 1, \ldots, n_j\}$. Now, instead of employing the standard normal approximation (16), the distribution of the test statistic under the null hypothesis is estimated from the data using the concept of random permutations. For this, let $(\mathbf{t}, \boldsymbol{\delta})^\pi \equiv \{(t_{ji}, \delta_{ji})^\pi : j = 0, 1; i = 1, \ldots, n_j\}$ denote a permuted version of the original data, meaning that the treatment indicator Z has been randomly shuffled and reassigned to the individual observations (t_{ji}, δ_{ji}) of the original data. We draw such samples B times (e.g. $B = 5000$) and calculate a test statistic similar to (15) for each of the permutation data sets:

$$Z^\pi = \frac{|\hat{\mu}_1^\pi(t^*) - \hat{\mu}_0^\pi(t^*)|}{\sqrt{\widehat{\mathrm{Var}}[\hat{\mu}_1^\pi(t^*)] + \widehat{\mathrm{Var}}[\hat{\mu}_0^\pi(t^*)]}} \tag{19}$$

We can then carry out the studentized permutation test

$$\varphi^\pi = \mathbf{1}\{|Z| > q_{1-\alpha}^\pi\} \tag{20}$$

where $q_{1-\alpha}^\pi$ denotes $(1-\alpha)$-quantile of the the permutation test statistics. Ditzhaus et al. (2023) show that the distribution of (19) is asymptotically equivalent to that of (15), motivating its usage for constructing a confidence interval similar to (18). Hence, the standard normal quantile $z_{1-\alpha/2}$ simply gets replaced by $q_{1-\alpha}^\pi$:

$$CI^\pi = \left[\hat{\mu}_1(t^*) - \hat{\mu}_0(t^*) \mp q_{1-\alpha}^\pi \sqrt{\widehat{\mathrm{Var}}[\hat{\mu}_1(t^*)] + \widehat{\mathrm{Var}}[\hat{\mu}_0(t^*)]}\right] \tag{21}$$

It should be noted that the testing problem (14) is two-sided and that the studentized permutation test as it is presented here makes use of this fact, exploiting the asymptotic symmetry of the test statistic (15). Nonetheless, this procedure can also be adapted to one-sided testing problems. Moreover, the test and confidence intervals obtained using asymptotic theory and using studentized permutation coincide as $n \to \infty$ (Ditzhaus et al. 2023).

When Horiguchi and Uno (2020) first introduced a permutation test for two-sample

RMST contrasts they extensively discussed how to handle scenarios in which for a permuted data set either $\hat{S}_0(t^*)$ or $\hat{S}_1(t^*)$ is not uniquely defined because the largest event time in that group is smaller than t^* and censored. Following Ditzhaus et al. (2023) and based on the results by Horiguchi and Uno (2020), we use the "horizontal extension" strategy in which we set $\hat{S}_j(t^*) = \hat{S}_j(t_{\max})$ in these cases, where, $t_{\max}$ denotes the largest event time within the respective group.

3.3 Pseudo-observations

3.3.1 General Concept

The idea of using *pseudo-observations* was first introduced by Andersen (2003) for applications to multi-state models and has later been contextualized to more general settings with time-to-event data (Andersen and Pohar Perme 2010). The motivation is to move away from hazard-based regression models for time-to-event data when we are actually interested in obtaining effect estimates for other quantities, e.g. for survival probabilities or restricted mean survival times. In this regard, hazard-based regression models may impede estimation, interpretation and statistical inference for such effects. For instance, if we were interested in the RMST difference between two treatment groups, a hazard-based regression model would need to be converted to the survival function first and then be (numerically) integrated to obtain an estimate of the RMST. Next, the estimation and uncertainty quantification of the (adjusted) RMST difference we are interested in would require further post-estimation steps (Sachs and Gabriel 2022). Moreover, these models might have underlying assumptions we would actually like to avoid such as the proportional hazards assumption. Using pseudo-observations we aim to circumvent such a procedure by estimating a generalized linear model (GLM) of the form

$$\mathbb{E}[V_i \,|\, \mathbf{x}_i] = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\,. \tag{22}$$

Here, $V_i = f(T_i)$ denotes some transformation of the original data reflecting the estimand we are interested in. For instance, $f(T_i \,|\, t^*) = \min(T_i,\, t^*)$ and the expectation thereof corresponds to the restricted mean survival time. The link function g needs to be chosen by the researcher and determines how the covariate effects $\boldsymbol{\beta}$ on $\mathbb{E}[V_i \,|\, \mathbf{x}_i]$ are interpreted. The problem is that V_i cannot be calculated for all observations $i = 1, \ldots, n$ due to right-censoring, therefore prohibiting the direct estimation of the model (22). This is where pseudo-observations become relevant: Instead of using the original time-to-event data as the response we calculate pseudo-observations in the following way (Andersen and Pohar Perme 2010; Sachs and Gabriel 2022):

$$P_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}\,. \tag{23}$$

Here, $\hat{\theta}$ denotes a well-behaved marginal estimator of $\theta = \mathbb{E}[V_i]$. This could be any estimator satisfying asymptotic efficiency but, usually, a nonparametric estimator is used. The advantages of these are that they do not exhibit any modeling assumptions and are usually fast to compute. Likewise, $\hat{\theta}_{-i}$ is the same kind of estimator but leaving out the i-th observation, i.e. using the i-th *jackknife sample* (Efron and Tibshirani 1993, Chapter 11). Following Andersen and Pohar Perme (2010), "the i-th pseudo-observation can be viewed as the contribution of the individual i to the [marginal] estimate $[\hat{\theta}]$ on the sample of size n". Thus, by calculating the pseudo-observations and using them as the response vector, we bypass the need to deal with censored observations. It is important to note, though, that the estimation of the model (22) proceeds using the pseudo-observations (23) for all individuals $i = 1, \ldots, n$, regardless of whether V_i could actually be computed directly or not. One theoretical justification for using pseudo-observations as a response variable in a GLM is their unbiasedness (given an unbiased marginal estimator $\hat{\theta}$) (Andersen and Pohar Perme 2010):

$$\begin{aligned} \mathbb{E}[P_i] &= \mathbb{E}[n\hat{\theta} - (n-1)\hat{\theta}_{-i}] \\ &= n\,\mathbb{E}[\hat{\theta}] - (n-1)\,\mathbb{E}[\hat{\theta}_{-i}] \\ &= n\theta - (n-1)\theta \\ &= \theta \end{aligned} \tag{24}$$

Moreover, it has been shown that the pseudo-observations P_i are asymptotically independent and identically distributed (Andersen and Pohar Perme 2010). However, one issue remains: Using pseudo-observations for the regression model (22) relies on the assumption that the censoring mechanism is independent of any covariates. If this is not the case any estimate obtained from these pseudo-observations will be biased (Andersen and Pohar Perme 2010). For such a situation there are two solutions available: First, if censoring depends on one or more categorical covariates, then the pseudo-observations may be calculated stratified by the different levels of these covariates and their interactions. This way, unbiasedness is retained but in comparison with independent censoring the standard errors will be inflated (Andersen and Pohar Perme 2010). If the covariates that impact the censoring mechanism are continuous, then the pseudo-observations can be calculated by estimating a model for the censoring mechanism and using inverse probability of censoring weighting methods (IPCW) (Binder et al. 2014; Overgaard et al. 2019). The latter method, however, will not be applied in this thesis.

Another important aspect to keep in mind is that the formulation of a classical generalized linear model $\mathbb{E}[Y_i \,|\, \mathbf{x}_i] = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$ is usually motivated and determined by (conditional) distributional assumptions about the response variable y_i (Fahrmeir et al. 2013, Chapter 5.4). For instance, if we had a binary response $y_i \in \{0, 1\}$ it would be natural to presume y_i to be binomially distributed. This would in turn imply the link

function g to be the canonical logit link function (Fahrmeir et al. 2013, Chapter 5.4). However, this approach is not reasonable for the pseudo-observations regression model (22) for two reasons: First, the pseudo-observations are not available per se but must be computed first and a particular probabilistic model would be hard to justify. Second, the motivation for using the regression model (22) in the first place is to obtain effect estimates $\boldsymbol{\beta}$ with a particular interpretation. For instance, we could use the model (22) to regress the covariates $\mathbf{x}_i$ on the survival probability and obtain associated effect estimates. If we wanted to estimate differences in survival probabilities between two treatment groups, the link function g would need to be specified as the identity link function (Sachs and Gabriel 2022) although the estimation of a survival *probability* might make the logit link function appear more intuitive. Due to this special situation, we need to employ quasi-likelihood methods instead of ordinary maximum-likelihood theory (Fahrmeir et al. 2013, Chapter 5.5). Hence, we avoid making a proper distributional assumption about the response variable but only need to make a correct specification of the expectation structure (22) together with a *working variance* structure σ_i^2. From this specification we can derive the quasi-score function, also known as *generalized estimating equation* (GEE)

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \frac{h'(\eta_i)}{\sigma_i^2}(y_i - \mu_i) \tag{25}$$

where $h = g^{-1}$, $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ and y_i gets replaced with the pseudo-observation P_i in our context. This estimating equation is similar to the score function of a likelihood-based model (Fahrmeir et al. 2013, Chapters 5.4 and 5.5). The key difference is that the working variance σ_i^2 is not determined by some distributional assumption but can be specified by the researcher as a function of the form $\sigma^2(\mu)$ (Fahrmeir et al. 2013, Chapter 5.5). The easiest option is to specify $\sigma_i^2 = \sigma^2$, i.e. choosing a constant variance (Sachs and Gabriel 2022). Other than that, the parameter estimates $\hat{\boldsymbol{\beta}}$ are similarly obtained by (numerically) finding the root of the generalized estimating equation $s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ (Fahrmeir et al. 2013, Chapter 5.5). Conducting statistical inference is also similar to how it would be done in a likelihood-based framework but details are devoted to Section 3.3.2.

3.3.2 Asymptotic Test

Although regression models based on pseudo-observations can be applied much more generally, we keep considering the two-sample setup depicted in Section 3. A possible formulation of a regression model (22) for the restricted mean survival time could then be

$$\mathbb{E}[\min(T, t^*)|Z] = \mu(t^* \mid Z) = \beta_0 + \beta_1 Z\,. \tag{26}$$

Since, here, we are interested in the RMST *difference* g^{-1} in (22) is simply the identity link. The model described in (26) can now be interpreted as follows: If $Z = 0$, i.e. the observation comes from the control group, we have $\mu(t^* \mid Z = 0) = \mu_0(t^*) = \beta_0$. Hence, β_0 reflects the RMST of the control group. On the other hand, if $Z = 1$ then $\mu(t^* \mid Z = 1) = \mu_1(t^*) = \beta_0 + \beta_1$. Putting this together, we have $\mu_1(t^*) - \mu_0(t^*) = \beta_0 + \beta_1 - \beta_0 = \beta_1$. Therefore, β_1 can be interpreted as the RMST difference between the two treatment groups.

As pointed out in Section 3.3.1, point estimates for the vector of regression coefficients $\boldsymbol{\beta}$ can be obtained by solving the generalized estimating equation (25). Succeeding inference and hypothesis tests then resemble those of generalized linear models (Fahrmeir et al. 2013, Chapter 5.4.2). Hence, for an estimate of a single parameter, here $\hat{\beta}_1$, we construct the test statistic

$$Z^{(\mathrm{PO})} = \frac{\hat{\beta}_1}{\widehat{\mathrm{se}}(\hat{\beta}_1)} \,. \tag{27}$$

Just like the test statistic (15), $Z_{(\mathrm{PO})}$ has, asymptotically, a standard normal distribution under the null hypothesis (cf. 16). As a result, we can obtain the following similar statistical test

$$\varphi^{(\mathrm{PO})} = \mathbf{1}\{|Z^{(\mathrm{PO})}| > z_{1-\alpha/2}\} \tag{28}$$

as well as a corresponding $1 - \alpha$ confidence interval

$$CI^{(\mathrm{PO})} = \left[\hat{\beta}_1 \mp z_{1-\alpha/2}\, \widehat{\mathrm{se}}(\hat{\beta}_1)\right] . \tag{29}$$

While the point estimation of the asymptotic method described in Section 3.1 and the one based on pseudo-observations usually yield virtually identical results, the estimation of the standard errors can lead to different conclusions. Using pseudo-observation regression models, the estimation of standard errors is based on the estimation of the covariance matrix $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$. If we were willing to assume that the working variance σ_i^2 was correctly specified we could use the inverse of the quasi-Fisher information matrix $\mathbf{F}(\hat{\boldsymbol{\beta}})$, the counterpart of the Fisher information matrix in maximum likelihood estimation, as an estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ (Fahrmeir et al. 2013, Chapter 5.5). However, we have already depicted that such an assumption is not reasonable. For instance, the support of the RMST ($\mathbb{R}^+$) is bounded from below. Therefore a model like (26) using an identity link is likely to exhibit a skewed distribution of the residuals, indicating heteroscedasticity. In addition, the property that the pseudo-observations are i.i.d. only holds *asymptotically* (cf. Section 3.3.1). Because of these reasons, the usage of a sandwich-type estimator of the covariance matrix is suggested (Fahrmeir et al. 2013, Chapter 5.5; Andersen 2003). Although the concept of a sandwich-type covariance matrix estimator can be formulated more generally (Zeileis 2006), we keep using the notation by Fahrmeir et al. (2013, Chapter 5.5) for the application to quasi-likelihood

models:

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{F}}^{-1}\hat{\mathbf{M}}\hat{\mathbf{F}}^{-1} \tag{30}$$

Here, we adopt the short-hand notation $\hat{\mathbf{F}} = \mathbf{F}(\hat{\boldsymbol{\beta}})$. $\hat{\mathbf{M}}$, on the other hand, is an empirical version of the "meat" matrix. Fahrmeir et al. (2013, Chapter 5.5) give a unique definition of $\hat{\mathbf{M}}$, but we shall leave it unspecified at the moment, allowing for different types of sandwich estimators. One class of such sandwich estimators are *heteroscedasticity consistent* (HC) estimators (Zeileis 2006) where the meat matrix is of the form $\mathbf{X}'\widehat{\boldsymbol{\Omega}}\mathbf{X}$, $\mathbf{X}$ being the design matrix of the regression model and $\widehat{\boldsymbol{\Omega}} = \text{diag}(w_1, \ldots, w_n)$ is a diagonal matrix of weights depending on the working residuals $r(y_i, \eta_i)$ (Zeileis 2006). Employing an HC-type covariance matrix estimator, we have in summary

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{F}}^{-1}\mathbf{X}'\widehat{\boldsymbol{\Omega}}\mathbf{X}\hat{\mathbf{F}}^{-1}\,. \tag{31}$$

What is now left is the specification of $\widehat{\boldsymbol{\Omega}}$. Various specifications are available, leading to different HC-type covariance matrix estimators (Zeileis 2004). However, since for classical linear models the HC3 covariance matrix estimator has shown to have the best performance overall (Long and Ervin 2000) and is the default option in corresponding software packages (Zeileis et al. 2020; Sachs and Gabriel 2022), we restrict our attention to this kind of covariance matrix estimator. Originally, it was introduced by MacKinnon and White (1985) for applications to the linear model but using it for generalized linear models is equally possible (Zeileis 2006). The definition of the weights w_i for the HC3 estimator is

$$w_i = \frac{\hat{\varepsilon}_i^2}{(1-h_i)^2}\,, \tag{32}$$

where $\hat{\varepsilon}_i = r(y_i, \eta_i)$ is the working residuals and h_i the leverage of the i-th observation, respectively. Using the diagonal elements of (31) for obtaining standard error estimates, we can now compute the test statistic (27) for conducting the statistical test (28) and for constructing the corresponding confidence interval (29).

3.3.3 Bootstrap Test

In comparison to the standard asymptotic method described in Section 3.1 the approach using pseudo-observations might have the advantage of obtaining more accurate estimates of standard errors in finite sample settings. When comparing it to the studentized permutation method presented in Section 3.2, however, it may be considered inferior apart from its flexibility to adapt it to situations beyond the testing problem (14). This is because inference is still based on the assumption of the test statistic (27) to be standard normally distributed under the null hypothesis whereas the studentized permutation test does not make this assumption. Therefore, we now intend to make a similar adaption to the approach using pseudo-observations.

In principle, we could just focus on β_1 from the model (26) and apply a studentized permutation approach similar to the one applied by Ditzhaus et al. (2023). However, as we argue that the main advantage of an approach based on pseudo-observations is its possibility to be extended to more general settings, e.g. one where we want to test the effect of a continuous covariate, we want any adaption to this method to retain that flexibility. Nonetheless, the basic idea of the approach proposed in the following is still similar to that of the studentized permutation test: Instead of making assumptions about the approximate distribution of the test statistic under the null hypothesis, in this case (27) following a standard normal distribution, and using this assumption even in small sample scenarios, we strive to estimate this distribution from the observed data directly. This means that we need to obtain parameter estimates and associated test statistics such as (27) as if the null hypothesis was true. For the studentized permutation test, this has been done by calculating the test statistic (19) using permuted versions $(\mathbf{t}, \boldsymbol{\delta})^\pi$ of the original data $(\mathbf{t}, \boldsymbol{\delta})$. Here, we slightly modify this notation, omitting the treatment group index j, but therefore including the matrix of covariates $\mathbf{X}$. Then, we denote the original data by $(\mathbf{t}, \boldsymbol{\delta}, \mathbf{X}) \equiv \{(t_i, \delta_i, \mathbf{x}_i') : i = 1, \dots, n\})$. In the simplest setting, $\mathbf{X}$ could just be an $n \times 1$ vector, e.g. representing the treatment assignment for each individual. Furthermore, let $(\mathbf{t}, \boldsymbol{\delta}, \mathbf{X})^b \equiv \{(t_i, \delta_i, \mathbf{x}_i')^b : i = 1, \dots, n\})$ denote a similar data set that has been obtained by case resampling with replacement, which we call a *bootstrap sample*. In a similar fashion to the studentized permutation test, we generate such bootstrap samples B times. For each of the B bootstrap samples, we now seek to calculate the pseudo-observations (23) and subsequently estimate a corresponding regression model (22). Finally, for any coefficient β of such a model, we can calculate the bootstrap test statistic

$$Z^b = \frac{|\hat{\beta}^b - \hat{\beta}|}{\widehat{\text{se}}(\hat{\beta}^b)}\,. \tag{33}$$

Whereas $\hat{\beta}$ denotes the estimate of β that has been obtained using the original data, all other components of (33) with b in the superscript are estimated based on the b-th bootstrap sample. The rationale of this formula is that "the estimate of β from the bootstrap samples should, on average, be equal to $\hat{\beta}$, at least asymptotically" (MacKinnon 2009) and therefore Z^b should mimic the distribution of (27) under the null hypothesis. Using these bootstrap samples, we can now conduct statistical inference similar as it has been done with the studentized permutation test. Therefore, a two-sided test for a given significance level $\alpha \in (0, 1)$ is given by

$$\varphi^b = \mathbf{1}\{|Z^{(\text{PO})}| > q^b_{1-\alpha}\} \tag{34}$$

with $q^b_{1-\alpha}$ being the $(1-\alpha)$-quantile of the bootstrap test statistics (33). Similarly, we can construct a $1-\alpha$ bootstrap-t confidence interval (Efron and Tibshirani 1993,

Chapter 12.5) based on that quantile:

$$CI^b = \left[\hat{\beta} \mp q^b_{1-\alpha}\, \widehat{\text{se}}(\hat{\beta})\right] \tag{35}$$

However, there is one obstacle to this approach hindering its practical feasibility: Due to the computation of the pseudo-observations, even the asymptotic test (28) already incorporates a resampling scheme. Applying a bootstrap on top of that implies a nested resampling procedure, which gets computationally more expensive (a) the larger the original sample is and (b) the more bootstrap samples we want to draw, i.e. the more accurate we want our statistical test in principle to be. Such a problem can become apparent quickly. For instance, if we had a sample of size $n = 50$ and wanted to draw $B = 1000$ bootstrap samples, then we would need to carry out $n \cdot B = 50000$ computations in total. One way to tackle this problem could be to embrace parallel computing capabilities. A more subtle approach may be to try to eliminate one of the two resampling levels entirely. This is indeed possible by using an approximate version of the pseudo-observations:

$$\begin{aligned} P_i &= n\hat{\theta} - (n-1)\hat{\theta}_{-i} \\ &= \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}_{-i}) \\ &\approx \hat{\theta} + n\frac{\partial \hat{\theta}}{\partial w_i} \end{aligned} \tag{36}$$

These approximations are based on the first-order influence function of $\hat{\theta}$ and are known as *infinitesimal jackknife* (IJ) pseudo-observations (Parner et al. 2023). It is required that the marginal estimator $\hat{\theta}$ can be written as a function of weights w_i attached to each individual $i = 1, \ldots, n$ in the data. As pointed out, the advantage over ordinary pseudo-observations obtained by using the jackknife samples is the computational speed. Despite this, the loss in numerical accuracy is negligible even for moderate sample sizes, making IJ pseudo-observations an attractive alternative to ordinary pseudo-observations (Parner et al. 2023).

Similar to the application of the studentized permutation test, using bootstrap resampling in combination with pseudo-observations we are likely to encounter situations in which the underlying marginal estimators of the RMST are not uniquely defined at t^* due to right-censoring. We employ the same simple strategy that has been used by Horiguchi and Uno (2020) and Ditzhaus et al. (2023) for the permutation methods, i.e. we extend the Kaplan-Meier estimate "horizontally" up to that time point.

4 Simulation Study

To evaluate and compare the empirical performance of the presented methods, in particular in settings with small to moderate sample sizes, we now set up a simulation study pursuing to cover a satisfactory range of different scenarios. In order to embed the design of this study into a structured framework, we make use of the ADEMP methodology introduced by Morris et al. (2019). ADEMP is an acronym translating to *Aims*, *Data-generating mechanisms*, *Estimands* (and other targets), *Methods*, and *Performance measures*. The concrete configuration of these aspects is described in the first subsection. Thereafter, we briefly comment on computational aspects regarding the implementation of the presented methods in statistical software on the one hand as well as the execution of the simulation study and its reproducibility. Lastly, we present the results of the study and discuss our findings.

4.1 Design

Aims

The aim of this simulation study is to investigate the performance of pseudo-observations methods for estimating and testing the RMST difference between two samples, e.g. a control and an experimental treatment group in a clinical trial, in comparison to alternative approaches. In particular, the focus is on settings with small to moderate sample sizes as all methods under investigation either directly or indirectly rely on large sample approximations based on asymptotic theory.

Data-generating mechanisms

The data-generating mechanisms are, in principle, adopted from the simulation study by Ditzhaus et al. (2023), including five factors being varied: The event time distributions, the censoring time distributions, the base allocation of the sample size, the sample multiplier as well as the effect size, i.e. the true RMST difference. However, we do not evaluate all of the scenarios constructed by Ditzhaus et al. (2023) but instead only concentrate on a subset of them. We selected this subset in a way that we think it covers a range of possibly heterogeneous scenarios, especially with respect to the event time distributions of the two populations. For the effect size $\Delta = \mu_1(t^*) - \mu_0(t^*)$ Ditzhaus et al. (2023) considered four values in total. Here, we only look at $\Delta \in \{0, 1.5\}$, reflecting scenarios under the null and under the alternative hypothesis, respectively. In addition, we investigate the following three of the nine pairs of event time distributions considered by Ditzhaus et al. (2023):

S1 Exponential distributions and proportional hazard alternatives: $T_0 \sim \text{Exp}(0.2)$ and $T_1 \sim \text{Exp}(\theta_\Delta)$

S7 Exponential vs piecewise Exponential: $T_0 \sim \text{Exp}(0.2)$ and T_1 with piecewise constant hazard function $\lambda_1(t) = 0.5 \cdot \mathbf{1}\{t \leq \theta_\Delta\} + 0.05 \cdot \mathbf{1}\{t > \theta_\Delta\}$

S8 Weibull distributions with crossing curves and shape alternatives: $T_0 \sim \text{Weib}(3, 1/8)$ and $T_1 \sim \text{Weib}(\theta_\Delta, 1/14)$

Here, θ_Δ denotes a generic parameter whose meaning depends on the respective context. It then gets calibrated according to the given effect size Δ and the restriction time t^*, the latter being fixed to $t^* = 10$. For S1, this corresponds to the rate parameter for the exponential distribution of the treatment group, whereas for S7 this is the time point at which the hazard of the treatment group changes. On the other hand, for S8, θ_Δ is the scale parameter for the Weibull distribution of the treatment group.[2] For the censoring distributions, in contrast, we keep using all levels of this factor examined by Ditzhaus et al. (2023):

C1 unequally Weibull distributed censoring (Weib, uneq): $C_0 \sim \text{Weib}(3, 1/18)$ and C_1 $\text{Weib}(0.5, 1/40)$

C2 equally uniformly distributed censoring (Unif, eq): $C_0 \sim \text{Unif}(0, 25)$ and $C_1 \sim \text{Unif}(0, 25)$

C3 equally Weibull distributed censoring (Weib, eq): $C_0 \sim \text{Weib}(3, 1/15)$ and $C_1 \sim \text{Weib}(3, 1/15)$

A visual representation of the pairs of both, the survival and the censoring distributions, is given in Figure 4.

Moreover, we utilize the same assumptions about the sample sizes and their allocations as Ditzhaus et al. (2023). In particular, we consider three base settings of how the samples are allocated between the two groups, i.e. $(n_0, n_1) \in \{(12, 18); (15, 15); (18, 12)\}$. To obtain different total sample sizes, these base allocations are multiplied with the integer K. In comparison to Ditzhaus et al. (2023), who considered $K \in \{1, 2, 4\}$, we add $K = 6$ to this set to incorporate scenarios for which we expect the asymptotic theory to take effect. The combination of all these factors, summarized in Table 1, leads to a total of 216 scenarios that are being evaluated, i.e. 108 scenarios for assessing the type I error control of the different methods ($\Delta = 0$) and 108 scenarios for analyzing the power of the tests ($\Delta = 1.5$).

For their simulation study evaluating the unstudentized permutation test, Horiguchi and Uno (2020) regenerated the simulated data whenever the RMST was inestimable

[2] In comparison to the parametrization of the Weibull distribution in (7), Ditzhaus et al. (2023) use a different parametrization where the scale parameter is $\sigma = \lambda^{-1}$. For the sake of consistency, we keep employing the parametrization as it is used in 7. Nonetheless, the distributional assumptions are equivalent to those in Ditzhaus et al. (2023).

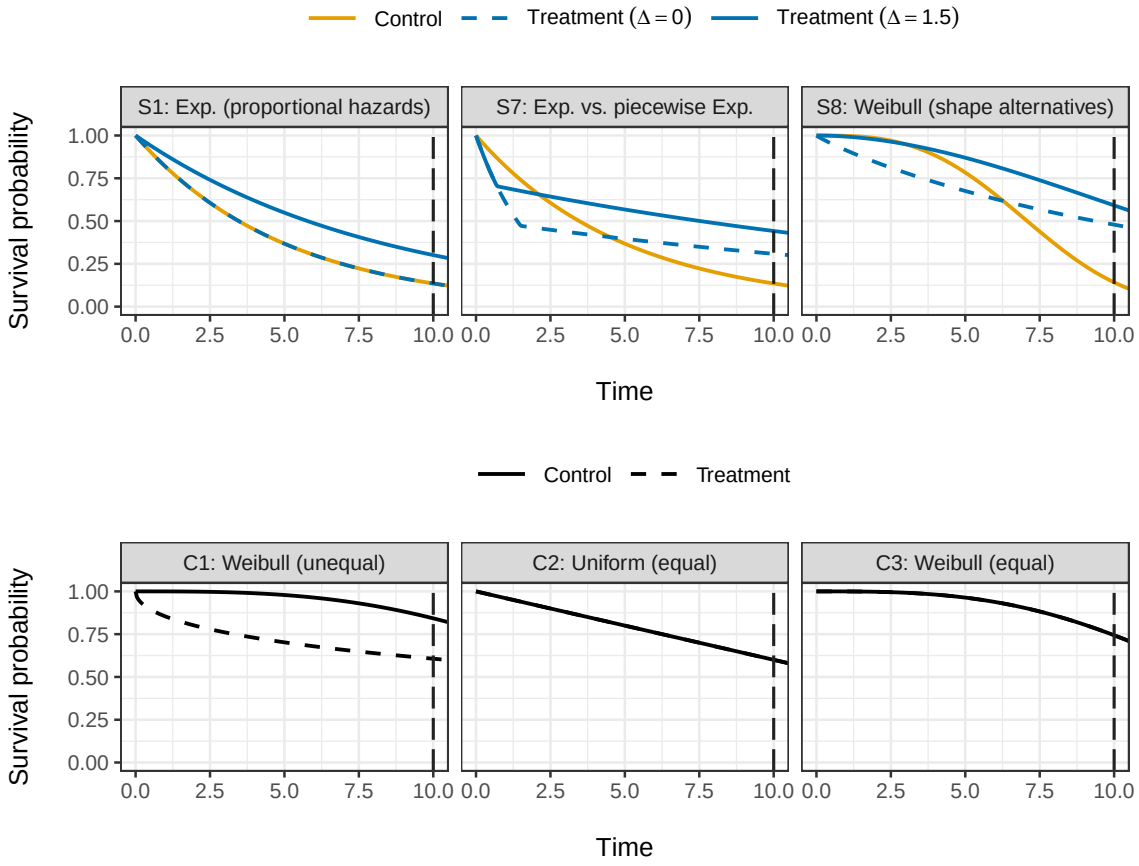

Figure 4 Simulation models for the event (top) and censoring (bottom) times. The vertical dashed line indicates the cutoff time point at $t^* = 10$ for which the RMST differences $\Delta = 0$ and $\Delta = 1.5$ hold, respectively..

Table 1 Factors and their levels for the data-generating mechanisms used in the simulation study .

Factor	Levels
Survival models	S1: Exponential distributions S7: Exponential and piecewise exponential distributions with crossing curves S8: Weibull distributions with crossing curves and shape alternatives
Censoring models	C1: unequal Weibull C2: equal uniform C3: equal Weibull
RMST difference $(\mu_1(t^*) - \mu_0(t^*))$	0; 1.5
Sample allocations	(12, 18); (15, 15); (18, 12)
Sample multipliers (K)	1; 2; 4; 6

for at least one of the two groups using a Kaplan-Meier-based estimator of the RMST. As described in Section 3.2, such a situation occurs when the largest event time is smaller than t^* and is censored. Ditzhaus et al. (2023) also adopted this procedure for their simulation study. As a consequence, we employ the same routine in our study.

Estimands and other targets

Although all of the methods involve the point estimation of a particular parameter, namely the two-sample RMST difference $\Delta = \hat{\mu}_1(t^*) - \hat{\mu}_0(t^*)$, they are all based on the Kaplan-Meier estimator of the survival function. Because of this, the point estimates of the different methods will be either exactly or nearly identical. The main distinction is how the statistical test of the null hypothesis in (14) is conducted. Therefore, this null hypothesis is the target of the simulation study.

Methods

All methods that will be assessed in this study have been presented and explained in detail in Section 3. Of primary interest, however, are the two proposed methods based on pseudo-observations, i.e. the one employing an asymptotic test and the other one using a bootstrap test. For the latter, we use IJ pseudo-observations as described in Section 3.3.3. In order to acquire additional information about how the usage of IJ instead of ordinary pseudo-observations might impact the performance of the bootstrap test, we also include an asymptotic test using these IJ pseudo-observations. However, these results are devoted to Appendix A. In this simulation study, the standard asymptotic test (17) has the role of the main reference method as it can be expected that, among all methods presented in this thesis, this is the one most commonly employed in practice, both by applied scientists (e.g. Manner et al. 2019) and methodological researchers (e.g. Dormuth et al. 2023). On the other hand, based on the results from their simulation study, we consider the studentized permutation method by Ditzhaus et al. (2023) as the current gold standard for conducting two-sample RMST tests. Other methods that were also investigated by Ditzhaus et al. (2023) and could have been included in our simulation study are the unstudentized permutation test by Horiguchi and Uno (2020) and the approach based on empirical likelihood ratios by Zhou (2021). Based on the results by Ditzhaus et al. (2023), in our opinion, these two methods cannot be assigned to any of the two aforementioned categories and therefore we do not cover them in our study.

Performance measures

Given the null hypothesis in (14) as the target of the simulation study and the original problem that the standard asymptotic test (17) has an inflated type I error rate, the type I error rate is also the primary performance measure. Corresponding secondary performance measures are the power and the coverage of the respective test and its associated confidence interval, respectively.

4.2 Computational Details

For each of the scenarios described in Section 4.1 we generated $N_{\text{sim}} = 5000$ data sets as it has been done by Ditzhaus et al. (2023). Similarly, we used $B = 2000$ resampling iterations, for both, the studentized permutation and the pseudo-observations bootstrap test. The nominal significance level α was set to 5%.

All computations were carried out using the `R` programming language in version 4.3.0 (R Core Team 2023) on the high-performance computing cluster of the GWDG in Göttingen.[3] The asymptotic test as well as the studentized permutation test were implemented by ourselves based on code supplied by Marc Ditzhaus. The code for the simulation study is hosted in a repository on GitLab and can be accessed upon request.

For the approaches based on pseudo-observations, we used the `rmeanglm()` function from the `{eventglm}` package (Sachs and Gabriel 2022). This function is essentially a wrapper around the `stats::glm()` function adapted to regression models for pseudo-observations. The arguments `model.censoring` and `formula.censoring` of that function permit to customize the computation of the pseudo-observations. While the latter is used to specify which variables should be controlled for when calculating the pseudo-observations, to the former we can pass custom "modules" (R functions) that carry out the actual calculation of the pseudo-observations before estimating the regression model. Although the `{eventglm}` package already provides a wide range of such modules, we programmed these functions by ourselves as well, since the "horizontal extension strategy" (see Section 3.2 and Section 3.3.3) does not work with these implementations. For the IJ pseudo-observations module, we used the `pseudo()` function from the `{survival}` package (Therneau 2023). Further direct and indirect package dependencies were required for running the simulation study, all of which were tracked and managed using the `{renv}` package (Ushey and Wickham 2023), contributing to the reproducibility of the results.

Another important consideration regarding the validity and reproducibility of the results is how we generated pseudo-random numbers. In our simulation study, we generated all random data sequentially using the Mersenne-Twister algorithm, therefore avoiding the need to worry about correlated data sets or using parallel random number generator (RNG) streams (Morris et al. 2019). Moreover, we used the `with_seed()` function from the `{withr}` package (Hester et al. 2022) for handling the RNG state within each simulation scenario. On the other hand, the application of the different statistical methods on the simulated data sets was done in parallel using the `{parallel}` package, which gets shipped with base R (R Core Team 2023). While we kept using

[3] https://gwdg.de/hpc/systems/scc/

the `with_seed()` function, we now used the L'Ecuyer-CMRG algorithm and parallel RNG streams for generating pseudo-random numbers. For the standard asymptotic test and those based on pseudo-observations this has no implications. Conversely, the studentized permutation method and the bootstrap test using pseudo-observations both rely on stochastic resampling, which is why proper handling of (parallel) random number generation was required.

4.3 Results

We present the results in the form of both, tables and figures. The former should provide all results in great detail whereas the latter shall serve as a way to grasp the overall patterns and conclusions faster.

4.3.1 Type I Error

Regarding the type I error rate, Figure 5 displays the results graphically aggregated by the total sample size across the two treatment groups. Additionally, Table 2 augments Figure 5 with detailed numerical results. In both, the table and the figure, we have depicted information about the 95% binomial confidence interval $[4.4\%, 5.6\%]$ around the nominal level $\alpha = 5\%$ derived from the fact that we have run $N_{\text{sim}} = 5000$ simulations for each scenario. Therefore, we can consider any of the statistical tests to have succeeded in terms of controlling the type I error for a given scenario, whenever the empirical type I error rate falls within that confidence interval.

The first thing we can observe is that, in line with the findings by Horiguchi and Uno (2020) and Ditzhaus et al. (2023), the standard asymptotic test indeed leads to too liberal test decisions across the majority of scenarios (see Figure 5). From the raw data in Table 2 we can conclude that it controlled the type I error in 35 out of 108 scenarios in total. This issue is particularly pronounced for small sample sizes ($K \in \{1, 2\}$). Noticeably, for $K = 1$, there is not a single instance for which the type I error was controlled by the asymptotic test. For $K = 2$ it controlled the type I error in 3 out of 27 scenarios, only. The performance of the test improves for $K = 4$ but even for this subset, the test failed in the majority of cases (10 out of 27 possible successes). When dealing with a total of 180 samples ($K = 6$) the test enhances further (22 successes) but still performs evidently worse than its competitors. Additionally, the type of allocation of the samples has a noticeable impact on the performance of the test: In settings with fewer samples in the treatment arm than in the control arm ($(n_0, n_1) = (18, 12)$), the asymptotic test successfully controlled the type I error in 8 out of 36 cases. In contrast, when the sample sizes were balanced ($(n_0, n_1) = (15, 15)$) or larger in the control arm

Table 2 Type I error rates of different methods in % (nominal level $\alpha = 5\%$). The values inside the binomial confidence interval $[4.4\%, 5.6\%]$ are printed bold .

		$N = K \cdot (12, 18)$				$N = K \cdot (15, 15)$				$N = K \cdot (18, 12)$			
Censoring	K	Asy	Perm	PO1	PO2	Asy	Perm	PO1	PO2	Asy	Perm	PO1	PO2
S1: Exponential distributions													
un. W.	1	7.0	**4.6**	**5.2**	**4.4**	7.2	**5.4**	5.8	**4.6**	8.3	5.9	6.4	**4.6**
	2	6.3	**4.8**	**5.2**	**4.9**	5.9	**5.1**	**5.1**	**4.8**	6.2	**5.1**	**5.2**	**4.6**
	4	**5.3**	**4.8**	**4.9**	**4.8**	**5.0**	**4.6**	**4.6**	**4.7**	5.7	**5.2**	**5.2**	**5.0**
	6	**5.0**	**4.5**	**4.5**	**4.5**	**5.2**	**5.0**	**5.0**	**5.0**	**5.1**	**4.9**	**4.8**	**4.8**
eq. U.	1	8.7	5.8	6.7	**5.1**	7.3	**4.8**	**5.5**	**4.4**	8.5	**5.5**	6.4	**4.9**
	2	6.4	**5.2**	**5.4**	**4.9**	5.9	**4.9**	**4.9**	**4.7**	6.4	**5.0**	**5.2**	**4.8**
	4	6.1	**5.4**	**5.4**	**5.3**	**5.4**	**4.8**	**4.8**	**4.9**	5.8	**5.4**	**5.3**	**5.2**
	6	**5.1**	**5.1**	**4.9**	**5.0**	5.7	**5.4**	**5.4**	**5.4**	**5.2**	**4.9**	**4.9**	**4.9**
eq. W.	1	7.9	**5.5**	6.1	**5.3**	7.2	**5.0**	**5.4**	**4.7**	7.7	**5.3**	6.1	**4.9**
	2	5.8	**4.9**	**5.0**	**5.0**	5.9	**5.1**	**5.2**	**5.0**	5.7	**4.7**	**4.8**	**4.5**
	4	5.8	**5.3**	**5.3**	**5.3**	**4.8**	**4.5**	**4.6**	**4.5**	5.7	**5.0**	**5.2**	**5.1**
	6	5.7	**5.5**	**5.5**	**5.5**	**5.4**	**5.3**	**5.2**	**5.2**	**5.1**	**4.8**	**4.8**	**4.8**
S7: Exponential and piecewise exponential distributions with crossing curves													
un. W.	1	6.5	3.8	**4.6**	3.7	6.9	**5.0**	**5.3**	3.9	8.3	6.2	6.3	4.0
	2	6.2	**4.9**	**5.3**	**4.9**	6.3	**5.1**	**5.1**	**4.6**	7.1	6.2	5.9	**5.0**
	4	**5.5**	**5.1**	**5.1**	**4.9**	5.7	**5.2**	**5.2**	**5.1**	6.2	**5.6**	5.7	**5.3**
	6	**5.1**	**4.8**	**4.9**	**4.7**	**5.5**	**5.1**	**5.2**	**5.1**	**5.4**	**5.1**	**5.0**	**4.7**
eq. U.	1	6.9	4.1	**5.0**	3.9	7.2	**5.2**	5.7	**4.6**	7.2	**4.8**	**5.4**	3.3
	2	6.5	**5.1**	**5.4**	**5.1**	6.5	**5.3**	**5.4**	**5.0**	6.2	**4.9**	**5.1**	**4.6**
	4	5.8	**5.1**	**5.3**	**5.1**	5.7	**5.1**	**5.1**	**5.1**	5.8	**5.2**	**5.3**	**5.0**
	6	**5.3**	**4.9**	**4.8**	**4.8**	**4.5**	4.2	4.2	4.3	**5.6**	**5.3**	**5.3**	**5.2**
eq. W.	1	7.4	**5.4**	6.1	**5.2**	6.7	**4.9**	**5.3**	**4.4**	8.2	6.1	6.5	3.8
	2	5.8	**4.7**	**5.0**	**4.8**	**5.5**	**4.7**	**4.8**	**4.5**	6.3	**5.3**	**5.3**	**4.8**
	4	**5.1**	**4.8**	**4.7**	**4.8**	**5.5**	**5.1**	**5.1**	**5.0**	5.9	**5.4**	**5.5**	**5.4**
	6	**5.1**	**4.9**	**4.9**	**4.9**	**5.6**	**5.2**	**5.4**	**5.5**	**5.5**	**5.3**	**5.3**	**5.3**
S8: Weibull distributions with crossing curves and shape alternatives													
un. W.	1	7.0	**4.6**	**5.2**	3.9	8.9	6.1	6.2	4.0	9.8	7.6	7.4	3.6
	2	**5.3**	**4.4**	**4.7**	**4.4**	6.4	**5.4**	**5.3**	**4.7**	7.4	6.1	5.9	**4.7**
	4	**5.5**	**4.8**	**4.9**	**4.9**	5.7	**5.5**	**5.4**	**5.1**	6.0	**5.5**	**5.5**	**4.9**
	6	**5.1**	**4.8**	**4.7**	**4.6**	**4.8**	**4.5**	**4.6**	**4.4**	5.9	**5.6**	**5.5**	**5.3**
eq. U.	1	7.4	**4.6**	**5.3**	4.1	7.3	**5.1**	**5.3**	3.7	8.8	6.3	6.7	3.3
	2	6.3	**5.1**	**5.3**	**4.9**	**5.6**	**4.7**	**4.7**	4.3	6.6	**5.5**	5.7	**4.5**
	4	5.7	**5.1**	**5.1**	**5.1**	**5.1**	**4.7**	**4.6**	**4.6**	5.8	**5.1**	**5.2**	**4.9**
	6	**5.5**	**5.2**	**5.2**	**5.2**	**5.6**	**5.2**	**5.2**	**5.0**	**5.4**	**5.1**	**5.0**	**4.8**
eq. W.	1	6.3	4.1	**4.5**	3.9	7.2	**5.4**	**5.5**	4.0	8.1	6.1	6.3	4.0
	2	6.3	**5.1**	**5.4**	**5.1**	6.0	**5.0**	**5.0**	**4.8**	6.3	**5.5**	**5.6**	**4.8**
	4	**5.4**	**4.8**	**4.9**	**4.9**	5.9	**5.4**	**5.5**	**5.2**	5.8	**5.3**	**5.3**	**5.2**
	6	5.8	**5.6**	**5.5**	**5.5**	6.2	6.0	6.0	5.9	**5.6**	**5.4**	**5.3**	**5.1**

Abbreviations: Asy, asymptotic test; Perm, studentized permutation test; PO1, pseudo-observations asymptotic; PO2, pseudo-observations bootstrap; un. W., unequal Weibull censoring; eq. U., equal uniform censoring; eq. W., equal Weibull censoring.

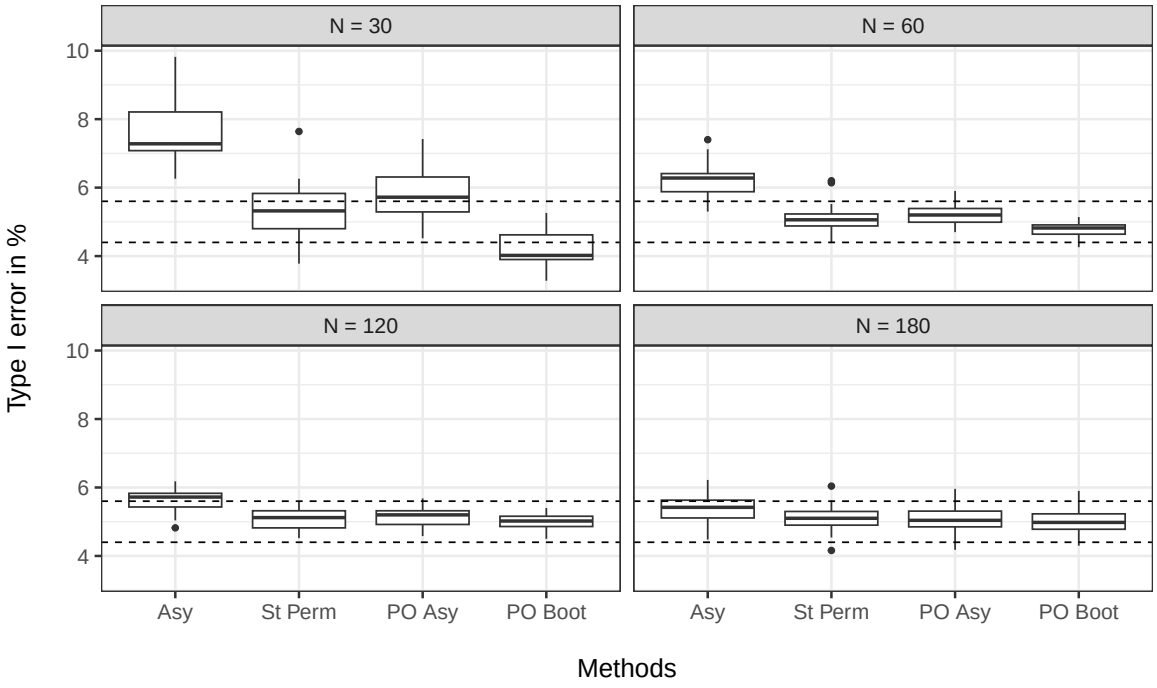

Figure 5 Type I error rates of different methods in % (nominal level $\alpha = 5\%$) aggregated by the total sample sizes ($N = n_0 + n_1$). The dashed lines depict the 95% binomial confidence interval $[4.4\%, 5.6\%]$.

($(n_0, n_1) = (12, 18)$) there were 14 and 13 successes, respectively. On the other hand, the performance of the test with respect to the type I error control seems to be less impacted by the distributional configurations of the survival and censoring models, overall. Considering these parameters marginally, the number of successes differs by 2 cases at most between the different settings.

Regarding the studentized permutation test, our own findings largely agree with those by Ditzhaus et al. (2023). Across all scenarios it even showed the best performance of all tests, being able to control the type I error in 93 out of 108 scenarios. Relative to the other methods, its performance particularly stands out for very small sample sizes ($K = 1$), where it controlled the type I error in 16 out of 27 instances. For larger sample sizes ($K \in \{2, 4, 6\}$), the performance is also decent with 25, 27 and 25 successes, respectively. As for the asymptotic test, the factor that influences the performance of the studentized permutation test the most besides the total sample size is the allocation of the samples. With the allocation $(n_0, n_1) = (18, 12)$ the test successfully controlled the type I error in 28 out of 36 scenarios in comparison to 33 ($(n_0, n_1) = (15, 15)$) and 32 ($(n_0, n_1) = (12, 18)$) occurrences. However, we can observe greater variations in the performance of the studentized permutation test regarding the different distributional settings of the survival times. In this regard, the best performance can be observed in settings with proportional hazards (S1, 34 out of 36 possible successes). For settings S7 (exponential and piecewise exponential distributions with crossing survival curves) and S8 (crossing Weibull survival curves with shape alternatives), on the other hand,

there are 30 and 29 successes, respectively. As for the censoring models, the conclusions are rather similar to those for the asymptotic test: While we cannot necessarily rule out a potential effect of the considered censoring distributions, the observed differences are comparatively small. For the censoring models C2 (equal, uniform) and C3 (equal, Weibull), the studentized permutation test succeeded 32 times. In the case of unequal Weibull censoring (C1), on the other hand, there are 29 successes. Finally, it is worth noticing that when the studentized permutation test does actually fail we cannot decisively attribute this failure to the test being too conservative or too liberal, overall (see Figure 5).

Moving to the first proposed method based on pseudo-observations and the corresponding asymptotic test, overall, we can conclude that it provides a valid alternative to the studentized permutation test when it comes to improving the type I error control relative to the standard asymptotic test (see Figure 5). In particular, it controlled the type I error successfully in 88 out of all 108 scenarios (see Table 2). For very small sample sizes ($K = 1$), it cannot entirely keep up with the studentized permutation test, however (see Figure 5). Nonetheless, for sample sizes larger than that ($K \in \{2, 4, 6\}$) the performance is almost similar. The concrete patterns of other factors impacting the performance of the pseudo-observations method are in line with those of the studentized permutation test and the standard asymptotic approach. For instance, the samples being more concentrated toward the control arm negatively impacts the performance of the test. In this case, the test controlled the type I error successfully in 24 out of 36 scenarios. In the other two instances, the test succeeded in 33 ($(n_0, n_1) = (12, 18)$) and 31 ($(n_0, n_1) = (15, 15)$) cases. Regarding the survival and censoring models, we can observe similar tendencies as for the studentized permutation test but they are less strong. The test performed best for survival model S1 (proportional hazards) but the differences to the other settings are less remarkable with only one success more than for the other two models. In terms of censoring, the method also performed worse in the case of unequal Weibull censoring (C1). But also in this regard, the difference in the number of successes sums up to 2, only. As opposed to the studentized permutation test, the failures of the asymptotic pseudo-observations approach can clearly be attributed to a slightly too liberal behavior of the test when the sample size is small (see Figure 5).

Lastly, we discuss the results of the pseudo-observations approach incorporating a bootstrap hypothesis test. In terms of the total number of scenarios in which this method was able to control the type I error, it improves the previously discussed pseudo-observations approach even further with 90 out of 108 such occurrences. However, it must be noted that for the most extreme scenario, in terms of the total sample size ($K = 1$), it actually performed slightly worse than the pseudo-observations approach employing an asymptotic test (12 compared to 13 successes). From Figure 5 we can conclude that this is due to a too conservative behavior of the test which gets more

pronounced the smaller the sample size is. Unlike the other tests, however, this method seems to be less impacted by how the samples are allocated. Precisely, we have 31 ($(n_0, n_1) = (12, 18)$), 29 ($(n_0, n_1) = (15, 15)$) and 30 ($(n_0, n_1) = (18, 12)$) successes. Conversely, the configuration of the survival distributions has a greater effect: While the method handled scenarios with proportional hazards (S1) extremely effectively (not a single failure in controlling the type I error), it had more trouble with crossing Weibull survival curves (S8). In the latter case, the method could control the type I error in 25 out of 36 cases. For model S7 (exponential and piecewise exponential distributions with crossing survival curves) this number corresponds to 29. As for the other methods, the influence of the different censoring models evaluated in this simulation study seems to be rather weak.

4.3.2 Power

In a similar fashion, the power values for the different methods and scenarios are displayed in Table 3 and Figure 6, respectively.

From Table 3 we can take away that the standard asymptotic test has the highest power in all scenarios among the methods investigated. However, having the discussion of the results in Section 4.3.1 in mind, this comes at the cost of an inflated type I error rate. Therefore, it is hard to assess the higher power of the asymptotic test as an advantage over the other methods. Other than that, the remaining methods can broadly be ranked in the following order (highest to lowest power): pseudo-observations asymptotic, studentized permutation, pseudo-observations bootstrap. Hence, the conservativeness of the bootstrap method that we have observed in Section 4.3.1 can also be recognized in terms of the power of the test. This conservativeness is most pronounced in settings with very small sample sizes ($K = 1$) as can be seen in Figure 6 as well as in the single rows of Table 3. There, we can also see that this behavior of the bootstrap method becomes more noticeable when there are fewer samples in the treatment arm than in the control arm ($(n_0, n_1) = (18, 12)$). Although this effect can also be observed for the other methods it is not as strong as for the bootstrap method. Nevertheless, even if the ranking of the methods is quite consistent across different scenarios, the absolute differences in power values between the respective methods are rather small and mostly of minor practical relevance. The only exceptions are the aforementioned scenarios with very small and unbalanced sample sizes.

Turning the focus towards the comparison of different parameter values of the data generating process, the strongest differences can be observed with respect to the different pairs of survival models. In Table 3 we can see that, regardless of the method considered, the power of the tests is the largest for model S8 (crossing Weibull survival curves

with shape alternatives) followed by model S1 (exponential distributions) and model S7 (exponential and piecewise exponential distributions with crossing survival curves). Considering all methods as well as all other parameters of the data generating process the lowest power value for this model is 26.1% (S8) in comparison to 12.0% (S1) and 9.5% (S7). The largest values, on the other hand, are 98.6% (S8), 83.2% (S1) and 73.9% (S7). From this, we can conclude that assumptions about the shapes of the survival curves play a critical role in planning a study and calculating the sample size for a clinical trial using the RMST difference as the effect measure. The type of censoring model only had a minor effect on the power of the tests. For models C1 (Weibull, unequal) and C2 (uniform, equal), the power values are quite similar, overall. For model C3 (Weibull, equal), the power is a little bit larger, in general.

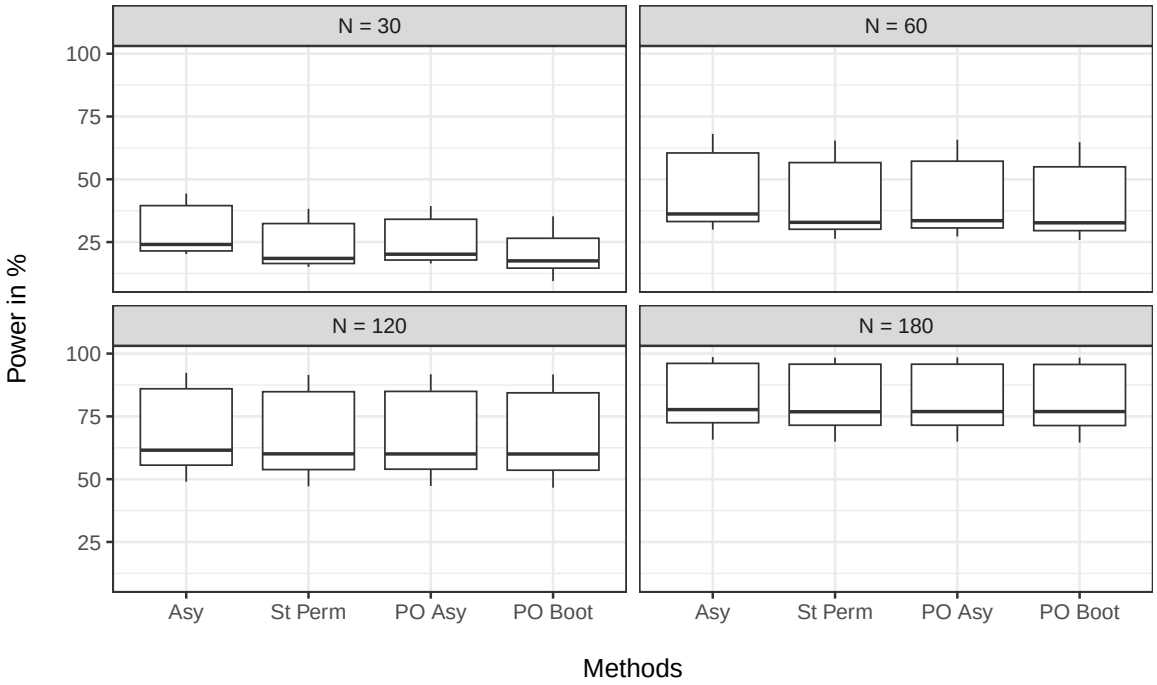

Figure 6 Power values of different methods in % (nominal level $\alpha = 5\%$) aggregated by the total sample sizes ($N = n_0 + n_1$).

4.3.3 Coverage

Finally, we discuss the empirical coverage rates of the confidence intervals associated with each method. As these results include both, null and alternative scenarios, a display in the form of a table is difficult. Therefore, the main results are presented in Figure 7 for a better overview. There, we distinguish not only between the total sample sizes (rows) but also between their allocations (columns). Similar to the results concerning the type I error rate in Section 4.3.1, we depict the 95% binomial confidence interval around the nominal coverage rate (95%) based on the 5000 simulation samples obtained.

Table 3 Rejection rates (power) of different methods in % (nominal level $\alpha = 5\%$) .

		$N = K \cdot (12, 18)$				$N = K \cdot (15, 15)$				$N = K \cdot (18, 12)$			
Censoring	K	Asy	Perm	PO1	PO2	Asy	Perm	PO1	PO2	Asy	Perm	PO1	PO2
S1: Exponential distributions													
un. W.	1	**25.1**	18.5	20.4	17.7	**23.2**	18.5	19.4	16.2	**23.0**	17.8	18.0	12.0
	2	**35.9**	32.4	33.5	32.7	**35.9**	32.9	32.8	32.2	**34.0**	30.8	31.2	29.8
	4	**60.9**	59.0	59.5	59.6	**59.3**	57.8	57.8	57.5	**57.0**	55.2	55.1	54.7
	6	**77.7**	76.8	76.9	76.9	**76.6**	75.8	75.9	75.9	**74.0**	72.8	72.7	72.5
eq. U.	1	**24.1**	18.8	20.2	17.5	**23.2**	17.8	18.8	15.9	**23.3**	17.2	18.7	13.6
	2	**35.1**	32.2	32.4	31.7	**37.5**	34.1	34.1	33.1	**36.2**	32.2	32.8	31.6
	4	**61.9**	60.2	60.4	60.4	**61.5**	60.1	60.0	60.0	**59.4**	57.6	57.9	57.3
	6	**76.5**	75.7	75.7	75.7	**78.1**	77.2	77.4	77.3	**75.6**	74.8	74.7	74.7
eq. W.	1	**24.8**	20.0	21.1	19.0	**24.9**	20.5	21.2	19.3	**24.9**	19.6	20.8	17.5
	2	**39.7**	36.5	37.0	36.2	**39.5**	36.9	37.2	36.5	**37.6**	34.6	35.0	34.0
	4	**64.5**	63.1	63.2	62.9	**64.6**	63.6	63.3	63.1	**64.1**	62.8	62.8	62.6
	6	**81.4**	80.6	80.7	80.7	**83.2**	82.4	82.5	82.4	**79.2**	78.2	78.4	78.4
S7: Exponential and piecewise exponential distributions with crossing curves													
un. W.	1	**21.4**	15.5	17.7	15.2	**20.8**	15.9	17.1	13.9	**20.7**	16.2	17.1	9.9
	2	**32.4**	29.0	30.0	29.3	**31.6**	28.6	29.0	28.2	**30.4**	27.4	27.8	26.0
	4	**54.2**	52.4	52.9	52.5	**52.7**	51.1	50.8	50.7	**49.0**	47.1	47.3	46.6
	6	**71.1**	70.1	70.3	70.2	**69.2**	68.2	68.4	68.1	**65.7**	64.9	65.0	64.6
eq. U.	1	**21.1**	16.0	17.7	14.8	**21.1**	16.3	17.7	14.5	**20.2**	15.2	16.5	9.5
	2	**33.0**	29.6	30.0	29.0	**32.1**	29.6	30.0	29.1	**30.0**	26.3	27.2	25.8
	4	**53.8**	52.5	52.4	52.2	**53.5**	52.0	52.3	51.8	**52.0**	50.0	50.4	49.8
	6	**70.1**	69.0	69.3	69.4	**70.2**	69.0	69.2	69.2	**67.1**	65.9	65.9	65.9
eq. W.	1	**21.8**	16.7	18.0	15.9	**21.5**	17.4	18.3	15.6	**20.5**	16.0	16.9	10.5
	2	**34.1**	31.3	31.5	30.8	**33.4**	30.7	31.2	30.3	**31.1**	28.0	28.3	26.7
	4	**57.2**	56.0	56.1	55.8	**57.2**	56.0	55.9	55.8	**52.9**	51.2	51.6	51.0
	6	**73.9**	73.2	73.3	73.3	**73.8**	73.1	73.1	72.9	**69.9**	68.7	69.1	68.7
S8: Weibull distributions with crossing curves and shape alternatives													
un. W.	1	**40.5**	32.9	35.5	31.3	**42.6**	36.4	37.2	31.6	**41.3**	35.4	35.8	26.9
	2	**62.2**	58.4	59.4	58.1	**62.3**	59.6	60.1	57.9	**59.6**	56.2	56.4	51.1
	4	**88.1**	86.9	87.2	87.2	**86.7**	85.7	85.9	85.2	**84.0**	83.2	83.1	82.2
	6	**96.6**	96.4	96.5	96.3	**96.5**	96.4	96.3	96.2	**94.4**	94.2	94.1	93.8
eq. U.	1	**37.9**	30.0	31.7	26.2	**39.8**	33.0	34.2	28.7	**39.2**	31.8	34.0	26.1
	2	**61.4**	57.1	57.9	56.6	**61.8**	58.1	58.6	57.0	**59.6**	55.5	56.5	53.3
	4	**86.7**	85.7	85.8	85.7	**87.4**	86.4	86.6	86.4	**85.3**	83.9	84.1	83.6
	6	**96.5**	96.2	96.2	96.2	**96.5**	96.3	96.3	96.3	**95.7**	95.4	95.3	95.1
eq. W.	1	**42.9**	36.4	37.7	33.7	**44.3**	38.3	39.0	35.3	**44.2**	38.0	39.3	32.8
	2	**67.5**	64.5	65.1	64.0	**68.1**	65.4	65.7	64.8	**64.6**	61.3	62.0	59.2
	4	**92.3**	91.5	91.8	91.7	**91.9**	91.4	91.5	91.3	**90.1**	89.3	89.4	89.0
	6	**98.6**	98.4	98.5	98.4	**98.0**	**98.0**	97.9	97.9	**97.4**	97.2	97.3	97.1

Note: Largest value per scenario is printed bold.
Abbreviations: Asy, asymptotic test; Perm, studentized permutation test; PO1, pseudo-observations asymptotic; PO2, pseudo-observations bootstrap; un. W., unequal Weibull censoring; eq. U., equal uniform censoring; eq. W., equal Weibull censoring.

In general, the results for the confidence intervals underline our previous findings for the type I error rate (Section 4.3.1) when it comes to comparing the different methods with each other. Consequently, the standard asymptotic method fails to provide a confidence interval that maintains the nominal coverage in most scenarios. More precisely, the empirical coverage of this confidence interval lies within the 95% binomial confidence interval $[94.4\%, 95.5\%]$ in 59 out of 216 cases, only. In Figure 7, we can see that these cases mainly correspond to scenarios with larger sample sizes ($K \in \{4, 6\}$). Still, even for $K = 6$ we can notice a tendency of this method of having a coverage below 95% (see Figure 7). As for the type I error, the studentized permutation method achieves the best performance in terms of the empirical coverage of the corresponding confidence interval with 184 out of 216 successes. Likewise, the pseudo-observations methods are competitive with 174 (asymptotic) and 175 (bootstrap) simulation scenarios for which the empirical coverage is close enough to the targeted one. The minor superiority of the studentized permutation method can again be attributed to the settings with very small sample sizes ($K = 1$), for which the pseudo-observations approaches cannot entirely keep up with the studentized permutation confidence interval.

Furthermore, Figure 7 highlights some of our previous findings regarding the study design, which are mostly independent of the chosen method. Hence, the nominal coverage of the confidence intervals is less frequently achieved when there are fewer subjects in the treatment arm than in the control arm ($(n_0, n_1) = (18, 12)$). Again, this effect is especially pronounced for cases with very small sample sizes ($K = 1$). In contrast, it does not seem to make much of a difference whether the samples are balanced between the two groups or if the number of samples predominates in the treatment arm. Nonetheless, this statement depends on the distributions of the event times that we are dealing with. Figure 8 shows the same results as Figure 7 from a different perspective. There, we can see that the described sample allocation effect is more distinct for model S8 (crossing Weibull survival curves with shape alternatives) than for the other two survival models. Still, we may conclude that for an actual study, we should seek a sample allocation that is either balanced or slightly concentrated toward the group for which a treatment effect is expected.

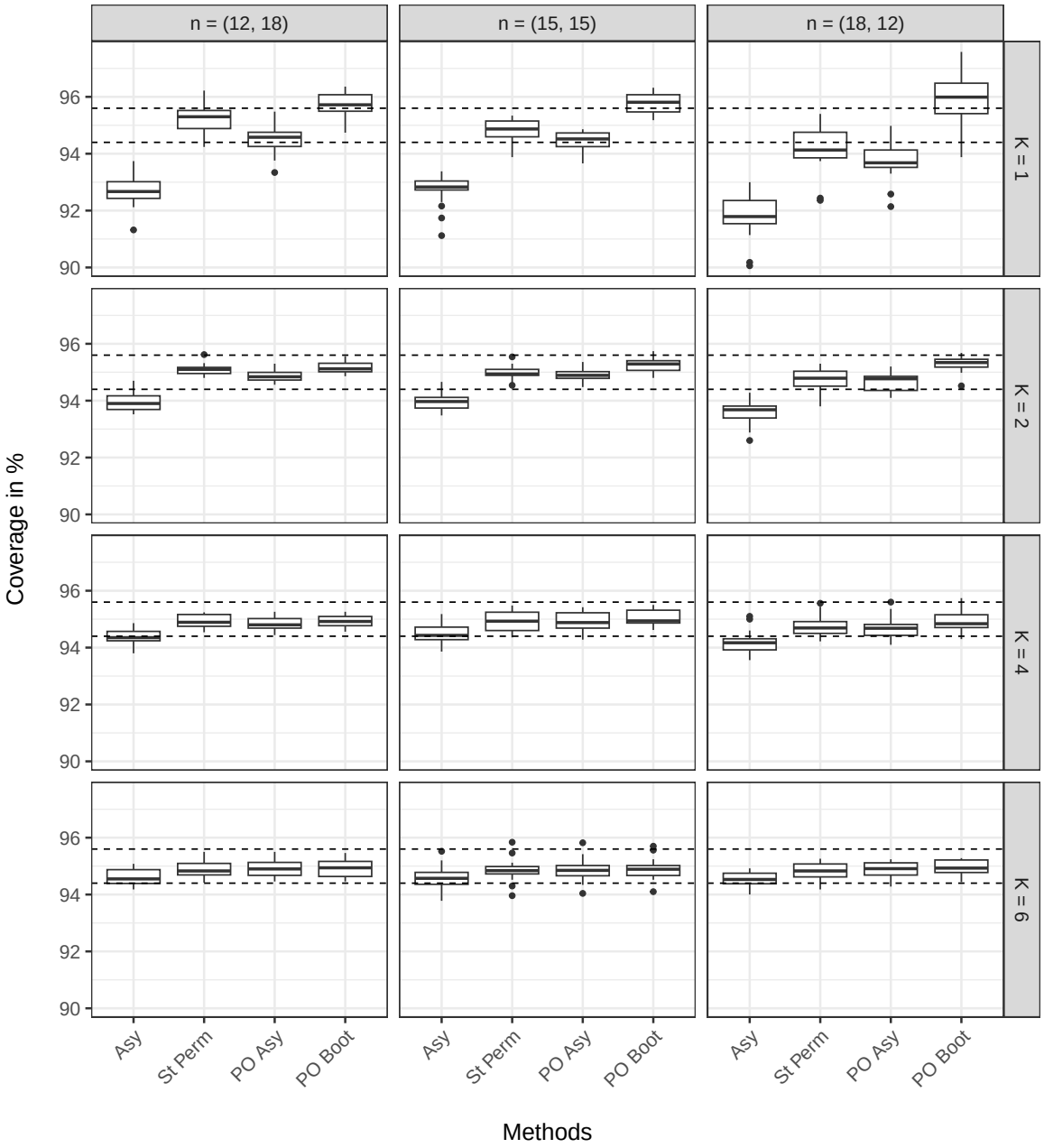

Figure 7 Confidence interval coverage of different methods in % (nominal level $\alpha = 5\%$) aggregated by sample allocations ((n_0, n_1)) and their multipliers (K). The dashed lines depict the 95% binomial confidence interval $[94.4\%, 95.6\%]$.

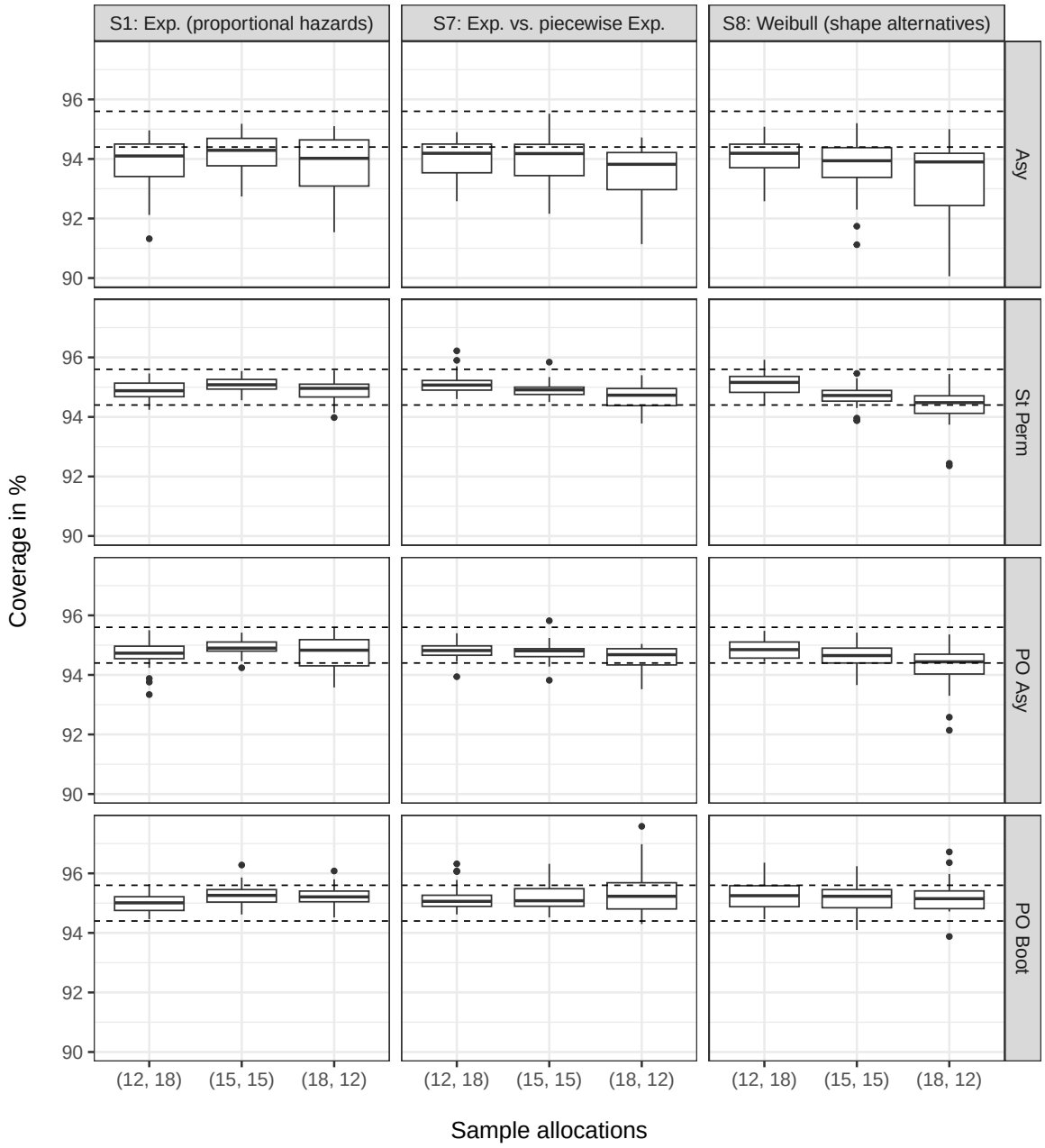

Figure 8 Confidence interval coverage for different sample allocations ((n_0, n_1)) in % (nominal level $\alpha = 5\%$) aggregated by survival models and different methods. The dashed lines depict the 95% binomial confidence interval $[94.4\%, 95.6\%]$.

5 Empirical Examples

It is considered good practice to complement statistical simulation studies with applications to real-world data sets, e.g. to illustrate if or to what extent the choice of method matters in practical applications (Friedrich and Friede 2024). Therefore, we want to apply the methods presented and compared in this thesis to different data sets from the medical literature and see how the results obtained by using these methods differ.

Moreover, we want to look into an aspect of the methods based on pseudo-observations that we have not covered in the simulation study, that is the possibility for directly adjusting for covariates. This facet is particularly interesting as it is known from other contexts that the incorporation of prognostic covariates into the analysis can increase the precision of treatment effect estimates and consequently increase the power of associated tests (Kahan et al. 2014). This is also true for analyses with time-to-event endpoints using the hazard ratio as the effect measure (Kahan et al. 2014) but there also exists some similar evidence for RMST differences already, though not based on pseudo-observations methods (Karrison and Kocherginsky 2018). Investigating this idea further might be especially interesting in view of the finding that the pseudo-observations bootstrap method could be too conservative in null scenarios and vice versa often had less power than its comparators in alternative scenarios. Of course, a systematic assessment of these thoughts requires further research and simulation studies. Nonetheless, we might be able to obtain an initial impression of this idea using exemplary data sets.

In the following, we compare the different methods for estimating and testing RMST differences examined in this thesis using three different data sets. Similarly as for the simulation study, we set the nominal significance level α to 5% and consider a two-sided testing problem. For both, the studentized permutation and the pseudo-observations bootstrap test, we use $B = 5000$ resampling iterations. Moreover, for each example, we consider three different cutoff time points t^* in order to highlight the dependence of the conclusions of this choice. The choice of these time points is made subjectively based on the maximum follow-up time per treatment group. Furthermore, the time points are spaced equidistantly going backward from the initially defined maximum cutoff time point $t^*_{\max}$. As Ditzhaus et al. (2023) noted in their illustration, we also want to emphasize that no adjustments for multiple testing are made. Thus, each combination of a data set, the cutoff time point, and the respective method must be viewed as if it was a single prespecified test although this is not the case.

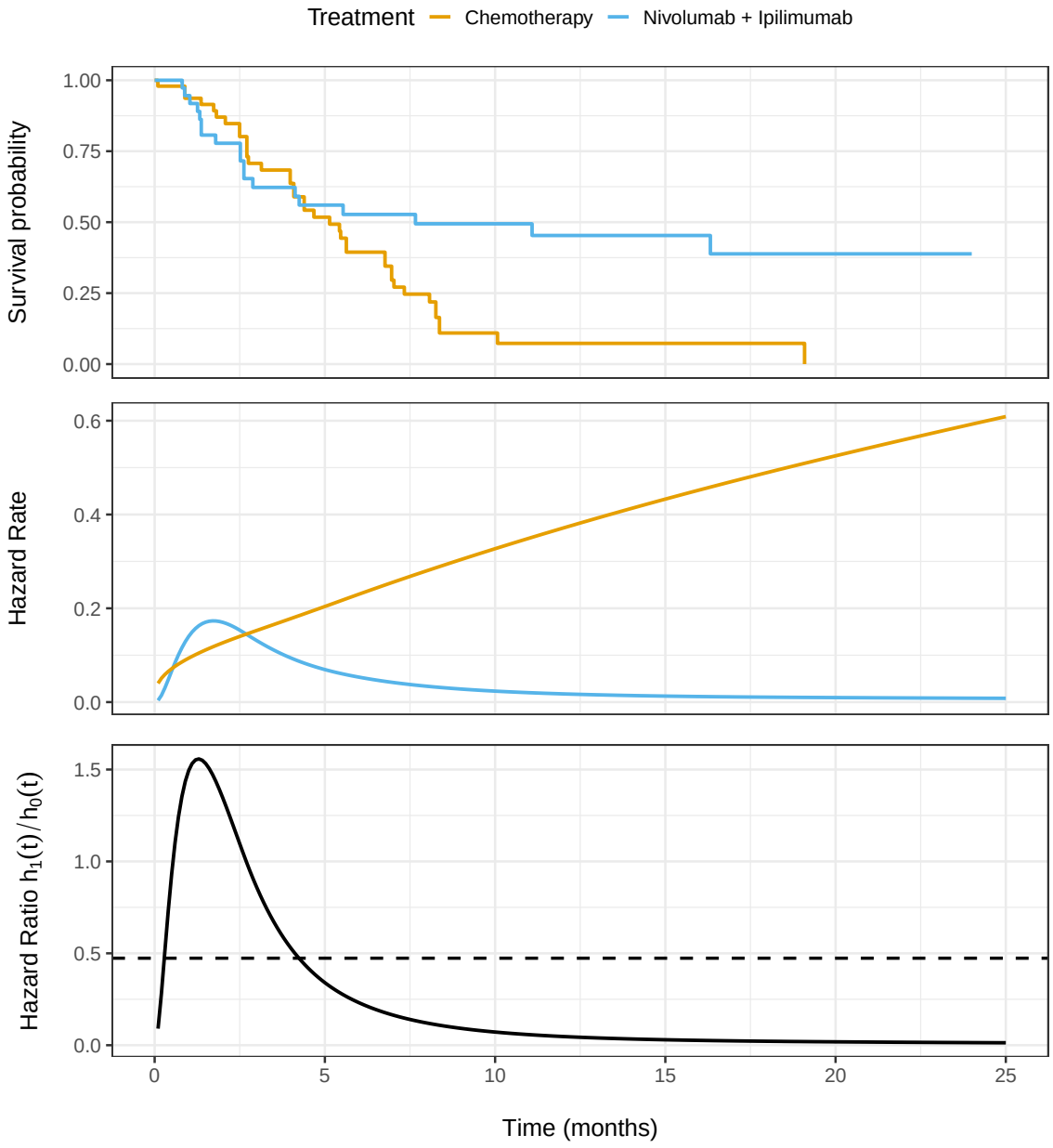

Figure 9 Estimated survival functions for the reconstructed data from Hellmann et al. (2018) obtained from the Kaplan-Meier estimator (survival function) and from using flexible parametric models (hazard functions). The dashed line in the bottom panel represents the constant hazard ratio estimated by the Cox model.

5.1 Hellmann et al. (2018)

The first example data set is the same as the one presented by Ditzhaus et al. (2023), which the authors have reconstructed from a Kaplan-Meier plot in Hellmann et al. (2018) using the algorithm by Guyot et al. (2012). In the study by Hellmann et al. (2018), the authors investigate the effect of nivolumab plus ipilimumab compared to chemotherapy for the treatment of non-small-cell lung cancer on progression-free survival. The original population in the study included 299 patients with a high tumor mutational burden. Besides the main study the authors also conducted some subgroup analyses. For one of these analyses, they stratified the sample based on the tumor PD-L1 (programmed death ligand 1) expression being less than or greater or equal to 1%. The subpopulation of patients with a PD-L1 expression of $< 1\%$ then consisted of 86 patients only, of which 48 were assigned to chemotherapy and 38 to nivolumab plus ipilimumab.

Figure 9 presents the reconstructed data set using the same methods that we have applied for producing Figure 2, i.e. estimating the survival functions using the Kaplan-Meier estimator and estimating the hazard functions using flexible parametric models with $(3, 2)$ degrees of freedom. Again, the constant hazard ratio estimated by the Cox proportional hazards model is also displayed as a dashed line in the bottom panel. As is often the case in oncology trials comparing immunotherapies with chemotherapy, we can observe a delayed effect of the nivolumab plus ipilimumab treatment. Hence, the hazard rate of the population receiving the nivolumab therapy is initially higher than that of the chemotherapy group but decreases afterward. On the other hand, the hazard rate of the chemotherapy group keeps increasing over time. As a consequence, the estimated survival curves of the two populations cross each other roughly after four months, meaning that before that time point, the survival probability of the chemotherapy group is generally higher than that of patients receiving the nivolumab treatment. After this time point, however, we can see a clear treatment effect in favor of the nivolumab therapy. The shape of the time-dependent hazard ratio also makes the proportional hazards assumption questionable. In fact, it can formally be rejected using the test for proportional hazards by Grambsch and Therneau (1994), resulting in a p-value of 0.01%.

Just as Ditzhaus et al. (2023) did, we now consider RMST-based analyses for the treatment effect of nivolumab plus ipilimumab against chemotherapy using the methods presented in this thesis. Likewise, we consider the cutoff values $t^* \in \{12, 15, 18\}$ months, for which the results are shown in Figure 10. The figure displays the point estimates for each method as well as their 95%-confidence intervals. In addition to that, Table 4 conveys the corresponding p-values for a more detailed comparison of the different methods augmented by the p-value of the log-rank test for testing the null hypothesis $S_1 = S_0$.

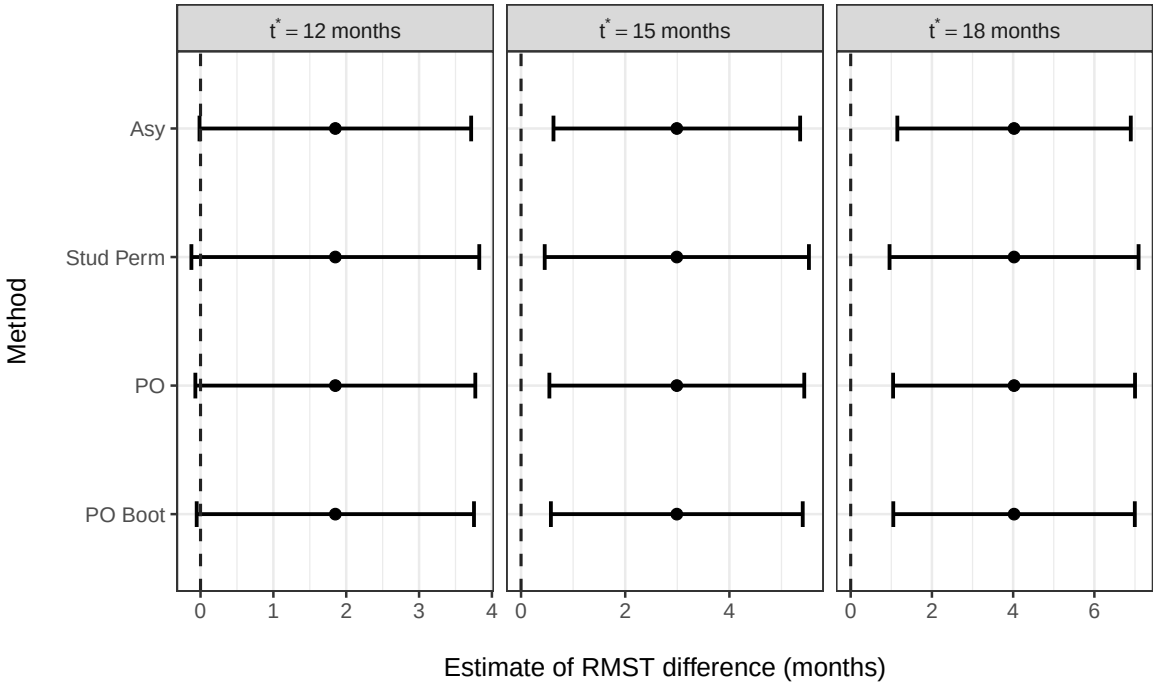

Figure 10 Point estimates of the RMST difference and 95%-confidence intervals for the reconstructed data from Hellmann et al. (2018). The dashed lines highlight an RMST difference of 0, i.e. no treatment difference.

Table 4 P-values in % for the reconstructed data from Hellmann et al. 2018 .

Method	$t^* = 12$ months	$t^* = 15$ months	$t^* = 18$ months
Tests for $\mu_1(t^*) = \mu_0(t^*)$			
Asy	5.2	**1.3**	**0.6**
Perm	6.7	**1.9**	**1.1**
PO1	5.9	**1.7**	**0.8**
PO2	5.6	**1.7**	**1.0**
Test for $S_1 = S_0$			
LR		**1.0**	

Note: Values smaller than or equal to 5% are printed bold.
Abbreviations: Asy, asymptotic test; Perm, studentized permutation test; PO1, pseudo-observations asymptotic; PO2, pseudo-observations bootstrap; LR, log-rank test.

As we have already noted in Section 4.1 and can be seen in Figure 10, the point estimates of the different methods are either exactly the same, as it is the case for the standard asymptotic and the studentized permutation method, or are nearly identical. Therefore, of greater interest are the confidence intervals and test decisions of the different methods. For $t^* = 12$ months, all methods just retain the null hypothesis of no treatment effect difference for a significance level of $\alpha = 5\%$. With regard to the standard asymptotic test, the different conclusion drawn here as opposed to Ditzhaus et al. (2023) is due to the fact that we have used Greenwood's variance estimator (12) instead of the Nelson-Aalen plug-in estimator (13). Among the methods compared, the studentized permutation test suggests the least evidence for concluding that there is a treatment effect followed by the bootstrap approach (see Table 4). For $t^* \in \{15, 18\}$, however, all methods reject the null hypothesis in favor of the alternative that, on average, the nivolumab plus ipilimumab therapy prolongs progression-free survival when comparing it to chemotherapy. Thus, in terms of a binary test decision, all methods lead to the same implications. Nonetheless, we can see that the confidence intervals of the studentized permutation method and those of the pseudo-observations methods are quite similar, whereas those of the standard asymptotic approach are a bit narrower. Therefore, concerning the quantification of uncertainty, the choice of the method does have an impact.

5.2 Edmonson et al. (1979)

The next example is from a study by Edmonson et al. (1979), in which patients suffering from ovarian cancer were randomized to either one of two treatments. The first treatment regimen consisted of cyclophosphamide only, whereas the other group received adriamycin in addition to cyclophosphamide. The data set is publicly available in the R `{survival}` package (Therneau 2023) as the "ovarian" data set.

Like before, the relevant estimated survival quantities are displayed in Figure 11. Looking at the estimated survival functions we might expect a positive treatment effect favoring the combination regimen. Yet, towards the end of the study, the differences in survival probabilities become smaller, making this deduction less clear. Regarding the proportional hazards assumption, the figures also suggest that this assumption should be made with caution. For instance, there is a sharp increase in the hazard rate of the experimental treatment group after ten months up until the 20th month, making the proportionality assumption questionable. However, we must consider that the sample size in this example is extremely small with 13 patients in each treatment arm, only. This may lead to highly variable estimates of the different survival quantities. The Grambsch-Therneau test yields a p-value of 10.15%, thus not providing any evidence against the null hypothesis that the proportional hazards assumption does actually hold.

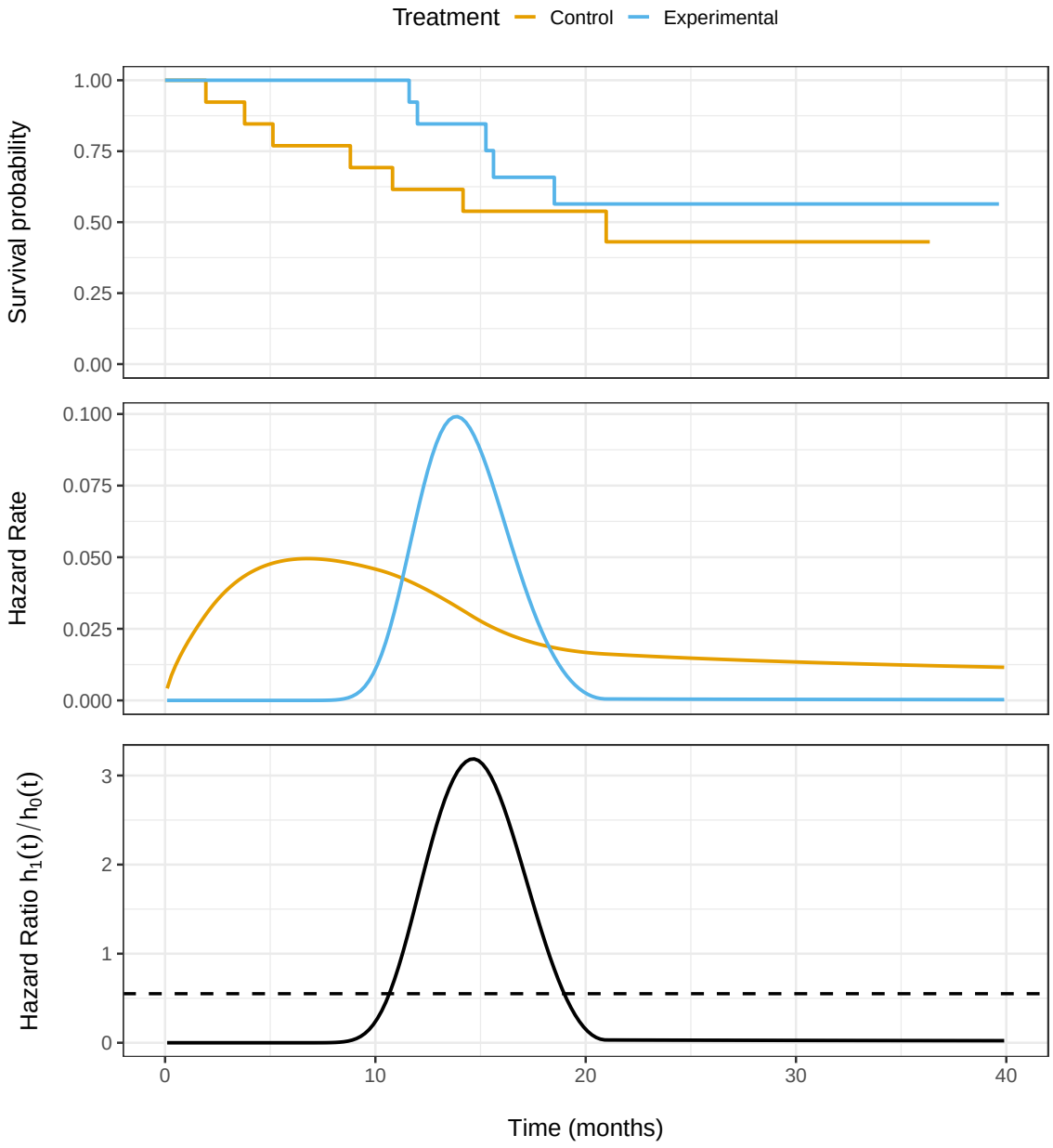

Figure 11 Estimated survival functions for the data from Edmonson et al. (1979) obtained from the Kaplan-Meier estimator (survival function) and from using flexible parametric models (hazard functions). The dashed line in the bottom panel represents the constant hazard ratio estimated by the Cox model.

Nonetheless, we keep considering RMST-based analyses, again choosing three different cutoff time points, here $t^* \in \{15, 20, 25\}$ months. Moreover, what makes this data set interesting is the fact that, in addition to the endpoint variables and the treatment assignment variable, it consists of additional covariates. For instance, it contains the age of the patients as well as their ECOG (Eastern Cooperative Oncology Group) performance score measured at baseline. The latter is an ordinal score describing the physical condition of the patient, ranging from 0 (no restrictions) to 5 (dead) (Oken et al. 1982). In the given example, however, only patients with a score of 1 or 2 were observed. Therefore, we can reduce this variable to a binary covariate, using the score 2 as the reference category. Ultimately, this means that, for the given example, we can compare not only four but six different methods for estimating the RMST difference between the treatment regimens. The two added methods result from adjusted versions of the pseudo-observations approaches using the aforementioned covariates. Corresponding results are presented in Figure 12 and Table 6 for the effect estimates and p-values, respectively.

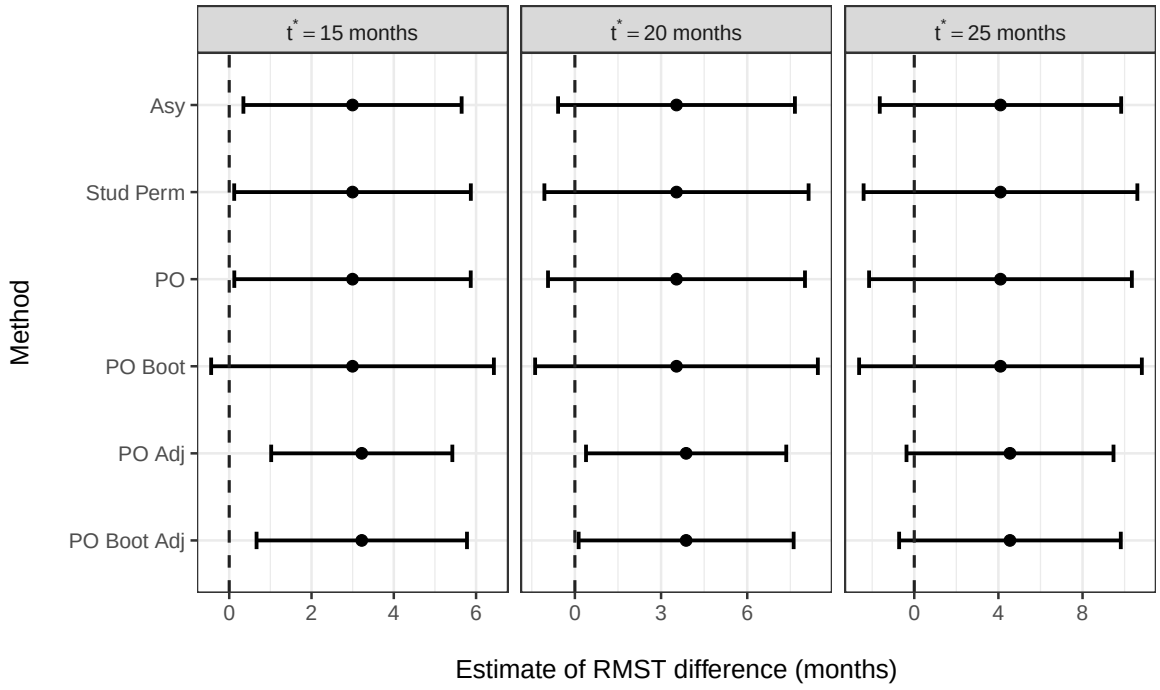

Figure 12 Point estimates of the RMST difference and 95%-confidence intervals for the data from Edmonson et al. (1979). The dashed lines highlight an RMST difference of 0, i.e. no treatment difference.

As previously, the point estimates of the unadjusted approaches are (nearly) identical, resulting in estimated RMST differences of about 3 ($t^* = 15$), 3.5 ($t^* = 20$) and 4 ($t^* = 25$) months. Regarding the uncertainty quantification and corresponding test decisions, the findings are a little more diverse than for the Hellmann example, however. Particularly, for $t^* = 15$ months we have that all of the unadjusted approaches except for the bootstrap method reject the null hypothesis of no RMST difference. There, the

Table 5 P-values in % for the data from Edmonson et al. 1979 .

Method	$t^* = 15$ months	$t^* = 20$ months	$t^* = 25$ months
Tests for $\mu_1(t^*) = \mu_0(t^*)$			
Asy	**2.7**	9.3	16.2
Perm	**4.6**	12.4	18.9
PO1	**4.1**	12.1	19.8
PO2	7.1	12.3	19.6
Adjusted tests for $\mu_1(t^*) = \mu_0(t^*)$			
PO1 Adj	**0.4**	**2.9**	7.0
PO2 Adj	**2.7**	**4.4**	7.9
Test for $S_1 = S_0$			
LR		30.3	

Note: Values smaller than or equal to 5% are printed bold.
Abbreviations: Asy, asymptotic test; Perm, studentized permutation test; PO1, pseudo-observations asymptotic; PO2, pseudo-observations bootstrap; LR, log-rank test.

standard asymptotic approach suggests the strongest evidence for this rejection (p-value of 2.7%). On the other hand, for $t^* \in \{20, 25\}$ all of the unadjusted methods retain this null hypothesis for the significance level $\alpha = 5\%$ (see Table 5), i.e. implying similar conclusions. The log-rank test also supports the statement that no treatment effect can be detected, knowing well, however, that it tests a different null hypothesis.

Of greater interest, however, are the results that we obtain using the pseudo-observations methods with which we adjust for additional covariates. As mentioned before, the adjusted methods include the age of the patients and their ECOG performance score as further covariates. In Figure 12 we can see that adjustment for covariates leads to both, a slightly larger treatment effect estimate as well as narrower confidence intervals associated with it. For $t^* = 20$ months, this implies that in contrast to all unadjusted methods the null hypothesis is rejected for $\alpha = 5\%$ (see Table 5). This even holds when using the bootstrap method that we have observed to be rather conservative in scenarios with such small sample sizes (Section 4.3.1). Nonetheless, for $t^* = 25$ months, the null hypothesis is also retained by the adjusted methods. However, the p-values are much smaller than those of the unadjusted methods and the fact that the confidence intervals are narrower remains.

5.3 Grana et al. (2002)

Lastly, we consider another study from the field of oncology by Grana et al. (2002). The authors conducted a controlled trial among 37 patients with high-grade glioma, which is a class of tumors concerning the central nervous system. All of the patients had received surgery and radiotherapy before the trial. The authors of the study were then interested in the effect of adjuvant intralesional radioimmunotherapy on two endpoints,

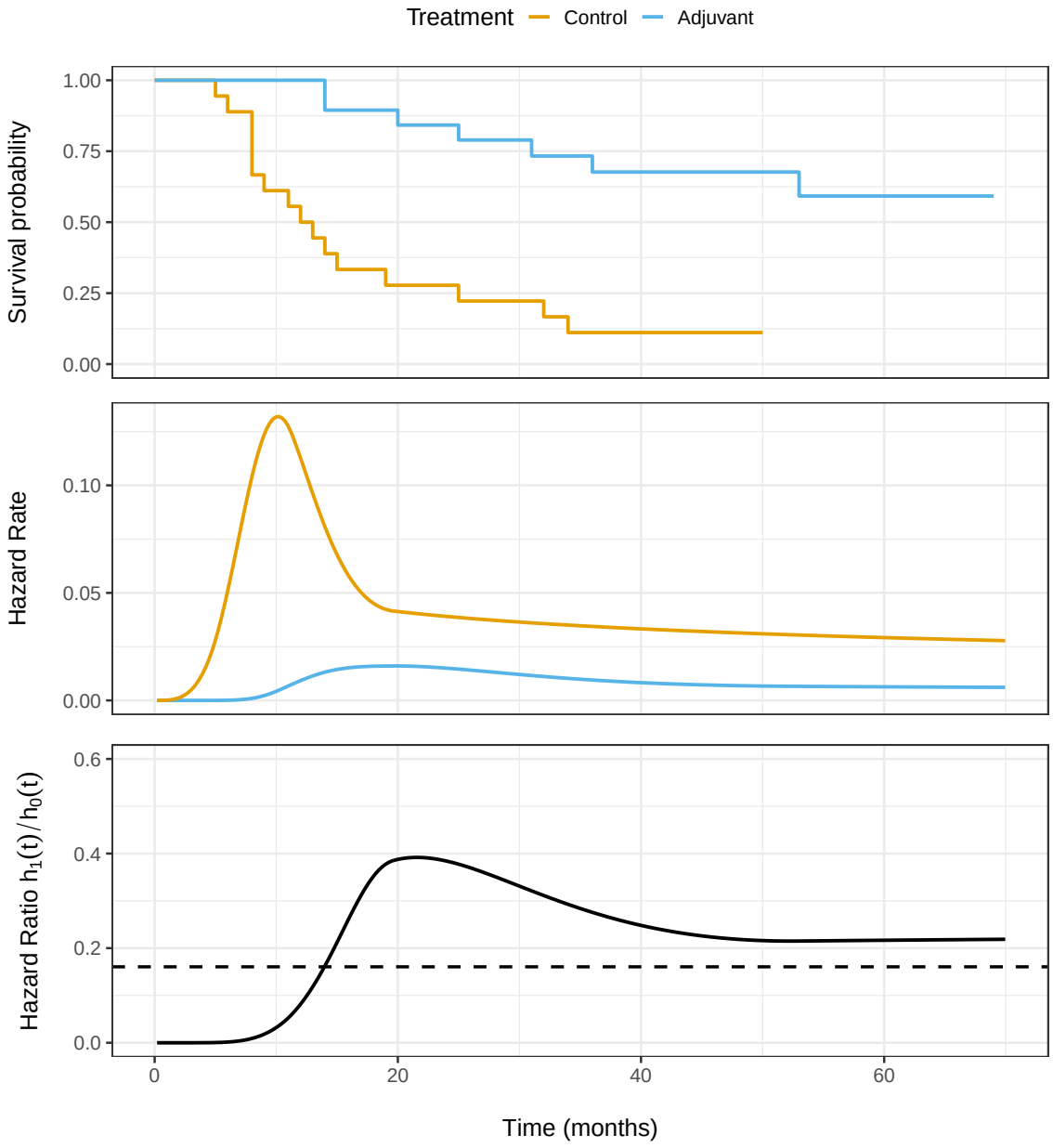

Figure 13 Estimated survival functions for the data from Grana et al. (2002) obtained from the Kaplan-Meier estimator (survival function) and from using flexible parametric models (hazard functions). The dashed line in the bottom panel represents the constant hazard ratio estimated by the Cox model.

disease-free survival and overall survival. Here, we consider the overall survival endpoint, only. The data were made publicly available by the authors in the form of tables in their paper. Besides the time-to-event endpoints and the treatment assignment, they also recorded further characteristics, including the age of the patients as well as the glioma type (grade III or glioblastoma). Furthermore, it is worth noticing that the treatment assignment was not randomized.

Again, the estimated survival and hazard functions of the two treatment regimens as well as the time-dependent hazard ratio are depicted in Figure 13. A visual inspection suggests to expect a treatment effect in favor of the adjuvant therapy. This impression is even more marked here than for the Edmonson example as the separation of the two survival curves is more consistent over time. Although there is still some variability in the estimated time-dependent hazard ratio it is much more stable than that from the Edmonson example. As a result, the p-value of the Grambsch-Therneau test is even larger with a value of 19.95%.

In their publication, Grana et al. (2002) analyzed the data separately for the two different glioma types using the log-rank test as the primary analysis tool. For the RMST-based analyses, we consider the whole data set instead for simplicity. With respect to the adjusted methods we then consider the glioma type as well as the age of the patients as control variables. As in the previous examples, we also compare the unadjusted methods with each other.

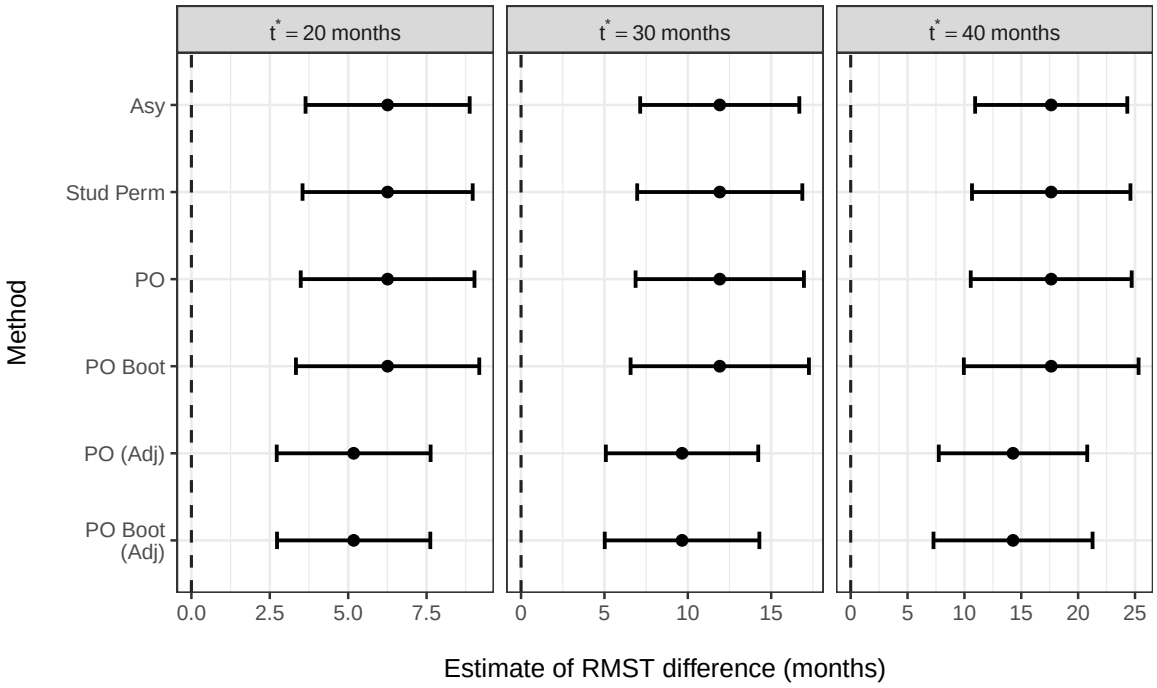

Figure 14 Point estimates of the RMST difference and 95%-confidence intervals for the data from Grana et al. (2002). The dashed lines highlight an RMST difference of 0, i.e. no treatment difference.

Regardless of the chosen cutoff time point ($t^* \in \{20, 30, 40\}$), in this example, all tests

Table 6 P-values in % for the data from Grana et al. 2002 .

Method	$t^* = 20$ months	$t^* = 30$ months	$t^* = 40$ months
Tests for $\mu_1(t^*) = \mu_0(t^*)$			
Asy	**<0.1**	**<0.1**	**<0.1**
Perm	**<0.1**	**<0.1**	**<0.1**
PO1	**<0.1**	**<0.1**	**<0.1**
PO2	**<0.1**	**<0.1**	**<0.1**
Adjusted tests for $\mu_1(t^*) = \mu_0(t^*)$			
PO1 Adj	**<0.1**	**<0.1**	**<0.1**
PO2 Adj	**0.2**	**0.1**	**<0.1**
Test for $S_1 = S_0$			
LR		**<0.1**	

Note: Values smaller than or equal to 5% are printed bold.
Abbreviations: Asy, asymptotic test; Perm, studentized permutation test; PO1, pseudo-observations asymptotic; PO2, pseudo-observations bootstrap; LR, log-rank test.

make the same decision, signifying a positive treatment effect of the adjuvant therapy (Table 6). As can be expected visually from Figure 13, the log-rank test does not deviate from this decision.

When taking a closer look at the point estimates and their confidence intervals in Figure 14 we can see two things. As for the previous two examples, the width of the confidence intervals varies slightly between the different methods. The patterns are consistent with our other findings, i.e. the standard asymptotic approach has the narrowest confidence interval and the bootstrap method has the widest one. Regarding the adjusted methods, there is a similar effect as we have seen in the Edmonson example. Hence, there is a shift in the treatment effect estimate and the width of the confidence intervals shrinks. Nonetheless, there are qualitative differences in this comparison. Previously, adjusting for covariates led to an effect estimate that has a larger magnitude. Here, the opposite is the case, i.e. covariate adjustment leads to a smaller effect estimate. While the confidence intervals obtained using covariate adjustment are smaller than those of the unadjusted methods, the differences are less remarkable in this example.

6 Conclusions

This thesis dealt with conducting statistical inference to compare two groups with a time-to-event endpoint under non-proportional hazards. We further focused on procedures for which the effect measure and the statistical test are aligned with each other. This led us to an assessment of the restricted mean survival time (RMST) as a population-level summary measure. We reviewed the literature about corresponding methods for estimating and testing such effects and found that the standard asymptotic

test based on a standard normal approximation suffers from an inflated type I error rate in finite sample settings. While there exist a couple of proposed methods dealing with this problem we focused on the studentized permutation test by Ditzhaus et al. (2023) since these authors demonstrated it to be superior to the competing approaches overall. In addition, we proposed two further methods for such scenarios based on pseudo-observations regression models. The first of these two approaches also uses an asymptotic test but the estimation of the standard error of the effect estimate is carried out differently. Here, the HC3 covariance matrix estimator is used, which we believed to be more robust in settings with rather small sample sizes based on results from previous research. Furthermore, we implemented a nonparametric bootstrap test for such regression models such that we would have an additional method that does not make a fixed assumption about the (asymptotic) distribution of the test statistic used for testing the effect. For this approach, we encountered the computational challenge of dealing with a nested resampling procedure with the calculation of the pseudo-observations on the first level and the bootstrap procedure on the second level. We have addressed this problem by replacing the usage of ordinary pseudo-observations with infinitesimal jackknife pseudo-observations, which can be computed much faster. After introducing the existing and our proposed approaches we set up a simulation study for empirically investigating and comparing them to each other. For this, we adapted certain characteristics from the simulation study by Ditzhaus et al. (2023), making the simulation study less subjective. Finally, we illustrated the application of all methods on empirical data sets from past clinical trials.

In summary, this thesis delivered two main contributions to existing research. On the one hand, we validated the results by Ditzhaus et al. (2023) regarding their proposed studentized permutation method for two-sample RMST-based tests, i.e. we could confirm that its operating characteristics are better than those of the standard asymptotic approach in a vast majority of different scenarios. On the other hand, we extended the repertoire of statistical methods for such tests for scenarios with moderate sample sizes by the two proposed methods based on pseudo-observations. For these, we could show that they provide further valid alternatives that should be preferred over the standard asymptotic test. The performance of both methods was only slightly inferior, overall, to that of the studentized permutation test. For the asymptotic pseudo-observations method this was due to a slightly too liberal behavior in settings with very small sample sizes, whereas the bootstrap approach was a little too conservative in these scenarios. Ditzhaus et al. (2023) mention that their studentized permutation method can directly be extended to other settings in survival analysis, e.g. to situations with competing risks where multiple causes for the event of interest must be considered and accounted for. They also refer to related summary measures such as window mean survival time (Paukner and Chappell 2021). The same arguments apply to the pseudo-observations

methods presented in this thesis and their implementation might even be considered more straightforward as existing software solutions can be applied directly.

Casting estimation and testing problems for the RMST into the framework of generalized linear models makes the pseudo-observations approaches attractive alternatives that should be considered further, both by practitioners and by methodological researchers in statistics. The most interesting feature of the pseudo-observations methods can be considered the ease of incorporating prognostic covariates into the analysis, which can increase the precision and power of the effect estimate and the statistical test, respectively. The demonstration on real-world data sets in Section 5 already gave a flavor to this idea. While for the pseudo-observations methods, we simply need to add the corresponding covariates into the regression model, the other approaches considered in this thesis would need to be modified, e.g. by means of stratification. However, even if we implemented such modifications to the other methods, we still might expect them to be less effective in taking advantage of the adjustment for covariates than the pseudo-observations approaches. For instance, taking a continuous covariate into account, using the pseudo-observations methods we can make full use of this information by simply incorporating it into the regression model. Using an approach based on stratification, on the other hand, would require us to more or less arbitrarily divide the continuous variable into discrete categories. Especially for situations with small sample sizes as considered in this thesis, this might lead to small subgroups and therefore imprecise and volatile effect estimates. These ideas could further be investigated and systematically evaluated by means of simulation studies in future research. Moreover, it would be interesting to see, whether covariate adjustment can diminish the conservativeness of the bootstrap method proposed in this thesis. One particular challenge in conducting such a simulation study would consist of setting up proper models for simulating survival data conditional on other covariates than the treatment indicator for a given RMST difference Δ. Other aspects and questions arising from these ideas are, for instance, the effects of incorporating non-prognostic covariates into the analysis or the misspecification of the functional form of an effect.

Another aspect that could be studied further is the application of other resampling procedures for pseudo-observations regression models as alternatives to the nonparametric bootstrap presented in this thesis. While the nonparametric bootstrap makes only few assumptions and is straightforward to implement, alternative resampling schemes may exhibit a better performance. One particular method we think of is the wild bootstrap (Liu 1988). Intuitively, we think that such an approach can also work for pseudo-observations regression models. However, theoretical considerations need to be made in this context. First, we need to keep in mind that we employ quasi-likelihood methods and therefore need to figure out which resampling methods are valid for these types of models. Second, it is not clear in how far the usage of pseudo-observations

alters or complicates things as opposed to situations in which the response vector is fully observed and used as-is. Besides potential improvements in terms of operating characteristics, what makes these ideas interesting is that they would be computationally much more efficient than the bootstrap procedure proposed in this thesis as we would avoid recalculating the pseudo-observations in each bootstrap iteration.

Lastly, we note that Munko et al. (2024) consider extensions of RMST-based tests to settings with more than two populations. They investigate how to conduct a test for the global null hypothesis of equal RMSTs across all groups as well as carrying out multiple contrast tests for the pairwise RMST differences. Similarly, the pseudo-observations approaches proposed in this thesis might be used and investigated for such problems.

References

Ambrogi, Federico, Simona Iacobelli, and Per Kragh Andersen (2022). "Analyzing differences between restricted mean survival time curves using pseudo-values". *BMC Medical Research Methodology* 22.1.

Andersen, Per Kragh (2003). "Generalised linear models for correlated pseudo-observations, with applications to multi-state models". *Biometrika* 90.1, pp. 15–27.

Andersen, Per Kragh and Maja Pohar Perme (2010). "Pseudo-observations in survival analysis". *Statistical Methods in Medical Research* 19.1, pp. 71–99.

Andersen, Per Kragh, Elisavet Syriopoulou, and Erik T. Parner (2017). "Causal inference in survival analysis using pseudo-observations". *Statistics in Medicine* 36.17, pp. 2669–2681.

Bardo, Maximilian, Cynthia Huber, Norbert Benda, Jonas Brugger, Tobias Fellinger, Vaidotas Galaune, Judith Heinz, Harald Heinzl, Andrew C Hooker, Florian Klinglmüller, Franz König, Tim Mathes, Martina Mittlböck, Martin Posch, Robin Ristl, and Tim Friede (2024). "Methods for non-proportional hazards in clinical trials: A systematic review". *Statistical Methods in Medical Research*.

Binder, Nadine, Thomas A. Gerds, and Per Kragh Andersen (2014). "Pseudo-observations for competing risks with covariate dependent censoring". *Lifetime Data Analysis* 20.2, pp. 303–315.

Collett, David (2015). *Modelling Survival Data in Medical Research*. 3rd ed. New York: Chapman and Hall/CRC.

Cox, D. R. (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2, pp. 187–220.

Ditzhaus, Marc, Menggang Yu, and Jin Xu (2023). "Studentized permutation method for comparing two restricted mean survival times with small sample from randomized trials". *Statistics in Medicine* 42.13, pp. 2226–2240.

Dormuth, Ina, Tiantian Liu, Jin Xu, Markus Pauly, and Marc Ditzhaus (2023). "A comparative study to alternatives to the log-rank test". *Contemporary Clinical Trials* 128.

Dormuth, Ina, Tiantian Liu, Jin Xu, Menggang Yu, Markus Pauly, and Marc Ditzhaus (2022). "Which test for crossing survival curves? A user's guideline". *BMC Medical Research Methodology* 22.1.

Edmonson, J. H., T. R. Fleming, D. G. Decker, G. D. Malkasian, E. O. Jorgensen, J. A. Jefferies, M. J. Webb, and L. K. Kvols (1979). "Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease". *Cancer Treatment Reports* 63.2, pp. 241–247.

Efron, Bradley and Robert J. Tibshirani (1993). *An Introduction to the Bootstrap*. Boston, MA: Springer US.

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang, and Brian Marx (2013). *Regression: Models, Methods and Applications*. Berlin, Heidelberg: Springer.

Freidlin, Boris and Edward L. Korn (2019). "Methods for Accommodating Nonproportional Hazards in Clinical Trials: Ready for the Primary Analysis?" *Journal of Clinical Oncology* 37.35, pp. 3455–3459.

Friedrich, Sarah and Tim Friede (2024). "On the role of benchmarking data sets and simulations in method comparison studies". *Biometrical Journal* 66.1.

Grambsch, Patricia M. and Terry M. Therneau (1994). "Proportional hazards tests and diagnostics based on weighted residuals". *Biometrika* 81.3, pp. 515–526.

Grana, C., M. Chinol, C. Robertson, C. Mazzetta, M. Bartolomei, C. De Cicco, M. Fiorenza, M. Gatti, P. Caliceti, and G. Paganelli (2002). "Pretargeted adjuvant radioimmunotherapy with Yttrium-90-biotin in malignant glioma patients: A pilot study". *British Journal of Cancer* 86.2, pp. 207–212.

Guyot, Patricia, AE Ades, Mario JNM Ouwens, and Nicky J. Welton (2012). "Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves". *BMC Medical Research Methodology* 12.1.

Hasegawa, Takahiro, Saori Misawa, Shintaro Nakagawa, Shinichi Tanaka, Takanori Tanase, Hiroyuki Ugai, Akira Wakana, Yasuhide Yodo, Satoru Tsuchiya, Hideki Suganami, and for the JPMA Task Force Members (2020). "Restricted mean survival time as a summary measure of time-to-event outcome". *Pharmaceutical Statistics* 19.4, pp. 436–453.

Hellmann, Matthew D., Tudor-Eliade Ciuleanu, Adam Pluzanski, Jong Seok Lee, Gregory A. Otterson, Clarisse Audigier-Valette, Elisa Minenza, Helena Linardou, Sjaak Burgers, Pamela Salman, Hossein Borghaei, Suresh S. Ramalingam, Julie Brahmer, Martin Reck, Kenneth J. O'Byrne, William J. Geese, George Green, Han Chang, Joseph Szustakowski, Prabhu Bhagavatheeswaran, Diane Healey, Yali Fu, Faith Nathan, and Luis Paz-Ares (2018). "Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden". *New England Journal of Medicine* 378.22, pp. 2093–2104.

Hester, Jim, Lionel Henry, Kirill Müller, Kevin Ushey, Hadley Wickham, and Winston Chang (2022). *withr: Run Code With Temporarily Modified Global State*. R package version 2.5.0.

Horiguchi, Miki and Hajime Uno (2020). "On permutation tests for comparing restricted mean survival time with small sample from randomized trials". *Statistics in Medicine* 39.20, pp. 2655–2670.

Kahan, Brennan C., Vipul Jairath, Caroline J. Doré, and Tim P. Morris (2014). "The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies". *Trials* 15.1.

Kalbfleisch, J. D. and Ross L. Prentice (2002). *The statistical analysis of failure time data*. 2nd ed. Wiley series in probability and statistics. Hoboken, N.J: J. Wiley.

Kaplan, E. L. and Paul Meier (1958). "Nonparametric Estimation from Incomplete Observations". *Journal of the American Statistical Association* 53.282, pp. 457–481.

Karrison, Theodore and Masha Kocherginsky (2018). "Restricted mean survival time: Does covariate adjustment improve precision in randomized clinical trials?" *Clinical Trials* 15.2, pp. 178–188.

Liu, Regina Y. (1988). "Bootstrap Procedures under some Non-I.I.D. Models". *The Annals of Statistics* 16.4, pp. 1696–1708.

Long, J. Scott and Laurie H. Ervin (2000). "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model". *The American Statistician* 54.3, pp. 217–224.

MacKinnon, James G and Halbert White (1985). "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties". *Journal of Econometrics* 29.3, pp. 305–325.

MacKinnon, James G. (2009). "Bootstrap Hypothesis Testing". In: *Handbook of Computational Econometrics*. John Wiley & Sons, Ltd, pp. 183–213.

Manner, David H., Chakib Battioui, Stefan Hantel, B. Nhi Beasley, Lee-Jen Wei, Mary Jane Geiger, J. Rick Turner, and Markus Abt (2019). "Restricted mean survival time for the analysis of cardiovascular outcome trials assessing non-inferiority: Case studies from antihyperglycemic drug development". *American Heart Journal* 215, pp. 178–186.

Morris, Tim P., Ian R. White, and Michael J. Crowther (2019). "Using simulation studies to evaluate statistical methods". *Statistics in Medicine* 38.11, pp. 2074–2102.

Munko, Merle, Marc Ditzhaus, Dennis Dobler, and Jon Genuneit (2024). "RMST-based multiple contrast tests in general factorial designs". *Statistics in Medicine*, pp. 1849–1866.

Oken, M. M., R. H. Creech, D. C. Tormey, J. Horton, T. E. Davis, E. T. McFadden, and P. P. Carbone (1982). "Toxicity and response criteria of the Eastern Cooperative Oncology Group". *American Journal of Clinical Oncology* 5.6, pp. 649–655.

Overgaard, Morten, Erik Thorlund Parner, and Jan Pedersen (2019). "Pseudo-observations under covariate-dependent censoring". *Journal of Statistical Planning and Inference* 202, pp. 112–122.

Parner, Erik T., Per Kragh Andersen, and Morten Overgaard (2023). "Regression models for censored time-to-event data using infinitesimal jack-knife pseudo-observations, with applications to left-truncation". *Lifetime Data Analysis* 29.3, pp. 654–671.

Paukner, Mitchell and Richard Chappell (2021). "Window mean survival time". *Statistics in Medicine* 40.25, pp. 5521–5533.

Peto, Richard and Julian Peto (1972). "Asymptotically Efficient Rank Invariant Test Procedures". *Journal of the Royal Statistical Society. Series A (General)* 135.2, pp. 185–207.

Quartagno, Matteo, Tim P Morris, Duncan C Gilbert, Ruth E Langley, Matthew G Nankivell, Mahesh KB Parmar, and Ian R White (2023). "A comparison of different population-level summary measures for randomised trials with time-to-event outcomes, with a focus on non-inferiority trials". *Clinical Trials*, pp. 594–602.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Robert, Caroline, Georgina V. Long, Benjamin Brady, Caroline Dutriaux, Michele Maio, Laurent Mortier, Jessica C. Hassel, Piotr Rutkowski, Catriona McNeil, Ewa Kalinka-Warzocha, Kerry J. Savage, Micaela M. Hernberg, Celeste Lebbé, Julie Charles, Catalin Mihalcioiu, Vanna Chiarion-Sileni, Cornelia Mauch, Francesco Cognetti, Ana Arance, Henrik Schmidt, Dirk Schadendorf, Helen Gogas, Lotta Lundgren-Eriksson, Christine Horak, Brian Sharkey, Ian M. Waxman, Victoria Atkinson, and Paolo A. Ascierto (2015). "Nivolumab in Previously Untreated Melanoma without *BRAF* Mutation". *New England Journal of Medicine* 372.4, pp. 320–330.

Royston, Patrick and Mahesh K. B. Parmar (2020). "A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome". *Trials* 21.1.

Royston, Patrick and Mahesh K. B. Parmar (2002). "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects". *Statistics in Medicine* 21.15, pp. 2175–2197.

Royston, Patrick and Mahesh K. B. Parmar (2011). "The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt". *Statistics in Medicine* 30.19, pp. 2409–2421.

Rufibach, Kaspar (2019). "Treatment effect quantification for time-to-event endpoints–Estimands, analysis strategies, and beyond". *Pharmaceutical Statistics* 18.2, pp. 145–165.

Sachs, Michael C. and Erin E. Gabriel (2022). "Event History Regression with Pseudo-Observations: Computational Approaches and an Implementation in R". *Journal of Statistical Software* 102, pp. 1–34.

Therneau, Terry M (2023). *survival: Survival Analysis*. R package version 3.5-6.

Uno, Hajime, Brian Claggett, Lu Tian, Eisuke Inoue, Paul Gallo, Toshio Miyata, Deborah Schrag, Masahiro Takeuchi, Yoshiaki Uyama, Lihui Zhao, Hicham Skali, Scott Solomon, Susanna Jacobus, Michael Hughes, Milton Packer, and Lee-Jen Wei (2014). "Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis". *Journal of Clinical Oncology* 32.22, pp. 2380–2385.

Ushey, Kevin and Hadley Wickham (2023). *renv: Project Environments*. R package version 1.0.0.

Zeileis, Achim (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimators". *Journal of Statistical Software* 11, pp. 1–17.

Zeileis, Achim (2006). "Object-oriented Computation of Sandwich Estimators". *Journal of Statistical Software* 16, pp. 1–16.

Zeileis, Achim, Susanne Köll, and Nathaniel Graham (2020). "Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R". *Journal of Statistical Software* 95, pp. 1–36.

Zhao, Lihui, Brian Claggett, Lu Tian, Hajime Uno, Marc A. Pfeffer, Scott D. Solomon, Lorenzo Trippa, and L. J. Wei (2016). "On the restricted mean survival time curve in survival analysis". *Biometrics* 72.1, pp. 215–221.

Zhou, Mai (2021). "Restricted mean survival time and confidence intervals by empirical likelihood ratio". *Journal of Biopharmaceutical Statistics* 31.3, pp. 362–374.

Appendices

A Additional Simulation Results

In the following, we present the simulation results for the pseudo-observations method using an asymptotic test that we have already presented in Section 4.3. We compare these results to those using the same asymptotic procedure but based on infinitesimal jackknife (IJ) pseudo-observations that we have used for the bootstrap method introduced in Section 3.3.3. The purpose is to get an impression of how the usage of IJ pseudo-observations instead of ordinary ones impacts the behavior of the testing procedure.

In general, the results suggest that the impact of using IJ pseudo-observations instead of ordinary pseudo-observations is rather low as the type I error rates and the power values are fairly similar. Looking at the plot of the coverage rates (Figure 15), however, we can see that for very small ($K = 1$) and unbalanced sample sizes there is a small impact. In this sense, the usage of ordinary pseudo-observations is still favorable whenever feasible. Nonetheless, the present results suggest that the differences in the operating characteristics between the asymptotic test using ordinary and the bootstrap test using IJ pseudo-observations shown in Section 4.3 can primarily attributed to the nonparametric bootstrap procedure.

Table 7 Type I error rates in % (nominal level $\alpha = 5\%$) of asymptotic tests using ordinary and infinitesimal jackknife pseudo-observations. The values inside the binomial confidence interval $[4.4\%, 5.6\%]$ are printed bold .

		$N = K \cdot (12, 18)$		$N = K \cdot (15, 15)$		$N = K \cdot (18, 12)$	
Censoring	K	PO	PO IJ	PO	PO IJ	PO	PO IJ
S1: Exponential distributions							
un. W.	1	**5.2**	**5.3**	5.8	5.8	6.4	6.5
	2	**5.2**	**5.4**	**5.1**	**5.2**	**5.2**	**5.4**
	4	**4.9**	**5.0**	**4.6**	**4.7**	**5.2**	**5.3**
	6	**4.5**	**4.6**	**5.0**	**5.0**	**4.8**	**4.9**
eq. U.	1	6.7	6.6	**5.5**	**5.5**	6.4	6.5
	2	**5.4**	**5.5**	**4.9**	**5.0**	**5.2**	**5.4**
	4	**5.4**	**5.5**	**4.8**	**4.9**	**5.3**	**5.4**
	6	**4.9**	**5.0**	**5.4**	**5.4**	**4.9**	**4.9**
eq. W.	1	6.1	6.1	**5.4**	**5.3**	6.1	5.9
	2	**5.0**	**5.0**	**5.2**	**5.2**	**4.8**	**4.8**
	4	**5.3**	**5.3**	**4.6**	**4.6**	**5.2**	**5.2**
	6	**5.5**	**5.5**	**5.2**	**5.2**	**4.8**	**4.8**
S7: Exponential and piecewise exponential distributions with crossing curves							
un. W.	1	**4.6**	**4.8**	**5.3**	**5.5**	6.3	6.6
	2	**5.3**	**5.4**	**5.1**	**5.3**	5.9	6.1
	4	**5.1**	**5.2**	**5.2**	**5.3**	5.7	5.8
	6	**4.9**	**4.9**	**5.2**	**5.3**	**5.0**	**5.0**
eq. U.	1	**5.0**	**5.0**	5.7	5.8	**5.4**	**5.4**
	2	**5.4**	**5.5**	**5.4**	**5.6**	**5.1**	**5.3**
	4	**5.3**	**5.3**	**5.1**	**5.2**	**5.3**	**5.5**
	6	**4.8**	**4.9**	4.2	4.2	**5.3**	**5.3**
eq. W.	1	6.1	6.1	**5.3**	**5.2**	6.5	6.4
	2	**5.0**	**5.0**	**4.8**	**4.8**	**5.3**	**5.3**
	4	**4.7**	**4.8**	**5.1**	**5.1**	**5.5**	**5.5**
	6	**4.9**	**4.9**	**5.4**	**5.4**	**5.3**	**5.3**
S8: Weibull distributions with crossing curves and shape alternatives							
un. W.	1	**5.2**	**5.5**	6.2	6.6	7.4	7.9
	2	**4.7**	**4.8**	**5.3**	**5.6**	5.9	6.1
	4	**4.9**	**5.0**	**5.4**	**5.4**	**5.5**	**5.5**
	6	**4.7**	**4.8**	**4.6**	**4.6**	**5.5**	**5.6**
eq. U.	1	**5.3**	5.8	**5.3**	5.7	6.7	7.2
	2	**5.3**	**5.4**	**4.7**	**4.8**	5.7	5.9
	4	**5.1**	**5.3**	**4.6**	**4.8**	**5.2**	**5.2**
	6	**5.2**	**5.3**	**5.2**	**5.3**	**5.0**	**5.1**
eq. W.	1	**4.5**	**4.6**	**5.5**	**5.6**	6.3	6.5
	2	**5.4**	**5.5**	**5.0**	**5.0**	**5.6**	**5.6**
	4	**4.9**	**4.9**	**5.5**	**5.5**	**5.3**	**5.3**
	6	**5.5**	**5.5**	6.0	6.0	**5.3**	**5.3**

Abbreviations: PO, ordinary jackknife pseudo-observations; PO IJ, infinitesimal jackknife pseudo-observations; un. W., unequal Weibull censoring; eq. U., equal uniform censoring; eq. W., equal Weibull censoring.

Table 8 Rejection rates (power) in % (nominal level $\alpha = 5\%$) of asymptotic tests using ordinary and infinitesimal jackknife pseudo-observations .

		$N = K \cdot (12, 18)$		$N = K \cdot (15, 15)$		$N = K \cdot (18, 12)$	
Censoring	K	PO	PO IJ	PO	PO IJ	PO	PO IJ
S1: Exponential distributions							
un. W.	1	20.4	20.9	19.4	20.0	18.0	19.0
	2	33.5	33.8	32.8	33.5	31.2	31.9
	4	59.5	59.7	57.8	58.0	55.1	55.7
	6	76.9	77.0	75.9	76.0	72.7	73.0
eq. U.	1	20.2	20.4	18.8	19.2	18.7	19.3
	2	32.4	33.1	34.1	34.6	32.8	33.7
	4	60.4	60.6	60.0	60.3	57.9	58.2
	6	75.7	75.9	77.4	77.5	74.7	74.9
eq. W.	1	21.1	21.1	21.2	21.3	20.8	20.8
	2	37.0	37.1	37.2	37.3	35.0	35.0
	4	63.2	63.3	63.3	63.4	62.8	62.8
	6	80.7	80.8	82.5	82.6	78.4	78.4
S7: Exponential and piecewise exponential distributions with crossing curves							
un. W.	1	17.7	18.1	17.1	17.6	17.1	17.7
	2	30.0	30.3	29.0	29.4	27.8	28.2
	4	52.9	52.9	50.8	51.1	47.3	47.6
	6	70.3	70.4	68.4	68.5	65.0	65.1
eq. U.	1	17.7	17.8	17.7	18.0	16.5	16.9
	2	30.0	30.4	30.0	30.4	27.2	27.5
	4	52.4	52.6	52.3	52.5	50.4	50.6
	6	69.3	69.4	69.2	69.4	65.9	66.1
eq. W.	1	18.0	17.9	18.3	18.3	16.9	16.9
	2	31.5	31.5	31.2	31.3	28.3	28.3
	4	56.1	56.1	55.9	55.9	51.6	51.6
	6	73.3	73.3	73.1	73.1	69.1	69.1
S8: Weibull distributions with crossing curves and shape alternatives							
un. W.	1	35.5	36.5	37.2	38.5	35.8	37.4
	2	59.4	59.9	60.1	60.7	56.4	57.3
	4	87.2	87.4	85.9	86.1	83.1	83.4
	6	96.5	96.5	96.3	96.4	94.1	94.2
eq. U.	1	31.7	32.8	34.2	35.3	34.0	35.3
	2	57.9	58.7	58.6	59.4	56.5	57.3
	4	85.8	86.1	86.6	86.7	84.1	84.3
	6	96.2	96.3	96.3	96.4	95.3	95.4
eq. W.	1	37.7	37.8	39.0	39.2	39.3	39.6
	2	65.1	65.2	65.7	65.9	62.0	62.1
	4	91.8	91.8	91.5	91.5	89.4	89.4
	6	98.5	98.5	97.9	98.0	97.3	97.3

Abbreviations: PO, ordinary jackknife pseudo-observations; PO IJ, infinitesimal jackknife pseudo-observations; un. W., unequal Weibull censoring; eq. U., equal uniform censoring; eq. W., equal Weibull censoring.

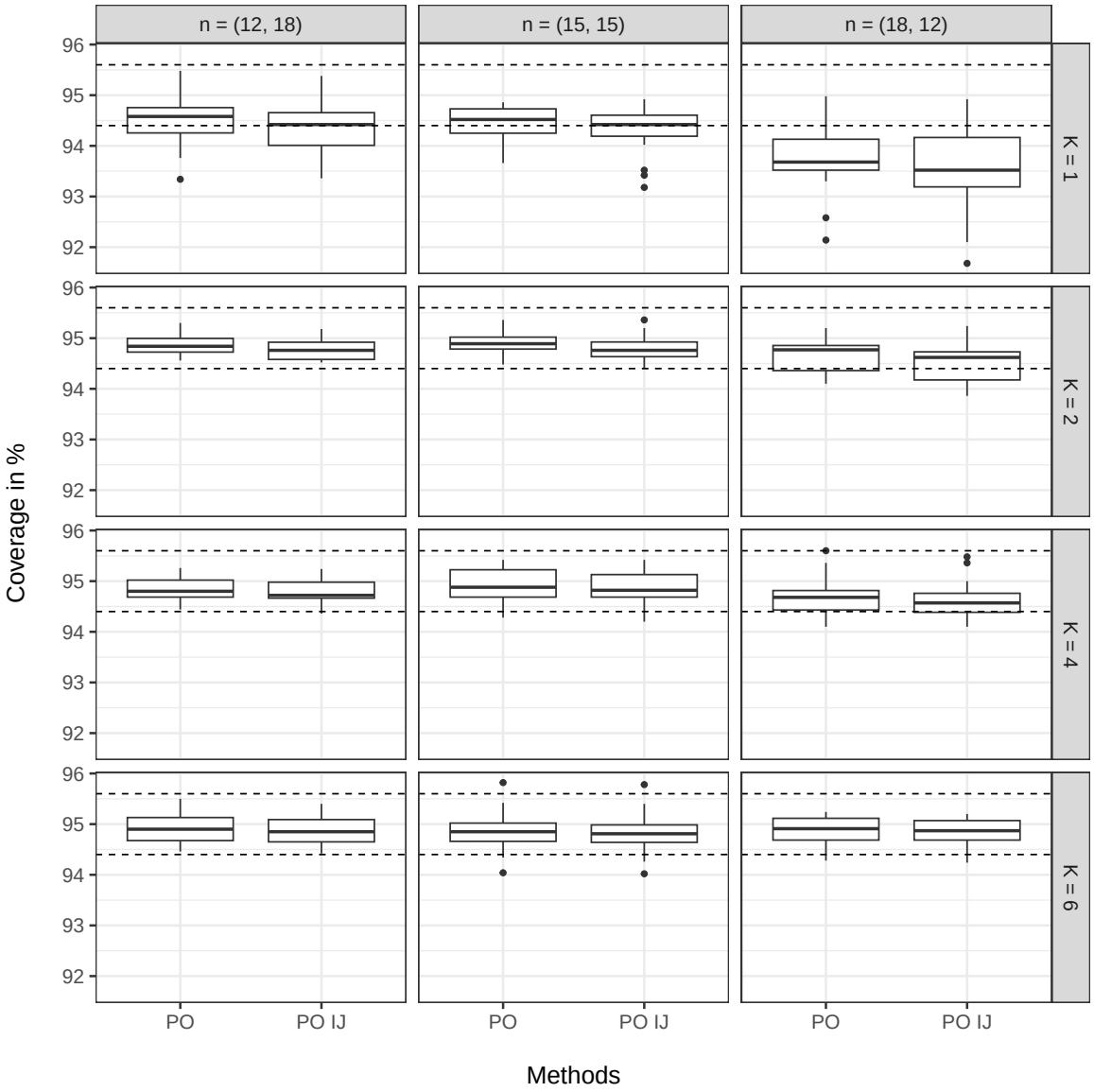

Figure 15 Confidence interval coverage of asymptotic methods using ordinary and infinitesimal jackknife pseudo-observations in % (nominal level $\alpha = 5\%$) aggregated by sample allocations ((n_0, n_1)) and their multipliers (K). The dashed lines depict the 95% binomial confidence interval $[94.4\%, 95.6\%]$.

Affidavit

I hereby declare that I have produced this work independently and without outside assistance, and have used only the sources and tools stated. I have clearly identified the sources of any sections from other works that I have quoted or given in essence. I have complied with the guidelines on good academic practice at the University of Göttingen. If a digital version has been submitted, it is identical to the written one. I am aware that failure to comply with these principles will result in the examination being graded "nicht bestanden", i.e. failed.

Göttingen, May 27, 2024

David Jesse