technische universität
dortmund

fakultät
statistik

Methods for Time-to-Event Data Analysis for Non-Proportional Hazard Settings

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät Statistik der Technischen Universität Dortmund

Vorgelegt von

Ina Dormuth

geboren in Dortmund

Dortmund, November 2024

Amtierender Dekan:
Prof. Dr. Philipp Doebler

Gutachter:
Prof. Dr. Markus Pauly (Technische Universität Dortmund)
Prof. Dr. Dennis Dobler (Technische Universität Dortmund)

Tag der Prüfung:
18. Dezember 2024

Abstract

The comparison of different treatment groups in terms of time to a specific event is paramount in clinical practice. Statistical methods to address such research questions are of fundamental importance. Hence, various tests and effect measures have been proposed to quantify such differences. One fundamental approach combines reporting the hazard ratio as an effect measure and the log-rank test to quantify the derivation. The main reason is that the log-rank test is optimal under proportional hazards. It is the most powerful test to detect differences between treatment arms in such scenarios. However, if we consider other kinds of hazard patterns, the power of the LR potentially drops drastically. At the same time, under non-proportional hazards, the hazard ratio changes over time and does not remain constant, making it challenging to interpret as an overarching measure of effect. Consequently, we need alternative approaches and recommendations on using them in settings where we can not assume proportional hazards.

This cumulative thesis is based on five publications dealing with non-proportional hazards. The first work summarizes the current state of two-group comparisons for time-to-event data under non-proportional hazards and compares the methods based on reconstructed real data. The second paper extends the comparison by including additional methods and reevaluating the approaches employing simulation studies. In the third manuscript, we revisit the average hazard ratio and the corresponding test statistic to derive a parametric and simulation-based sample size approach. The fourth manuscript extends the multi-directional weighted log-rank test to multiple testing problems, resulting in a new approach allowing for more than two treatment groups. Finally, the fifth publication incorporates the multi-directional log-rank test into an adaptive design, resulting in a highly flexible approach for situations with little prior knowledge.

Acknowledgments

I want to thank my supervisor, Markus, for the embracing combination of support and freedom he has given me. Thank you for letting me explore my own ways while ensuring I had the opportunities and resources needed to succeed. Your mentorship has been invaluable.

I am also deeply thankful to my second advisor, Dennis, who provided new perspectives and fruitful discussions.

I would also like to thank the wonderful people at Charité Berlin, where I spent two enriching months. Special thanks to Geraldine, Frank, and Caro, who made my stay productive and enjoyable and contributed significantly as co-authors on several of my PhD papers. Your collaboration and insights have been incredibly beneficial.

A heartfelt thanks goes to Jin. Our collaboration on two of my PhD papers was a great start to my PhD. I am thankful that I could visit you in Shanghai after all the Emails we wrote, and your hospitality and expertise greatly enhanced my research.

I am also grateful to Menggang, another co-author. Your input provided a valuable perspective to my work.

I want to express my deepest appreciation and pay tribute to Marc, who was an incredible source of support during my academic journey. As a former postdoc in our group and later a Junior Professor in Magdeburg, Marc's guidance, especially in the early stages of my PhD, was instrumental. His dedication and encouragement profoundly impacted my work, and his influence will always be remembered. His passing deeply saddens me, and I will forever be grateful for the time and wisdom he shared with me.

To my colleagues, friends, and colleagues who became friends, thank you for standing by me through all the ups and downs, enduring my complaints, and constant encouragement. Your patience and understanding have been a source of strength and comfort throughout this process.

Finally, I would like to acknowledge the use of Grammarly. This online writing assistance tool helped improve the clarity and coherence of the manuscript by providing suggestions on grammar, punctuation, and style.

List of Publications

This cumulative thesis is based on the following five manuscripts:

Article 1: **Dormuth, I.**, Liu, T., Xu, J., Yu, M., Pauly, M., & Ditzhaus, M. (2022). Which test for crossing survival curves? A user's guideline. BMC Medical Research Methodology, 22(1), 34.

Contribution of the author:
All authors were involved in the planning of the study. Ina Dormuth conducted the literature review from which she searched, reconstructed and treated the data in R. Furthermore, Ina Dormuth prepared the first draft of the publication, which was then jointly polished by all authors.

Article 2: **Dormuth, I.**, Liu, T., Xu, J., Pauly, M., & Ditzhaus, M. (2023). A comparative study to alternatives to the log-rank test. Contemporary Clinical Trials, 128, 107165.

Contribution of the author:
All of the authors were involved in the study's planning. Ina Dormuth planned, implemented, and conducted the simulation study. Furthermore, Ina Dormuth prepared the first draft of the publication, which was then jointly polished by all authors.

Article 3: **Dormuth, I.**, Pauly, M., Rauch, G., & Herrmann, C. (2024). Sample Size Calculation Under Nonproportional Hazards Using Average Hazard Ratios. Biometrical Journal, 66(6).

Contribution of the author:
All of the authors were involved in the publication's planning. Ina Dormuth planned, implemented, and conducted the simulation study. In collaboration with Carolin Herrmann, she developed the mathematical theory and prepared the first draft of the manuscript. All authors then jointly polished the manuscript.

Article 4: **Dormuth, I.**, Herrmann, C., Konietschke, F., Pauly, M., Wirth, M. & Ditzhaus, M. (2024). Single CASANOVA? Not in multiple comparisons. arXiv preprint arXiv:2410.21098.

Contribution of the author:
All of the authors were involved in the planning of the publication. Ina Dormuth derived the methodology from this. Furthermore, she planned, implemented, and conducted the simulation study. Ina Dormuth also prepared the first draft of the manuscript with support from the other authors. All authors then jointly polished the manuscript.

Article 5: Danzer M.F. & **Dormuth, I.** (2024). Adaptive weight selection for time-to-event data under non-proportional hazards. arXiv preprint arXiv:2409.15145.

Contribution of the author:
Both authors were involved in the publication's planning. Ina Dormuth implemented the weighted log-rank test, while Moritz Fabian Danzer implemented the adaptive procedure and the simulation study and derived the theory behind the adaptive procedure. Both authors worked jointly on the manuscript.

Further publications during my PhD:

(i) Time-to-event related

(1) **Dormuth, I.**, Allignol, A., Dobler, D., Schumann, F. & Pauly, M. (2024). Confidence Intervals for Left-Truncated Data: How Reliable are Confidence Intervals in Case of Left-Truncation? An Analysis Motivated from Pregnancy Outcome Probabilities. submitted.

We revisited estimating pregnancy outcome probabilities from observational cohorts considering left-truncated data and competing risks, such as live birth, induced abortion, or spontaneous abortion. To address unstable estimates caused by small sample sizes at early time points, we investigated more accurate methods for constructing confidence intervals using simulation studies and real-world data.

Contribution of the author:
Ina Dormuth conducted the simulation study, evaluated the results in the publication, and polished the manuscript.

(2) Thurow, M., **Dormuth, I.**, Sauer, C., Ditzhaus, M., & Pauly, M. (2023). How to simulate realistic survival data? A simulation study to compare realistic simulation models. arXiv preprint arXiv:2308.07842.

We used reconstructed benchmark data sets to simulate realistic clinical trial data. We considered different simulation approaches and evaluated their performance based on accuracy measures and runtime.

Contribution of the author:
Ina Dormuth supervised the Master thesis, which was the groundwork of this publication. She helped plan the scope of the work and polished the final manuscript. Furthermore, she supervised the development of the corresponding R-package.

(ii) Other topics

(1) Rutinowski, J., Franke, S., Endendyk, J., **Dormuth, I.**, Roidl, M., & Pauly, M. (2024). The self-perception and political biases of ChatGPT. Human Behavior and Emerging Technologies.

We analyzed the political biases and self-perceptions of OpenAI's ChatGPT. We derived results indicating a progressive bias through repeated completion of various established tests, including the political compass and G7-specific questionnaires. Tests of personality traits show that ChatGPT views itself as open, agreeable, and non-malicious, with an ENFJ Myers-Briggs type and minimal dark characteristics.

Contribution of the author:
Ina Dormuth was involved in the experimental planning, data evaluation, and polishing of the final draft of the manuscript.

(2) Sauer, C., Lange, F. J. D., Thurow, M., **Dormuth, I.** & Boulesteix, A.-L. (2024). Towards more practically relevant and neutral comparative simulation studies: Illustrating chances, challenges, and possible implementations of simulation settings based on real data. In preparation.

We compared existing simulation and real-data-based studies regarding their strengths and weaknesses to derive general recommendations on enhancing practical relevance and reducing researcher bias. Multiple real-world data sets could offer more realistic, generalizable results and shift the focus from hypothetical to actual data-driven scenarios, complementing traditional simulations exploring broader parameter ranges.

Contribution of the author:
Ina Dormuth was involved in the publication's planning and provided, together with Maria Thurow, the code for the simulation study of the survival example.

Contents

Part I

Introduction

1 Motivation

When planning a clinical trial, various decisions must be made. These include the definition of endpoints of interest. In time-to-event analysis, these are usually progression-free, event-free, or overall survival (Rufibach, 2019). Before starting the trial, we must define the estimator of interest and, based on that, the corresponding sample size calculation approaches and tests (Proschan et al., 2006). The choices depend on the underlying data and the respective research question. Hazard ratios, Schoenfeld's formula, and the log-rank test are standard tools in time-to-event analysis (Ananthakrishnan et al., 2021). The above approaches assume proportional hazards (PH), where the hazard ratio remains constant over time. Under this assumption, the log-rank test and the corresponding sample size calculation methods, such as those proposed by Schoenfeld (1981) and Freedman (1982), are widely used due to their optimal power properties. However, the assumption of proportional hazards is often violated in real-world data, leading to a significant loss of power and incorrect conclusions if the traditional methods are used indiscriminately (Lin et al., 2020). One primary reason for the continued use of the log-rank test, even when the PH assumption does not hold, is its simplicity and the availability of its implementation in various statistical software (e.g., R, Stata). However, the violation of the PH assumption is not uncommon. Treatment effects that vary over time, such as those seen in immunotherapy, where long-term benefits follow a high initial risk, frequently result in non-proportional hazards (Mick and Chen, 2015; Alexander et al., 2018; Ananthakrishnan et al., 2021; Rahman et al., 2019). A study by Trinquart et al. (2016) found that the PH assumption was violated in nearly 25% of analyzed phase III oncology studies. Kristiansen (Kristiansen, 2012) conducted a survey revealing that 70% of studies with crossing survival curves used the log-rank test, not considering derivation from the PH assumption.

Royston and Parmar (2020) published a simulation study comparing nine methods implemented in Stata and showed a preference for modified weighted log-rank tests Fleming et al. (1987); Royston and Parmar (2014); Lee (1996); Karrison (2016). However, the authors did not include scenarios for crossing hazards. Another overview was given by Lin et al. (2020), focusing on combined weighted Kaplan-Meier and weighted log-rank tests. The authors conclude that as long as we do not have prior knowledge, the MaxCombo test showed the most robust behavior among the tests under consideration (Lin et al., 2020). The most extensive study regarding crossing hazards was given by Li et al. (2015), who compared 21 tests designed to handle crossing hazards. They stated that the two-stage test by Qiu and Sheng (2008) or the test by Kraus (2009) are the most suitable among the studied tests. A general overview of existing methods and

recommendations regarding trial design was created by Ananthakrishnan et al. (2021) without numerical comparison. None of the mentioned reviews considered new results on projection type, sample space partition, or area under the survival curve tests (Brendel et al., 2014; Ditzhaus and Friedrich, 2020; Gorfine et al., 2020; Liu et al., 2020).

Most of the publications above present a new approach and a simulation study investigating the performance of various methods. Boulesteix et al. (2013) argue that an independent comparison should prioritize the evaluation rather than introducing a new method. They conclude that such conduct maintains reasonable neutrality and ensures a rational study design and assessment. Friedrich and Friede (2024) revisit the issue and emphasize the importance of choosing the comparison approach. Most manuscripts on statistical methodology rely on simulation studies and typically include a single real-world data set to motivate and illustrate the investigated method. In machine learning, especially within supervised learning, a prevalent approach involves comparing the performance of methods using benchmark data sets. Friedrich and Friede (2024) argue that both approaches come with their strengths and weaknesses and should be combined.

In this thesis's first and second papers, we extend the existing publications by providing neutral comparisons, including the best performers from existing simulation studies and more recent approaches, and combining artificial data and real-world data comparison approaches (Dormuth et al., 2022, 2023). Therefore, we conducted an extensive simulation study and supported the analysis with a broad real-world data analysis. Our simulation study covers 20 representative scenarios, including four null scenarios, four scenarios with PH, four with non-PH (excluding crossing structures), and eight with an emphasis on crossing hazards. Since most procedures exhibit good properties for large samples, our study focuses on small to moderate sample sizes. Since real-world survival data is usually unavailable, we also reconstructed published phase III clinical trial patient data employing the algorithm by Guyot et al. (2012). Therefore, we conducted a PubMed search with fixed search criteria. The search aimed to identify studies with crossing survival curves and published numbers at risk at various time points. We reviewed 1,400 recent clinical oncology papers and selected studies that met our criteria. Our final selection was further restricted by insufficient information, such as the lack of reported numbers at risk in nearly 30% of the papers.

The hazard ratio varies over time and is not constant under non-PH conditions. Hence, it is difficult to interpret as a global effect measure. Thus, several alternative effect measures have been proposed. These include the restricted mean survival time (RMST), the Mann-Whitney effect, the relative time, or the average hazard ratio (AHR). The RMST, for example, evaluates the difference between groups by comparing the areas under the Kaplan-Meier survival curves up to a specific time point (Royston and Parmar, 2011). The Mann-Whitney effect measures the probability that a randomly chosen subject from one group will survive longer than a randomly selected subject from another group (Koziol and Jia, 2009). The relative time, introduced by Phadnis and Mayo (2021), compares the time points at which a specified percentage of events

has occurred in each group. Lastly, the average hazard ratio extends the traditional hazard ratio by incorporating a time-dependent weighting function to account for non-PH (Kalbfleisch and Prentice, 1981).

The availability of corresponding statistical methods and software for sample size calculation under non-PH scenarios is crucial. Currently, the technique by Lakatos (1986) is frequently used for sample size calculation under non-PH, but it requires detailed knowledge of the survival curves under the alternative hypothesis (Phadnis and Mayo, 2021). Various other methods, such as those developed by Zhao et al. (2016) and Xiong and Wu (2017), provide sample size calculations tailored to specific survival models and non-PH conditions but have not yet gained widespread acceptance. This might be due to the lack of user-friendly software implementations.

Despite these advancements, there remains a need for practical and accessible sample size calculation tools for recently suggested effect measures under non-PH conditions. For instance, the R package `SSRMST` provides functions for sample size calculations based on RMST (Uno et al., 2014) and `npsurvSS` offers tools for various survival metrics, including RMST and weighted log-rank tests (Yung and Liu, 2020). The Wilcoxon-Mann-Whitney test has also been extended for right-censored data to test the Mann-Whitney effect and average hazard ratio in specific cases (Brückner and Brannath, 2017; Dobler and Pauly, 2018; Rauch et al., 2018).

The third manuscript aims to provide comprehensive guidance for sample size calculation in clinical trials using the average hazard ratio as the primary effect measure(Dormuth et al., 2024b). Therein, we compare different sample size calculation methods and assess the statistical power of corresponding test statistics under various simulation scenarios. Additionally, we demonstrate the importance and impact of method selection on a real data example.

Many trials comprise more than two groups (treatment arms) or have a factorial structure, posing unique statistical methodological challenges. The primary research question in such trials is manifold but usually subject to whether any of the arms differ from the others. Flexible multiple comparison procedures that can be adapted to the question of interest are essential in modern survival analysis. Until now, most existing methods use pairwise multiple log-rank tests that are adjusted for multiplicity with a particular correction (e.g., Bonferroni) (Logan et al., 2005). However, they may lack efficiency because they either do not consider the correlation across the multiple test statistics or are subject to somewhat restrictive dependency assumptions (Gao et al., 2008). This aspect is taken into account by so-called multiple contrast test procedures (MCTPs, usually conducted as maximum tests), which are valid for arbitrary correlations of the test statistics and use the correlation within the multiplicity adjustment. There exist several MCTPs for various endpoints (means, proportions, Mann-Whitney effects) (Bretz et al., 2001; Schaarschmidt et al., 2009; Hasler and Hothorn, 2008; Konietschke et al., 2013; Blanche et al., 2022; Munko et al., 2024). Most of the approaches, however, do not apply to time-to-event data. When considering multiple groups, the PH assumption becomes more unlikely, and the demand for flexible testing procedures increases.

Therefore, the fourth paper aims to close this gap and introduce powerful and flexible multiple contrast tests for multiple group comparisons (Dormuth et al., 2024a). We evaluate these new approaches using an adjusted log-rank test, an adjusted multiple directional log-rank test, and a maximum test of multiple weighted log-rank tests as competitors. In extensive simulation studies, we assess their performance regarding type-I error control and global and local power. Our findings reveal that our novel approaches do not fully exploit the familywise error rate and are not among the most powerful methods in any of the considered simulation scenarios. We could observe similar behavior in our real-world data example comparing seven treatment groups of multiple myeloma patients.

In addition to the uncertainty caused by the type of effect, we often have to deal with uncertainty regarding the effect size. For the latter, adaptive designs are widely used as a suitable approach to tackle this challenge (Bauer et al., 2016; Wassmer and Brannath, 2016). In adaptive designs, we can perform interim analyses at various time points. At each interim analysis, it is possible to terminate the study, e.g., due to futility or safety reasons, or to adapt the further study design. Specifically, this adaptive approach enables us to re-evaluate the initial assumptions concerning the effect size established at the study's outset and make necessary adjustments. In such multi-stage designs, the log-rank test is often the test of choice for time-to-event endpoints. More robust approaches to violating the PH assumption have been proposed in recent years. These include weighted log-rank test (Hasegawa, 2016) and MaxCombo (Ghosh et al., 2022) approaches.

We extend these concepts by proposing a systematic selection procedure based on the available information for the optimal test statistic (Danzer and Dormuth, 2024). Based on the preceding results regarding the performance of the mdir test, we have chosen to select this method as a first-stage test, as it is designed to address significant uncertainties concerning the type of effect. Then, we use the information available up to the interim analysis to pick a weighted log-rank test. We select the test with the highest conditional power. To improve our decisions, we employed the Royston and Parmar (2002) model to extrapolate the survival data. They propose to use a natural cubic spline transformation of the survival function to model the survival curve beyond the current time horizon. In our real-data example, we could show that our new approach has the potential to salvage a trial that would otherwise conclude with an ambiguous outcome. The extensive simulation studies revealed that our novel procedures outperform the two-stage log-rank test under non-PH while maintaining reasonable power and PH.

The structure of this thesis is as follows: Chapter 2 provides the methodological background. Chapter 3 summarizes each of the five publications, highlighting the novel aspects of each manuscript and its contribution to the respective research field. The thesis concludes with a discussion in Chapter 4, followed by the full-length papers.

2 Statistical Methods

This chapter outlines the general methodological framework of the techniques discussed in this thesis. For clarity, symbols are used consistently across different topics.

We consider $k \geq 2$ treatment groups, each consisting of n_j independent subjects ($j = 1, \ldots, k$), where each subject has event time T_{ji} and a right-censoring time C_{ji} ($i = 1, \ldots, n_j$). The random variables $T_{ji} \sim F_j$ and $C_{ji} \sim G_j$ follow continuous distribution functions and are assumed to be independent. Based on the cumulative distribution function F_j, we obtain the survival function $S_j(t) = 1 - F_j(t)$. Both can be estimated using the Kaplan-Meier estimator, in the following denoted as $\widehat{S}_j(t)$ and $\widehat{F}_j(t)$, respectively. We define $X_{ji} = \min(T_{ji}, C_{ji})$ as the observed time and $\delta_{ji} = I(X_{ji} = T_{ji})$ as the event indicator, where $I(\cdot)$ denotes the indicator function. The cumulative hazard rate functions are defined as $A_j(t) = \int_0^t (1 - F_j(s))^{-1} dF_j(s)$. Let $N_j(t) = \sum_{i=1}^{n_j} I\{X_{ji} \leq t, \delta_{ji} = 1\}$ represent the number of observed events in group j up to time t, and define $Y_j(t) = \sum_{i=1}^{n_j} I\{X_{ji} \geq t\}$ as the number of individuals at risk just before time t in the same group. These quantities allow us to define the Nelson–Aalen estimator for A_j, given by $\widehat{A}_j(t) = \int_0^t \frac{I\{Y_j(s)>0\}}{Y_j(s)} \mathrm{d}N_j(s)$ ($j = 1, \ldots, k$; $t \geq 0$). The corresponding pooled quantities are $N = \sum_{j=1}^k N_j$ and $Y = \sum_{j=1}^k Y_j$, leading to the pooled Nelson–Aalen estimator $\widehat{A}(t) = \int_0^t \frac{I\{Y(s)>0\}}{Y(s)} \, \mathrm{d}N(s)$ ($t \geq 0$).

With these definitions, we then describe the two-sided weighted log-rank test for $k = 2$ as follows:

$$T(w) = \left(\frac{n}{n_1 n_2}\right)^{1/2} \int_0^\infty w\{\widehat{F}(t-)\} \frac{Y_1(t) Y_2(t)}{Y_1(t) + Y_2(t)} \left\{ \mathrm{d}\widehat{A}_2(t) - \mathrm{d}\widehat{A}_1(t) \right\}. \tag{2.1}$$

Here, $t \mapsto \widehat{F}(t-)$ is the left-continuous version of $\widehat{F}$ and w is a continuous weight function. Under the null hypothesis, the two group test statistic asymptotically follows a χ^2 distribution (Klein and Moeschberger, 2006).

Fleming and Harrington (2013) examined a subclass of continuous weights w, including the classic log-rank weight. These weights can generally emphasize late, early, or central times. Ditzhaus and Friedrich (2020) introduced a weight designed for crossing hazard alternatives. The classical log-rank test and a log-rank test for early differences are implemented in the `survival` package (Therneau and Grambsch, 2021) in R.

2.1 Alternative Two-Sample Tests

While the log-rank test is still widely used even under non-proportional hazards (non-PH), several novel approaches have been proposed that are specifically tailored to handle non-PH data. A straightforward extension is the above-mentioned weighted log-rank test that allows using distinct weights to address non-proportionality. The choice of the weight function depends on the alternative of interest, resulting in a test that has high power against the specific alternatives (Klein and Moeschberger, 2006). An example is the Peto-Peto test, which favors early events (Legrand, 2021). Without prior knowledge regarding the survival distributions, the choice of the weight function is arbitrary and can result thus in sub-optimal power performance (Li et al., 2015).

Qiu and Sheng (2008) proposed a two-stage testing procedure as a solution. In the first stage, they employ the unweighted log-rank test. If they cannot reject the null hypothesis, they use an independent test sensitive to crossing in the second stage. The decision bounds are adjusted to control the overall Type-I error rate. The method is implemented in the R package `TSHRC` (Sheng et al., 2019).

While the two-stage test is specifically designed for PH and crossing hazard patterns, the group of omnibus tests claims to be powerful for multiple alternative hypotheses. Gorfine et al. (2020) presented an omnibus permutation test based on a sample space partition. The authors offer two test statistics based on Pearson's chi-square or the likelihood ratio. The critical values are obtained through a resampling approach designed for censored data. The tests are implemented in the R package `KONPsurv` (Schlesinger and Gorfine, 2020). To overcome the challenge of weight choice, combination approaches of several weighted log-rank test statistics were introduced (Ditzhaus and Friedrich, 2020; Brendel et al., 2014; Ditzhaus and Pauly, 2019; Lin et al., 2020). The multi-directional log-rank (mdir) test combines several weighted log-rank tests into a joint Wald-type statistic (Ditzhaus and Friedrich, 2020). For the two-sided testing problem, critical values can be obtained precisely by a limit distribution or be approximated via permutation. The two-sided mdir test and a one-sided version are implemented in the R package `mdir.logrank` (Ditzhaus and Friedrich, 2018). Another straightforward approach to combining multiple weighted log-rank tests is to use their maximum as the test statistic of interest (Lee, 2007; Lin et al., 2020). The critical value of the so-called MaxCombo test can then be determined as the maximum of a centered multivariate normal distribution.

An alternative route is considering novel survival estimates besides the hazard ratio to derive test statistics. One concept to tackle the requirement of PH is to employ a time-flexible weight function to the hazard function. Such a weight function allows us to take the variation over time into account (Rauch et al., 2018). This results in an interpretable effect measure, the average hazard ratio (AHR), even when the PH assumption is violated. The corresponding Wald-type test statistic can be derived using plug-in estimators (Brückner and Brannath, 2017; Kalbfleisch and Prentice, 1981). Although this method is not available on CRAN anymore, the R code is available

on GitHub in the `AHR` repository (Brückner, 2018). The well-known Mann–Whitney effect is a particular case of the AHR (Dormuth et al., 2024b). Dobler and Pauly (2018) introduced a bootstrap and a permutation-based test for the null hypothesis of "tendentiously equal survival curves." This is less restrictive than testing for equality in survival. Besides hazard ratios, the restricted mean survival time (RMST) is often used as an effect measure in survival analysis. We can interpret it as the mean event-free survival time up to a predefined time point (Kim et al., 2017). The corresponding test hypotheses are then formulated based on the difference in RMST. Such tests are also valid for assessing the equality of survival functions, as equal survival functions imply equal RMSTs. Conversely, situations where the RMSTs are equal but the survival functions differ occur. Based on the RMST, various methods exist to obtain a test statistic or critical values, e.g., through resampling or asymptotic theory (Uno et al., 2014; Tian et al., 2018). These approaches are implemented in R packages: `surv2sample` (Tian et al., 2017) and `survRM2` (Uno et al., 2020). Royston and Parmar (2016) extend the permutation-based RMST approach by combining it with a Cox test. The authors empirically derive the null distribution as an incomplete beta distribution for the minimum p-value. This approach is implemented in the `stctest` function in STATA. A limitation of RMST-based methods is that they can be misleading when survival curves cross, as the effects can partially offset each other. The area between curves (ABC) is an alternative effect measure introduced to tackle this problem (Liu et al., 2020). Its test statistic is based on an L1-distance of the two Kaplan-Meier curves (Liu et al., 2020). Although this method is not yet available on CRAN, the R code is available on GitHub in the `RBT4TCSC` repository (Liu, 2020).

The multi-directional log-rank test

Ditzhaus and Friedrich (2020) promoted the multi-directional log-rank (mdir) test to test $H_0 : \{A_1 = A_2\}$. This method was originally introduced by Brendel et al. (2014) and allows the combination of multiple weighted log-rank tests as defined in (2.1) into one test statistic.

For $m = 2$ weights, the studentized quadratic form is given by:

$$S = (T(w^{(1)}), T(w^{(2)}))\, \hat{\Sigma}^{-}\, (T(w^{(1)}), T(w^{(2)}))^{\top},$$

where $(T(w^{(1)}), T(w^{(2)}))^{\top}$ indicates the transposed vector of $(T(w^{(1)}), T(w^{(2)}))$ and $\hat{\Sigma}^{-}$ represents the Moore-Penrose inverse of the empirical covariance matrix of $(T(w^{(1)}), T(w^{(2)}))$. Under the null hypothesis, the test statistic follows a χ^2 distribution (Ditzhaus and Friedrich, 2020). Additional weights can be combined to cover a broader range of alternative hypotheses. Ditzhaus and Friedrich (2020) also proposed a permutation-based alternative and Ditzhaus and Pauly (2019) provided a one-sided version of the test. We use the R package `mdir.logrank` (Ditzhaus and Friedrich, 2018).

Tests based on the average hazard ratio

The average hazard ratio (AHR) generalizes the well-known hazard ratio using a time-dependent flexible weighting function (Brückner and Brannath, 2017). For $k = 2$ groups, the weighted

AHR for group j is defined as:

$$\theta_j(G) = -\int_0^\infty \frac{\lambda_j(t)}{\lambda_1(t) + \lambda_2(t)} G(dt)$$

with G a decreasing weight function. Since we can only estimate survival functions S_j up to a finite time point, we truncate the weight function G at $L < \sup\{u : S_j(u)C_j(u) > 0\}$, ensuring $\int_0^\infty dG_L(t) = -1$. In the following, we employ the definitions by Brückner and Brannath (2017) and fix the shape parameter of the weight function to 1 as described by Rauch et al. (2018). Then, the representation of the AHR equals the well-known Mann–Whitney effect (Dobler and Pauly, 2018). We can estimate our weight function by $\hat{G}(t, L) = \hat{S}_1(t, L)\hat{S}_2(t, L)$, with $\hat{S}_i(t, L)$ denoting the Kaplan-Meier estimator of S with an observed time t smaller or equal to the pre-defined constant L.

The truncated AHR is

$$\theta_1(G_L) = \frac{x_1}{1 - G(L)},$$

where $x_1 = -\int_0^L S_1(t)S_2(dt)$. We estimate the AHR with reference-group 1 without loss of generality at time $t \leq L$ by

$$\hat{\theta}_1(t) = \frac{-\int_0^L \hat{S}_1(t, s)\hat{S}_2(t, ds)}{1 - \hat{G}(t, L)}.$$

To test for equality, we investigate the hypotheses of interest $H_0 : \theta_1 = 0.5$ versus $H_1 : \theta_1 \neq 0.5$ (Rauch et al., 2018). Based on the above definitions, the Wald-type test statistic is then defined as

$$Z = \frac{\sqrt{n}(\hat{\theta}_1 - 0.5)}{\sqrt{\hat{v}_\theta}} \sim \mathcal{N}(0, 1),$$

the variance estimator $\hat{v}_\theta$ can be found in Brückner and Brannath (2017) or in the Supplementel Material of Dormuth et al. (2024b). The test can be extended to k-sample problems and is implemented in the R package `AHR` (Brückner, 2018).

2.2 Alternative k-Sample Tests

Different research hypotheses arise whenever $k > 2$ treatment groups are involved in a clinical trial. On a global level, we can be interested in whether there is a difference between any of the treatments under consideration $\mathcal{H}_0 : A_1 = \ldots = A_k$ vs. $\mathcal{H}_1 : A_1 \neq \ldots \neq A_k$. Various tests can be employed to test the global null hypothesis, such as analysis of variance (ANOVA) based methods (Konietschke et al., 2013). Ditzhaus et al. (2021) introduced the CASANOVA approach as an ANOVA-based testing procedure for null hypothesis formulated in terms of cumulative hazards (see below for more details). In the end, we are often interested in the local level, e.g.,

the exact treatments that are different from each other $H^0_{j_1 j_2} : A_{j_1} = A_{j_2}$ vs. $H^1_{j_1 j_2} : A_{j_1} \neq A_{j_2}$ (Konietschke et al., 2012). Such research questions require different testing approaches.

Multiple Testing

Statistical tests are typically designed to limit the risk of a Type-I error when testing a single null hypothesis $\mathcal{H}_0$. However, when multiple hypotheses are tested simultaneously, the likelihood of false positive conclusions increases as the number of tests grows (Bretz et al., 2016). Various measures can be taken when extending the concept of the Type I error rate to situations involving multiple tests. We focus on the familywise error rate (FWER), which is the probability of wrongly rejecting at least one true null hypothesis. It should be noted that the FWER becomes more strict with an increasing number of hypotheses (Bretz et al., 2016).

When dealing with multiplicity, the standard frequentist approach to the analysis of clinical trial data may require adjusting the Type I error rate, e.g., with Bonferroni (Logan et al., 2005). Single-step or stepwise procedures can be incoherent under certain conditions, meaning the rejection of the global null hypothesis must not result in a rejection of a local hypothesis and vice versa. This may raise difficulties in interpreting the results (Bretz et al., 2016). In recent years, numerous researchers have introduced multiple contrast test procedures (MCTPs), typically performed as maximum tests. These methods are robust to arbitrary correlations among test statistics and incorporate such correlations within the multiplicity adjustment across various outcomes, including means, proportions, and Mann-Whitney effects (Bretz et al., 2001; Schaarschmidt et al., 2009; Hasler and Hothorn, 2008; Konietschke et al., 2013; Blanche et al., 2022).

CASANOVA

Ditzhaus et al. (2021) propose a test for general factorial designs. Therefore, they introduce a two-way design with factors B (with b levels) and C (with c levels). Set $k = b \cdot c$ and decompose the group index j into $j = (j_B, j_C)$, where $j_B = 1, \ldots, b$ and $j_C = 1, \ldots, c$. The null hypothesis of interest is $\mathcal{H}_0 : \mathbf{HA} = \mathbf{0}_d, \quad \mathbf{A} = (A_1, \ldots, A_k)^\top$ and $\mathbf{H} \in \mathbb{R}^{q \times k}$ is a contrast matrix and fulfilling $\mathbf{H1_k} = \mathbf{0}_q$ with $\mathbf{1_k}$ and $\mathbf{0}_q$ denoting vectors of ones and zeros, respectively.

They claim that in analysis-of-variance settings, it is useful to work with the projection matrix $\mathbf{V} = \mathbf{H}^\top(\mathbf{H}\mathbf{H}^\top)^-\mathbf{H}$, as it describes the same null hypothesis as $\mathbf{H}$. But unlike $\mathbf{H}$, $\mathbf{V}$ is unique, symmetric and idempotent. Employing the standard counting process notation introduced before, they obtain $Z_j(w) = n^{1/2} \int_0^\infty w(t) d\hat{A}_j(t)$ with $w(t) = \tilde{w}\{\hat{F}(t-)\} \frac{Y_1(t) \ldots Y_k(t)}{nY(t)^{k-1}}$, $(t \geq 0)$ and $\mathbf{Z}(w) = (Z_1(w), \ldots, Z_m(w))$. A combination of weights is used to detect different alternatives, leading to a joint Wald-type statistic.

$$S(w) = (\mathbf{V}\mathbf{Z}(w)^\top (\mathbf{V}\widehat{\mathbf{\Sigma}}(w)\mathbf{V})^- \mathbf{V}\mathbf{Z}(w).$$

Since the entries of $\mathbf{Z}(w)$ are highly dependent, the covariance matrix estimator $\widehat{\Sigma}(w)$ is not a

simple diagonal matrix. The test statistic is asymptotically χ^2 distributed under the null hypothesis and certain conditions, allowing for an asymptotically exact test $\phi = I\{S > \chi^2_{f,\alpha}\}$, with $f = m \cdot \text{rank}(\mathbf{V})$. For better finite sample performance, a permutation strategy is recommended. The method is implemented in the R package `GFDsurv` (Ditzhaus et al., 2022).

2.3 Adaptive-designs

Adaptive designs allow for interim analyses at multiple time points (Bauer et al., 2016). Such designs provide the flexibility to terminate a trial early, either with or without rejection of the null hypothesis, or adjust the further course of the trial. We focus on two-stage designs with one interim and one final analysis. Testing at multiple time points generally increases the significance level and requires thus corresponding adjustments.

We define a test statistic and a corresponding p-value in each stage (p_1 and p_2 in two-stage designs). These two p-values should fulfill the p-cloud property to define the adaptive design as a combination function $C\colon [0,1]^2 \to [0,1]$ (Brannath et al., 2012). This p-value combination function must be non-decreasing in p_1 and p_2 and continuous in p_2. Various combination function candidates have been proposed in the literature (Fisher, 1970; Bauer, 1989; Lehmacher and Wassmer, 1999). For any such combination function, we can write the significance level α in terms of stage-wise boundaries α_0 and α_1 as $\alpha = \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 I_{\{C(p_1,p_2)\leq c\}} dp_1\, dp_2$. Hence, we stop the trial due to rejection of H_0 either if $p_1 \leq \alpha_1$ in the first stage or if we stop for futility if $p_1 \geq \alpha_0$. H_0 can be rejected at the second stage if $C(p_1, p_2) \leq c$. All common approaches for combination functions are implemented R package `rpact` (Wassmer and Pahlke, 2023).

2.4 Royston-Parmar splines

As described earlier, the pooled and group-specific survival functions S, S_j can be estimated non-parametrically by Kaplan-Meier estimators. Such a non-parametric estimation can be inefficient compared to a parametric estimation approach if the distribution of T or T_{ji} lies within the parametric family assumed for the estimation process. Additionally, a parametric approach allows extrapolation of the survival curve beyond the time horizon in the data since the estimated parameters directly specify a distribution on $[0, \infty)$.

Despite the potential efficiency, parametric approaches can be criticized for being too restrictive in terms of the shape of the distribution. Hence, essential characteristics of the survival mechanism could potentially not be captured (Latimer and Adler, 2022). For instance, estimation within the family of exponential distributions for several treatment groups directly leads to the assumption of PH.

Royston and Parmar (2002) introduced a flexible approach for extrapolating survival curves. It is based on natural cubic splines with a transformation of the survival function S, given by a link function $g : [0, 1] \to \mathbb{R}$, and is modeled by

$$g(S(t; \mathbf{z})) = g(S_{\text{baseline}}(t)) + \boldsymbol{\beta}^\top \mathbf{z} = s(x; \phi) + \boldsymbol{\beta}^\top \mathbf{z},$$

with $S(t; \mathbf{z})$ the survival distribution given covariates $\mathbf{z}$ and corresponding regression coefficients $\boldsymbol{\beta}^\top$ and $x = \log(t)$. For p internal knots, we denote the cubic spline $s : \mathbb{R} \times \mathbb{R}^{p+2} \to \mathbb{R}$ parameterized by $\phi \in \mathbb{R}^{p+2}$.

The choice of p and the link function can be based on the Akaike Information Criterion. An informal choice based on the appearance of the fitted survival curves is also suggested (Royston and Parmar, 2002; Royston et al., 2011). Simulation studies show that correct knot placement is not crucial, as the splines are flexible enough for large enough p (Rutherford et al., 2015).

2.5 How to Compare Them? Simulation and Real-World Data

Method comparisons are crucial when it comes to deriving recommendations for applied researchers. These overviews are designed to facilitate the choice of an appropriate method. Whenever we look for the "best" approach for a specific process, one major challenge is to create a fair study that reflects the real application case well. Friedrich and Friede (2024) describe and compare simulation studies and benchmarking data sets regarding their strengths and weaknesses for such fair comparisons. Their final recommendation is to use both to get the best of both worlds.

Simulation Studies

Simulation studies allow us to examine the behavior of methods based on synthetic data having known properties. We can compare multiple methods to find the best performer or assess method robustness when assumptions are violated (Friedrich and Friede, 2024). In the survival setting, we must make assumptions regarding the survival and the censoring distribution. Bender et al. (2005) emphasize the importance of investigating various distributions to capture various survival data. In the following publications, we categorize the data-generating distributions into three groups based on their hazard patterns: (i) PH, (ii) non-PH but non-crossing, and (iii) crossing hazards (see Figure 2.1). While the exponentially distributed survival times are always proportional to each other, distributions like a log-normal distribution can generate non-PH.

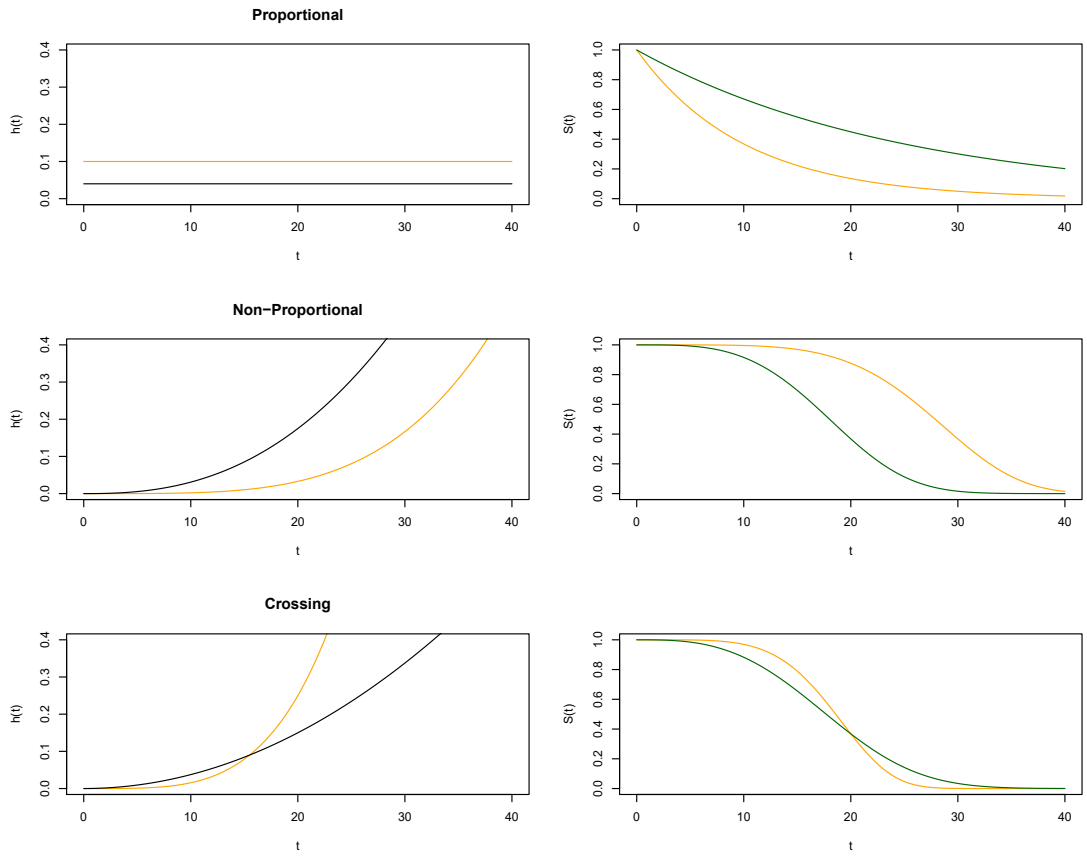


Figure 2.1: Examples of different hazard patterns and the corresponding survival curves.

Thus, the choice of survival distributions is crucial. In synthetic simulation studies, other parameters besides the survival distribution can be controlled. In our context, this includes censoring distributions and probability, sample sizes, group allocations, effect size, and more. These controlled data allow us to generalize our results and derive recommendations on when to use specific methods. While these advantages make simulation studies very useful, artificial data can not fully capture real-world behavior (Friedrich and Friede, 2024).

Real-world Data and Data Reconstruction

Due to the modeling limitations of synthetic data, real-world data is still crucial to fully evaluate the benefits of novel methodology (Friedrich and Friede, 2024). When dealing with clinical data, especially survival data, real-world data availability is limited due to ethical restrictions. One way to obtain real-world data while avoiding these restrictions is by employing a data reconstruction algorithm. Various approaches have been proposed for survival data (Guyot et al., 2012; Liu et al., 2021; Zhao et al., 2022). One of the most applied is the one introduced by Guyot et al.

(2012). It requires three types of input: (i) a high-quality resolution Kaplan-Meier curve, (ii) the number at risk at multiple time points, and (iii) the total number of events for each group. The algorithm is still usable in cases where the number at risk and the total number of events are unavailable. Nevertheless, applying the algorithm to data where all information is available is recommended. To process the Kaplan-Meier curves, various tools can be used; we employed WebPlotDigitizer (Rohatgi, 2019) to obtain the corresponding coordinates. In the first step, we estimate the number at risk without censoring and the number censored in the intervals given by the time points with a number at risk reported. For the censoring distribution, we assume an even distribution of censoring times on each interval. This assumption is a limitation of the algorithm since dependent censoring is common in clinical trials (Emura and Chen, 2018). Another limitation of the approach is the unavailability of covariate information. The algorithm estimates the number of events at each extracted coordinate using the inverted Kaplan-Meier estimates. An iterative procedure readjusts the number of censored observations until it fits the reported number at risk. If the total number of events is available, we compare the estimated number of events with the reported and readjust if required. The resulting data is individual patient data, including survival time, the censoring indicator, and a group indicator. Friedrich and Friede (2024) do not only recommend taking both approaches into account but recommend a straightforward combination as described, e.g., in Thurow et al. (2023).

3 Summary of the Articles

3.1 Article 1: Which test for crossing survival curves? A user's guideline

The first article overviews existing approaches for two sample comparisons of survival data under non-proportional hazards. This research field is vital since the proportional hazards assumption is often violated in real life, as Kristiansen (2012) observed. Nevertheless, the often-used log-rank test loses the power to detect differences between groups to an unknown extent. Furthermore, our study is crucial for two reasons. First, the comparisons provided by the authors of a new approach might be (unintentionally) biased towards their method. Second, the original publications are addressed to other statisticians and rarely to practitioners. This work addresses both issues by providing a more accessible presentation of the methods and an independent comparison.

To identify pertinent methods, we conducted an extensive literature review updating the comparison performed by Li et al. (2015). We determined eleven approaches that can be broadly categorized into four groups: (i) Log-rank test and its weighted variants, (ii) Two-stage test, (iii) Omnibus tests, and (iv) Tests based on the area under the survival curve. Subsequently, we defined search criteria for eligible PubMed data sets to avoid bias towards specific data sets. These criteria included the necessity for crossing survival curves, non-significant log-rank tests, and non-informative censoring. The obtained data sets were then checked for compatibility with the reconstruction algorithm from Guyot et al. (2012). We screened 1,400 publications and ultimately reconstructed eighteen datasets that met all the criteria. To ensure the quality of the reconstruction, we recalculated all estimands provided in the publication. We applied all identified methods to the reconstructed real-world data sets and evaluated whether a difference could be detected.

The log-rank test never rejected the null hypothesis, which was consistent with the original studies. The Peto-Peto test identified survival differences in four studies, while restricted mean survival (RMST)-based tests found differences in up to five studies. Omnibus tests, particularly those by Gorfine et al. (2020) and Ditzhaus and Friedrich (2020), detected the most differences, with Ditzhaus and Friedrich's test rejecting the null hypothesis in eight cases. These findings highlight the power of omnibus tests, especially when the proportional hazards assumption is violated. Nevertheless, final recommendations should only be given based on simulation studies.

To summarize, the first article established the foundation for this cumulative thesis. Building on this work, more extensive simulation studies were conducted to support the observed results, and extensions of the most promising approaches were developed.

3.2 Article 2: A comparative study to alternatives to the log-rank test

Since method comparison on real-world data has disadvantages such as an unknown ground truth or limited data availability, additional simulation studies are crucial to provide final methodological recommendations (Friedrich and Friede, 2024). Simulation studies based on artificial data allow us to investigate the control of the Type I error and the power of the approaches. Thus, we aimed to quantify the log-rank test's power loss in various settings to derive the impact of using it under non-proportional hazards.

We conducted an extensive simulation study including twenty different survival settings. These settings include four examples under the null, four proportional, four non-proportional but non-crossing, and eight crossing hazards settings. The survival distributions included in this survey follow the recommendations by Bender et al. (2005). For each setting, we considered various sample sizes, censoring rates, and censoring distributions, resulting in 800 simulation scenarios. Here, we also investigated the impact of the number of chosen weights on the mdir test to derive recommendations.

Most tests generally adhere to the 0.05 significance level regarding the type I error. However, the MaxCombo and the permutation-based RMST test are consistently conservative, while the other RMST and area between curves (ABC) tests tend to be more liberal in some scenarios. Overall, all tests effectively control the type I error. Regarding power, we observed different results depending on the underlying hazard patterns. The log-rank test consistently exhibited the highest power across all configurations in proportional settings. Even for smaller sample sizes, the power of the log-rank test is consistently high, while power loss was noticeable for the other tests. We could observe that increased censoring had a less pronounced effect on power, and the choice of censoring distribution seems to have no impact. For non-PH, non-crossing scenarios, test power varied more between different settings. The log-rank test maintained high power but was not universally the most powerful test. The Peto-Peto test showed high power for early differences in survival functions, and omnibus tests and the two-stage (TS) test were also powerful and more robust against violations of the PH assumption. The log-rank, Peto-Peto, and RMST-based tests had significantly lower power for crossing hazard scenarios than the other methods. High censoring improved power for some tests (e.g., Peto-Peto, RMST), while others lost power regardless of censoring levels. The test by Gorfine et al. (2020), Lin et al. (2020), and Ditzhaus and Friedrich (2020) demonstrated consistently high power. Overall scenarios show that the test by Ditzhaus and Friedrich (2020) is usually more powerful with two weights instead

of three or four.

Based on the simulation study results, we can derive some recommendations: The RMST-based test is not recommended for crossing hazard scenarios. Omnibus tests such as MaxCombo, KONP, and mdir seem robust and could be recommended for various scenarios. Without prior knowledge about the data, we recommend using the mdir test with the two default weights provided in the R package Ditzhaus and Friedrich (2018).

In summary, this paper extends and supports the results obtained in our first study. We were able to derive some advice on when to use or not to use which statistical test. Furthermore, it inspired future work on extensions of the mdir test, alternative effect measures, and related tests.

3.3 Article 3: Sample size calculation under non-proportional hazards using average hazard ratios

This project aimed to guide sample size calculation for clinical trials using the average hazard ratio (AHR) as the primary endpoint and a corresponding test statistic. We introduced a simulation-based sample size calculation approach (AHRsim) and an asymptotic sample size calculation approach (AHRasymp). To evaluate their performance, we compared them to the Schoenfeld formula (SF) and the log-rank simulation-based approach (LRsim) by comparing their required sample sizes and resulting statistical power under various settings.

Extensive simulation studies were conducted to assess different survival times and censoring distributions. The study considered six scenarios with proportional and non-proportional (non-crossing and crossing) hazards, setting the target power to 0.8 and type I error to 0.05. Sample sizes varied significantly across methods and scenarios, with Schoenfeld often yielding smaller sizes in proportional hazard settings. Under non-proportional hazards, sample sizes varied more, especially for the asymptotic average hazard ratio approach. Mis-specifications led to changing power levels, with no approach being consistently robust. Larger censoring rates notably decreased power for Schoenfeld and log-rank methods. Non-proportional hazard scenarios, particularly with crossing hazards, posed greater challenges, often leading to computational limitations for Schoenfeld and log-rank approaches. The asymptotic and simulation-based average hazard ratio methods showed more stability under non-proportional hazards but were not immune to power fluctuations.

Finally, we applied the four sample size calculation approaches to a reconstructed data example. For the real-world example, the AHR-based test exhibits more power than the log-rank test, leading to significantly smaller required sample sizes for follow-up studies. Additionally, the results of the LR test align with the original publication and are not significant. The AHR-based test, on the other hand, delivers significant results. The respective effect measures point in

different directions, emphasizing the impact of the choice of testing procedure especially under non-PH.

One limitation of our work is that it focuses on a specific weight function of the AHR. However, the proposed approach can also be applied to other weight functions within the same family (for details, see the Methods section). Furthermore, while the approach can be extended to unbalanced group designs, we did not investigate the performance of the new methods under these circumstances.

Both the simulation study and the real data example highlight that different sample size calculation approaches yield widely varying results depending on the data situation. This step in clinical trial planning holds substantial relevance for both statistical and economic outcomes. Our findings underscore the importance of appropriate sample size calculation approaches and inference methods, especially when non-proportional hazards are assumed, where no universally recommended effect measure exists. Overall, it is recommended to consider different approaches while incorporating all available information.

3.4 Article 4: Single CASANOVA? Not in multiple comparisons

In this work, we introduced three new approaches to tackle multiple comparisons for survival data under potentially non-proportional hazards. For the first method, we extended the CASANOVA approach introduced by Ditzhaus et al. (2021) to a multiple testing setting. Therefore, we derived a maximum test that allowed local and global test decisions (multiCASANOVA). In the first case, Wald-type statistics of multiple weighted log-rank tests are used to derive a decision for each contrast. Maximum tests have the property of consistent test decisions between local and global tests, achieved using the maximum of these Wald-type statistics for the global test. Since the distribution of the obtained multiCASANOVA test is unknown, we employed resampling approaches to derive our test decisions. As wild bootstrap procedures have shown promise in the survival setup, we use two discrete wild bootstrap approaches to evaluate their robustness, depending on the weight functions considered. Additionally, we introduced a straightforward maximum test of different weighted log-rank tests (multiWeightedLR). The local test uses the maximum single-weighted log-rank tests, while the global test uses their maximum. With the properties of maximum tests, we can assume a multivariate normal distribution to obtain critical values. These new approaches are compared to Bonferroni-adjusted versions of the unweighted log-rank and mdir test.

We designed a simulation study that included a proportional, a non-proportional, but not crossing, a crossing, and a mixed scenario with various parameter settings. We investigated the behavior of all methods for the two most common contrast matrix types, Tukey and Dunnett, and for different sample sizes. Under the null hypothesis of equal survival, all approaches control the familywise

error rate well for a group size of $n = 100$ independent of the choice of contrast matrix. The new approaches tend to be slightly conservative compared to the Bonferroni-adjusted tests. In most cases, the adjusted LR test seems to have the lowest variability among the tests. It drastically loses power under the crossing hazard scenario, while the other approaches maintain a reasonable power. Considering all four scenarios and the various parameter combinations, the adjusted mdir test tends to be the most robust in power. The real data example of seven treatment groups for multiple myeloma patients supports these findings.

One limitation of our work is the conservative behavior of the novel approaches. In future work, we plan to investigate the causes and improve the FWER exploitation, e.g., by incorporating the closed testing procedure as in Blanche et al. (2022). At the same time, we are interested in the properties of other maximum tests, such as a maximum of mdir tests, and aim to study these further to enhance the impact of our work. Researching the effects of the number of groups under consideration on each test would be another future research path. While we limited our work to two contrast matrix types, all approaches can easily be extended to other contrast matrices.

Our findings emphasize the importance of looking for alternatives to the log-rank test in multiple testing situations. We have demonstrated that multiple contrast tests offer a promising approach to address multiple testing problems in survival analysis. The presented methods exhibit high robustness across various settings and can potentially become powerful tools in this field.

3.5 Article 5: Adaptive weight selection for time-to-event data under non-proportional hazards

The fifth paper considers another source of uncertainty besides the hazard pattern: determining the effect size. We explore adaptive trial designs as a solution to address these uncertainties, allowing for adjustments based on interim analyses. Such designs enable decisions like stopping the trial early, modifying the sample size, or altering the analysis schedule.

The paper expands the use of multi-stage test statistics in adaptive designs and emphasizes the limitations of the log-rank test in non-proportional hazard scenarios. We propose using weighted log-rank tests, particularly the multi-directional log-rank (mdir) test, to improve power in such settings. This approach enhances robustness in terms of the effect size and the type of effect in the analysis. We present a framework consisting of a two-stage adaptive design. The first stage employs different choices of mdir tests. We employed Royston-Parmer splines (Royston and Parmar, 2002) to extrapolate the survival data, allowing us to use the accumulated data to select the most appropriate single-weighted log-rank test that maximizes conditional power. We then continue in the second stage with the chosen single-weighted log-rank test.

We assess our adaptive selection procedure's type I error rate compliance and power within an extensive simulation study, including varying sample sizes and balanced group sizes. We include

seven different survival distributions for the power comparisons, each chosen so that a specific weighted log-rank test will be optimal. These include PH together with late and early effect scenarios. We perform an interim analysis at five years and a final analysis at eight years, with participants enrolling uniformly up to six years and only administrative censoring applied. In the first stage, we evaluate seven different combination tests and eight differently weighted log-rank tests. We then consider nine candidate Royston-Parmar spline models, selecting the one with the highest AIC. In the second stage, we choose one of the eight differently single-weighted log-rank tests based on conditional power calculations.

Our simulations reveal light inflation of the type I error that can be associated with group sequential designs without any adaptation. Hence, our framework does not introduce any further inflation. While we could observe that the adaptive selection procedure can close the power gap between one- and two-stage designs, the test selection procedure picks adequate but not optimal tests. Overall, this two-stage procedure performs better than traditional log-rank tests in non-proportional hazard settings and offers more flexibility than single-stage combination tests. While it maintains efficiency in proportional hazard settings, the design allows sample size adjustments and interim decision-making. The real data example illustrates the potential of weight adaptation. While the single-stage log-rank test could not reject the hypothesis of equal distributions of overall survival, the adaptive procedures and particularly our new framework resulted in significant results.

As for all weight-based testing procedures, our framework's effectiveness depends on the pre-chosen weights. Another limitation of our work lies within the non-optimal test selection in the second stage, which the extrapolation approach might cause. As adaptive designs in general, our approach does not allow for informed interim decisions if the survival curves of the groups do not separate until after the interim analysis and is thus less suited for very late effects. To address some of these limitations, we aim to investigate additional weight functions, different extrapolation techniques, and alternative decision-making measures, such as predictive power. Another research path is the evaluation of other combination approaches for the stage-wise p-values.

We have demonstrated that use cases with high uncertainty benefit from flexible and robust adaptive methods. We enhanced this flexibility and robustness by embedding various weighted tests and their combinations with a test selection procedure. Thus, we see the potential of our approach in clinical trials with limited prior knowledge.

4 Discussion and Outlook

This thesis consists of five manuscripts tackling the challenge of non-proportional hazards in time-to-event analysis. The assumption of proportional hazards (PH) is (implicitly) made whenever the unweighted log-rank test is applied. In reality, this assumption is often violated, and until now, no standard practice has been established for that.

The first publication emphasized the necessity of non-PH approaches and provided a first assessment of their performance. Scanning published clinical oncology trials for suitable data raised awareness of good practices in publishing survival data and how to analyze it. Furthermore, it provided an unbiased overview of state-of-the-art methods that are more robust to violations of the PH assumption. The log-rank test failed to reject the null hypothesis in all cases, aligning with the original findings. Among the remaining tests, the multi-directional log-rank (mdir) approach introduced by Brendel et al. (2014); Ditzhaus and Friedrich (2020) found the most differences, with other omnibus and RMST-based tests showing varying success. This paper was focused on pointing out the importance of alternative approaches in time-to-event analysis. Still, no final recommendations could be derived solely from (reconstructed) real-world data.

This motivated the second paper to evaluate the performance of promising non-PH approaches on artificial data. We executed extensive simulation studies covering PH, non-PH, and crossing hazard scenarios with multiple censoring distributions and sample sizes. The results of the simulation study favored the mdir test again. While the test seems robust toward multiple alternatives, no sample size approach or effect measure has been introduced yet. Such methodologies are necessary to facilitate applying new testing procedures in practice and provide future research opportunities. One research opportunity identified through the first two publications was the potential of the mdir test. Its robustness and flexibility are favorable properties that could benefit more complex study designs, including multiple testing or adaptive designs.

In clinical trials with more than one treatment group, non-PH are even more frequent. Thus, the availability of non-PH approaches becomes crucial. Since the mdir approach has shown desirable properties, Dormuth et al. (2024a) adapted the concept for multiple testing problems in various ways. Ditzhaus et al. (2021) introduced the extension of the mdir concept to factorial designs by deriving the Cumulative Aalen Survival Analysis-of-Variance (CASANOVA) test. We expanded this approach to the multiple-contrast testing problem. The novel so-called multiCASANOVA approach provides (asymptotically valid) local and global test decisions based on a wild bootstrap procedure. Furthermore, we introduced a novel maximum test of single-weighted log-rank tests.

Both approaches control the familywise error rate (FWER) but show conservative behavior. In future research, we want to focus on exploiting the FWER more with these approaches. Therefore, we plan on combining our current study with closed testing procedures as described in Blanche et al. (2022).

The following publication introduces the mdir test to an adaptive design approach with adaptive weight selection. Our framework considered a two-stage design with one interim and one final analysis. While we focused on various mdir candidates in the first stage, we selected a specific single weighted log-rank test in the second stage. Therefore, we used Royston and Parmar (2002) splines to facilitate the weight choosing. We chose the best test according to conditional power. Both extrapolation and the selection criterion allow investigation of alternative approaches such as a penalized version of the selected approach (Liu et al., 2018) or predictive power(Spiegelhalter et al., 1986). One primary limitation is rooted in the high flexibility of the approach because it comes with the cost of loss of efficiency. We, therefore, recommend using our approach primarily in situations where there is little initial knowledge about the effect size or type.

One major step when planning a clinical trial is the sample size calculation based on pre-defined quantities. Under non-PH, various approaches for sample size calculation have been proposed (Phadnis and Mayo, 2021). While the mdir test has many promising properties, no corresponding effect measure has been introduced up to this point, hampering sample size calculation for the approach. To tackle the topic of sample size calculation, we focused on alternative effect measures. Taking the general criticism of the limited interpretability of hazard ratios under non-PH into account while considering the limited power performance of restricted mean survival approaches in the first two publications of this thesis, we evaluated an alternative effect measure, the average hazard ratio (AHR). While the effect measure and corresponding test have been introduced (Rauch et al., 2018; Brückner and Brannath, 2017), no in-depth consideration of sample size calculation approaches for AHR-based testing has been provided. We derived an asymptotic and a simulation-based sample size approach for AHR-based analysis. For their evaluation, we conducted simulation studies, including different survival time and censoring distributions as well as comparisons with two common sample size calculation approaches in survival analysis. One limitation of the presented framework is its focus on one specific weight function for the AHR. However, extending other weight functions of the same family is straightforward. Furthermore, more extensive simulation studies would be required to derive final recommendations on when to use AHR-based approaches. Nevertheless, we were able to underline the importance of considering different methods for sample size estimation while incorporating all available information. A future path of research could include Bayesian approaches to employ more available information (Kunzmann et al., 2021).

This paragraph discusses the general limitations of this thesis and some ideas for addressing them in future research. The first two publications focused on neutral comparisons of existing methods. One of our current projects (Sauer et al., 2024) aims to assess how neutral these comparisons were measured by state-of-the-art frameworks designed to improve neutrality in

such studies. Furthermore, the results of these two publications provide new research directions. One possible extension would be comparing sample size estimation procedures for non-PH scenarios. Considering the increasing popularity of the maxCombo test (Lin et al., 2020) together with its low power in our simulation study (Dormuth et al., 2023), we want to investigate the reasons behind this behavior further to derive more precise recommendations on when to use the test. After identifying the most promising candidates, further research is required to facilitate the application of those methods. Consequently, closer cooperation with practitioners and the industry would be beneficial. We are confident that we can find a multiple testing procedure with better power performance than the adjusted mdir test. Therefore, we plan to investigate different modifications to the presented methods, such as a maximum mdir test and the incorporation of the closed testing procedure. Further research on the adaptive weight selection approach would include different weight functions and extrapolation approaches. We plan to evaluate different combination methods for the stagewise p-values. Finally, we aim to present a framework for sample size re-estimation within our proposed approach. The AHR (sample size) approach showed promising properties in the real-world data example. Further performance investigation, especially in more extensive comparisons, would be necessary to derive final recommendations on when to use the approach. Ideally, future research involves considering different weighting functions and how they impact the sample size calculation and the test's power performance. Generally, all presented approaches do not consider covariates, and adjustment for those would be a beneficial addition to the presented methods.

Overall, the five works combined in this thesis extended the survival analysis landscape by (i) providing a neutral overview of existing alternatives for the two-sample log-rank test, (ii) exploring alternative effect measures to the hazard ratio, (iii) deriving new methodology from promising approaches identified by the comparisons in the first two publications. This thesis adds to the vital research field of non-PH, aiming toward finding a gold standard when the assumption of PH is potentially violated.

Bibliography

Alexander, B. M., Schoenfeld, J. D., and Trippa, L. (2018). Hazards of hazard ratios-deviations from model assumptions in immunotherapy. *The New England journal of medicine*, 378(12):1158–1159.

Ananthakrishnan, R., Green, S., Previtali, A., Liu, R., Li, D., and LaValley, M. (2021). Critical review of oncology clinical trial design under non-proportional hazards. *Critical reviews in oncology/hematology*, 162:103350.

Bauer, P. (1989). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie*, 20(4):130–148.

Bauer, P., Bretz, F., Dragalin, V., König, F., and Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine*, 35(3):325–347.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.

Blanche, P., Dartigues, J.-F., and Riou, J. (2022). A closed max-t test for multiple comparisons of areas under the roc curve. *Biometrics*, 78(1):352–363.

Boulesteix, A.-L., Lauer, S., and Eugster, M. J. (2013). A plea for neutral comparison studies in computational sciences. *PloS one*, 8(4):e61562.

Brannath, W., Gutjahr, G., and Bauer, P. (2012). Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association*, 107(498):824–832.

Brendel, M., Janssen, A., Mayer, C.-D., and Pauly, M. (2014). Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, 41(3):742–761.

Bretz, F., Genz, A., and A. Hothorn, L. (2001). On the numerical availability of multiple comparison procedures. *Biometrical Journal*, 43(5):645–656.

Bretz, F., Hothorn, T., and Westfall, P. H. (2016). *Multiple comparisons using R*. Chapman & Hall/CRC Press, Boca Raton, Fla.

Brückner, M. (2018). AHR. `https://github.com/cran/AHR`. Accessed: 2024-06-13.

Brückner, M. and Brannath, W. (2017). Sequential tests for non-proportional hazards data. *Lifetime data analysis*, 23:339–352.

Danzer, M. F. and Dormuth, I. (2024). Adaptive weight selection for time-to-event data under non-proportional hazards. *arXiv preprint arXiv:2409.15145*.

Ditzhaus, M., Dobler, D., Pauly, M., Steinhauer, P., and Munko, M. (2022). *GFDsurv: Tests for Survival Data in General Factorial Designs*. R package version 0.1.1.

Ditzhaus, M. and Friedrich, S. (2018). *mdir.logrank: Multiple-Direction Logrank Test*. R package version 0.0.4.

Ditzhaus, M. and Friedrich, S. (2020). More powerful logrank permutation tests for two-sample survival data. *Journal of Statistical Computation and Simulation*, 90(12):2209–2227.

Ditzhaus, M., Genuneit, J., Janssen, A., and Pauly, M. (2021). CASANOVA: Permutation inference in factorial survival designs. *Biometrics*, pages 1–13.

Ditzhaus, M. and Pauly, M. (2019). Wild bootstrap logrank tests with broader power functions for testing superiority. *Computational Statistics & Data Analysis*, 136:1–11.

Dobler, D. and Pauly, M. (2018). Bootstrap- and permutation-based inference for the Mann–Whitney effect for right-censored and tied data. *TEST*, 27(3):639–658.

Dormuth, I., Herrmann, C., Konietschke, F., Pauly, M., Wirth, M., and Ditzhaus, M. (2024a). Single casanova? not in multiple comparisons. *arXiv preprint arXiv:2410.21098*.

Dormuth, I., Liu, T., Xu, J., Pauly, M., and Ditzhaus, M. (2023). A comparative study to alternatives to the log-rank test. *Contemporary Clinical Trials*, 128:107165.

Dormuth, I., Liu, T., Xu, J., Yu, M., Pauly, M., and Ditzhaus, M. (2022). Which test for crossing survival curves? a user's guideline. *BMC medical research methodology*, 22(1):1–7.

Dormuth, I., Pauly, M., Rauch, G., and Herrmann, C. (2024b). Sample size calculation under nonproportional hazards using average hazard ratios. *Biometrical Journal*, 66(6):e202300271.

Emura, T. and Chen, Y.-H. (2018). *Analysis of survival data with dependent censoring: copula-based approaches*, volume 450. Springer.

Fisher, R. A. (1970). Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer.

Fleming, T. R. and Harrington, D. P. (2013). *Counting processes and survival analysis*, volume 625. John Wiley & Sons.

Fleming, T. R., Harrington, D. P., and O'sullivan, M. (1987). Supremum versions of the log-rank and generalized wilcoxon statistics. *Journal of the American Statistical Association*, 82(397):312–320.

Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statistics in medicine*, 1(2):121–129.

Friedrich, S. and Friede, T. (2024). On the role of benchmarking data sets and simulations in method comparison studies. *Biometrical Journal*, 66(1):2200212.

Gao, X., Alvo, M., Chen, J., and Li, G. (2008). Nonparametric multiple comparison procedures for unbalanced one-way factorial designs. *Journal of Statistical Planning and Inference*, 138(6):2574–2591.

Ghosh, P., Ristl, R., König, F., Posch, M., Jennison, C., Götte, H., Schüler, A., and Mehta, C. (2022). Robust group sequential designs for trials with survival endpoints and delayed response. *Biometrical Journal*, 64(2):343–360.

Gorfine, M., Schlesinger, M., and Hsu, L. (2020). K-sample omnibus non-proportional hazards tests based on right-censored data. *Statistical methods in medical research*, 29(10):2830–2850.

Guyot, P., Ades, A., Ouwens, M. J., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC medical research methodology*, 12:1–13.

Hasegawa, T. (2016). Group sequential monitoring based on the weighted log-rank test statistic with the fleming–harrington class of weights in cancer vaccine studies. *Pharmaceutical Statistics*, 15(5):412–419.

Hasler, M. and Hothorn, L. A. (2008). Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(5):793–800.

Kalbfleisch, J. D. and Prentice, R. L. (1981). Estimation of the average hazard ratio. *Biometrika*, 68(1):105–112.

Karrison, T. G. (2016). Versatile tests for comparing survival curves based on weighted log-rank statistics. *The Stata Journal*, 16(3):678–690.

Kim, D. H., Uno, H., and Wei, L.-J. (2017). Restricted mean survival time as a measure to interpret clinical trial results. *JAMA cardiology*, 2(11):1179–1180.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Konietschke, F., Bösiger, S., Brunner, E., and Hothorn, L. A. (2013). Are multiple contrast tests superior to the anova? *The International Journal of Biostatistics*, 9(1):63–73.

Konietschke, F., Hothorn, L. A., and Brunner, E. (2012). Rank-based multiple test procedures and simultaneous confidence intervals. *Electronic Journal of Statistics*, 6.

Koziol, J. A. and Jia, Z. (2009). The concordance index c and the mann–whitney parameter pr (x> y) with randomly censored data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(3):467–474.

Kraus, D. (2009). Adaptive Neyman's smooth tests of homogeneity of two samples of survival data. *Journal of Statistical Planning and Inference*, 139(10):3559–3569.

Kristiansen, I. (2012). Prm39 survival curve convergences and crossing: a threat to validity of meta-analysis? *Value in health*, 15(7):A652.

Kunzmann, K., Grayling, M. J., Lee, K. M., Robertson, D. S., Rufibach, K., and Wason, J. M. (2021). A review of bayesian perspectives on sample size derivation for confirmatory trials. *The American Statistician*, 75(4):424–432.

Lakatos, E. (1986). Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clinical Trials*, 7(3):189–199.

Latimer, N. R. and Adler, A. I. (2022). Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. *BMJ Medicine*, 1(1).

Lee, J. W. (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*, pages 721–725.

Lee, S.-H. (2007). On the versatility of the combination of the weighted log-rank statistics. *Computational statistics & data analysis*, 51(12):6557–6564.

Legrand, C. (2021). *Advanced survival models*. Chapman and Hall/CRC.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290.

Li, H., Han, D., Hou, Y., Chen, H., and Chen, Z. (2015). Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One*, 10(1):e0116774.

Lin, R. S., Lin, J., Roychoudhury, S., Anderson, K. M., Hu, T., Huang, B., Leon, L. F., Liao, J. J., Liu, R., Luo, X., et al. (2020). Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. *Statistics in Biopharmaceutical Research*, 12(2):187–198.

Liu, N., Zhou, Y., and Lee, J. J. (2021). Ipdfromkm: reconstruct individual patient data from published kaplan-meier survival curves. *BMC medical research methodology*, 21(1):111.

Liu, T. (2020). Rbt4tcsc. `https://github.com/LTTGH/RBT4TCSC`. Accessed: 2024-06-13.

Liu, T., Ditzhaus, M., and Xu, J. (2020). A resampling-based test for two crossing survival curves. *Pharmaceutical Statistics*, 19(4):399–409.

Liu, X.-R., Pawitan, Y., and Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical methods in medical research*, 27(5):1531–1546.

Logan, B. R., Wang, H., and Zhang, M.-J. (2005). Pairwise multiple comparison adjustment in survival analysis. *Statistics in medicine*, 24(16):2509–2523.

Mick, R. and Chen, T.-T. (2015). Statistical challenges in the design of late-stage cancer immunotherapy studies. *Cancer immunology research*, 3(12):1292–1298.

Munko, M., Ditzhaus, M., Dobler, D., and Genuneit, J. (2024). RMST-based multiple contrast tests in general factorial designs. *Statistics in Medicine*, 43(10):1849–1866.

Phadnis, M. A. and Mayo, M. S. (2021). Sample size calculation for two-arm trials with time-to-event endpoint for nonproportional hazards using the concept of relative time when inference is built on comparing weibull distributions. *Biometrical Journal*, 63(7):1406–1433.

Proschan, M. A., Lan, K. G., and Wittes, J. T. (2006). *Statistical monitoring of clinical trials: a unified approach*. Springer Science & Business Media.

Qiu, P. and Sheng, J. (2008). A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):191–208.

Rahman, R., Fell, G., Ventz, S., Arfé, A., Vanderbeek, A. M., Trippa, L., and Alexander, B. M. (2019). Deviation from the proportional hazards assumption in randomized phase 3 clinical trials in oncology: prevalence, associated factors, and implications. *Clinical Cancer Research*, 25(21):6339–6345.

Rauch, G., Brannath, W., Brückner, M., and Kieser, M. (2018). The Average Hazard Ratio – A Good Effect Measure for Time-to-event Endpoints when the Proportional Hazard Assumption is Violated? *Methods of information in medicine*, 57(03):089–100.

Rohatgi, A. (2019). Webplotdigitizer: Version 4.4. `https://automeris.io/WebPlotDigitizer`.

Royston, P., Lambert, P. C., et al. (2011). *Flexible parametric survival analysis using Stata: beyond the Cox model*, volume 347. Stata press College Station, TX.

Royston, P. and Parmar, M. K. (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):2175–2197.

Royston, P. and Parmar, M. K. (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):2409–2421.

Royston, P. and Parmar, M. K. (2014). An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials*, 15(1):1–10.

Royston, P. and Parmar, M. K. (2016). Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Medical Research Methodology*, 16:1–13.

Royston, P. and Parmar, M. K. (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*, 21:1–17.

Rufibach, K. (2019). Treatment effect quantification for time-to-event endpoints–estimands, analysis strategies, and beyond. *Pharmaceutical statistics*, 18(2):145–165.

Rutherford, M. J., Crowther, M. J., and Lambert, P. C. (2015). The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation*, 85(4):777–793.

Sauer, C., Lange, F. J. D., Thurow, M., Dormuth, I., and Boulesteix, A.-L. (2024). Towards more practically relevant and neutral comparative simulation studies: Illustrating chances, challenges, and possible implementations of simulation settings based on real data. *In preparation*.

Schaarschmidt, F., Biesheuvel, E., and Hothorn, L. A. (2009). Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials. *Journal of Biopharmaceutical Statistics*, 19(2):292–310.

Schlesinger, M. and Gorfine, M. (2020). *KONPsurv: KONP Tests: Powerful K-Sample Tests for Right-Censored Data*. R package version 1.0.3.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68(1):316–319.

Sheng, J., Qiu, P., and Geyer, C. J. (2019). *TSHRC: Two Stage Hazard Rate Comparison*. R package version 0.1-6.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled clinical trials*, 7(1):8–17.

Therneau, T. M. and Grambsch, P. M. (2021). *survival: Survival Analysis*.

Thurow, M., Dormuth, I., Sauer, C., Ditzhaus, M., and Pauly, M. (2023). How to simulate realistic survival data? a simulation study to compare realistic simulation models. *arXiv preprint arXiv:2308.07842*.

Tian, L., Fu, H., Ruberg, S. J., Uno, H., and Wei, L.-J. (2018). Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*, 74(2):694–702.

Tian, L., Uno, H., and Horiguchi, M. (2017). *surv2sampleComp: Inference for Model-Free Between-Group Parameters for Censored Survival Data*. R package version 1.0-5.

Trinquart, L., Jacot, J., Conner, S. C., and Porcher, R. (2016). Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, 34(15):1813–1819.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., et al. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of clinical Oncology*, 32(22):2380.

Uno, H., Tian, L., Horiguchi, M., Cronin, A., Battioui, C., and Bell, J. (2020). *survRM2: Comparing Restricted Mean Survival Time*. R package version 1.0-3.

Wassmer, G. and Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*, volume 301. Springer.

Wassmer, G. and Pahlke, F. (2023). *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*. R package version 3.3.4.

Xiong, X. and Wu, J. (2017). A novel sample size formula for the weighted log-rank test under the proportional hazards cure model. *Pharmaceutical statistics*, 16(1):87–94.

Yung, G. and Liu, Y. (2020). Sample size and power for the weighted log-rank test and kaplan-meier based tests with allowance for nonproportional hazards. *Biometrics*, 76(3):939–950.

Zhao, J. J., Syn, N. L., Tan, B. K. J., Yap, D. W. T., Teo, C. B., Chan, Y. H., and Sundar, R. (2022). Kmsubtraction: reconstruction of unreported subgroup survival data utilizing published kaplan-meier survival curves. *BMC medical research methodology*, 22(1):93.

Zhao, L., Claggett, B., Tian, L., Uno, H., Pfeffer, M. A., Solomon, S. D., Trippa, L., and Wei, L. (2016). On the restricted mean survival time curve in survival analysis. *Biometrics*, 72(1):215–221.

Part II

Publications

Aritcle 1

Dormuth, I., Liu, T., Xu, J., Yu, M., Pauly, M., & Ditzhaus, M. (2022). Which test for crossing survival curves? A user's guideline. BMC medical research methodology, 22(1), 34.

Dormuth *et al.*
BMC Medical Research Methodology (2022) 22:34
https://doi.org/10.1186/s12874-022-01520-0

BMC Medical Research Methodology

RESEARCH Open Access

Which test for crossing survival curves? A user's guideline

Ina Dormuth[1*], Tiantian Liu[2], Jin Xu[3], Menggang Yu[4], Markus Pauly[1] and Marc Ditzhaus[1]

Abstract

Background: The exchange of knowledge between statisticians developing new methodology and clinicians, reviewers or authors applying them is fundamental. This is specifically true for clinical trials with time-to-event endpoints. Thereby, one of the most commonly arising questions is that of equal survival distributions in two-armed trial. The log-rank test is still the gold-standard to infer this question. However, in case of non-proportional hazards, its power can become poor and multiple extensions have been developed to overcome this issue. We aim to facilitate the choice of a test for the detection of survival differences in the case of crossing hazards.

Methods: We restricted the review to the most recent two-armed clinical oncology trials with crossing survival curves. Each data set was reconstructed using a state-of-the-art reconstruction algorithm. To ensure reproduction quality, only publications with published number at risk at multiple time points, sufficient printing quality and a non-informative censoring pattern were included. This article depicts the *p*-values of the log-rank and Peto-Peto test as references and compares them with nine different tests developed for detection of survival differences in the presence of non-proportional or crossing hazards.

Results: We reviewed 1400 recent phase III clinical oncology trials and selected fifteen studies that met our eligibility criteria for data reconstruction. After including further three individual patient data sets, for nine out of eighteen studies significant differences in survival were found using the investigated tests. An important point that reviewers should pay attention to is that 28% of the studies with published survival curves did not report the number at risk. This makes reconstruction and plausibility checks almost impossible.

Conclusions: The evaluation shows that inference methods constructed to detect differences in survival in presence of non-proportional hazards are beneficial and help to provide guidance in choosing a sensible alternative to the standard log-rank test.

Keywords: Survival analysis, Time-to-event outcome, Crossing, Non-proportional hazards, Oncology, Log-rank test, Restricted-mean survival

Background

Time-to-event studies are the paramount studies in clinical practice. Typical examples are two-armed trials providing a reliable comparison of the efficacy and safety of two treatments. Statistical methods that infer a potential difference in survival are of fundamental importance [1]. Among methods designed to compare the overall survival of two groups, the log-rank test (LR) is still the most used [2]. Beyond a certain resistance to statistical innovations [3], there is also a theoretical reason: The LR is optimal in case of proportional hazards (PH) [4]. In other words, if the hazard functions of the two groups are proportional, the LR is the most powerful method to detect differences between them. However, this changes completely for other kinds of hazard patterns, in particular for crossing

*Correspondence: Ina.dormuth@tu-dortmund.de
[1] TU Dortmund University, Joseph-von-Fraunhofer-Straße 2-4, 44221 Dortmund, Germany
Full list of author information is available at the end of the article

hazards and the rejection rates of the LR drop significantly. The alarming observation of Kristiansen [5], who reviewed 175 studies in five renowned journals, is that the LR was applied in 70% of the cases despite crossing survival curves. These crossings can occur e.g. in oncology when comparing tumor dissection versus radiation strategies due to different time-dependent effects.

Consequently, several methods have been and are still proposed to tackle non-PH situations. However, due to the speed of research and the number of new methods, the exchange of knowledge is a challenge. Therefore, Ananthakrishnan et al. [6] recently provided a critical review on methods in the presence of possible non-PHs and their limitations and advantages. While they give detailed information regarding the assumptions and the context, they do not provide any numerical evaluation of the methods. We include here state of the art tests with the aim of providing biostatisticians, physicians and reviewers with a condensed overview of suitable methods for non-PH settings that are implemented in the open statistical software R. These methods not only show good results in various simulation studies but also on real data.

Methods

There are several papers that develop alternatives to the LR in case of non-PH or even crossing hazards. Treating them all would go far beyond the scope of this work. Hence, we focused our comparisons on standard methods that performed well in other simulation studies and more recent ones that were not yet included in extensive evaluations. Here, all analyses are conducted using the free and open-source software R [7] (except for the test introduced by Royston [8]).

Fortunately, the paper by Li et al. [9] already provides a review on methods for crossing hazards up to 2014. Based on extensive simulation studies they recommend two procedures: First, Neyman's smooth test proposed by Kraus [10]. This test is not considered further since the corresponding R package was removed recently. Second, a two-stage procedure (2ST) that is based on the LR and a crossing-hazards test is proposed (see the Supplement for more details.). The test is described by Qiu and Sheng [11] and implemented in the R package *TSHRC* [12].

Further methods have been developed since 2014. We have included the most relevant ones into our study. For example, Gorfine et al. [13] presented two omnibus permutation tests based on a sample space partition, which showed promising results in non-PH situations. These are either based on test statistics of Pearson's chi square (KONP chi) or likelihood-ratio type (KONP llr) and are available in the R package *KONPsurv* [14]. They compared their new approach with the well-established test of Yang and Prentice [15], which belongs to the class of weighted log-rank tests and employs adaptive weights. Since Gorfine et al. [13] could show in simulations that their new tests are more powerful in the studied non-PH settings, the Yang and Prentice test is not included in our comparison. Another idea starts with the class of weighted LR. This class is long known and includes the LR as well as the common Peto-Peto test (PP). Recently, a flexible combination of several weighted LRs into one test procedure was proposed [16–18]. It is based upon a combination of alternatives and carried out as a permutation procedure. Recently, it has been implemented in the R package *mdir.log-rank* [19]. The multiple-direction log-rank test (mdir) combines several weighted log-rank tests into one joint Wald-type statistic, which can be interpreted as a projection on a large alternative space spanned by pre-chosen weights. The latter ensures that mdir has not only a reasonable power in the directions of the chosen weights (e.g. for PHs or a specific crossing curve situation) but also in the directions of any linear combination of the pre-chosen weights. Moreover, the weights are allowed to be data-dependent. Another approach that combines multiple weighted log-rank tests is the MaxCombo test (MaxCombo). Different to mdir, the final test statistic is the maximum over standardized weighted LR tests [20]. We used the same list of weights as proposed in the description of the *nphsim* package [21]. We refer to the supplement for specific as well as technical details on all methods. Besides HR, the restricted mean survival time (RMST) can be used to quantify the difference between two survival curves [22]. It describes the mean event-free survival time up to a pre-defined time point τ. Hypothesis tests constructed using the RMST examine whether the RMST difference between groups is zero. This test is also valid to test equality of two survival functions, since equal survival functions imply equal RMST. Unfortunately, it is possible to observe situations where the RMSTs are equal but the survival functions are not. This has to be kept in mind while using RMST-based tests. We consider three RMST-based proposals: The first two utilize the group-wise RMST differences as test statistic and either calculate p-values based on resampling (RMST1) or obtained using asymptotic theory (RMST2) [23, 24]. The former is provided by the R package *surv2sample* [25] while the latter can be computed with the function *rmst2* in *survRM2* [26]. Eventually, Royston and Parmar [27, 28] propagate a test combining a Cox test and a permutation-based RMST test (coxRMST). The test by Royston and Parmar is only available in STATA using the *stctest* function. Finally, we consider a test based on an integrated L_1-distance of the two Kaplan-Meier curves as test statistic. It can be interpreted as the area between curves (ABC) and was introduced in Liu et al. [29]. It has not been

implemented in R yet and was thus coded by ourselves according to the author's descriptions. The code can be found in the supplements.

A detailed description of all eleven tests and corresponding test statistics can be found in the Supplement. Furthermore, a simple example in R is given in the Supplement. Below we will compare them based upon different studies. To this end, we reconstruct data from published Kaplan-Meier curves using the algorithm developed by Guyot et al. [30] and deriving the data from the curves with the freely available *Webplotdigitizer* [31].

Results

Eligibility screening and data extraction

Our study was motivated by the work of Matabuena and Padilla [32] which includes three oncology studies with crossing Kaplan Meier (KM) curves. We subsequently performed a PubMed screening of recent oncology studies with similar patterns. To ensure these patterns, the search matched *((Phase 3) OR (phase III)) OR (Kaplan-Meier) OR (Kaplan Meier))* for Cancer and Humans were used. To categorize them, multiple criteria listed in Fig. 1 were defined to identify relevant studies on PubMed. 1400 of the most recent papers (status from Oct 5, 2020) on clinical oncology were searched for crossing survival curves with published number at risk at multiple time points. More details can be found in eTable 1 in the online Supplement. The executed LR test had to be non-significant and the two arms should only cross one or two times. To ensure a good reconstructibility, a sufficient number of events and high quality of the curves as well as non-informative censoring over time were required. In the end, the reconstruction algorithm of Guyot et al. [30] was applied to fifteen publications that met these requirements and the three studies discussed in the paper of Matabuena and Padilla [32]. Beyond insufficient information (e.g., almost 30% of the publications did not report the number at risks) another reason for the final small number of publications can be publication bias since non-significant results are less often reported.

Data reconstruction

The individual patient data from the three studies found in Matabuena and Padilla [32] and the fifteen other studies under consideration [33–50] were reconstructed using the algorithm introduced by Guyot et al. [30]. To assess the quality of reconstruction, the reported key statistics (median survival and HR with confidence interval) published in each paper were recalculated and compared to the original values (see Table 1).

Comparison of tests for proportional hazards and crossing hazards

The reconstructed individual patient data were then used to compare the different testing approaches. For all resampling-based methods, the number of iterations was set to 5000 and for all RMST procedures the parameter τ was set to 90% of the minimum of the largest censored or uncensored time among the arms [51]. The results are listed in Table 2.

It can be observed that the LR test never succeeds to reject the null hypothesis of equal survival in both groups at the 5% level. This leads to the exact same conclusion as in the eighteen published studies. The PP is designed to find early differences [52]. It succeeds in revealing an inequality in survival for four of the eighteen studies under consideration [33, 40, 45, 47]. Let us next consider the three RMST tests. These do not rely on the assumption of PHs but are also not specifically designed to detect crossings [53]. The resampling-based (RMST1) and the distribution-based version (RMST2) reject the null hypothesis in three cases [33, 34, 40], while the combined test (coxRMST) rejects the null hypothesis in five cases [33, 39, 40, 45, 47]. These findings support the analyses of Royston et al. [54]. The six remaining tests are all omnibus tests with different properties. The two tests by Gorfine et al. [13]. (KONP chi and KONP llr) find differences in survival in the same six cases [33, 34, 41, 42, 45, 47]. The omnibus test by Ditzhaus and Friedrich [17] (mdir) can reject the null hypothesis in eight out of eighteen cases [33, 37, 39–41, 45, 47]. The two-stage procedure (2ST) detects differences in five out of eighteen data sets [33, 40, 41, 45, 47]. The ABC has significant results for the same five studies as the two-stage test [33, 40, 41, 45, 47]. The MaxCombo test leads to p-values smaller than 0.05 for seven of the eighteen data sets [34, 39–42, 45, 47]. In these specific data examples, the test by Ditzhaus and Friedrich [14] is the test that detects the most differences. These results are consistent with those of Li et al. [9], Gorfine et al. [13] and Royston and Parmar [28] who also indicated that omnibus tests have greater power when deviating from the proportional hazards assumption. Evaluation of the methods' performance

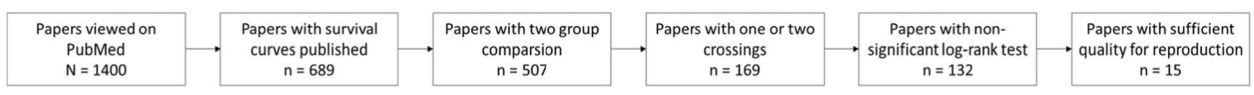


Fig. 1 Flow chart of papers under consideration

Table 1 Assessment of data reconstruction quality

Publication	MS G1	MS G2	HR [CI]
Bang et al. (2020) [37]	5.80 (5.88)	4.30 (4.44)	0.83 [0.53, 1.31] (0.82 [0.52, 1.29])
Becker et al. (2020) [39]	not defined	6.00 (6.21)	5.50 (5.51)
Bellmunt et al. (2017) [45]	3.30 (3.24)	2.10 (2.08)	0.98 [0.81, 1.19] (0.93 [0.77, 1.13])
Cortes et al. (2019) [46]	4.90 (4.94)	4.70 (4.72)	0.63 (0.62)
Ferris et al. (2016) [41]	2.00 (2.02)	2.30 (2.29)	0.89 [0.70,1.13] (0.89 [0.70,1.14])
Fradet et al. (2019) [47]	3.30 (3.35)	2.10 (2.16)	0.96 [0.79, 1.16] (0.92 [0.77, 1.11])
Godfrey et al. (2018) [36]	–	–	1.40 [0.54, 3.61] (1.40 [0.53, 3.69])
Golan et al. (2019) [38]	18.90 (18.90)	18.10 (18.10)	0.91 [0.56, 1.46] (0.88 [0.55, 1.42])
Hammel et al. (2019) [35]	21.20 (21.36)	6.00 (5.93)	0.72 [0.41, 1.27] (0.72 [0.42, 1.24])
Jones et al. (2020) [34]	26.00 (26.00)	20.0 (18.80)	0.59 [0.34, 1.05] (0.58 [0.33, 1.02])
Jones et al. (2018) [33]	15.10 (15.08)	8.10 (8.02)	0.72 [0.45, 1.17] (0.71 [0.44, 1.15])
Kotani et al. (2019) [48]	8.60 (8.62)	8.00 (8.02)	0.74 [0.48, 1.14] (0.72 [0.47, 1.11])
Kreuzer et al. (2020) [50]	19.40 (19.40)	20.90 (21.30)	1.22 [0.60, 2.47] (1.26 [0.62, 2.56])
Lu et al. (2018) [40]	4.63 (4.68)	4.23 (4.33)	0.78 [0.60, 1.00] (0.74 [0.55, 1.01])
Malone et al. (2020) [49]	–	–	0.66 [0.41, 1.07] (0.68 [0.42, 1.10])
Motzer et al. (2015) [42]	4.60 (4.46)	4.40 (4.07)	0.88 [0.75, 1.03] (0.87 [0.98, 1.34])
Mukai et al. (2019) [44]	27.90 (27.90)	16.60 (16.60)	0.55 [0.23, 1.29] (0.55 [0.23, 1.29])
Toxopeus et al. (2018) [43]	–	–	1.02 [0.75, 1.39] (1.01 [0.75, 1.39])

Quality of data reconstruction regarding the published median survival (MS) in group 1 and 2 (G1 and G2), the hazard ratio (HR) with 95% confidence intervals (CI). For each study the published statistics are given with the corresponding statistics of the reconstructed data in parentheses. Three studies did not report MS (—) and two did not provide confidence intervals

Table 2 *P*-values of the different tests applied to the reconstructed individual patient data of each publication

Publication	LR	PP	RMST1	RMST2	coxRMST	KONP_chi	KONP_llr	Mdir	2ST	ABC	MaxCombo
Bang et al. (2020) [37]	0.37	0.07	0.11	0.12	0.11	0.14	0.15	**0.03**	0.06	0.13	0.1
Becker et al. (2020) [39]	0.09	0.47	0.22	0.22	**0.02**	0.14	0.09	**0.02**	0.27	0.15	**0.04**
Bellmunt et al. (2017) [45]	0.49	**0.03**	0.38	0.38	**0.003**	**< 0.001**	**< 0.001**	**< 0.001**	**0.03**	**0.002**	**< 0.001**
Cortes et al. (2019) [46]	0.19	0.24	0.23	0.24	0.28	0.40	0.36	0.41	0.87	0.29	0.56
Ferris et al. (2016) [41]	0.33	0.84	0.23	0.23	0.25	**0.01**	**0.009**	**0.02**	**0.04**	**0.03**	**< 0.001**
Fradet et al.(2019) [47]	0.40	**0.02**	**0.04**	**0.04**	**0.009**	**< 0.001**	**< 0.001**	**< 0.001**	**0.03**	**0.001**	**< 0.001**
Godfrey et al. (2018) [36]	0.49	0.48	0.58	0.59	0.63	0.18	0.20	0.75	0.90	0.44	0.74
Golan et al. (2019) [38]	0.61	0.78	9.74	0.75	0.50	0.58	0.59	0.61	0.22	0.61	0.66
Hammel et al. (2019) [35]	0.22	0.35	0.62	0.62	0.33	0.19	0.19	0.16	0.27	0.38	0.09
Jones et al. (2020) [34]	0.05[a]	0.11	0.14	0.14	0.07	**0.02**	**0.02**	0.12	0.41	0.11	**0.04**
Jones et al. (2018) [33]	0.17	**0.03**	**0.02**	**0.02**	**0.05**	**0.03**	**0.04**	**0.009**	**0.05**	**0.03**	0.05[a]
Kotani et al. (2019) [48]	0.14	0.24	0.38	0.38	0.20	0.48	0.48	0.28	0.45	0.45	0.17
Kreuzer et al. (2020) [50]	0.53	0.25	0.07	0.08	0.17	0.28	0.28	0.10	0.10	0.07	0.27
Lu et al. (2018) [40]	0.06	**0.007**	**0.02**	**0.02**	**0.02**	0.07	0.07	**0.007**	**0.04**	**0.01**	**0.01**
Malone et al. (2020) [49]	0.11	0.12	0.13	0.13	0.17	0.08	0.09	0.28	0.57	0.12	0.22
Motzer et al. (2015) [42]	0.07	0.51	0.13	0.13	0.11	**0.03**	**0.03**	**0.01**	0.26	0.08	**< 0.001**
Mukai et al. (2019) [44]	0.17	0.22	0.15	0.17	0.26	0.22	0.25	0.33	0.53	0.16	0.44
Toxopeus et al. (2018) [43]	0.91	0.84	0.75	0.75	0.35	0.37	0.36	0.15	0.11	0.56	0.34

[a] Only 0.05 due to rounding down

Bold values indicate *p*-values smaller than the 5% type-I error level

under PHs reveals that almost all of the approaches reject the null hypothesis when the LR does (for details see the Supplement). In future simulation studies, the performance of the tests and their extensions to multi-arm settings will be further evaluated [13, 55–57].

Discussion

To assess efficacy of two treatments the LR is generally regarded as the gold standard. The LR is optimal in terms of power under the PH assumption but can lose sufficient power in non-PH situations. The results of our PubMed analysis, however, show that there are many situations, where the LR is used in case of non-PH. At the same time, several alternatives are presented, which succeed to detect differences where the LR fails. The majority of these tests are available in statistical software (R). Hence, their execution is almost as user-friendly as calculating the LR. To furtherfacilitate their application, we provide minimal examples on how to use the implemented R functions in the supplement.

To exemplify the different implications, we reconstructed individual patient data from eighteen recent oncology trials that met the eligibility criteria of our analysis. In particular, high quality KM plots with sufficient information were necessary for the reconstruction algorithm. Based on these eighteen studies we compared the test decisions of eleven different testing procedures. It turns out, that the LR alternatives can exhibit power to identify differences between groups. Omnibus approaches, which have high power against several alternatives (such as PH and crossings in case of the mdir test), turned out to be particularly suitable for this purpose (see the Supplement for additional information regarding PH performance).

Limitations

One of the main limitations of this kind of study is the dependence on the selection of data sets. To make a clear statement regarding the quality of the individual procedures in a direct comparison, extensive simulation studies are necessary. These are part of our own ongoing research. Nevertheless, it can be said that the LR cannot reject the null hypothesis in real situations involving non-proportional hazards included in this paper, while various omnibus tests are able to do so. Furthermore, the data used here are reconstructed individual patient data and thus does not have the same quality as the original data. While many properties of the data such as non-proportionality are conserved, the biggest reconstruction issue is the assumption of uniformly distributed censoring times. However, the assessment of the reconstruction quality turned out to be very satisfying.

Recommendations for reviewers

Regarding the insights of our investigation, attention in the reviewing process of study reports should be paid to

(1) the appropriate choice of the statistical method. Especially when the PH assumption cannot be justified in advance, e.g. by a preliminary study, alternatives to the LR should be considered. Due to multiplicity issues, we do not advocate the common practice of pre-testing the PH assumption. Instead, we suggest directly applying a procedure which can detect survival curve differences in PH as well as non-PH settings, such as the methods presented in this paper.
(2) the quality of the data presentation and the report of all relevant information. This includes, in particular, the table of the number at risks at multiple time points, which was not reported in almost 30% of the reviewed publications. These tables and all relevant information can be easily accessed through each common statistical software and should be provided in every study report. They are mandatory for a reliable assessment of the results and, moreover, facilitate a secondary analysis, e.g. for meta-analysis studies, by reconstructing the original data in a reasonable quality [25].

Conclusion

We conclude that in case of non-PH, the choice of a suitable test procedure is relevant and the LR is not always the best choice. Therefore, we recommend to use all prior information available and to consider more options to test for differences in survival than just the LR. In terms of study design there are still some limitations since not all of the tests are used for sample size estimation and some tests are not freely available in R (see the Supplements for more information). Finally, we recommend using omnibus tests such as the mdir test for inference when no prior information on the pattern of hazards is available.

Abbreviations

2ST: Two-stage test; ABC: Area between curves; coxRMST: Combined Cox and permutation based RMST test; IPD: Individual patient data; KM: Kaplan-Meier; KONP chi: K-sample omnibus non-proportional hazards test with chi-square test statistic; KONP llr: K-sample omnibus non-proportional hazards test with log likelihood ratio type test statistic; LR: Log-rank test; mdir: Multiple-direction log-rank test; PH: Proportional hazards; PP: Peto-Peto test; RMST: Restricted mean suvival time.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01520-0.

Additional file 1.

Acknowledgements
The authors are grateful to the editor, the associate editor and the two referees for their valuable feedback and suggestions that improved the quality of the paper.

Authors' contributions
All of the authors were involved in the planning of the study. Ina Dormuth conducted the literature review from which she searched, reconstructed and treated the data in R. This initial step was jointly supervised by Dr. Marc Ditzhaus and Prof. Dr. Markus Pauly. Dr. Tiantian Liu provided the R-Code for the ABC-method, which is not available as an R-package yet and participated in writing the methods section. Furthermore, Ina Dormuth prepared the first draft of the publication, which was then jointly polished by all authors. Prof. Dr. Jin Xu and Prof. Dr. Menggang Yu gave final notes for improvement. We followed the same procedure for the revision. The author(s) read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. Marc Ditzhaus and Markus Pauly were supported by German Research Foundation Grant No PA 2409/5–1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
We state that all methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]TU Dortmund University, Joseph-von-Fraunhofer-Straße 2-4, 44221 Dortmund, Germany. [2]Technion – Israel Institute of Technology, Haifa, Israel. [3]East China Normal University, Shanghai, China. [4]University of Wisconsin-Madison, Madison, USA.

Received: 24 June 2021 Accepted: 18 January 2022
Published online: 30 January 2022

References

1. Fleming TR, Lin DY. Survival analysis in clinical trials: past developments and future directions. Biometrics. 2000;56(4):971–83. https://doi.org/10.1111/j.0006-341X.2000.0971.x.
2. Kleinbaum DG, Klein M. Survival Analysis, vol. 3: Springer; 2010.
3. Sharpe D. Why the resistance to statistical innovations? Bridging the communication gap. Psychol Methods. 2013;18(4):572–82. https://doi.org/10.1037/a0034177.
4. Fleming TR, Harrington DP. Counting Processes and Survival Analysis. Wiley; 2011.
5. Kristiansen I. PRM39 survival curve convergences and crossing: a threat to validity of meta-analysis. Value Health. 2012;15(7):A652.
6. Ananthakrishnan R, Green S, Previtali A, Liu R, Li D, LaValley M. Critical review of oncology clinical trial design under non-proportional hazards. Crit Rev Oncol Hematol. 2021;162:103350. https://doi.org/10.1016/j.critrevonc.2021.103350.
7. R Core Team. R: A Language and Environment for Statistical Computing. 2020.
8. Royston P. A combined test for a generalized treatment effect in clinical trials with a time-to-event outcome. Stata J Promot Commun Stat Stata. 2017;17(2):405–21. https://doi.org/10.1177/1536867X1701700209.
9. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical inference methods for two crossing survival curves: a comparison of methods. PLoS One. 2015;10(1):1–18.
10. Kraus D. Adaptive Neyman's smooth tests of homogeneity of two samples of survival data. J Stat Plan Inference. 2009;139(10):3559–69.
11. Qiu P, Sheng J. A two-stage procedure for comparing hazard rate functions. J R Stat Soc Ser B Stat Methodol. 2008;70(1):191–208.
12. Sheng J, Qiu P, Geyer CJ. TSHRC: Two Stage Hazard Rate Comparison. 2019. https://CRAN.R-project.org/package=TSHRC. Accessed 25 Oct 2021.
13. Gorfine M, Schlesinger M, Hsu L. K-sample omnibus non-proportional hazards tests based on right-censored data. ArXiv Prepr ArXiv; 2019. p. 190105739.
14. Schlesinger M, Gorfine M. KONPsurv: KONP Tests: Powerful K-Sample Tests for Right-Censored Data.; 2020. https://CRAN.R-project.org/package=KONPsurv. Accessed 25 Oct 2021
15. Yang S, Prentice R. Improved logrank-type tests for survival data using adaptive weights. Biometrics. 2010;66(1):30–8.
16. Brendel M, Janssen A, Mayer CD, Pauly M. Weighted Logrank permutation tests for randomly right censored life science data: weighted logrank permutation tests. Scand J Stat. 2014;41(3):742–61. https://doi.org/10.1111/sjos.12059.
17. Ditzhaus M, Friedrich S. More powerful logrank permutation tests for two-sample survival data. ArXiv180705504 Math Stat. 2018. http://arxiv.org/abs/1807.05504. Accessed 6 May 2020
18. Ditzhaus M, Pauly M. Wild bootstrap logrank tests with broader power functions for testing superiority. Comput Stat Data Anal. 2019;136:1–11.
19. Ditzhaus M, Friedrich S. Mdir.Logrank: Multiple-Direction Logrank Test.; 2018. https://CRAN.R-project.org/package=mdir.logrank. Accessed 25 Oct 2021
20. Lee SH. On the versatility of the combination of the weighted log-rank statistics. Comput Stat Data Anal. 2007;51(12):6557–64.
21. Wang Y, Wu H, Anderson KM, Roychoudhury S, Hu T, Liu H. NPHSIM: simulation and power calculations for time-to-event clinical trials; 2017. R package version 0.1.1.9000.
22. Kim DH, Uno H, Wei LJ. Restricted mean survival time as a measure to interpret clinical trial results. JAMA Cardiol. 2017;2(11):1179–80.
23. Tian L, Fu H, Ruberg SJ, Uno H, Wei LJ. Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations: efficiency of two sample tests via the restricted mean survival time. Biometrics. 2018;74(2):694–702. https://doi.org/10.1111/biom.12770.
24. Uno H, Claggett B, Tian L, et al. Moving beyond the Hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol. 2014;32(22):2380–5. https://doi.org/10.1200/JCO.2014.55.2208.
25. Tian L, Uno H, Horiguchi M. Surv2sampleComp: Inference for Model-Free Between-Group Parameters for Censored Survival Data. https://rdrr.io/cran/surv2sampleComp/man/surv2sample.html. Accessed 25 Oct 2021.
26. Uno H, Tian L, Horiguchi M, Cronin A, Battioui C, Bell J. SurvRM2: Comparing Restricted Mean Survival Time; 2020. https://CRAN.R-project.org/package=survRM2. Accessed 25 Oct 2021.
27. Royston P, Parmar MKB. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. BMC Med Res Methodol. 2016;16(1):16. https://doi.org/10.1186/s12874-016-0110-x.
28. Royston P, Parmar MK. A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome. Trials. 2020;21(1):1–17. https://doi.org/10.1186/s13063-020-4153-2.
29. Liu T, Ditzhaus M, Xu J. A resampling-based test for two crossing survival curves. Pharm Stat. 2020;19(4):399–409.
30. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol. 2012;12(1):9. https://doi.org/10.1186/1471-2288-12-9.
31. WebPlotDigitizer - Extract data from plots, images, and maps. https://automeris.io/WebPlotDigitizer/. Accessed Oct 25 2021

32. Matabuena M, Padilla OHM. Energy distance and kernel mean embeddings for two-sample survival testing. arXiv preprint arXiv:1912.04160, 2019.
33. Jones RL, Demetri GD, Schuetze SM, et al. Efficacy and tolerability of trabectedin in elderly patients with sarcoma: subgroup analysis from a phase III, randomized controlled study of trabectedin or dacarbazine in patients with advanced liposarcoma or leiomyosarcoma. Ann Oncol. 2018;29(9):1995–2002. https://doi.org/10.1093/annonc/mdy253.
34. Jones RH, Casbard A, Carucci M, et al. Fulvestrant plus capivasertib versus placebo after relapse or progression on an aromatase inhibitor in metastatic, oestrogen receptor-positive breast cancer (FAKTION): a multicentre, randomised, controlled, phase 2 trial. Lancet Oncol. 2020;21(3):345–57. https://doi.org/10.1016/S1470-2045(19)30817-4.
35. Hammel P, Kindler HL, Reni M, et al. Health-related quality of life in patients with a germline BRCA mutation and metastatic pancreatic cancer receiving maintenance olaparib. Ann Oncol. 2019;30(12):1959–68. https://doi.org/10.1093/annonc/mdz406.
36. Godfrey AL, Campbell PJ, MacLean C, et al. Hydroxycarbamide plus aspirin versus aspirin alone in patients with essential Thrombocythemia age 40 to 59 years without high-risk features. J Clin Oncol. 2018;36(34):3361–9. https://doi.org/10.1200/JCO.2018.78.8414.
37. Bang Y, Li C, Lee K, et al. Liposomal irinotecan in metastatic pancreatic adenocarcinoma in Asian patients: subgroup analysis of the NAPOLI-1 study. Cancer Sci. 2020;111(2):513–27. https://doi.org/10.1111/cas.14264.
38. Golan T, Hammel P, Reni M, et al. Maintenance Olaparib for germline *BRCA* -mutated metastatic pancreatic Cancer. N Engl J Med. 2019;381(4):317–27. https://doi.org/10.1056/NEJMoa1903387.
39. Becker H, Pfeifer D, Ihorst G, et al. Monosomal karyotype and chromosome 17p loss or TP53 mutations in decitabine-treated patients with acute myeloid leukemia. Ann Hematol. 2020;99(7):1551–60. https://doi.org/10.1007/s00277-020-04082-7.
40. Lu S, Chen Z, Hu C, et al. Nedaplatin plus docetaxel versus cisplatin plus docetaxel as first-line chemotherapy for advanced squamous cell carcinoma of the lung — a multicenter, open-label, randomized, Phase III Trial. J Thorac Oncol. 2018;13(11):1743–9. https://doi.org/10.1016/j.jtho.2018.07.006.
41. Ferris RL, Blumenschein G, Fayette J, et al. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. N Engl J Med. 2016;375(19):1856–67. https://doi.org/10.1056/NEJMoa1602252.
42. Motzer RJ, Escudier B, McDermott DF, et al. Nivolumab versus Everolimus in advanced renal-cell carcinoma. N Engl J Med. 2015;373(19):1803–13. https://doi.org/10.1056/NEJMoa1510665.
43. Toxopeus E, van der Schaaf M, van Lanschot J, et al. Outcome of patients treated within and outside a randomized clinical trial on neoadjuvant Chemoradiotherapy plus surgery for esophageal Cancer: extrapolation of a randomized clinical trial (CROSS). Ann Surg Oncol. 2018;25(8):2441–8. https://doi.org/10.1245/s10434-018-6554-y.
44. Mukai H, Shimizu C, Masuda N, et al. Palbociclib in combination with letrozole in patients with estrogen receptor–positive, human epidermal growth factor receptor 2–negative advanced breast cancer: PALOMA-2 subgroup analysis of Japanese patients. Int J Clin Oncol. 2019;24(3):274–87. https://doi.org/10.1007/s10147-018-1353-9.
45. Bellmunt J, de Wit R, Vaughn DJ, et al. Pembrolizumab as second-line therapy for advanced urothelial carcinoma. N Engl J Med. 2017;376(11):1015–26. https://doi.org/10.1056/NEJMoa1613683.
46. Cortes JE, Heidel FH, Hellmann A, et al. Randomized comparison of low dose cytarabine with or without glasdegib in patients with newly diagnosed acute myeloid leukemia or high-risk myelodysplastic syndrome. Leukemia. 2019;33(2):379–89. https://doi.org/10.1038/s41375-018-0312-9.
47. Fradet Y, Bellmunt J, Vaughn DJ, et al. Randomized phase III KEYNOTE-045 trial of pembrolizumab versus paclitaxel, docetaxel, or vinflunine in recurrent advanced urothelial cancer: results of >2 years of follow-up. Ann Oncol. 2019;30(6):970–6. https://doi.org/10.1093/annonc/mdz127.
48. Kotani D. Retrospective cohort study of trifluridine/tipiracil (TAS-102) plus bevacizumab versus trifluridine/tipiracil monotherapy for metastatic colorectal cancer, vol. 9; 2019.
49. Malone S, Roy S, Eapen L, et al. Sequencing of androgen-deprivation therapy with external-beam radiotherapy in localized prostate Cancer: a phase III randomized controlled trial. J Clin Oncol. 2020;38(6):593–601. https://doi.org/10.1200/JCO.19.01904.
50. Kreuzer KA, Furman RR, Stilgenbauer S, et al. The impact of complex karyotype on the overall survival of patients with relapsed chronic lymphocytic leukemia treated with idelalisib plus rituximab. Leukemia. 2020;34(1):296–300. https://doi.org/10.1038/s41375-019-0533-6.
51. Tian L, Jin H, Uno H, et al. On the empirical choice of the time window for restricted mean survival time. Biometrics. 2020;76(4):1157–66. https://doi.org/10.1111/biom.13237.
52. Legrand C. Advanced survival models: CRC Press; 2021.
53. Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the Hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. J Clin Oncol. 2016;34(15):1813–9. https://doi.org/10.1200/JCO.2015.64.2488.
54. Royston P. Combined test versus logrank/Cox test in 50 randomised trials. Trials. 2019;10:1–10.
55. Chen Z, Huang H, Qiu P. Comparison of multiple hazard rate functions. Biometrics. 2016;72(1):39–45.
56. Ditzhaus M, Genuneit J, Janssen A, Pauly M. CASANOVA: Permutation inference in factorial survival designs. Biometrics. 2021;1–13.
57. Chen Z, Huang H, Qiu P. An improved two-stage procedure to compare hazard curves. J Stat Comput Simul. 2017;87(9):1877–86.

Publisher's Note

Article 2

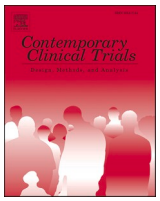
Dormuth, I., Liu, T., Xu, J., Pauly, M., & Ditzhaus, M. (2023). A comparative study to alternatives to the log-rank test. Contemporary Clinical Trials, 128, 107165.

Contemporary Clinical Trials 128 (2023) 107165

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial



A comparative study to alternatives to the log-rank test

Ina Dormuth [a,*], Tiantian Liu [b], Jin Xu [c], Markus Pauly [a,d], Marc Ditzhaus [e]

[a] Department of Statistics, TU Dortmund University, Dortmund, Germany
[b] Technion – Israel Institute of Technology, Haifa, Israel
[c] East China Normal University, Shanghai, China
[d] Research Center Trustworthy Data Science and Security, UA Ruhr, Dortmund, Germany
[e] Department of Mathematics, Otto von Guericke University Magdeburg, Magdeburg, Germany

ARTICLE INFO

Keywords:
Survival analysis
Crossing hazards
Non-proportional hazards
Simulation study
Log-rank

ABSTRACT

Background: Studies to compare the survival of two or more groups using time-to-event data are of high importance in medical research. The gold standard is the log-rank test, which is optimal under proportional hazards. As the latter is no simple regularity assumption, we are interested in evaluating the power of various statistical tests under different settings including proportional and non-proportional hazards with a special emphasis on crossing hazards. This challenge has been going on for many years now and multiple methods have already been investigated in extensive simulation studies. However, in recent years new omnibus tests and methods based on the restricted mean survival time appeared that have been strongly recommended in biometric literature.

Methods: Thus, to give updated recommendations, we perform a vast simulation study to compare tests that showed high power in previous studies with these more recent approaches. We thereby analyze various simulation settings with varying survival and censoring distributions, unequal censoring between groups, small sample sizes and unbalanced group sizes.

Results: Overall, omnibus tests are more robust in terms of power against deviations from the proportional hazards assumption.

Conclusion: We recommend considering the more robust omnibus approaches for group comparison in case of uncertainty about the underlying survival time distributions.

1. Introduction

The distributional comparison of two populations with censored time-to-event data is one of the most common inferential problems in survival analysis. The log-rank test is used as a standard tool in many medical or clinical studies. It is known to be optimal under the assumption of proportional hazards (PH). However, this assumption is often not met in reality due to various forms of derivation such as crossing hazards, or early/late differences in survival curves. Kristiansen [1] conducted a survey revealing that in 70% of studies with crossing survival curves the log-rank test was used even though this leads to loss in power. Furthermore, Trinquart et al. [2] revisited 54 phase III oncology studies from five leading medical journals (New England Journal of Medicine, Lancet, Lancet Oncology, Journal of Clinical Oncology, Journal of the American Medical Association) and found that for almost a fourth of the comparisons the proportional hazard assumption was rejected. Non-proportionality as severe as crossing can appear when the treatment effects change over time. A common example is seen in immunotherapy which bears an early high risk but a long-term benefit [3,4]. Thus, the question on how to deal with non-proportional hazards is of high interest and has been investigated by many authors. For example, Royston and Parmar [5] published a simulation study comparing nine methods implemented in Stata and showed a preference for modified weighted log-rank tests [6–9]. However, they did not include the situation of crossing hazards. Another overview was given by Lin et al. [10], focusing on combined weighted Kaplan–Meier and weighted log-rank tests. They conclude that as long as we do not have prior knowledge the MaxCombo test showed the most robust behavior among the tests under consideration [10]. Perhaps the most extensive study regarding crossing hazards was given in Li et al. [11] who compared 21 tests designed to handle crossing hazards. They stated that the two-stage test by Qiu and Sheng [12] or the test by Kraus

* Corresponding author.
E-mail address: ina.dormuth@tu-dortmund.de (I. Dormuth).

https://doi.org/10.1016/j.cct.2023.107165
Received 9 November 2022; Received in revised form 17 March 2023; Accepted 20 March 2023
Available online 25 March 2023
1551-7144/

[13] are the most suitable among the studied tests. A general overview of existing methods and recommendations regarding trial design was created by Ananthakrishnan et al. [14] without numerical comparison. None of the mentioned reviews considered new results on projection type, sample space partition or area under the survival curve tests [15–18]. Recently, some of the new procedures have shown considerable power advantages in illustrative data analyses [19].

We therefore enrich these investigations by comparing the best performers from the above already existing simulation studies with more recent approaches. Our comprehensive simulation study covers 20 representative scenarios including four null scenarios, four scenarios with PH, four scenarios with non-PH (excluding crossing structures) and eight scenarios with a special emphasis on crossing hazards. Since most procedures exhibit good properties for large samples, our study focuses on small to moderate sample sizes. The present simulation study is intended to demonstrate the advantages of alternative methods. As a side effect, this shall motivate further research on further aspects (e.g. sample size calculation approaches, superiority) of these methods that will make their application easier for clinical practice. In the next section we will review more details on the tests under study and their implementation. Afterwards, the different simulation and parameter settings are presented alongside with the results of the simulation study. The utility of the tests is further evaluated using reconstructed data from a phase III clinical trial with moderate sample size. The findings are then discussed and conclusions are drawn, particularly focusing on the tests' power.

2. Methods

Multiple approaches to test the hypothesis of two equal survival functions have been developed. For ease of presentation, we categorize them in four groups and review the main ideas of the recommended ones in each group. Details on the methods can be found in the cited literature as well as the extended methods section in the Supplement.

2.1. Log-rank test and its weighted variants

The standard to compare two survival functions S_1 and S_2 is the log-rank test (LR) [20]. It belongs to the class of weighted log-rank tests [21] that use the difference between the expected and observed number of events to derive a test statistic. These tests differ in the weight functions that they are employing. For instance, the log-rank test gives the same weight to all event times. Therefore, it is optimal under proportional hazards. The Peto-Peto test (PP) uses the Kaplan–Meier estimator $\widehat{S}(t)$ of the survival function as weight, which leads to a test that is more sensitive to early differences [22]. Various approaches to compute sample sizes for log-rank tests have been introduced, with Schoenfeld's formula being the most popular [23].

In reality, due to the lack of prior information about the survival behavior of comparing populations, any mismatch of weight (or test) selection and difference in the true survival functions will lead to sub-optimal power performance [11].

2.2. Two-stage test

The two-stage (TS) method introduced by Qiu and Sheng [12] provides a solution to the weight selection problem in dealing with possible non-PH situations. The procedure gets its name from the sequential testing approach. More specifically, it conducts the standard log-rank test in the first stage. If the LR test does not reject the null hypothesis, an asymptotically independent test for crossing hazards is carried out. It is shown to be efficient with good adaptation and reliable in power performance under both PH and non-PH situations [12,18]. The approach was extended to the k-sample case employing asymptotically independent tests [24].

2.3. Omnibus tests

Another remedy to avoid potentially sub-optimal power performance is to use an omnibus test that does not have any inclination of the alternative hypothesis.

The mdir test proposed by Brendel et al. [15] and revisited by Ditzhaus and Friedrich [16] uses a quadratic form-type statistic in multiple weighted LR statistics to cover broader alternatives. The test has high power for all alternatives corresponding to the chosen weights and combinations thereof. The test should be used especially when no prior information is available, because with prior knowledge a weighted test with only one suitable weight would have a higher power. A notable feature of the mdir test is that its permuted version allows to handle small sample cases with satisfactory type-I error and power performance [15,16]. The mdir test was extended to handle the one-sided testing problem as well as factorial designs [25,26]. A procedure for sample size calculation does not exist yet.

The class of maximum weighted log-rank tests bears a different approach to combine multiple weighted log-rank tests. Here, multiple test statistics with different weights are considered and the final test statistic is defined as the maximum over all of them. The MaxCombo test (MC) proposed by Lin et al. [10] combines four weighted log-rank tests with Flemming-Harrington type weights targeting difference in survival functions with PH, late difference, middle difference, and early difference, respectively. An iterative sample size calculation approach was provided by Roychoudhury et al. [27]. The test can also be used for one-sided hypotheses. Although the MC test combines different weighted log-rank tests, it is still relevant which weights are initially chosen. The mdir test offers the advantage that the weights are additionally combined linearly which allows covering more alternatives. This could be a possible explanation for the fact that the mdir test can successfully reject the null hypothesis in contrast to the MC test.

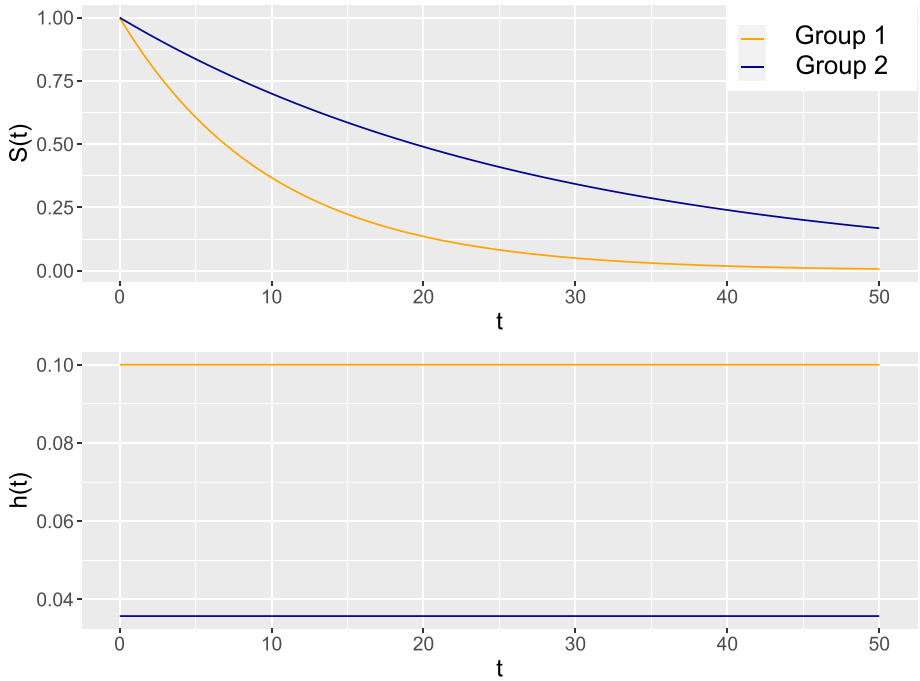
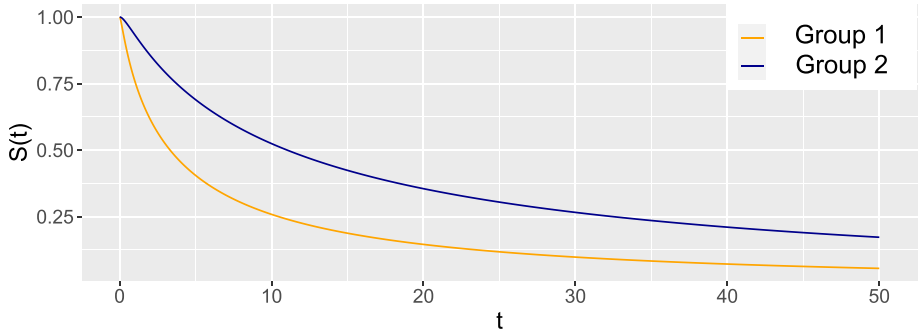
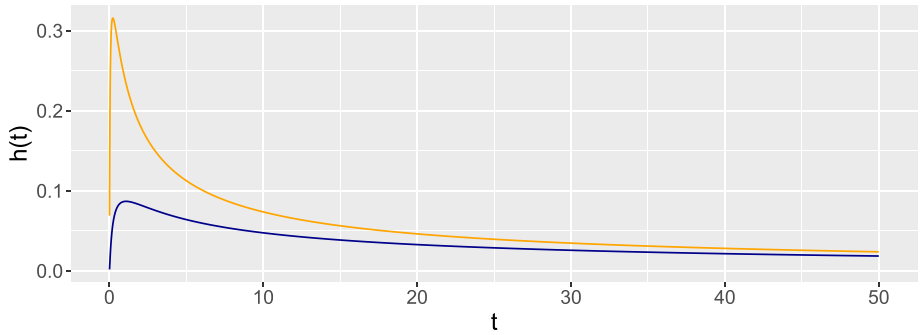
Gorfine et al. [17] introduced K-sample omnibus non-proportional hazards (KONP) tests based on sample space partition that also tackles right censored data. P-values are obtained employing a censoring-friendly permutation procedure. The provided tests are based on two different test statistics, namely the log-likelihood ratio (KONP_llr) and the chi-squared test statistic (KONP_chi). Extensive simulation studies [17] showed that the choice of test statistic does not influence the performance. Hence, we only consider the KONP_chi test in our study.

2.4. Tests based on the area under the survival curve

Tests based on restricted mean survival times (RMST) are often advocated in the context of crossing hazards [28–30,2]. The RMST can be interpreted as the mean of event-free survival time up to τ, where τ is a pre-defined time till which the truncated mean is of interest. In practice, τ is recommended to be 90% of the minimum of the largest censored or uncensored event-time in the two groups [31]. The RMST-based test enjoys the merit of easy interpretation and is distribution free [29]. Moreover, it can be used to test superiority or non-inferiority. For a better type-I error control, Horiguchi and Uno [32] proposed an RMST permutation approach, which is, however, only valid under exchangeability. The latter implies, among others, equal censoring patterns and, thus, do not meet practical reality. As Ditzhaus et al. [33] pointed out, this problem can be repaired by permuting the studentized statistic leading also to valid permutation-based confidence intervals.

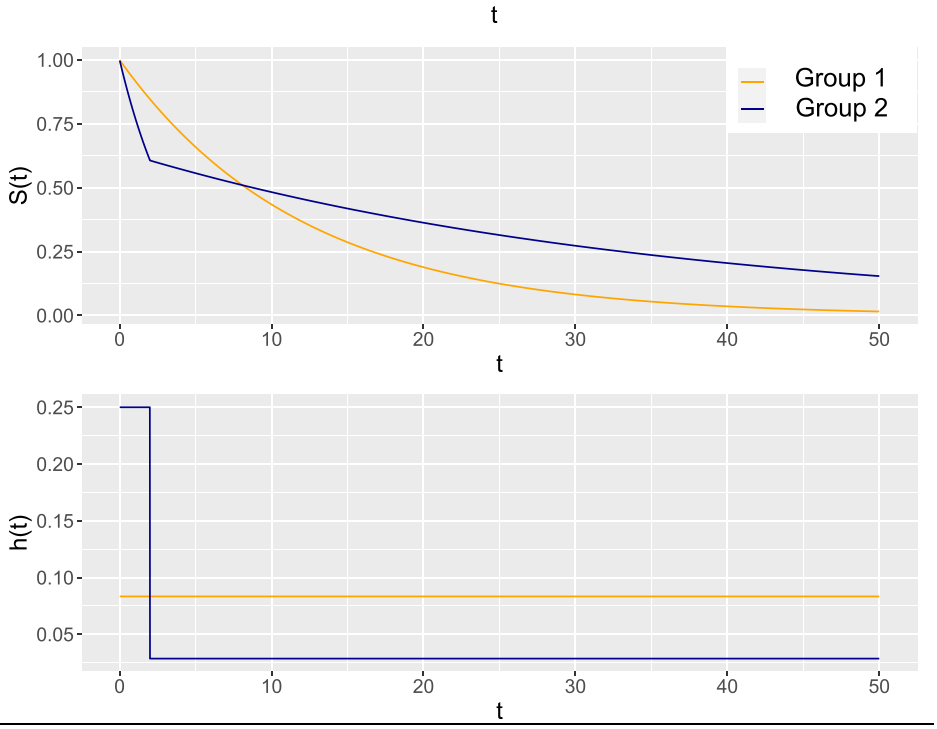
The test proposed by Liu et al. [18] aims to detect crossing survival curves based on the area between the curves (ABC). It can capture the alternative of two crossing survival functions that produce the same RMST. The test obtains its p-value by (group-wise) bootstrapping, which allows different censoring distributions between groups. This test is shown to be more powerful than other distance-based tests such as the modified Kolmogorov–Smirnov test [34] and the generalized Cramér-von Mises test [35]. Since the test statistic quantifies the difference in absolute value, it cannot be used for superiority or non-inferiority testing.

Table 1
Three exemplary scenarios. One scenario with proportional hazards (PH3), one with non-proportional and non-crossing hazards (NPH4) and one with crossing hazards (C3).

Scenario	CDF	Survival and hazard curves
PH3	$F_1(t) = \exp(0.1)$ $F_2(t) = \exp(1/28)$	
NPH4	$F_1(t) = \text{logN}(1.2, 1.7)$ $F_2(t) = \text{logN}(2.4, 1.3)$	 

(continued on next page)

Table 1 (*continued*)

Scenario	CDF	Survival and hazard curves
C3 [11]	$F_1(t) = \exp(1/12)$ $F_2\left(t\right) = \begin{cases} \exp(0.25) & t \leqslant 2 \\ \exp(1/35) & t > 2 \end{cases}$	

3. Simulation study

To evaluate the performance of the presented methods, we employed extensive Monte Carlo simulations for different scenarios and settings. We simulated data for two groups under exponential, Weibull, Gompertz and log-normal distributions. Thus, we follow well-established recommendations on the choice of survival distributions for simulation studies [36].

3.1. Scenarios

We considered four null scenarios, each with a different distribution function. For alternatives, we considered (i) four scenarios with proportional hazards, (ii) four scenarios with non-proportional and non-crossing hazards, and (iii) eight scenarios with crossing hazards. The concrete survival and hazard functions can be found in the Supplement, see Tables S2–S6 therein. For each scenario we vary the group sizes (from 20 to 100), the censoring rates (from 0% to 60%) and the censoring distributions (uniform, exponential) as listed in Table S1 in the Supplements. Thus, we studied 20(scenarios) $\times$ 5(sample sizes) $\times$ 4 (censoring rates) $\times$ 2(censoring distributions) $=$ 800 different settings. We list three exemplary scenarios in Table 1.

For each setting 5,000 replications were performed. Throughout, we set the type I error level to be 0.05. The actual type-I error and power were estimated by the rejection rates. For 2 out of 800 scenarios (all with small sample sizes) the KONP test fails to provide a result. In these cases, the power was set to NA. In these situations, the combination of simulation scenario, very small sample size and censoring distribution lead to an observed censoring rate of 1. Throughout, we used R 4.0.0 [37] for all simulations.

3.2. Implementation details

The LR as well as the PP can be called in R using the function `survdiff` from the `survival`[38] package. The concrete execution depends on the choice of rho (rho $= 0$ for LR and 1 for PP). The R package `TSHRC`[39] contains the implementation of the TS test via the function `twostage`. The mdir is included in the R package `mdir.logrank`[40]. Later, we refer to the test, mdir-x, where 'x' stands for the number of weights considered. For the MC we use the weights proposed by Lin et al. [10] and its implementation in the R package `nphsim`[41]. The KONP is implemented in the R package `KONPsurv`[42]. The packages provide tests based on two different test statistics, namely the log-likelihood ratio and the chi-squared test statistic. Since the authors did not detect any difference in performance we only consider the chi-squared test statistic (KONP). An RMST-based test for two group comparisons is given in the R package `survRM2`[43]. The function used here is `rmst2`, where we need to define a truncation time tau. The published R code for the ABC test is provided on Github (https://github.com/LTTGH/RBT4TCSC). For both tests, τ was set to 90% of the minimum of largest censored or uncensored event-time in two groups [31].

3.3. Results

Here we only report the results under uniform censoring. The results under the exponential censoring are similar and provided in the Supplement.

3.3.1. Type-I-error

Fig. 1 compares the type I errors obtained by eleven tests (significance level $\alpha = 0.05$) under the four considered scenarios. These employ different distributions that are commonly used in survival analysis [36]. One boxplot summarizes 40 data points (see Table S1 in the Supplement for the exact numbers) representing the size of the test for a specific parameter constellation based on 5,000 simulation runs. The red-dotted lines display the binomial confidence intervals for the type-I error. For most of the tests it can be seen that the type-I error is usually within the red-dotted lines, implying a reasonable derivation from the significance

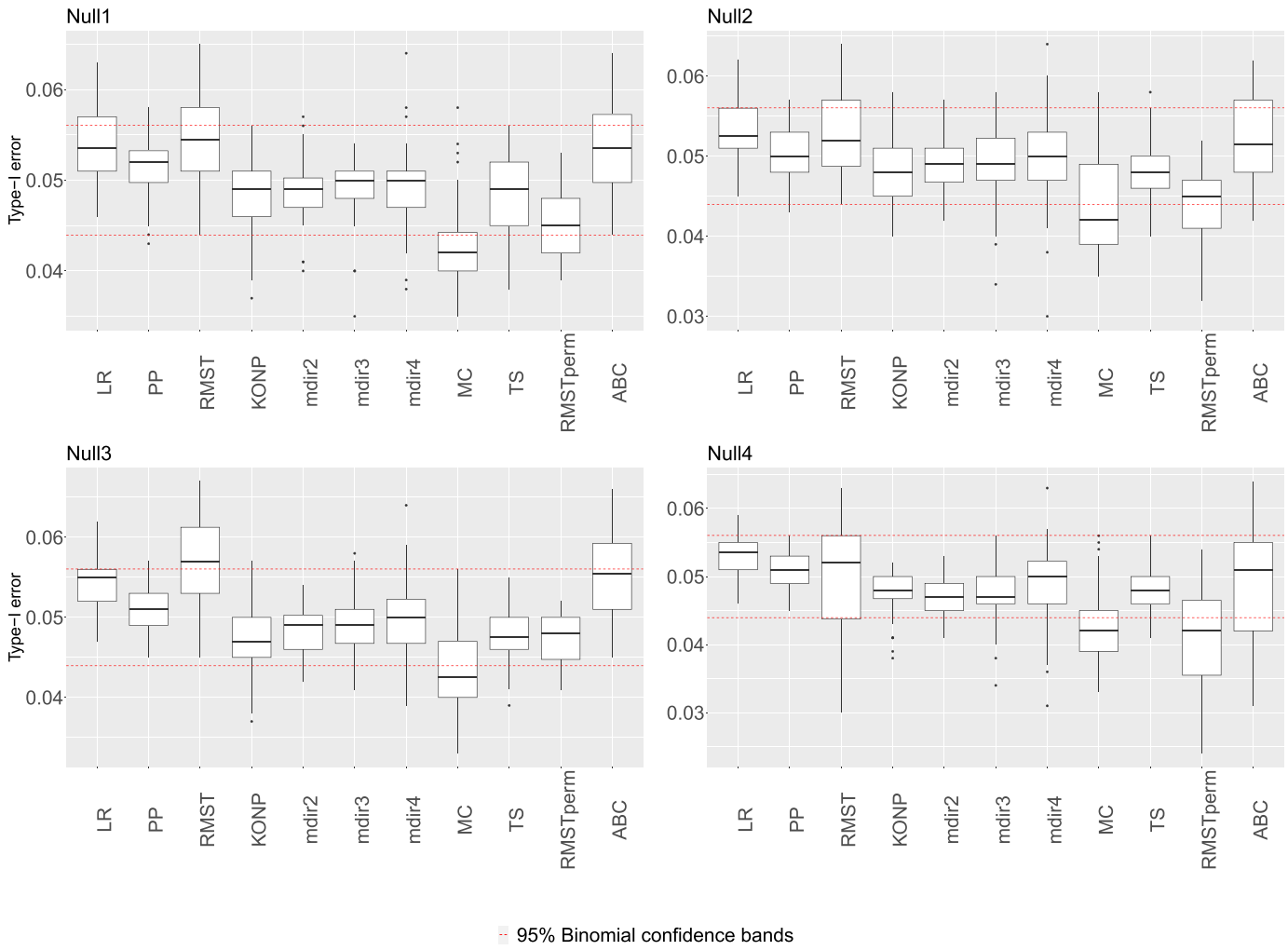


Fig. 1. Type-I errors of the eleven pairwise tests: LR log-rank test, PP peto-peto test, RMST restricted mean survival based test, KONP k-sample omnibus non-proportional hazards test, mdir2 mdir test with two weights, MC maxcombo test, TS two-stage test, RMSTperm RMST-based test with permuted studentized test statistic, ABC area between curves based test. The dotted line represents the corresponding 95% binomial interval [0.044, 0.056].

level of 0.05. However, the MC and the RMSTperm test are relatively conservative in all settings. Moreover, in the Null1 and Null3 scenario (top and bottom left) the RMST- and ABC-tests exhibit a rather liberal behavior. Nevertheless, all tests seem to control the type-I error reasonably well.

3.3.2. Power

We here summarize our findings for the 20 different alternative scenarios. For ease of presentation, we only display three representative scenarios: one for each category of possible relationships between hazards. Each row of Fig. 2 represents a different scenario and the columns display different censoring rates.

PH settings. The proportional settings are all generated using either the exponential or Weibull distribution. As expected, the LR test has the highest power over all settings and parameter combinations. In the first row of Fig. 2 we observe that the other tests also provide good power for the different settings. However, with smaller sample sizes, they exhibit a noticeable power loss compared to the log-rank test. This effect is less pronounced for increased censoring. Similar patterns appear in all other scenarios with proportional hazards, see Figs. S1–S4 in the Supplement.

Non-PH and non-crossing settings. Depending on the setting, the power of the tests varies significantly. In our example setting (see Table 1) depicted in the second row of Fig. 2, the log-rank test has still high power for multiple parameter settings and also high censoring. At the same time, it does not dominate across all parameter combinations anymore. In the NPH4 scenario (second row in Table 1), the Peto-Peto test yields high power as the two survival functions have significant early difference. Analyzing Figs. S5–S8, the omnibus tests as well as the TS test are similarly powerful. Hence, under moderate violation of the PH assumption without crossing in hazard functions, the LR test can still be used while the omnibus is more robust against power loss.

3.3.3. Crossing hazards settings

In the setting with crossing hazards (third row Fig. 2), we observe a drastically lower power for the LR, the PP and the RMST based tests compared to all other methods under consideration. This pattern is present for low censoring as well as higher censoring. Nevertheless, high censoring leads to improvement in terms of power for some of the tests such as PP and RMST. A similar behavior can be observed for the ABC test: While the power for large groups drops slightly, the power derived from smaller data sets is higher than in the low censoring setting. The other tests under consideration lose power with high censoring regardless of the group sizes. Considering the other seven scenarios with crossing hazards available in the Supplement (Figs. S9–S16), we see that in four of the eight scenarios the LR test is among the three tests with the lowest power over different censoring and sample size settings. The PP test performs comparably in seven out of eight scenarios and the RMST based test results in low power for all scenarios. However, the RMST based test often seems to have higher power for crossing scenarios with higher censoring. This behavior can be attributed to the fact that higher

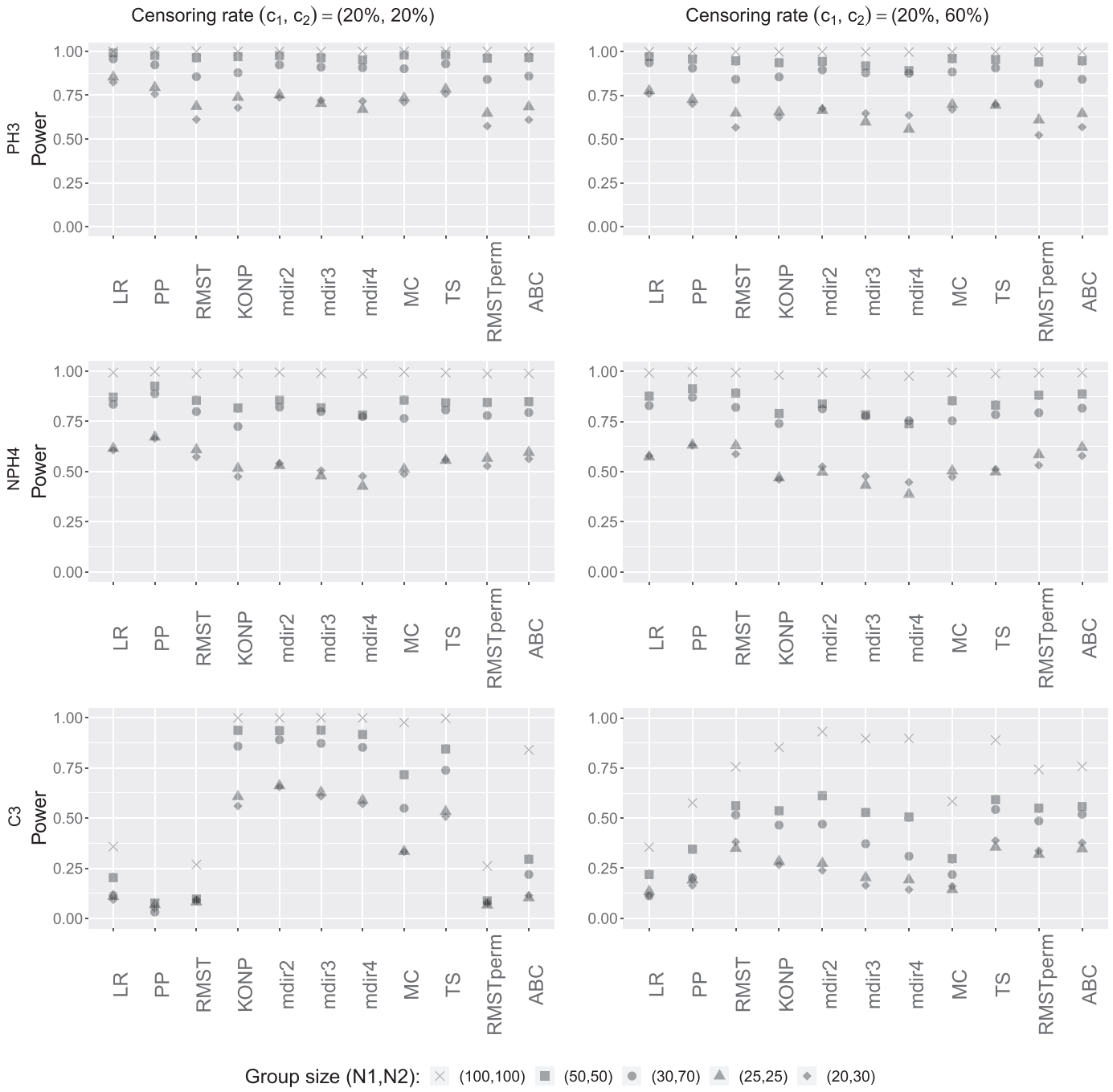


Fig. 2. Power of the eleven pairwise tests (significance level $\alpha = 0.05$) in representative scenarios under uniform censoring. PH3 proportional hazards scenario three, NPH4 non-proportional hazards scenario four, C3 crossing scenario three. LR log-rank test, PP peto-peto test, RMST restricted mean survival based test, KONP k-sample omnibus non-proportional hazards test, mdir2 mdir test with two weights, MC maxcombo test, TS two-stage test, RMSTperm RMST-based test with permuted studentized test statistic, ABC area between curves based test.

censoring rates often mask the crossing pattern of hazard rates. Consistently high power across the various scenarios is evident for the KONP test, the MC and the mdir tests. The TS test appears to be powerful for some scenarios but is less robust in terms of power than the omnibus tests. Finally, the ABC test has decent power for most of the scenarios but is no competitor for the omnibus tests and the TS test. Regarding the choice of weights for the mdir test, it can be seen that except for one crossing scenario the mdir2 test including the log-rank and crossing weight is as powerful as or more powerful than the mdir3 or mdir4. Hence, we would recommend using the mdir test with these two weights only.

In summary, it has to be further investigated why the MC test is so conservative (Fig. 1). A reasonable assumption is the small sample size in the groups. Taking the conservative behavior into account, we can assume that larger sample sizes might also lead to higher power in the alternative scenarios. The results of Lin et al. [10] support this assumption. They considered much larger sample sizes starting from 300 and did not observe a similar behavior. The results show that it is adequate to include two different weighted LR statistics in the mdir test. The RMST-based test cannot be recommended in situations with crossing hazards. Globally, we do recommend the use of omnibus tests such as MC, KONP or mdir when no prior knowledge is available. They show robust power behavior for proportional, non-proportional and crossing scenarios.

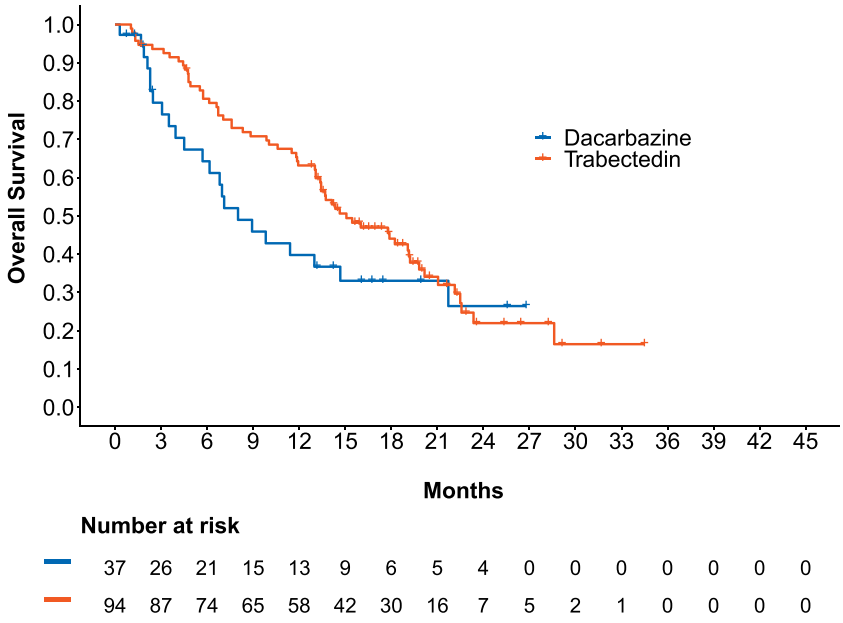


Fig. 3. Reconstructed Kaplan–Meier curves of overall survival in elderly patients with sarcoma from Jones et al. [44].

Table 2
P-values of the eleven pairwise tests (nominal level $\alpha = 0.05$) applied to the reconstructed data from Jones et al. [44]. LR log-rank test, PP peto-peto test, RMST restricted mean survival based test, KONP k-sample omnibus non-proportional hazards test, mdir2 mdir test with two weights, MC maxcombo test, TS two-stage test, RMSTperm permutation-based studentized RMST test, ABC area between curves based test.

LR	PP	RMST	KONP	mdir2	mdir3	mdir4	MC	TS	RMSTperm	ABC
0.17	**0.03**	**0.02**	**0.04**	**0.01**	**0.03**	0.05	0.06	**0.05**	0.24	**0.03**

4. Real data example

We evaluate the performance of the considered tests on real data from a clinical trial for 131 elderly patients with advanced liposarcoma or leiomyosarcoma, where the overall survival functions under two treatments (Dacarbazine and Trabectedin) show a clear cross at a late time period [44].

As in Dormuth et al. [19], we reconstructed the patient-level data using the state-of-the-art reconstruction algorithm by Guyot et al. [45]. The Kaplan–Meier plot using the reconstructed data is shown in Fig. 3. It is apparent that the Kaplan–Meier curves depart from each other early but converge at later times. The quality of the reconstructed data is examined to be sufficiently satisfactory (see TableS27 in the Supplement).

The p-values for eight considered tests are listed in Table 2. All the considered methods except LR and MC succeed to reject the null hypothesis at significance level 0.05 with mdir2 giving the strongest note. Although the MC test combines different weighted log-rank tests, it is still relevant to consider which weights are initially chosen. This also holds for the mdir test but which has the additionall advantage that all linear combinations of the chosen weights are implicitly taken into account. In this way, the mdir procedure is more robust when the optimal weight was not chosen. This could be a possible explanation for the rejection of the mdir test in contrast to the MC test.

5. Discussion

We investigated the type-I error and power of the gold-standard log-rank test and various recent tests that are recommended as alternatives in case of potential proportional hazards violation. To this end we conducted an extensive simulation study including 20 representative scenarios. In the null settings most tests respected the type-I error. Only the MC test appears to be more conservative for small sample sizes. Regarding power, the simulation study indicates that the log-rank test does not experience a drastic loss compared to the other methods in case of non-proportional and non-crossing scenarios. However, if crossings are present, the power difference among the tests is much more pronounced with good performances for all omnibus tests. In fact, especially the KONP and mdir omnibus tests show a more stable power over all scenarios. Regarding the mdir test, no advantage in inclusion of more weights for the mdir test was found in our settings. We therefore recommend using the default setting of the test if no prior knowledge is available.

Limitations of our study. We only investigated the methods' performance for small to moderate sample sizes. In this paper we only discussed the two-sided testing problem since not all procedures have existing versions for testing superiority or non-inferiority. Here, we see further research potential as this kind of one-sided testing only exist for the weighted LR, the RMST-based and the mdir tests [14,25]. A similar statement holds for *k*-sample or more general ANOVA settings [17,26]. We therefore recommend to also investigate and compare their mathematical properties, e.g. relative efficiencies.

In order for the recommended procedures to find their way into biostatistical practice, methods for accurate sample size calculation are needed. Otherwise, inclusion in study protocols as well as the ability to draw sufficiently powered conclusions can not be supported. Furthermore, statistical significance alone does not always corroborate clinical relevance. More interpretable statistical measures such as compatible confidence interval for meaningful estimands/parameters [46,32] are necessary in addition to the tests' decisions. Furthermore, we recommend to also investigate and compare their mathematical properties, e.g. in terms of relative efficiencies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within thehttps://uaruhr.de. The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359. Markus Pauly was supported by German Research Foundation Grant No PA 2409/5–1. Marc Ditzhaus was supported by German Research Foundation Grant No DI 2906/1–2. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.cct.2023.107165.

References

[1] I. Kristiansen, PRM39 Survival curve convergences and crossing: a threat to validity of meta-analysis, Value Health 15 (7) (2012) A652.
[2] L. Trinquart, J. Jacot, S.C. Conner, R. Porcher, Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials, J. Clin. Oncol. 34 (15) (2016) 1813–1819.
[3] R. Mick, T.T. Chen, Statistical challenges in the design of late-stage cancer immunotherapy studies, Cancer Immunol. Res. 3 (12) (2015) 1292–1298.
[4] B.M. Alexander, J.D. Schoenfeld, L. Trippa, Hazards of hazard ratios-deviations from model assumptions in immunotherapy, N. Engl. J. Med. 378 (12) (2018) 1158–1159.
[5] P. Royston, M.K.B. Parmar, 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) (2020) 1–17.
[6] T.R. Fleming, D.P. Harrington, M. O'sullivan, Supremum versions of the log-rank and generalized Wilcoxon statistics, J. Am. Stat. Assoc. 82 (397) (1987) 312–320.
[7] P. Royston, M.K. Parmar, An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect, Trials 15 (1) (2014) 1–10.
[8] J.W. Lee, Some versatile tests based on the simultaneous use of weighted log-rank statistics, Biometrics 721–725 (1996).
[9] T.G. Karrison, Versatile tests for comparing survival curves based on weighted log-rank statistics, Stata J. 16 (3) (2016) 678–690.
[10] R.S. Lin, J. Lin, S. Roychoudhury, et al., Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis, Stat. Biopharm. Res. 12 (2) (2020) 187–198.
[11] H. Li, D. Han, Y. Hou, H. Chen, Z. Chen, Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods, PLoS One 10 (1) (2015).
[12] P. Qiu, J.A. Sheng, Two-stage Procedure for Comparing Hazard Rate Functions, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 70 (1) (2008) 191–208.
[13] D. Kraus, Adaptive Neyman's Smooth Tests of Homogeneity of Two Samples of Survival Data, J. Stat. Plann. Inference 139 (10) (2009) 3559–3569.
[14] R. Ananthakrishnan, S. Green, A. Previtali, R. Liu, D. Li, M. LaValley, Critical Review of Oncology Clinical Trial Design under Non-Proportional Hazards, Crit. Rev. Oncol./Hematol. 162 (2021), 103350.
[15] M. Brendel, A. Janssen, C.D. Mayer, M. Pauly, Weighted Logrank Permutation Tests for Randomly Right Censored Life Science Data, Scand. J. Stat. 41 (3) (2014) 742–761.
[16] M. Ditzhaus, S. Friedrich, More powerful logrank permutation tests for two-sample survival data, J. Stat. Comput. Simul. 90 (12) (2020) 2209–2227.
[17] M. Gorfine, M. Schlesinger, L. Hsu, K-sample omnibus non-proportional hazards tests based on right-censored data, Stat. Methods Med. Res. 29 (10) (2020) 2830–2850.
[18] T. Liu, M. Ditzhaus, Xu J.A Resampling-based Test for Two Crossing Survival Curves, Pharm. Stat. (2020).
[19] I. Dormuth, T. Liu, J. Xu, M. Yu, M. Pauly, Which Test for Crossing Survival Curves? A User's Guide, BMC Med. Res. Methodol. (2022). Accepted.
[20] R. Singh, K. Mukhopadhyay, Survival analysis in clinical trials: Basics and must know areas, Perspect. Clin. Res. 2 (4) (2011) 145.
[21] T.R. Fleming, D.P. Harrington, Counting Processes and Survival Analysis, John Wiley & Sons, 2011.
[22] C. Legrand, Advanced Survival Models, CRC Press, 2021.
[23] D.A. Schoenfeld, Sample-size formula for the proportional-hazards regression model, Biometrics (1983) 499–503.
[24] Z. Chen, H. Huang, P. Qiu, Comparison of Multiple Hazard Rate Functions, Biometrics 72 (1) (2016) 39–45.
[25] M. Ditzhaus, M. Pauly, Wild bootstrap logrank tests with broader power functions for testing superiority, Comput. Stat. Data Anal. 136 (2019) 1–11.
[26] M. Ditzhaus, J. Genuneit, A. Janssen, M. Pauly, CASANOVA: Permutation inference in factorial survival designs, Biometrics (2021).
[27] S. Roychoudhury, K.M. Anderson, J. Ye, P. Mukhopadhyay, Robust design and analysis of clinical trials with nonproportional hazards: A straw man guidance from a cross-pharma working group, Stat. Biopharm. Res. (2021) 1–15.
[28] P. Royston, M.K. Parmar, The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt, Stat. Med. 30 (19) (2011) 2409–2421.
[29] P. Royston, M.K. Parmar, Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials with a Time-to-Event Outcome, BMC Med. Res. Methodol. 13 (1) (2013) 152.
[30] D.H. Kim, H. Uno, L.J. Wei, Restricted mean survival time as a measure to interpret clinical trial results, JAMA Cardiol. 2 (11) (2017) 1179–1180.
[31] L. Tian, H. Jin, H. Uno, et al., On the empirical choice of the time window for restricted mean survival time, Biometrics 76 (4) (2020) 1157–1166.
[32] M. Horiguchi, H. Uno, On permutation tests for comparing restricted mean survival time with small sample from randomized trials, Stat. Med. 39 (20) (2020) 2655–2670.
[33] M. Ditzhaus, M. Yu, J. Xu, Studentized permutation method for comparing restricted mean survival times with small sample from randomized trials. arXiv preprint arXiv:2102.10186, 2021.
[34] T.R. Fleming, J.R. O'Fallon, P.C. O'Brien, D.P. Harrington, Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data, Biometrics 607–625 (1980).
[35] M. Schumacher, Two-Sample Tests of Cramér–von Mises-and Kolmogorov–Smirnov-Type for Randomly Censored Data, Int. Stat. Rev./Revue Internationale de Statistique (1984) 263–281.
[36] R. Bender, T. Augustin, M. Blettner, Generating Survival Times to Simulate Cox Proportional Hazards Models, Stat. Med. 24 (11) (2005) 1713–1723.
[37] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020.
[38] T.M. Therneau, A Package for Survival Analysis in R, 2021. R package version 3.2-10.
[39] J. Sheng, P. Qiu, C.J. Geyer, TSHRC: Two Stage Hazard Rate Comparison, 2019. R package version 0.1-6.
[40] M. Ditzhaus, S. Friedrich, mdir.logrank: Multiple-Direction Logrank Test, 2018. R package version 0.0.4.
[41] Y. Wang, H. Wu, K. Anderson, S. Roychoudhury, T. Hu, H. Liu, nphsim: Non proportional hazards sample size and simulation, 2017. R package version 0.1.1.9000.
[42] M. Schlesinger, M. Gorfine, KONPsurv: KONP Tests: Powerful K-Sample Tests for Right-Censored Data, 2020. R package version 1.0.3.
[43] H. Uno, L. Tian, A. Cronin, C. Battioui, M. Horiguchi, survRM2: Comparing Restricted Mean Survival Time, 2017. R package version 1.0-2.
[44] R. Jones, G. Demetri, S. Schuetze, et al., Efficacy and Tolerability of Trabectedin in Elderly Patients with Sarcoma: Subgroup Analysis from a Phase II, Randomized Controlled Study of Trabectedin or Dacarbazine in Patients with Advanced Liposarcoma or Leiomyosarcoma, Ann. Oncol. 29 (9) (2018) 1995–2002.
[45] P. Guyot, A. Ades, M.J. Ouwens, N.J. Welton, Enhanced Secondary Analysis of Survival Data: Reconstructing the Data from Published Kaplan-Meier Survival Curves, BMC Med. Res. Methodol. 12 (1) (2012) 9.
[46] D. Dobler, M. Pauly, Bootstrap-and permutation-based inference for the Mann-Whitney effect for right-censored and tied data, Test 27 (3) (2018) 639–658.

Article 3

Dormuth, I., Pauly, M., Rauch, G., & Herrmann, C. (2024). Sample Size Calculation Under Nonproportional Hazards Using Average Hazard Ratios. Biometrical Journal, 66(6)

Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

Sample Size Calculation Under Nonproportional Hazards Using Average Hazard Ratios

Ina Dormuth[1] | Markus Pauly[1,2] | Geraldine Rauch[3,4] | Carolin Herrmann[3]

[1]Department of Statistics, TU Dortmund University, Dortmund, Germany | [2]Research Center Trustworthy Data Science and Security, UA Ruhr, Dortmund, Germany | [3]Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, Berlin, Germany | [4]Technical University Berlin, Berlin, Germany

Correspondence: Ina Dormuth (ina.dormuth@tu-dortmund.de)

Received: 5 October 2023 | **Revised:** 18 March 2024 | **Accepted:** 28 April 2024

Funding: This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the University Alliance Ruhr (https://uaruhr.de). Moreover, the work of Markus Pauly was supported by an Individual DFG Research Project.

Keywords: effect measure | hazard ratio | log-rank test | sample size | simulation study | survival analysis | time-to-event data

ABSTRACT

Many clinical trials assess time-to-event endpoints. To describe the difference between groups in terms of time to event, we often employ hazard ratios. However, the hazard ratio is only informative in the case of proportional hazards (PHs) over time. There exist many other effect measures that do not require PHs. One of them is the average hazard ratio (AHR). Its core idea is to utilize a time-dependent weighting function that accounts for time variation. Though propagated in methodological research papers, the AHR is rarely used in practice. To facilitate its application, we unfold approaches for sample size calculation of an AHR test. We assess the reliability of the sample size calculation by extensive simulation studies covering various survival and censoring distributions with proportional as well as nonproportional hazards (N-PHs). The findings suggest that a simulation-based sample size calculation approach can be useful for designing clinical trials with N-PHs. Using the AHR can result in increased statistical power to detect differences between groups with more efficient sample sizes.

1 | Background

Time-to-event endpoints are common outcomes of interest, for example, in oncological trials. When two or more groups are compared, proportional hazards (PH) are often assumed. Under this assumption, hazard ratios are meaningful effect measures and the log-rank test is optimal in terms of power. Even when the assumption of PH is violated, the log-rank test and corresponding sample size calculation approaches (e.g., Schoenfeld 1981; Freedman 1982) are often used, even though the Schoenfeld formula has only been derived under PH settings. The violation of the PH assumption can influence the power of the test severely for detecting a difference (Lin et al. 2020). One reason for the popularity of the log-rank test under nonproportional hazards (N-PHs), however, is its easy application due to its implementation in various software (R, Stata, etc.). There are multiple reasons for non-PH. In terms of treatment analysis, they can occur when the treatment effects depend on time. A classic example of such a pattern is immunotherapy, which is known for high risk in the early stages but a long-term benefit (Alexander, Schoenfeld, and Trippa 2018; Ananthakrishnan et al. 2021; Mick and Chen 2015). For example, Trinquart et al. (2016) analyzed 54 phase III oncology studies published in five journals. They found that in almost 25% of the studies the PH assumption did not hold.

There are multiple other effect measures that stay interpretable under deviations of the assumption of PH such as the restricted mean survival, the Mann–Whitney effect, the relative time or the

average hazard ratio (AHR) (Royston and Parmar 2011; Mann and Whitney 1947; Phadnis and Mayo 2021; Kalbfleisch 1981). The restricted mean survival time (RMST) describes the average event-free survival up to a specific time point (Royston and Parmar 2013). More precisely, the difference between two groups is evaluated by the difference of the areas under the respective Kaplan–Meier curves up to that specific time point. The Mann–Whitney effect for right-censored data (e.g., Koziol and Jia 2009) describes the probability that a randomly drawn subject of one group survives longer than a randomly drawn subject of another group. Phadnis and Mayo (2021) suggested the concept of relative time in the N-PHs setting. The relative time describes the ratio of times of the intervention and control group where, in each group, the same specific percentage of individuals has experienced an event. Another effect measure is the AHR (Kalbfleisch 1981), which can be seen as an extension of the Mann–Whitney effect. Its core is to extend the idea of hazard ratios by using a time-dependent flexible weighting function. Unlike the nonweighted hazard ratios, the AHRs stay interpretable under non-PH since they incorporate the influence of time through their weighting function. Uno and Horiguchi (2023) and Phadnis and Mayo (2021) provide a detailed overview of recent approaches regarding sample size calculation under non-PH.

For practical application with regard to clinical trials, it is essential to have reasonable statistical methods and software to plan and evaluate the respective trial. This especially includes a corresponding statistical test and sample size calculation approach. Currently, the sample size calculation approach by Lakatos (1986) is mostly used under the assumption of non-PH. The approach, however, does require a good knowledge of the survival curves under the alternative (Phadnis and Mayo 2021).

Regarding different effect measures, there have been some developments focusing on sample size calculation. Tests based on the restricted mean survival time (Royston and Parmar 2016; Tian et al. 2018; Uno et al. 2014) have been provided with a sample size calculation approach (Royston 2018). A function to calculate sample sizes for the RMST for superiority and noninferiority testing problems is provided in the R package `SSRMST` (Horiguchi and Uno 2017; Uno et al. 2015). Also, the relative time approach is accompanied by a sample size calculation approach (Phadnis and Mayo 2021). In this approach, Weibull distributed survival times are used and the relative time refers to the ratio of times for which a prespecified percentage of samples in each of the groups has had the event of interest. Macros for SAS exist according to the authors (Phadnis and Mayo 2021) but the approach cannot be applied in cases of crossing survival functions. We would also like to point out that the R package `npsurvSS` (Yung and Liu 2019) provides a variety of codes for sample size and power calculations in regard to t-year survival, pth percentile survival (Yung and Liu 2020), restricted mean survival time and the weighted log-rank test. Similarly, there exists the Wilcoxon–Mann–Whitney test for the Mann–Whitney effect as well as extended test versions for time-to-event data (Brückner and Brannath 2017; Dobler and Pauly 2018, 2020; Rauch et al. 2018), which can also be used to test the AHR in special cases. In general, Kalbfleisch (1981), Schemper, Wakounig, and Heinze (2009), and Brückner and Brannath (2017) suggested test statistics for the AHR. To our knowledge, other than with regard to the log-rank test (Cortés Martínez et al. 2021; Hasegawa 2014), no generally applicable sample size calculation approach was suggested for the AHR so far. This could be the main reason why the AHR is not frequently applied (Rauch et al. 2018). A PubMed study in March for AHRs resulted in only four clinical trials reporting the effect measure. However, already Schemper (1992) reviewed alternatives to the Cox PH model and emphasized the advantages of the AHR as an effect measure. Its application by a weighted Cox regression is analyzed in detail in Schemper, Wakounig, and Heinze (2009). For its application, the R package `coxphw` (Dunkler et al. 2018) may be used, which consists of the implementation of the AHR and provides a sample size calculation approach based on Schoenfeld's formula (Schoenfeld 1981) for PH. Moreover, Rauch et al. (2018) compare the two most common estimators for AHRs with the log-rank test in a Monte Carlo simulation with various settings. It was shown that both approaches outperform the log-rank test in terms of power for N-PH scenarios and that the classical hazard ratio is not always well defined.

Our aim is to provide guidance for sample size calculation in a clinical trial that employs the AHR as the primary endpoint. We assess the effectiveness of these methods by comparing their required sample sizes and resultant statistical power under variations from the original assumptions. Therefore, we first describe the setup, the utilized standardized test statistic as well as a simulation-based sample size calculation and an asymptotic sample size calculation approach. Afterwards, we provide extensive simulation studies where we evaluate different survival time and censoring distributions as well as group allocation ratios. Finally, we illustrate the sample size calculation approaches based on a data example and discuss our findings.

2 | Methods

We consider the standard two-sample survival setup with survival times T_ℓ and censoring times C_ℓ $(\ell = 1,\dots,n)$. The survival function is given by $S_i(t) = \mathbb{P}(T_\ell > t | Z_\ell = i)$, and the survival function of the respective censoring times in group i is given by $K_i(t)$ at time t, $i \in \{0,1\}$. The observed time is denoted by $Y_\ell = \min(T_\ell, C_\ell)$ with the corresponding censoring indicator $\delta_\ell = \mathbb{1}_{T_\ell \leq C_\ell}$ for individual ℓ. Let $Z_\ell \in \{0,1\}$ indicate the group association of individual ℓ and $n = n_0 + n_1$ be the total sample size. We are interested in comparing the two groups with respect to their survival times. In the following, we assume $n_0 = n_1$ with $n_1/n \to \eta \in (0,1)$. Next, we briefly present the AHR and a resulting test statistic. Based on these quantities, we then derive a simulation-based and an asymptotic sample size calculation approach. In what follows, we do not require the PH assumption.

2.1 | The AHR and a Corresponding Test Statistic

The AHR extends the idea of the well-known hazard ratio by using a time-dependent flexible weighting function. To introduce it, recall that the hazard rate in each group i is defined as

$$\lambda_i(s) = \lim_{h \to 0} \frac{P(s \leq T_\ell < s + h | T_\ell \geq s, Z_\ell = i)}{h},$$

and describes the instantaneous risk of experiencing an event at a time s knowing individual ℓ did not experience an event

before. With the hazard rates of the two groups, we then define the weighted AHR as Kalbfleisch (1981)

$$\theta_i(G) = -\int_0^\infty \frac{\lambda_i}{\lambda_0 + \lambda_1}(t)\, G(dt), \tag{1}$$

where G is a decreasing weight function and the integral is defined as generalized Lebesgue–Stieltjes integral for signed measures. Here, we simply have $-\int \dots G(dt) = \int \dots H(dt)$ for $H = -G$. In order to obtain time-dependent weights, we consider a class of weight functions depending on the shape of the survival time distributions S_0 and S_1 in the two groups:

$$G_{\overline{\alpha}}(t) = S_0^{\overline{\alpha}}(t) S_1^{\overline{\alpha}}(t),\ t > 0,\ \overline{\alpha} > 0.$$

The function $G_{\overline{\alpha}}$ is decreasing in t and fulfills $\int_0^\infty dG_{\overline{\alpha}}(t) = -1$. As $\theta_0(G_{\overline{\alpha}}) + \theta_1(G_{\overline{\alpha}}) = 1$, it suffices to consider one of the two hazard ratios (θ_1 or θ_0) (Brückner and Brannath 2017). In the following, we set $\overline{\alpha} = 1$ since then Equation (1) equals the well-known Mann–Whitney effect $P(T_1 > T_0)$. Note that we can only estimate the survival functions S_i consistently up to a finite time point. Therefore, the weight function G is truncated at a constant time point

$$L < \sup\{u : S_i(u) C_i(u) > 0\},$$

where the right-hand side describes the supremum of the support of survival function S_i and censoring function C_i in group i. With consistent estimators up to L, the truncated weight function is given by

$$G_L(t) = \frac{G(t)}{1 - G(L)} \mathbb{1}_{t \le L}.$$

The denominator $1 - G(L)$ is a normalization factor, which ensures $\int_0^\infty dG_L(t) = -1$. The truncation time point L is chosen in advance. The truncated AHR is then defined as

$$\theta_1(G_L) = \frac{x_1}{1 - G(L)}, \tag{2}$$

where

$$x_1 = \int_0^L \lambda_1(t) G(dt) = -\int_0^L S_0(t) S_1(dt).$$

Estimators $\hat{\theta}_i(t), i \in \{0, 1\}$, for the AHR at time $t \le L$, are given by

$$\hat{\theta}_1(t) = \frac{-\int_0^L \hat{S}_0(t, s) \hat{S}_1(t, ds)}{1 - \hat{G}(t, L)}. \tag{3}$$

The estimators $\hat{S}_i$ of the survival functions are chosen appropriately, for example, by Kaplan–Meier estimators, and the estimators $\hat{G}$ for the weight function are given by

$$\hat{G}(t, L) = \hat{S}_0(t, L) \hat{S}_1(t, L).$$

Note that the notation $\hat{S}_i(t, L)$ indicates that the observed time t that is supposed to be smaller or equal to the constant L.

In this setting, we are interested in comparing two groups with respect to their survival, that is, we consider the testing problem

$$H_0 : \theta_1 = 0.5 \text{ versus } H_1 : \theta_1 \neq 0.5.$$

To explain the form of the test statistic, We note that

$$\sqrt{n}(\hat{\theta}_1(t) - \theta_1(t)) \sim N(0, \hat{v}_\theta). \tag{4}$$

holds as, for example, shown by Brückner and Brannath (2017). It thus seems natural to consider the test statistic

$$Z = \frac{\sqrt{n}(\hat{\theta}_1 - \theta_1)}{\sqrt{\hat{v}_\theta}} \sim \mathcal{N}(0, 1). \tag{5}$$

Note that the dependence of t was omitted for readability aspects. A formula for the variance $\hat{v}_\theta$ is given in Equation (1) in the Supporting Information. Dobler and Pauly (2018) provide a representation of the variance formula based on counting processes.

2.2 | Asymptotic Sample Size Calculation

One strategy for calculating the sample size is to take advantage of the asymptotic normality of the test statistic from Equation (5). Let us, therefore, consider a test problem of the null hypothesis H_0 against the alternative hypothesis H_1 with a (an approximatively) standard normally distributed test statistic under H_0. In general, the goal is to determine a sample size such that for an alternative hypothesis H_g at a given effect g a power of $1 - \beta$ is attained while the type I error rate is controlled by α. In case the difference between the two groups is described by a normally distributed test statistic with expected value κ and variance of 1, it follows from the definition of type I and type II error rates that

$$\kappa = z_{1-\alpha/2} + z_{1-\beta}, \tag{6}$$

for a two-sided level α test with z_α denoting the α-quantile of the standard normal distribution (cf. Kieser 2020).

Assume we can infer a minimal clinically relevant effect size $\tilde{\theta}$ together with a variance estimate $\tilde{v}_\theta$ based on medical knowledge and/or literature. Note that in certain cases, especially the variance estimates need to be inferred from simulation studies. Then, $\kappa = \sqrt{n} \cdot (\tilde{\theta} - 0.5)/\sqrt{\tilde{v}_\theta}$. Consequently, the required total number of patients is set to the smallest integer fulfilling

$$n \ge \left(\frac{(z_{1-\alpha/2} + z_{1-\beta}) \cdot \sqrt{\tilde{v}_\theta}}{\tilde{\theta} - 0.5} \right)^2. \tag{7}$$

We obtain $\tilde{\theta}$ and $\tilde{v}_\theta$ by simulating 10,000 datasets using the assumed survival distribution. Then, we estimate the parameters for each dataset by plugging in their mean in the sample size formula. This sample size approach will be referred to as *AHRasymp* as it depends on the asymptotic normality of the standardized estimator stated in Equation (5).

2.3 | Simulation-Based Sample Size Calculation

While analytical approaches for calculating the sample size do not always lead to a solvable problem, a simulation-based approach can always be applied. This is especially advantageous if a complex study design or data distribution is present. Rauch, Schüler, and Kieser (2017) presented such an approach for a specific Weibull setting. Our simulation-based sample size calculation, called *AHRsim*, generalizes their work with respect to arbitrary data distributions and is computationally more efficient and robust. Schematically, AHRsim can be described as follows:

1. Decide on a true event data distribution (e.g., $F_1(t) = \text{Weibull}(0.6, 8)$ and $F_2(t) = \text{Weibull}(0.6, 4)$ or $F_1(t) = \text{Lognormal}(1.2, 1.7)$ and $F_2(t) = \text{Lognormal}(2.4, 1.3)$, accrual duration, minimal follow-up time and censoring distribution and generate the underlying data.
2. Specify the number of simulation runs n_{sim}. For every simulated dataset, we provide the calculation of the AHR and corresponding test statistic.
3. The power is then given by n_{sig}/n_{sim}, where n_{sig} is the number of corresponding p values falling below a prespecified significance-level threshold α.
4. This leads us to the actual sample size calculation procedure: For a given event-time distribution with specified accrual time and minimal follow-up time, starting values for the sample sizes in the intervention and control group and censoring time distribution, calculate the corresponding power as stated above. We start with sample size values that do not exceed the desired power value $1 - \beta$ (e.g., 80%). We enlarge the sample sizes per group in a stepwise manner, and we stop the sample size increase when a power of at least $1 - \beta$ is reached for the first time.

To reduce the run time of the AHRsim approach, a smart choice of the starting value for the sample size in step four is crucial. Since we calculate the sample size based on Schoenfeld (SF) and the approximation approach anyway, we decided to use this information. After observing noticeable differences between these sample sizes, we implemented a more robust approach for the starting value. For details, we refer to the Supporting Information.

3 | Simulation Study

For the evaluation of the two suggested sample size approaches and a comparison with Schoenfeld's (1983) formula as well as a simulation-based log-rank approach, we performed extensive simulation studies for different scenarios. In the subsequent sections, our focus is exclusively directed toward event-time-driven analyses. As such, our simulations involve the generation of event times and censoring times, excluding the consideration of enrollment. Furthermore, we evaluate the power of the AHR-based and log-rank tests, respectively.

The simulation study is structured into two different parts with two distinct goals:

(G1) Calculation of simulation-based sample sizes as well as asymptotically derived sample sizes and their comparison, also with respect to sample sizes obtained from Schoenfeld's formula and the simulation-based log-rank approach. To this end, we simulate different survival time distributions and censoring time distributions.

(G2) Evaluation of misspecified survival time distributions for all four sample size approaches. Here, data of the determined sample sizes with other survival and censoring time distributions are generated (for an overview of the variations, see Tables S1–S12 in the Supporting Information), and the respective power values are calculated and compared to the initially intended value $1 - \beta$.

3.1 | Scenarios and Parameter Values

In the following, we set the target power to $1 - \beta = 0.8$ and the type I error to $\alpha = 0.05$. The simulation was performed with the software R (version 4.1.2) R Core Team (2022) and $n_{sim} = 10,000$. For the simulation-based sample size calculation, we set L to 90% of the minimum of the highest censored or uncensored event times in both groups (Tian et al. 2020). The AHR is calculated using the R package AHR (Brueckner 2018). We consider six settings: two with proportional, two with nonproportional but noncrossing, and two with crossing hazards (see Table 1). Each scenario is divided into two censoring categories, denoted by an additional "a" and "b." Here, "a" represents a lower level (20% in each group), and "b" is a higher level of nonadministrative censoring (20% in group one and 40% in group two). These settings represent the initial assumptions we make about our data in order to derive the sample size. Based on these assumptions, we derive the required parameters for each of the four sample sizes approaches. To quantify the robustness of the methods in terms of power, we examine various deviations from these assumptions, ranging from simple deviations for only one parameter up to combinations of deviations for several parameters. The alternative survival distributions were chosen to have the same hazard ratio as the initial curves (see Table 2). Note that the hazard ratio has no meaningful interpretation as an effect measure under N-PHs and is only used as an easy comparison criterion for scenario selection. The concrete details on the deviations are displayed in Tables S1–S12 of the Supporting Information.

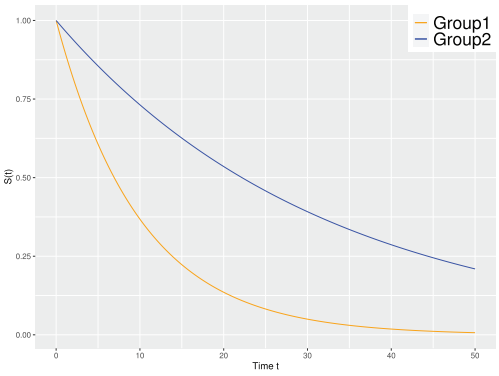
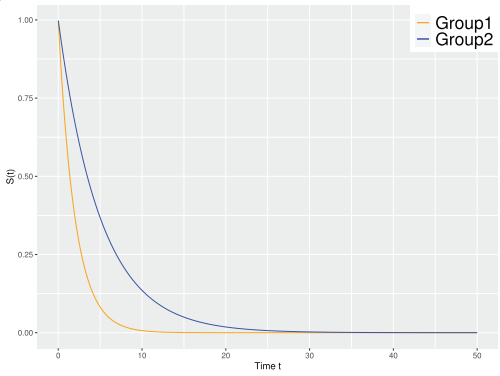
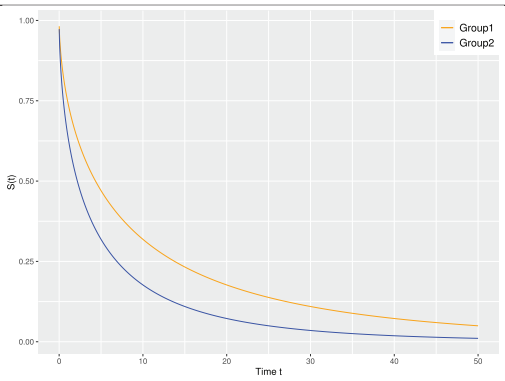
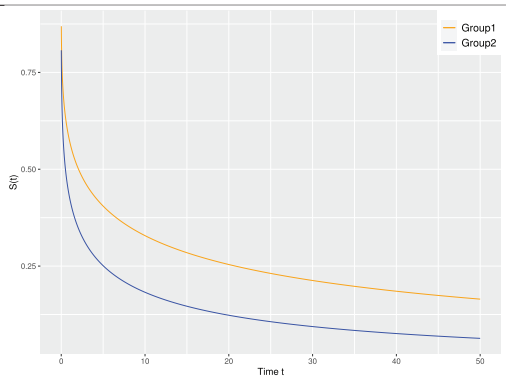
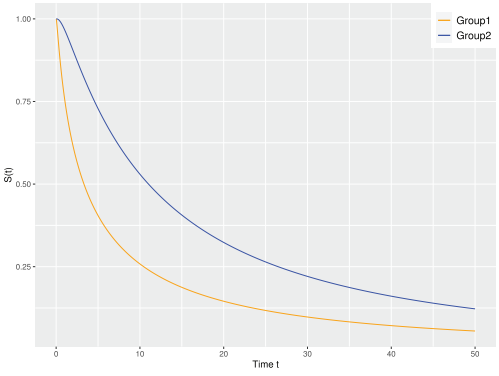
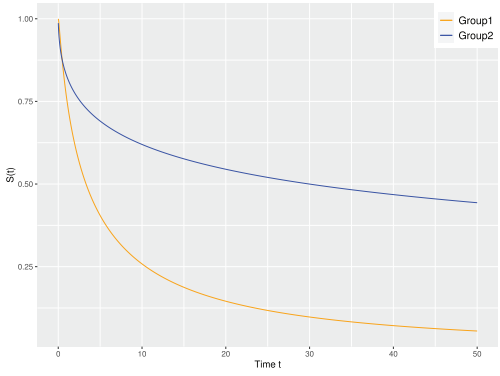
3.2 | Results

We discuss the results individually for the different hazard relationships under consideration in terms of the two simulation goals.

3.2.1 | G1: Comparison of Different Sample Size Calculation Approaches

The calculated sample sizes according to the four approaches (Schoenfeld, simulation-based log-rank as well as asymptotic and simulation-based AHR approaches) among all scenarios can be found in Table 3. As expected in the PH settings (S1a, S1b, S2a, S2b), we observe that sample sizes retrieved via the classical Schoenfeld approach are smaller than sample sizes based on the other three approaches. Moreover, we observe that the calculated sample sizes obtain rather similar

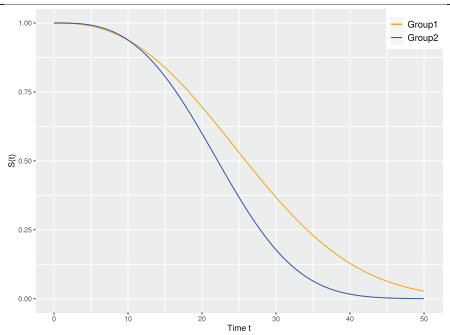
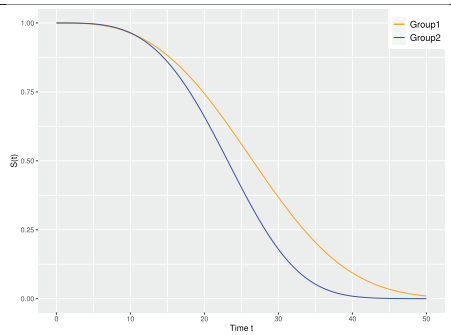
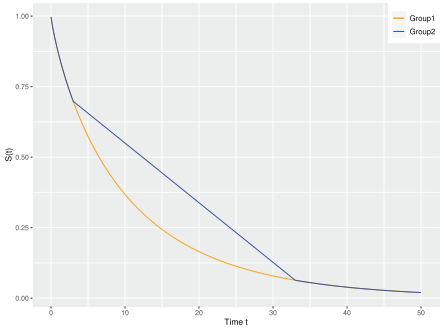
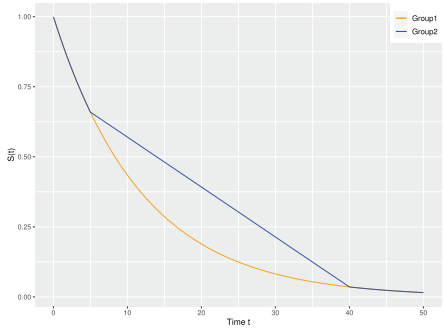
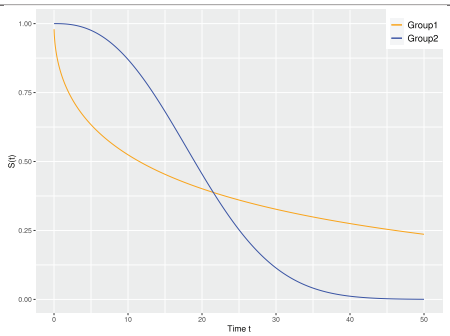
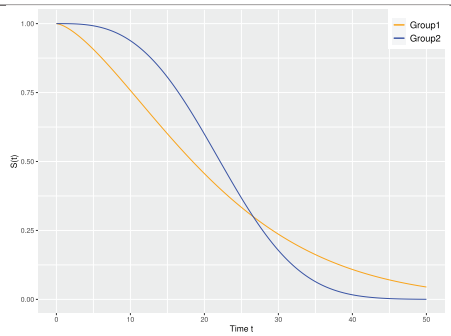
TABLE 1 | Simulation scenarios and the alternative distribution considered. The hazard ratio (HR) given is the mean based on 100 datasets with $n_0 = n_1 = 150$. The alternative scenario (alt) was chosen to have the same HR as the initial scenario.

Scenario	Original survival curves	Alternative survival curves
S1 (proportional) **HR**: ~ 0.41	 $F_1(t) = Exp(0.1)$ $F_2(t) = Exp(1/25)$	 $F_1(t) = Exp(1/2)$ $F_2(t) = Exp(1/5)$
S2 (proportional) **HR**: ~ 1.52	 $F_1(t) = Weibull(0.6, 8)$ $F_2(t) = Weibull(0.6, 4)$	 $F_1(t) = Weibull(0.3, 7)$ $F_2(t) = Weibull(0.3, 1.7)$
S3 (non-proportional) **HR**: ~ 0.41	 $F_1(t) = Lognormal(1.2, 1.7)$ $F_2(t) = Lognormal(2.4, 1.3)$	 $F_1(t) = Lognormal(1.2, 1.7)$ $F_2(t) = Lognormal(3.4, 3.6)$

values only within some specific PH settings (S1a and S1b) but they can also deviate drastically (n_AHRsim and n_SF in S2b). Across all considered scenarios, the LRsim obtains larger sample sizes than the SF approach. Furthermore, the LRsim approach comes along with computational limitations in S6b.

In the N-PHs setting, the crossing hazard scenarios (S5a, S5b, S6a, S6b) tend to have larger sample sizes than the noncrossing hazard scenarios (S3a, S3b, S4a, S4b) when considering n_SF and n_LRsim. For the other two sample size calculation approaches that observation does not hold. Furthermore, in almost all N-PHs scenarios, the sample sizes retrieved from the AHRsim

TABLE 2 | Simulation scenarios and the alternative cumulative density function (CDF) considered. The hazard ratio (HR) given is the mean based on 100 datasets with $n_0 = n_1 = 150$. The alternative scenario (alt) was chosen to have the same HR as the initial scenario..

Scenario	CDF (Visualization of the survival curves)	Alternative CDF (Visualization of the other survival curves)
S4 (non-proportional) **HR**: ~ 1.57	 $F_1(t) = Weibull(2.5, 30)$ $F_2(t) = Weibull(3, 25)$	 $F_1(t) = Weibull(3, 30)$ $F_2(t) = Weibull(3.5, 25.7)$
S5[39] (crossing) **HR**: ~ 0.71	 $F_1(t) = Weibull(0.849, 10)$ $F_2(t) = \begin{cases} Weibull(0.849, 10) & t \leq 3 \\ Unif(3, 33) & 3 < t \leq 33 \\ Weibull(0.849, 10) & t > 33 \end{cases}$	 $F_1(t) = Weibull(1, 12)$ $F_2(t) = \begin{cases} Weibull(1, 12) & t \leq 5 \\ Unif(5, 40) & 5 < t \leq 40 \\ Weibull(1, 12) & t > 40 \end{cases}$
S6 (crossing) **HR**: ~ 0.74	 $F_1(t) = Weibull(0.5, 24)$ $F_2(t) = Weibull(2.5, 22)$	 $F_1(t) = Weibull(1.5, 23.5)$ $F_2(t) = Weibull(3, 25)$

sample size calculation approach are smaller than those from the AHRasymp approach. It is notable that the sample sizes in the crossing hazards settings S6a and S6b are of very different scales compared to the Schoenfeld and LRsim-based sample sizes (i.e., up to a factor of ~40 when comparing the sample sizes through SF and AHRsim in S6b). Moreover, we note that in some cases (e.g., S4 for AHRsim and AHRasymp), the sample sizes become smaller even though the censoring rate increases (compare the lower with upper scenario in each of the scenario boxes). One explanation is that censoring can distort the estimated (average) hazard form.

3.2.2 | G2: Assessing the Influence of Misspecified Distributional Assumptions in Terms of Statistical Power

The obtained power for all deviations from the initial assumptions can be found in the Supporting Information

TABLE 3 | Total sample sizes according to the different calculation approaches for each simulation scenario. SF (Schoenfeld approach), LRsim (simulation-based log-rank approach), AHRsim (simulation-based average hazard ratio approach), and AHRasymp (asymptotic average hazard ratio approach); — indicates that simulations were not executed due to computational limitations.

Scenario	n_SF	n_LRsim	n_AHRasymp	n_AHRsim
S1a	48	54	52	54
S1b	48	56	52	56
S2a	216	236	260	334
S2b	214	312	320	414
S3a	78	80	48	38
S3b	70	78	48	36
S4a	136	160	368	432
S4b	146	202	348	224
S5a	356	360	382	342
S5b	292	324	382	330
S6a	2034	—	264	228
S6b	12,826	—	350	310

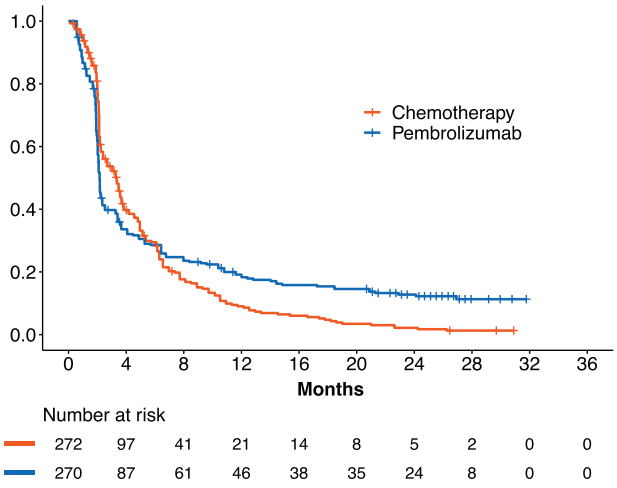


FIGURE 1 | Reconstructed individual patient data based on the UC study by Fradet et al. (2019).

(see Tables S1–S12). We here only summarize our findings. First, it needs to be noted that the power values in the original settings (cf. A1 in Tables S1–S12 in the Supporting Information) are not always equal to 0.8, and hence, the power values for the deviating settings (A2–A23) are to be considered deviations from those power starting values. Especially the AHRasymp approach deviates more often from the 0.8 in scenario A1 than the other sample size calculation approaches (e.g., 0.66 in S2a and S2b as well as 0.86 in S3a and S3b). Moreover, there are scenarios (S6a and S6b) in which the power simulations for SF and LRsim were not possible due to computational limitations. For AHRasymp and AHRsim, it was always possible to retrieve the respective power values. In specific alternative settings, the simulations yielded no results due to too early censoring. However, this was always consistent across all four approaches within one setting. Regarding the power performance under misspecified distributional assumptions for the survival time, none of the approaches seems to be robust across all deviations (e.g., power values of 0.40–0.80 compared to 0.66 (AHRasymp), 0.51–0.81 compared to 0.69 (SF), 0.61–0.91 compared to 0.80 (LRsim), and 0.55–0.91 compared to 0.81 (AHRsim) in scenario S2b). However, specific approaches in certain scenarios seem relatively stable regarding power performance under misspecified distributional assumptions (e.g., LRsim and AHRsim in S1a, AHRsim in S2a). Furthermore, it can be noted that under most PH (S1–S2a), the alternative scenarios do not influence the power values as much as in the N-PH scenarios (e.g., 0.78–0.99 (SF), 0.77–0.98 (LRsim), 0.63–0.87 (AHRasymp), 0.53–0.81 (AHRsim) in S3b). Larger censoring rates have the most considerable impact on the power decrease of the SF and LRsim sample size calculation approach for several PH settings (cf. A11 in S1a–S2a). Also, under exponentially distributed censoring times, the alternative scenarios do not influence the power values considerably in S1a–S2a. In the following, we focus on the N-PHs scenarios and first restrict ourselves to the noncrossing hazards scenarios (S3a, S3b, S4a, S4b). Here, the misspecifications in survival time distributions have a big impact on the power values of all four sample size calculation approaches, such as 0.25 for AHRasymp in S4b, 0.51 for AHRsim in S3b, 0.38 for SF in S4b, and 0.42 for LRsim in S4b. Overall, there is a poor performance regarding misspecifications in S4a and especially in S4b for all four sample size calculation approaches. Under crossing hazards (S5a, S5b, S6a, S6b), it first must be noted that the SF and LRsim approach was not calculated in all settings (i.e., S6a and S6b) due to computational limitations. In the other crossing hazards settings, the two AHR-based sample size calculation approaches tend to be more robust (S5a: power range of 0.75–0.87 for SF, 0.76–0.89 for LRsim, 0.78–0.82 for AHRsim, 0.81–0.85 for AHRaysmp; S5b: 0.62–0.87 for SF, 0.58–0.85 for LRsim, 0.61–0.81 for AHRsim, 0.66–0.85 for AHRasymp). Also, in scenarios where the sample sizes cannot be calculated for the SF and LRsim approaches, the two AHR-based sample size approaches seem relatively stable regarding deviations from the prespecified assumptions.

4 | Illustrative Data Example

When planning clinical trials, one often has preliminary studies that allow for more detailed planning of the new trial. We consider a trial to evaluate the long-term safety and efficacy outcomes of pembrolizumab compared to chemotherapy in patients with advanced urothelial cancer (UC) that progressed after platinum-based chemotherapy (Fradet et al. 2019). We assume that Figure 1 shows the survival time curves of such a preliminary study.

Since the figure alone is not sufficient to compute sample sizes, we have reconstructed the individual patient data from the original publication figure using the reconstruction algorithm by Guyot et al. (2012) and webPlotDigitizer (Rohatgi 2019). The algorithm does result in time-to-event data including event time and a censoring indicator. From these reconstructed patient data, we then estimated the hazard ratio and AHR as well as the variance. Using SF, the LRsim, the AHRsim, and the AHRasymp approach, we computed the required sample size based on the given values. When using SF and LRsim we imply PH. In the

TABLE 4 | Sample sizes obtained by SF (Schoenfeld formula) and AHRsim (simulation-based AHR approach) based on reconstructed patient data from Fradet et al. (2019).

Sample size approach	Pembrolizumab	Chemotherapy
SF	2761	2964
LRsim	2777	2798
AHRsim	462	465
AHRasymp	514	552

TABLE 5 | *p* values obtained by the two-sided log-rank test and the AHR (average hazard ratio) test as well as the corresponding hazard ratio or AHR, respectively, based on reconstructed patient data from Fradet et al. (2019).

Test statistic	*p* value	Effect
log-rank	0.400	0.924
AHR	**0.041**	1.269

Significant values are given in bold.

methods section, we described how to generate data based on distributional assumptions, in order to calculate the required sample size. To avoid relying on assumed survival distributions in this example, we utilized case resampling with 1000 iterations for censored data (Thurow et al. 2023). The given example shows a case of crossing survival curves; hence, the implied assumption of PH is not correct. Thus, one could expect a large necessary sample size using the Schoenfeld formula and the LRsim approach, which is not recommended for non-PH settings. The AHR-based approaches, on the other hand, do not require this assumption and are, hence, suitable for the underlying survival distributions. This results in higher power for this illustrative example and, hence, a smaller required sample size. The obtained sample sizes are listed in Table 4. We recalculated the *p* value for the log-rank test and the AHR-based test using the reconstructed data of sample site 542 (cf. Table 5).

It can be seen that the reconstructed data reflects the *p* value of the log-rank test quite well (*p* value of 0.31 in the original publication by Fradet et al. 2019) and is not significant. In comparison, the AHR-based test yields a significant result. Furthermore, the corresponding effect measures point in different directions: while the AHR indicates an increased risk of Prembrolizumab, the log-rank test suggests a small favorable effect. This shows that the choice of the testing procedure matters especially in non-PH situations. It should be noted that even though the original study does not reach the required sample sizes for a power of 0.8, the AHR-based test can reject the null hypothesis. To investigate this further, we did a small simulation using case resampling with 1000 iterations regarding the power of the test with the given sample sizes (Davison and Hinkley 1997; Thurow et al. 2023). From the 1000 sampled datasets and corresponding test decisions, we could derive a power of 0.58 for the given sample sizes. The R-code for this data example is provided in the Supporting Information.

5 | Discussion

The objective of our work was to offer guidance on determining an appropriate sample size for a clinical trial utilizing the AHR as the primary endpoint. We evaluated the effectiveness of the simulation-based sample size calculation approach (AHRsim) and the asymptotic sample size calculation approach (AHRasymp) together with the Schoenfeld formula (SF) as well as the log-rank (LR) simulation-based approach (LRsim) by comparing their required sample sizes and resulting statistical power, considering potential deviations from the original assumptions. To accomplish this, we initially presented the setup involving the test statistic, along with simulation-based and asymptotic approaches for sample size calculation. Subsequently, we conducted extensive simulation studies to assess different survival time and censoring distributions as well as variations in group allocation ratios. Finally, we applied the four different sample size calculation approaches to a specific data example and deliberated on our findings.

One limitation of our work is that we solely focused on one specific weight function of the AHR. However, the suggested approach can also be applied to other weight functions of the same family (for details, see the Methods section). Another limitation of our study relates to the specific scenarios considered for each category (proportional, nonproportional, and crossing hazards), where the results did not always coincide. While we observed certain tendencies, it is important to emphasize that conducting extensive simulation studies is still recommended. Such studies would enable the exploration of specific behaviors and rules regarding power and sample size. Furthermore, it is crucial to acknowledge that a high percentage of censorship can obscure the true behavior of hazard rates and consequently potentially result in underestimations of the required sample size. Therefore, cautious interpretation of sample size estimations is warranted, particularly in cases where a substantial proportion of censorship is anticipated.

The Schoenfeld method exhibits several downsides under scenarios with N-PHs. One of them is that the SF method has been derived under PH settings. In detail, it is particularly nonrobust when hazards cross, resulting in the requirement for hugely enlarged sample sizes compared to the alternative approaches relying on the AHR. However, the SF method can still be valuable for obtaining starting values in simulation approaches, when dealing with N-PHs.

Even though the AHRsim method demonstrates sometimes less power in certain settings of the PH settings, it should be kept in mind that not every sample size approach is applicable in every situation and the AHRsim approach provides flexibility in this regard. Overall, across various scenarios, the violation of distributional assumptions results in a slightly smaller loss of power for the AHRsim method compared to the SF and log-rank methods in many nonproportional and crossing hazard situations. Also, in scenarios where the SF and LRsim approaches cannot be applied any longer due to computational limitations, the AHR-based sample size approaches are still applicable. Moreover, the real-world example illustrates that in certain situations, the AHR-based test exhibits more power than the log-rank test. Consequently, employing the simulation-based approach would

lead to significantly smaller sample sizes for follow-up studies. In comparison, the asymptotic approach of the AHR method (AHRasymp) faces problems in certain scenarios, as previously observed in studies by Dobler and Pauly (2020), including low power even when distributional assumptions are met.

Both the simulation study and the real data example emphasize that different sample size calculation approaches yield widely varying results depending on the data situation. Hence, this step in the planning of clinical trials holds substantial relevance in terms of both statistical and economic outcomes. These findings underscore the importance of conducting preliminary simulation studies to enable the selection of the appropriate sample size calculation approach as well as inference method. This especially holds when N-PHs are assumed, where no generally recommended effect measure exists. Overall, it is recommended to consider different approaches while incorporating all available information. In future work, it could be of interest to also incorporate Bayesian approaches (Chen, Ibrahim, and Chu 2011; Chen et al. 2015; Chi and Ibrahim 2006; Kim, Park, and Kim 2011; Xu, Psioda, and Ibrahim 2022) in order to exploit more of the available information.

Acknowledgments

This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the University Alliance Ruhr (https://uaruhr.de). The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359. Moreover, the work of Markus Pauly was supported by an Individual DFG Research Project. We would like to thank the Expert Reviewers for taking the time and effort necessary to review the manuscript. We sincerely appreciate all valuable comments and suggestions, which helped us to improve the quality of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available in the Supporting Information of this article.

Open Research Badges

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

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

References

Alexander, B. M., J. D. Schoenfeld, and L. Trippa. 2018. "Hazards of Hazard Ratios-Deviations From Model Assumptions in Immunotherapy." *New England Journal of Medicine* 378, no. 12: 1158–1159.

Ananthakrishnan, R., S. Green, A. Previtali, R. Liu, D. Li, and M. LaValley. 2021. "Critical Review of Oncology Clinical Trial Design Under Non-Proportional Hazards." *Critical Reviews in Oncology/Hematology* 162: 103350.

Brückner, M., and W. Brannath. 2017. "Sequential Tests for Non-Proportional Hazards Data." *Lifetime Data Analysis* 23, no. 3: 339–352.

Brueckner, M. 2018. *Ahr*. https://github.com/cran/AHR.

Chen, L. M., J. G. Ibrahim, and H. Chu. 2011. "Sample Size and Power Determination in Joint Modeling of Longitudinal and Survival Data." *Statistics in Medicine* 30, no. 18: 2295–2309.

Chen, Q., D. Zeng, J. G. Ibrahim, M.-H. Chen, Z. Pan, and X. Xue. 2015. "Quantifying the Average of the Time-Varying Hazard Ratio via a Class of Transformations." *Lifetime Data Analysis* 21: 259–279.

Chi, Y.-Y., and J. G. Ibrahim. 2006. "Joint Models for Multivariate Longitudinal and Multivariate Survival Data." *Biometrics* 62, no. 2: 432–445.

Cortés, Martínez, J., R. B. Geskus, K. Kim, and G. G. Melis. 2021. "Using the Geometric Average Hazard Ratio in Sample Size Calculation for Time-to-Event Data With Composite Endpoints." *BMC Medical Research Methodology* 21, no. 1: 1–14.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press.

Dobler, D., and M. Pauly. 2018. "Bootstrap- and Permutation-Based Inference for the Mann–Whitney Effect for Right-Censored and Tied Data." *TEST* 27, no. 3: 639–658.

Dobler, D., and M. Pauly. 2020. "Factorial Analyses of Treatment Effects Under Independent Right-Censoring." *Statistical Methods in Medical Research* 29, no. 2: 325–343.

Dunkler, D., M. Ploner, M. Schemper, and G. Heinze. 2018. "Weighted Cox Regression Using the R Package Coxphw." *Journal of Statistical Software* 84, no. 2: 1–26.

Fradet, Y., J. Bellmunt, D. Vaughn, et al. 2019. "Randomized Phase III KEYNOTE-045 Trial of Pembrolizumab Versus Paclitaxel, Docetaxel, or Vinflunine in Recurrent Advanced Urothelial Cancer: Results of > 2 Years of Follow-Up." *Annals of Oncology* 30, no. 6: 970–976.

Freedman, L. S. 1982. "Tables of the Number of Patients Required in Clinical Trials Using the Logrank Test." *Statistics in Medicine* 1, no. 2: 121–129.

Gorfine, M., M. Schlesinger, and L. Hsu. 2020. "K-Sample Omnibus Non-Proportional Hazards Tests Based on Right-Censored Data." *Statistical Methods in Medical Research* 29, no. 10: 2830–2850.

Guyot, P., A. Ades, M. J. Ouwens, and N. J. Welton. 2012. "Enhanced Secondary Analysis of Survival Data: Reconstructing the Data From Published Kaplan–Meier Survival Curves." *BMC Medical Research Methodology* 12, no. 1: 9.

Hasegawa, T. 2014. "Sample Size Determination for the Weighted Log-Rank Test With the Fleming–Harrington Class of Weights in Cancer Vaccine Studies." *Pharmaceutical Statistics* 13, no. 2: 128–135.

Horiguchi, M., and H. Uno. 2017. *SSRMST: Sample Size Calculation Using Restricted Mean Survival Time*. R package version 0.1.1.

Kalbfleisch, J. D. 1981. "Estimation of the Average Hazard Ratio." *Biometrika* 68, no. 1: 105–112.

Kieser, M. 2020. *Methods and Applications of Sample Size Calculation and Recalculation in Clinical Trials*. Cham, Switzerland: Springer.

Kim, Y., J. K. Park, and G. Kim. 2011. "Bayesian Analysis for Monotone Hazard Ratio." *Lifetime Data Analysis* 17: 302–320.

Koziol, J. A., and Z. Jia. 2009. "The Concordance Index C and the Mann–Whitney Parameter Pr (X > Y) With Randomly Censored Data." *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 51, no. 3: 467–474.

Lakatos, E. 1986. "Sample Size Determination in Clinical Trials With Time-Dependent Rates of Losses and Noncompliance." *Controlled Clinical Trials* 7, no. 3: 189–199.

Lin, R. S., J. Lin, S. Roychoudhury, et al. 2020. "Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards: A Comparative Analysis." *Statistics in Biopharmaceutical Research* 12, no. 2: 187–198.

Mann, H. B., and D. R. Whitney. 1947. "On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other." *The Annals of Mathematical Statistics* 18, no. 1: 50–60.

Mick, R., and T.-T. Chen. 2015. "Statistical Challenges in the Design of Late-Stage Cancer Immunotherapy Studies." *Cancer Immunology Research* 3, no. 12: 1292–1298.

Phadnis, M. A., and M. S. Mayo. 2021. "Sample Size Calculation for Two-Arm Trials With Time-to-Event Endpoint for Nonproportional Hazards Using the Concept of Relative Time When Inference is Built on Comparing Weibull Distributions." *Biometrical Journal* 63, no. 7: 1406–1433.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rauch, G., W. Brannath, M. Brückner, and M. Kieser. 2018. "The Average Hazard Ratio – A Good Effect Measure for Time-to-Event Endpoints When the Proportional Hazard Assumption is Violated?" *Methods of Information in Medicine* 57, no. 3: 89–100.

Rauch, G., S. Schüler, and M. Kieser. 2017. *Planning and Analyzing Clinical Trials With Composite Endpoints*. Cham, Switzerland: Springer.

Rohatgi, A. 2019. *Webplotdigitizer: Version 4.4*. https://automeris.io/WebPlotDigitizer.

Royston, P. 2018. "Power and Sample-Size Analysis for the Royston–Parmar Combined Test in Clinical Trials With a Time-to-Event Outcome." *Stata Journal* 18, no. 1: 3–21.

Royston, P., and M. K. 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, no. 19: 2409–2421.

Royston, P., and M. K. Parmar. 2013. "Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials With a Time-to-Event Outcome." *BMC Medical Research Methodology* 13, no. 1: 1–15.

Royston, P., and M. K. Parmar. 2016. "Augmenting the Logrank Test in the Design of Clinical Trials in Which Non-Proportional Hazards of the Treatment Effect may be Anticipated." *BMC Medical Research Methodology* 16, no. 1: 1–13.

Schemper, M. 1992. "Cox Analysis of Survival Data With Non-Proportional Hazard Functions." *Journal of the Royal Statistical Society: Series D (The Statistician)* 41, no. 4: 455–465.

Schemper, M., S. Wakounig, and G. Heinze. 2009. "The Estimation of Average Hazard Ratios by Weighted Cox Regression." *Statistics in Medicine* 28, no. 19: 2473–2489.

Schoenfeld, D. 1981. "The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions." *Biometrika* 68, no. 1: 316–319.

Schoenfeld, D. A. 1983. "Sample-Size Formula for the Proportional-Hazards Regression Model." *Biometrics* 39, no. 2: 499–503.

Thurow, M., I. Dormuth, C. Sauer, M. Ditzhaus, and M. Pauly. 2023. "How to Simulate Realistic Survival Data? A Simulation Study to Compare Realistic Simulation Models." Preprint, submitted August 15, 2023. https://doi.org/10.48550/arXiv:2308.07842.

Tian, L., H. Fu, S. J. Ruberg, H. Uno, and L.-J. Wei. 2018. "Efficiency of Two Sample Tests via the Restricted Mean Survival Time for Analyzing Event Time Observations." *Biometrics* 74, no. 2: 694–702.

Tian, L., H. Jin, H. Uno, et al. 2020. "On the Empirical Choice of the Time Window for Restricted Mean Survival Time." *Biometrics* 76, no. 4: 1157–1166.

Trinquart, L., J. Jacot, S. C. Conner, and R. Porcher. 2016. "Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials." *Journal of Clinical Oncology* 34, no. 15: 1813–1819.

Uno, H., B. Claggett, L. Tian, et al. 2014. "Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis." *Journal of Clinical Oncology* 32, no. 22: 2380.

Uno, H., and M. Horiguchi. 2023. "Ratio and Difference of Average Hazard With Survival Weight: New Measures to Quantify Survival Benefit of New Therapy." *Statistics in Medicine* 42, no. 7: 936–952.

Uno, H., J. Wittes, H. Fu, et al. 2015. "Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies." *Annals of Internal Medicine* 163, no. 2: 127–134.

Xu, J., M. A. Psioda, and J. G. Ibrahim. 2022. "Bayesian Design of Clinical Trials Using Joint Models for Longitudinal and Time-to-Event Data." *Biostatistics* 23, no. 2: 591–608.

Yung, G., and Y. Liu. 2019. *npsurvSS: Sample Size and Power Calculation for Common Non-Parametric Tests in Survival Analysis*. R package version 1.0.1.

Yung, G., and Y. Liu. 2020. "Sample Size and Power for the Weighted Log-Rank Test and Kaplan–Meier Based Tests With Allowance for Nonproportional Hazards." *Biometrics* 76, no. 3: 939–950.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Article 4

Dormuth, I., Herrmann, C., Konietschke, F., Pauly, M., Wirth, M. & Ditzhaus, M. (2024). Single CASANOVA? Not in multiple comparisons. arXiv preprint arXiv:2410.21098.

SINGLE CASANOVA? NOT IN MULTIPLE COMPARISONS

Ina Dormuth[1*], Carolin Herrmann[2], Frank Konietschke[3], Markus Pauly[1,4], Matthias Wirth[5,6], and Marc Ditzhaus[†7]
[1]Department of Statistics, TU Dortmund University, Dortmund, North Rhine-Westphalia, 44227, Germany
[2]Mathematical Institute, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[3]Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, 10117 Berlin, Germany
[4]Research Center Trustworthy Data Science and Security, UA Ruhr, 44227 Dortmund
[5] Department of Hematology, Oncology and Cancer Immunology, Charité - Universitätsmedizin Berlin, Berlin, Germany
[6] Department of General, Visceral and Pediatric Surgery, University Medical Center Göttingen, Göttingen, Germany.
[7]Department of Mathematics, Otto von Guericke University Magdeburg, Magdeburg, Germany

ABSTRACT

When comparing multiple groups in clinical trials, we are not only interested in whether there is a difference between any groups but rather the location. Such research questions lead to testing multiple individual hypotheses. To control the familywise error rate (FWER), we must apply some corrections or introduce tests that control the FWER by design. In the case of time-to-event data, a Bonferroni-corrected log-rank test is commonly used. This approach has two significant drawbacks: (i) it loses power when the proportional hazards assumption is violated [1] and (ii) the correction generally leads to a lower power, especially when the test statistics are not independent [2]. We propose two new tests based on combined weighted log-rank tests. One as a simple multiple contrast tests of weighted log-rank tests and one as an extension of the so-called CASANOVA test [3]. The latter was introduced for factorial designs. We propose a new multiple contrast test based on the CASANOVA approach. Our test promises to be more powerful under crossing hazards and eliminates the need for additional p-value correction. We assess the performance of our tests through extensive Monte Carlo simulation studies covering both proportional and non-proportional hazard scenarios. Finally, we apply the new and reference methods to a real-world data example. The new approaches control the FWER and show reasonable power in all scenarios. They outperform the adjusted approaches in some non-proportional settings in terms of power.

***K*eywords** Multiple contrast tests, Non-proportional hazards, Survival analysis, Weighted log-rank test

1 Introduction

Time-to-event or survival analysis is essential across medical research, engineering, and social sciences. Trials often involve multiple groups (treatment arms) or factorial designs, creating unique statistical challenges. The primary research focuses not merely on whether any arms differ but specifically identifying which groups show differences. Thus, traditional global test procedures like ANOVA-type methods, which test null hypotheses of equal hazard ratios or cumulative hazard rate functions, are often inadequate [4, 3]. Instead, flexible multiple comparison procedures are crucial in modern data analysis. Current approaches typically employ pairwise multiple log-rank tests with adjustments for multiplicity (e.g., Bonferroni correction) [5], but these methods can lack efficiency due to restrictive assumptions about the correlation structure of test statistics [6, 7]. In recent years, many researchers developed *multiple contrast test procedures* (MCTPs) along with simultaneous confidence intervals (SCIs) (usually conducted as maximum tests), which are valid for arbitrary correlations of the test statistics and use the correlation within the multiplicity adjustment for various endpoints (means, proportions, Mann-Whitney effects) [8, 9, 10, 2, 11]. Munko et al. [12] introduced a restricted mean survival time (RMST)-based multiple contrast tests for time-to-event data. Since the RMST should not be employed under crossing hazards [13, 14], we aim to close this gap and introduce a powerful and flexible MCTP for analyzing survival data with crossing hazards.

The log-rank test is one of the most prominent test procedures in survival analysis. The method is well known to be optimal when the proportional hazards (PH) assumption is met but significantly loses power otherwise [14]. Even though the problem is fairly well known, a substantial amount of investigators of (clinical) trials still ignore the issue and publish their findings upon log-rank tests in leading high-quality peer-review journals even when the assumption is violated [15, 16, 14]. For the analysis of two independent samples, weighted log-rank tests and their combinations comprise a great alternative to the classical log-rank test and are beneficial in non-proportional hazards models [17, 18, 19, 20]. Ditzhaus and Friedrich [20] propose a Wald-type test of multiple weight functions within a single multivariate test. Which weight function to choose depends on the alternative of interest and cannot be recommended in a general way. However, the test does not provide information on which weight function appears most powerful. For the analysis of more than two samples and factorial designs, Ditzhaus et al. [3] extended these procedures to the *Cumulative Aalen Survival Analysis-of-Variance* (CASANOVA) method. In principle, they are global ANOVA-based tests (quadratic forms) and can be used to estimate and test main and interaction effects in general factorial designs. Estimating and testing user-specific contrasts are impossible, limiting their application in statistical practice. To overcome these shortcomings, we propose a novel flexible MCTP. Extensive simulation studies indicate that the test is more powerful under non-proportional hazards and eliminates the need for additional p-value correction. The remainder of the paper is organized as follows. The second section introduces the main statistical methods employed in the analyses. The third section describes the simulation setup and the corresponding results. The following section applies the methods of interest to a real-world data example. The final conclusions are drawn in section five, together with future research questions.

2 Set up

Multiple contrasts are faced in many research questions related to time-to-event endpoints. Applying separate tests without adjusting for multiple testing increases the likelihood of false discoveries and inflated error rates. In the following, we present different well-established statistical methods for an underlying multiple contrast problem with time-to-event endpoints, as well as our newly developed method based on a combination of multi-directional log-rank tests and the concept of maximum tests.

Statistical model

First, we define the underlying statistical model. Therefore, we consider a study design involving $k \geq 2$ groups (treatment arms) of n_j independent subjects, each with time-to-event data T_{ji} and right-censoring time C_{ji}. The statistical model considered here can be summarized by mutually independent positive random variables $T_{ji} \sim F_j, \quad \text{and} \quad C_{ji} \sim G_j, \quad j = 1, \ldots, k;\ i = 1, \ldots, n_j$, where F_j and G_j are both continuous distribution functions, respectively. Furthermore, let $X_{ji} = \min(T_{ji}, C_{ji})$ denote the observed time and $\delta_{ji} = I(X_{ji} = T_{ji})$ the censoring status with $I(\cdot)$ being the indicator function. The statistical model considered here does not entail any parameters but rather the survival distributions that could be used to define reasonable treatment effects. The cumulative hazard rate function for group j is defined by

$$A_j(t) = \int_0^t (1 - F_j(x))^{-1} dF_j(x), t \geq 0,\ j = 1, \ldots, k. \tag{1}$$

We further assume non-zero sized groups by $n_j/n \to \kappa_j \in (0, 1)$ as $\min(n_j : j = 1, \ldots, k) \to \infty$ and we exclude the case of only censored values within one group by assuming that $0 < F_j(t) < 1$ and $0 < G_j(t) < 1\ \forall j = 1, ..., k$ and some $t > 0$.

Multiple null hypotheses

The cumulative hazard rate function of treatment arm j called $A_j(t)$, summarizes the total accumulated risk of experiencing the event that has been gained by progressing to time t. No difference (i.e., no effect) between treatment arms j_1 and j_2 with $j_1 \neq j_2$, corresponds to $A_{j_1}(t) \equiv A_{j_2}(t)$, or, equivalently, $A_{j_1}(t) - A_{j_2}(t) \equiv 0$. In the several sample problems, let $\boldsymbol{H} \in \mathbb{R}^{q \times k}$ be a contrast matrix satisfying $\boldsymbol{H}\mathbf{1}_{\boldsymbol{k}} = \mathbf{0}_q$ with $\mathbf{1}_{\boldsymbol{k}}$ and $\mathbf{0}_q$ denoting vectors of ones and zeros, respectively. We denote the entries of $\boldsymbol{H}$ as h_{j_1,j_2}. For ease of presentation, we focus on the two-sample problem. Here, the most prominent matrices are the ones of Dunnett- and Tukey-type. The entries are composed of a single -1 and 1, indicating the two sample comparisons of interest. We define the corresponding index sets $I_{\text{Dunnett}} = \{(1, 2), \ldots, (1, k)\}$ and $I_{\text{Tukey}} = \{(1, 2), \ldots, (1, k), (2, 3), \ldots, (2, k), \ldots, (k - 1, k)\}$. In the following, we will indicate the position in the matrix or vector by the corresponding indices j_1 and j_2, for example, $(-1, 1, 0, \ldots, 0) = \boldsymbol{h}_{1,2}$ for $j_1 = 1$ and $j_2 = 2$.

The hypotheses we seek to infer are expressed in relation to the cumulative hazard rate functions as follows:

$$\begin{aligned}\mathcal{H}_0 &: \boldsymbol{H}\boldsymbol{A} = \mathbf{0}_q, \quad \boldsymbol{A} = (A_1, \ldots, A_k)^\top, \\ H_0^{j_1 j_2} &: \{\boldsymbol{h}_{j_1 j_2}\boldsymbol{A} = 0\},\ (j_1, j_2) \in I,\end{aligned}$$

with $\boldsymbol{A}^\top$ denoting the transposed vector of $\boldsymbol{A}$ and I being either I_{Dunnett} or I_{Tukey}. In general, the contrast matrix selection depends on the specific question of interest underlying the analysis.

3 Statistical Tests

Adjusted log-rank

As a reference method, we consider the Bonferroni adjusted log-rank test. Therefore, we define the Bonferroni-adjusted significance level $\alpha_{\text{Bonferroni}} = \alpha/q$ where α is the original significance level and q is the number of comparisons. The Bonferroni adjustment for multiple comparisons in a survival setting is a standard procedure in clinical settings, as discussed in Logan et al. (2005) [5]. The authors have provided a comprehensive description and suggested various methods for adjusting the number of comparisons.

We define the weighted log-rank test as a generalization of the classical log-rank test. Therefore, we employ the conventional counting process notation. Let $N_j(t) = \sum_{i=1}^{n_j} I\{X_{ji} \leq t, \delta_{ji} = 1\}$ represent the cumulative number of observed events within group j up to time t. Furthermore, we introduce $Y_j(t) = \sum_{i=1}^{n_j} I\{X_{ji} \geq t\}$, which denotes the number of individuals at risk just before time t in group j. These counting processes enable us to define the Nelson-Aalen estimator for A_j as $\widehat{A}_j(t) = \int_0^t \frac{I\{Y_j(s)>0\}}{Y_j(s)} \mathrm{d}N_j(s)$ for $j = 1, \ldots, k$ and $t \geq 0$.

Then, the weighted log-rank statistic for testing the local null hypothesis $H_0^{j_1 j_2} : \{\boldsymbol{h}_{j_1 j_2}\boldsymbol{A} = \mathbf{0}\} = \{A_{j_1} = A_{j_2}\}$ can be defined as [17]:

$$\begin{aligned}T_{j_1,j_2}(w) &= T(w, h_{j_1,j_2}) \\ &= \Big(\frac{n}{n_{j_1} n_{j_2}}\Big)^{1/2} \int_0^\infty w\{\widehat{F}_{j_1,j_2}(t-)\} \frac{Y_{j_1}(t) Y_{j_2}(t)}{Y_{j_1}(t) + Y_{j_2}(t)} \mathrm{d}(h_{j_1,j_2}\widehat{\boldsymbol{A}}(t)) \\ &= \Big(\frac{n}{n_{j_1} n_{j_2}}\Big)^{1/2} \int_0^\infty w\{\widehat{F}_{j_1,j_2}(t-)\} \frac{Y_{j_1}(t) Y_{j_2}(t)}{Y_{j_1}(t) + Y_{j_2}(t)} \Big\{\mathrm{d}\widehat{A}_{j_2}(t) - \mathrm{d}\widehat{A}_{j_1}(t)\Big\}.\end{aligned}$$

Here, $\widehat{F}_{j_1,j_2}(t-)$ represents the left-continuous version of the estimator $\widehat{F}_{j_1,j_2}$, and w is a continuous weight function and $\widehat{\boldsymbol{A}} = (\widehat{A}_1, \ldots, \widehat{A}_k)^\top$. Fleming and Harrington [18] examined a specific subclass of weights w given by $w(t) = t^r(1-t)^g$ $(r, g \in \mathbb{N}_0)$. For instance, when $r = g = 0$, the log-rank test is obtained. We derive the individual p-values for the tests from the χ^2 distribution and compare them to $\alpha_{\text{Bonferroni}}$. To make a global statement, we compare the minimal p-value among all local tests to the adjusted significance level.

For practical implementation, we utilize the R package `survival` and its function `survdiff`. [21]

Adjusted mdir

In the context of our specific objectives, we are interested in more robust testing procedures towards multiple alternatives. For two group comparisons, the multi-directional log-rank test has been proposed as a combination procedure of different weighted log-rank tests [19, 20]. The test assumes the equality of survival under the null hypothesis, with the choice of weights determining the alternative hypothesis. We are particularly interested in weights that intersect the x-axis, such as $w(t_i) = 1 - 2t_i$ as they are specifically designed to address crossing hazard alternatives.

By default, the R package `mdir.logrank` [22] implements a combination of the log-rank weight $w^{(1)} \equiv 1$ and this crossing weight. Dormuth et al. [14] showed that this default set of weights seems to be robust against multiple alternatives. Nevertheless, if desired, additional weights can be combined to cover more alternative hypotheses. With these two weights, the local test statistic takes a studentized quadratic form:

$$S_{j_1 j_2} = (T_{j_1 j_2}(w^{(1)}), \ldots, T_{j_1 j_2}(w^{(m)}))\, \hat{\Sigma}^-_{j_1 j_2}\, (T_{j_1 j_2}(w^{(1)}), \ldots, T_{j_1 j_2}(w^{(m)}))^\top.$$

The entries of $\hat{\Sigma}_{j_1 j_2} \in \mathbb{R}^{m \times m}$ are given by

$$(\hat{\Sigma}_{j_1 j_2})_{p,s} = \frac{n}{n_{j_1} n_{j_2}} \int_{[0,\infty)} w^{(s)}\{\hat{F}_{j_1 j_2}(t-)\} w^{(p)}\{\hat{F}_{j_1 j_2}(t-)\} \frac{Y_{j_1}(t) Y_{j_2}(t)}{Y_{j_1}(t) + Y_{j_2}(t)} d\hat{A}_{j_1 j_2}(t), \quad (p, s = 1, ..., m)$$

with $\hat{A}_{j_1 j_2}$ the pooled Nelson-Aalen estimator of groups j_1 and j_2. $\hat{\Sigma}^{-}_{j_1 j_2}$ represents the Moore-Penrose inverse of the empirical covariance matrix of the weighted log-rank tests. For linearly independent weights $w^{(1)}, \ldots, w^{(m)}$ fulfilling the assumptions of [3] (continuous and of bounded variation), the test statistic $S_{j_1 j_2}$ can be assumed to be χ^2_m distributed under the null hypothesis. Ditzhaus and Friedrich [20] also proposed a permutation-based approach.

Again, we employ the Bonferroni adjusted significance level $\alpha_{\text{Bonferroni}}$ to compare to the obtained local p-values. Analogously to the adjusted log-rank test procedure, we obtain the global test decision by comparing the smallest p-value to the adjusted significance level.

MultiWeightedLR

Knowing that maximum tests are a common approach for multiple testing problems [23], a straightforward extension of the weighted log-rank test is to use the maximum over them and exploit the covariance structure between the different tests. We use the same weights as for the adjusted mdir approach without combining them in a quadratic form. Instead, we consider each weighted test individually. After calculating the corresponding covariance matrix, we take the maximum of all weighted test statistics as our global maximum test statistic. Mathematically, we write:

$$T_{\max} = \max_{r \in \{1,\ldots,m\}, (j_1,j_2) \in I} (T_{j_1,j_2}(w^{(r)}))$$

For the local testing problem we focus on $T^{j_1 j_2}_{\max} = \max_{(j_1,j_2)\in I}(T_{j_1,j_2}(w^{(r)}))$. Similar to the proof of Theorem 2 in [3], it can be shown that the vector $(T_{j_1,j_2}(w^{(r)}))_{r,j_1,j_2}$ is, under regularity conditions, asymptotic multivariate normally distributed with expected value $\mathbf{0}_{m\cdot q}$ and covariance matrix $\widehat{\Sigma} \in \mathbb{R}^{m\cdot q \times m\cdot q}$. We thus take the equicoordinate $(1-\alpha)$-quantile [4] of this distribution as critical value to obtain the MultiWeightedLR test in the statistic $T_{\max}$.

multiCASANOVA

Ditzhaus et al. [3] proposed the CASANOVA (Cumulative Aalen Survival Analysis-of-Variance) approach for general factorial designs with right-censored time-to-event data. The core idea of the method is an extension of weighted log-rank tests to the factorial design setup. Therefore, they expanded the combination approach of weighted log-rank tests (mdir) for the two-sample scenario to the general factorial survival designs implemented in the R package `GFDsurv` [24]. For further information we refer to Ditzhaus et al. [3]

We aim to extend the CASANOVA approach to allow the estimation and testing of user-specific contrasts in a multiple testing framework. Compared to the aforementioned approaches, the main difference is that we consider pooled quantities over all groups, not only the two groups of interest. To this end, we define a local test statistic for contrast $\boldsymbol{h}_{j_1,j_2}$ as

$$\begin{aligned}
\tilde{T}_{j_1,j_2}(w) &= \tilde{T}(w, \boldsymbol{h}_{j_1,j_2}) \\
&= \Big(\frac{n}{n_{j_1} n_{j_2}}\Big)^{1/2} \int_0^\infty w\{\widehat{F}(t-)\} \frac{Y_{j_1}(t) Y_{j_2}(t)}{Y(t)} \mathrm{d}(\boldsymbol{h}_{j_1,j_2} \widehat{\boldsymbol{A}}(t)) \\
&= \Big(\frac{n}{n_{j_1} n_{j_2}}\Big)^{1/2} \int_0^\infty w\{\widehat{F}(t-)\} \frac{Y_{j_1}(t) Y_{j_2}(t)}{Y(t)} \Big\{\mathrm{d}\widehat{A}_{j_2}(t) - \mathrm{d}\widehat{A}_{j_1}(t)\Big\},
\end{aligned}$$

where $\widehat{F}(t-)$ represents the left-continuous version of the pooled estimator $\widehat{F}$ and $Y(t)$ is the total number of individuals at risk over all groups. As in the adjusted mdir test, we combine several weights (still for one single contrast) by considering the corresponding quadratic form given by

$$C_{j_1,j_2} = (\tilde{T}_{j_1,j_2}(w^{(1)}), ..., \tilde{T}_{j_1,j_2}(w^{(m)}))\widehat{\mathrm{Cov}}^{-}_{j_1,j_2} (\tilde{T}_{j_1,j_2}(w^{(1)}), ..., \tilde{T}_{j_1,j_2}(w^{(m)}))^\top,$$

where the inner matrix is defined by

$$(\widehat{\text{Cov}}_{j_1 j_2})_{p,s} = \frac{n}{n_{j_1} n_{j_2}} \int_{[0,\infty)} w^{(s)}\{\hat{F}(t-)\} w^{(p)}\{\hat{F}(t-)\} \frac{Y_{j_1}(t) Y_{j_2}(t)}{Y(t)} d\hat{A}(t), \quad (p, s = 1, ..., m)$$

and $\widehat{\text{Cov}}^{-}_{j_1,j_2}$ represents its Moore-Penrose inverse. Similar to the maximum approach within MultiWeightedLR, we now consider the maximum of these Wald-type statistics over all contrasts of interest as the global test statistic

$$C_{\max} = \max_{(j_1,j_2)\in I} (C_{j_1,j_2}).$$

Note that we did not take the maximum over the different weights as those are already incorporated within the quadratic forms.

We use the common wild bootstrap approach for counting processes in time-to-event analyses [25, 26] to approximate the limiting distribution. Therefore, we consider independent and identically distributed variables $G_1, \ldots, G_{n_j}$ with $E(G_i) = 0$ and $\text{Var}(G_i) = 1$. We obtain the wild bootstrap version of the Nelson-Aalen estimator by:

$$\widehat{A}^*_j(t) = \int_0^t \frac{I\{Y_j(s) > 0\}}{Y_j(s)} \mathrm{d} \left(\sum_{i=1}^{n_j} G_i N_j(s) \right).$$

By using $\widehat{A}^*_j(t)$ instead of $\widehat{A}_j(t)$ to derive $\tilde{T}_{j_1,j_2}$ and thus C_{j_1,j_2} we obtain their wild bootstrap versions $\tilde{T}^*_{j_1,j_2}$ and $C^*_{j_1,j_2}$, respectively. Since counting processes are discrete, we opt for discrete distributions for the G_i. We focus on two common choices: (i) the Rademacher distribution [27], and (ii) the centered Poisson distribution [28]. This results in two different wild bootstrap quantiles depending on the distribution of choice: q^*_α the α-quantile of $C_{\max}$ given our data (X_{ji}, δ_{ji}). Then we obtain the global test decision by evaluating $C_{\max} > q^*_\alpha$ and the local test decisions by $C_{j_1,j_2} > q^*_\alpha$.

4 Simulation Study

We conducted an extensive simulation study in R 4.4.0 [29] to evaluate the rejection rate and the power performance of the candidate methods.

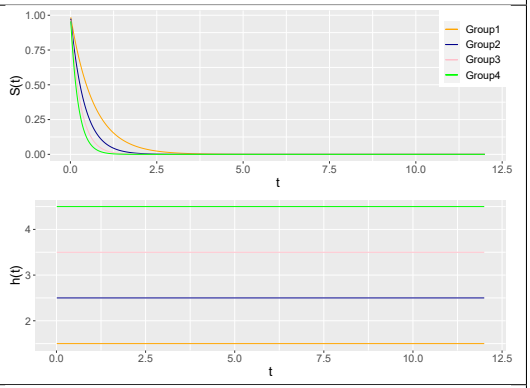
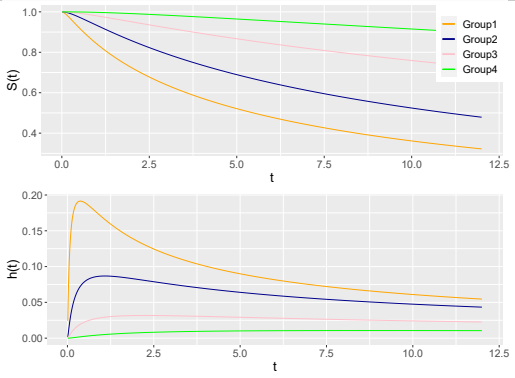
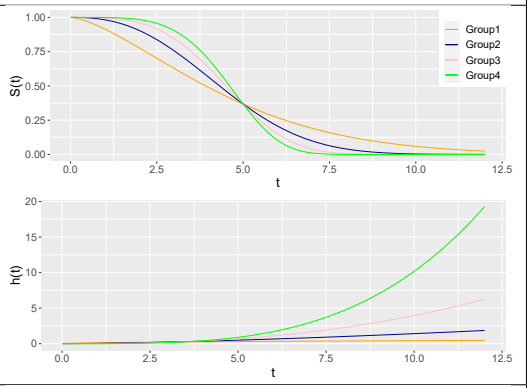
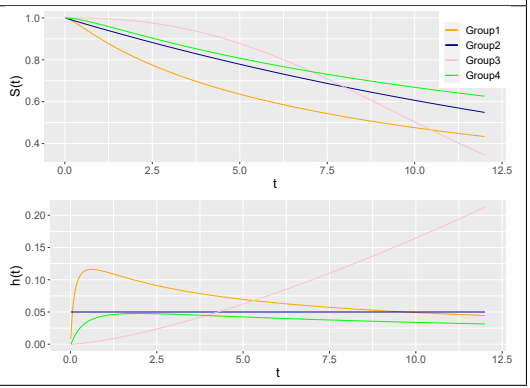
Simulation Setup

We simulated data for $k = 4$ groups considering the Tukey- and Dunnett-type contrast matrices. We considered four scenarios, each with different distribution functions. Each represents a specific case of hazard relationships such as (i) proportional hazards, (ii) non-proportional and non-crossing hazards, (iii) crossing hazards, and (iv) a mixed scenario. The specific survival functions are presented in Table 1. We set the group size for each scenario to 100; the censoring rates vary between 0% and 30% with uniform censoring. The work of [14] indicated that the choice of censoring distribution does not have a major impact on the performance of statistical tests. Considering all possible combinations of censoring, survival distributions, and contrast matrices, we end up with a total of 4(scenarios)x1120 parameter combinations = 4480 different settings. This is because we only considered the combination of different survival time distributions for the individual groups, but we did not consider the order. This means that S_1, S_1, S_1, S_2 is the same combination as S_1, S_1, S_2, S_1. For the Tukey-type contrast matrices, we considered every possible comparison for $k = 4$ that results in six tests. For the Dunnett-type contrast matrices, we compared the first group to all the other groups, resulting in 3 contrasts.

$10,000$ simulation runs with 1000 resampling iterations were performed for each setting. The global level of significance was set to 0.05 throughout.

Single CASANOVA? Not in multiple comparisons

Table 1: Simulation scenarios.

Scenario	CDF	Visualization of the survival and hazard curves
Prop	$F_1(t) = \textit{Exponential}(1.5)$ $F_2(t) = \textit{Exponential}(2.5)$ $F_1(t) = \textit{Exponential}(3.5)$ $F_2(t) = \textit{Exponential}(4.5)$	
NProp	$F_1(t) = \textit{Lognormal}(1.7, 1.7)$ $F_2(t) = \textit{Lognormal}(2.4, 1.6)$ $F_3(t) = \textit{Lognormal}(3.5, 1.7)$ $F_4(t) = \textit{Lognormal}(4.5, 1.6)$	
Cross	$F_1(t) = \textit{Weibull}(1.5, 5)$ $F_2(t) = \textit{Weibull}(2.5, 5)$ $F_3(t) = \textit{Weibull}(3.5, 5)$ $F_4(t) = \textit{Weibull}(4.5, 2.4)$	
Mix	$F_1(t) = \textit{Lognormal}(2.3, 1.7)$ $F_2(t) = \textit{Exponential}(0.05)$ $F_3(t) = \textit{Weibull}(2.4, 11.7)$ $F_4(t) = \textit{Lognormal}(3, 1.6)$	

Simulation results under the null hypothesis

Figure 1 illustrates the familywise error rate (FWER) for all survival scenarios for the different contrast matrix types. We set the α-level to 5%. The dashed lines represent the binomial confidence interval $[4.57\%, 5.43\%]$.

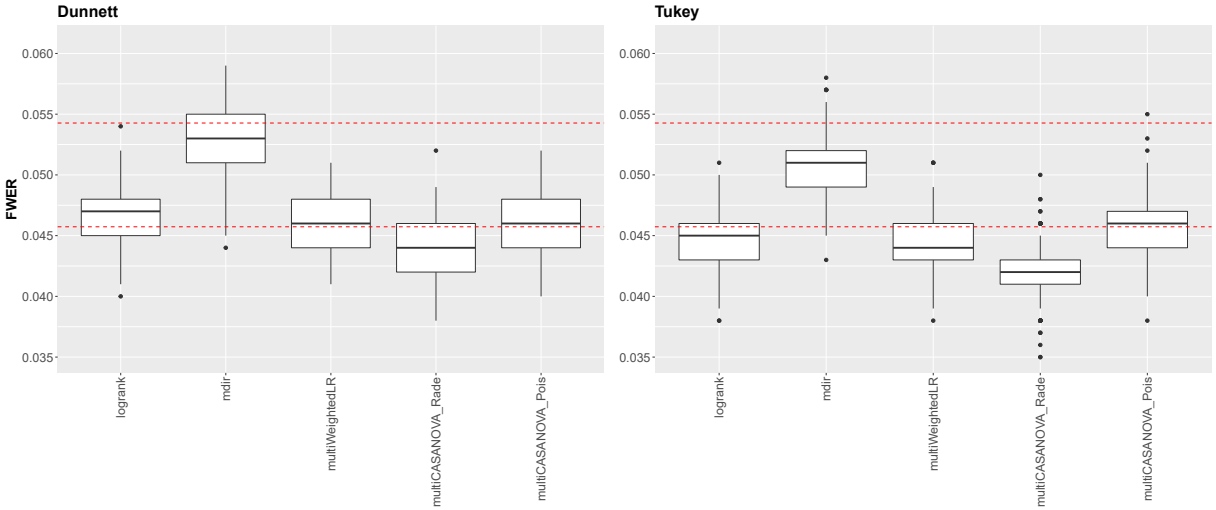


Figure 1: FWER under $\mathcal{H}_0$ for all settings for the Dunnett-type (left) and Tukey-type (right) contrast matrices. The dashed lines represent the borders of the binomial confidence interval $[4.57\%, 5.43\%]$

For both contrast matrices, almost all methods control the FWER well. The adjusted mdir is the only test that is a little liberal when comparing the median to the global α-level of 5% for the Dunnett-type matrix. The new multiple-testing approaches are more conservative than the adjusted approaches, especially for the Tukey-type contrast matrices, with the multiWeightedLR being the most conservative.

Simulation results under the alternative hypothesis

We focused on the local decisions under the alternative hypothesis to assess the power. Figures 2 and 3 illustrate the rejection rates when different survival distributions are present. Each figure consists of four subfigures, one for each scenario. It should be noted that a higher number of tests decreases the power of each local hypothesis. This property is visible in the plots showing generally higher power for the Dunnett plots than the Tukey plots. Besides that, the tests behave similarly for both contrast matrix types. The adjusted log-rank test is the most powerful in the setting with proportional hazards, while the other tests perform equally well. Under non-proportional but non-crossing hazards, all tests have a high power, with the new approaches yielding a slightly lower variability. In the crossing scenario, the log-rank test loses power drastically due to violating the PH assumption. The four approaches designed for nPH data have high power, with the multiWeightedLR being slightly less powerful than the other three tests. In the mixed setting, the adjusted mdir performs best in terms of power, followed by the three methods introduced in this paper. The log-rank test has the highest variability and the lowest median power.

The rejection rates for the local tests with no difference in survival are depicted in Figures 5 and 6 in the Supplemental Material. Overall, the rejection rates among the approaches are similar, with lower rejection rates for the Tukey-type contrast matrices. Additionally, the power for each local test is provided in the tables in the Supplemental Material.

The adjusted mdir test performs best for two of the four settings considered. Considering that it showed slightly liberal behavior under the null hypothesis, these results should be interpreted carefully. The methods introduced in this publication yield robust results regarding power among the different scenarios. The adjusted log-rank test loses power dramatically in the scenario with crossing hazards.

Single CASANOVA? Not in multiple comparisons

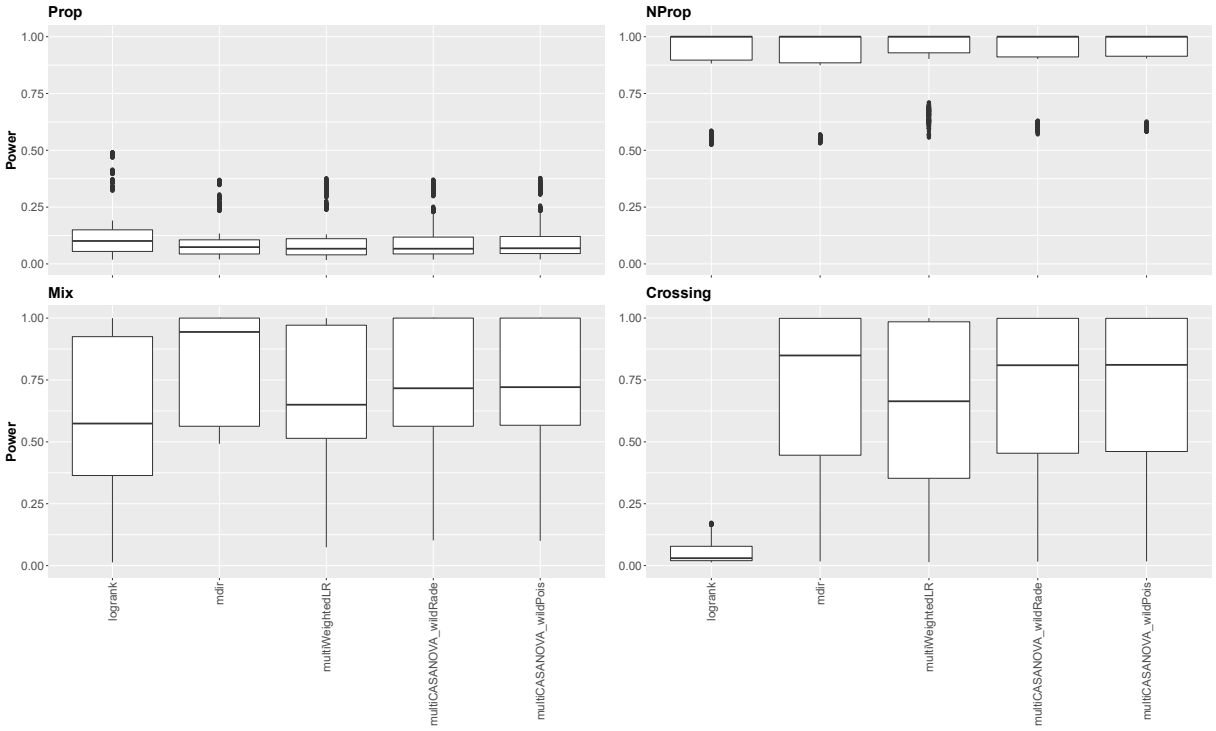


Figure 2: Local power over all tests under the alternative for Dunnett-type contrasts for all four scenarios (each boxplot contains 1136 data points).

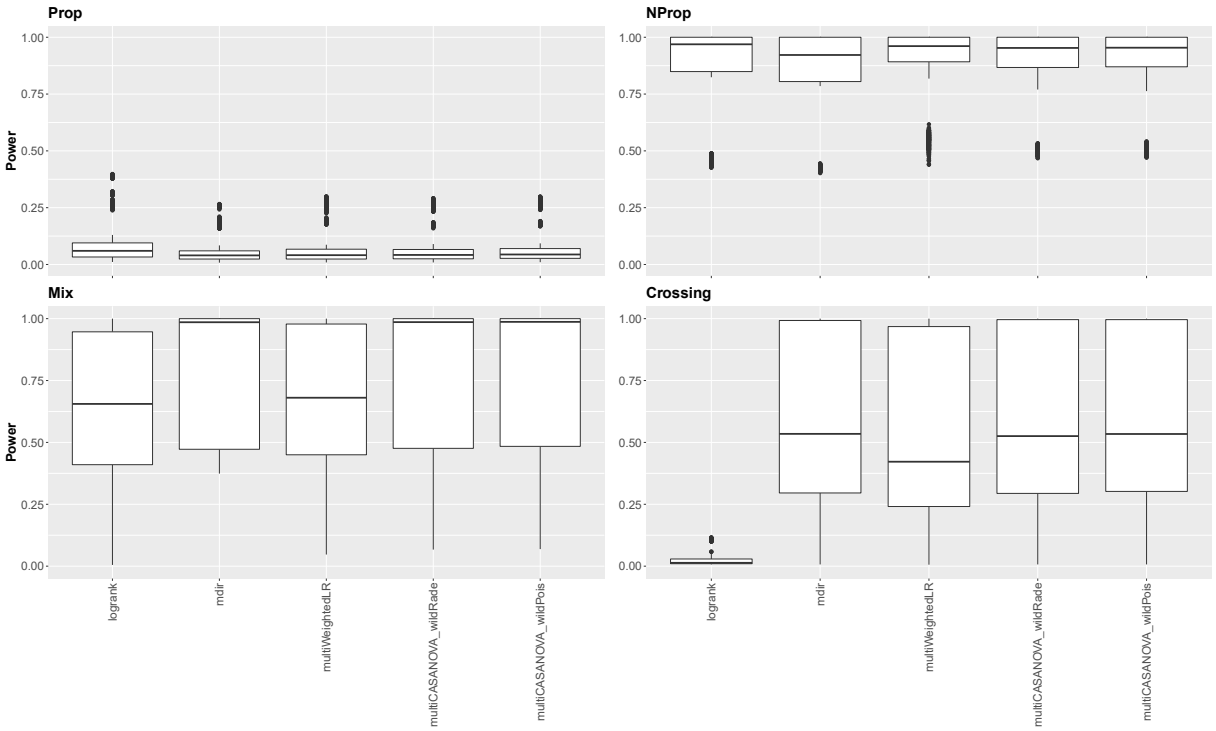


Figure 3: Local power over all tests under the alternative for Tukey-type contrasts for all four scenarios (each boxplot contains 2016 data points).

In the Appendix in Figure 7-9, we present an additional analysis of the behavior of the different tests concerning the FWER and power for smaller sample sizes ($n = 50$). The results indicate that the multiWeightedLR approach exhibits an inflated FWER, likely due to the normal approximation. In contrast, both multiCASANOVA bootstrap approaches maintain strong control over the family-wise error rate and consistently deliver good results in terms of power. The adjusted LR and mdir test still control the FWER but show increased variability in terms of power.

5 Illustrative Data Example

To illustrate the novel approaches on real-world data, we used publicly available data from the CoMMpass study (dbGaP accession: phs000748.v4.p3). This study is designed to associate clinical outcome with genetic profiles and contains longitudinal clinical and molecular data from multiple myeloma (MM) patients. Based on the transcriptional profile and the expression level of biologically relevant core machinery that plays a vital role in the stress response [30], we clustered MM patients into seven groups. Figure 4 shows the Kaplan-Meier curves of the seven different groups. We assume that we are interested in comparing every group with one another and consider a Tukey design. The significance level was set to $\alpha = 0.05$ with a total of 21 tests. The corrected significance level is thus $\alpha_{\text{Bonferroni}} = 0.0024$.

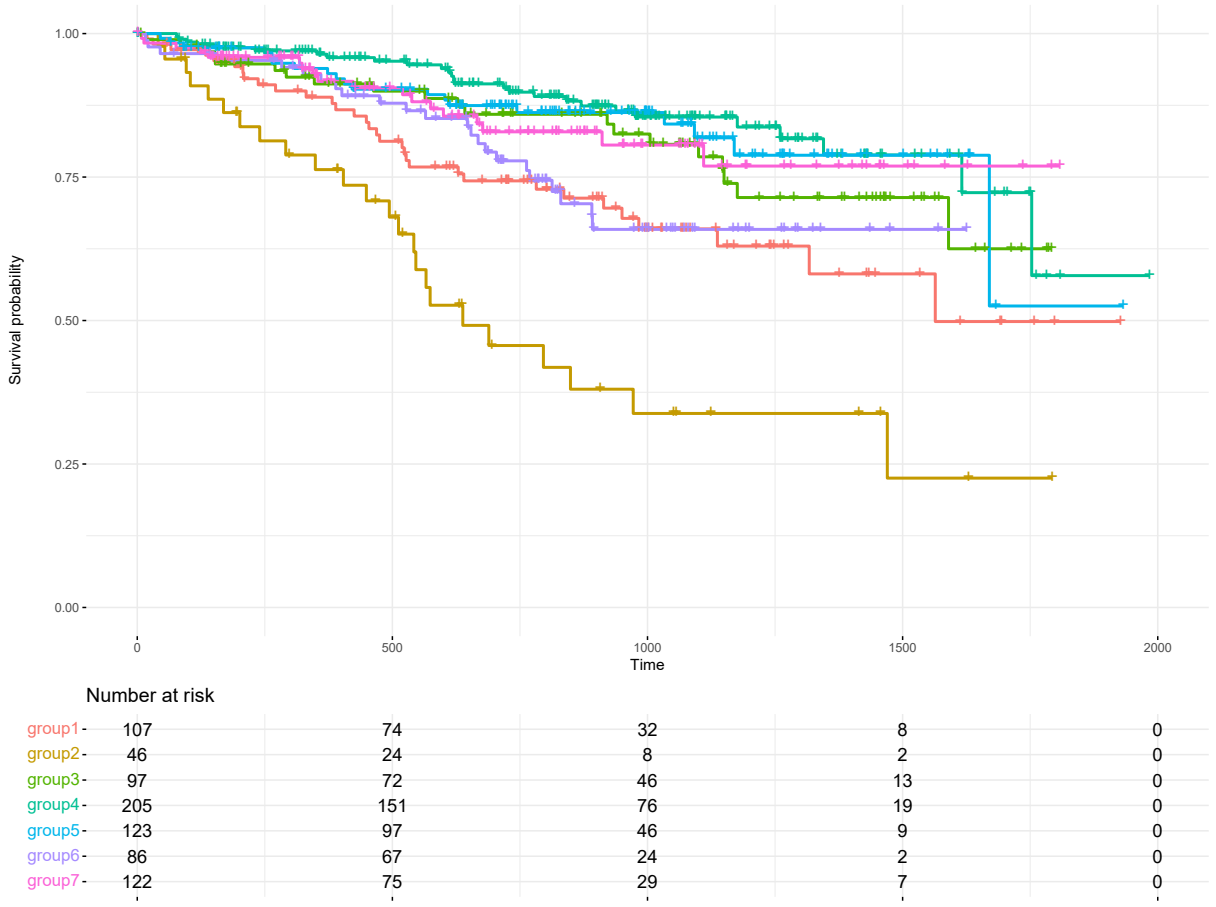


Figure 4: Kaplan-Meier plot of the seven treatment groups of patients with multiple myeloma (MM).

By examining the survival curves, we anticipate that the methods will identify a significant overall difference among the groups. Specifically, we expect group 2 to differ from the other groups. However, we do not expect to see any differences among groups 3, 4, and 5. We applied all testing procedures described in this paper to investigate these premises. For all approaches combining multiple weighted log-rank tests, we included the $w^{(1)} = 0$ and $w^{(2)} = 1 - 2\widehat{S}(t_i)$. We set the number of resampling iterations to 1000 for all resampling-based approaches.

The detailed results are listed in the Supplemental Material Table 3. The adjusted log-rank and mdir test detected six significant differences between groups, while the three new methods only detected five (see Table 2). The found differences are consistent among the methods. All tests found the pair-wise differences between groups two and three, four, five, and seven, as well as between groups one and four, to be significant. A significant result for the comparison between groups two and six was only found by the adjusted LR test and the adjusted mdir test.

Table 2: Number of rejected local null hypotheses for each of the tests

	logrank	**mdir**	**multiWeightedLR**	**multiCASANOVA_wildRade**	**multiCASANOVA_wildPois**
Nbr. of rejected $H_0^{j_1 j_2}$	6	6	6	5	5

All tests could reject the global hypothesis of any difference between groups as well. In summary, we could show that in the case of a real-world application, the results are consistent with the results of the adjusted log-rank test.

6 Discussion

We explored various statistical methods for addressing multiple contrast problems with time-to-event endpoints, including traditional and newly developed approaches. To assess the approaches' performance, we compared the Family-Wise Error Rate (FWER) control and the power performance of these methods under different survival scenarios. The results of our simulation study and real-world data application provide valuable insights into the strengths and limitations of each approach.

Most methods maintain adequate control of the FWER. The adjusted mdir test exhibited a slightly liberal behavior, particularly for Dunnett-type contrasts. This deviation suggests that while the adjusted mdir test might be powerful, it occasionally exceeds the acceptable error rate, which warrants caution in its interpretation under null conditions. On the other hand, the multiWeightedLR and multiCASANOVA methods were generally more conservative, particularly for Tukey-type contrast matrices. This conservativeness could imply a lower risk of Type I errors but may come at the cost of reduced statistical power.

Under alternative hypotheses, the power analysis revealed notable differences in the tests' performance depending on the survival scenario. For proportional hazards, the adjusted log-rank test showed the highest power, outperforming the other methods. Under non-proportional and non-crossing hazards, we could observe high power among all tests, with the new approaches showing slightly lower variability. The robustness of these methods suggests that they are suitable choices when proportional hazards are not guaranteed. The log-rank test's power decreased drastically in the specific case of crossing hazards. In contrast, the four approaches specifically designed for non-proportional hazards (adjusted mdir, multiWeightedLR, and the two multiCASANOVA variants) maintained high power, confirming their utility in these settings. Finally, the adjusted mdir test outperformed other methods in the mixed scenario, achieving the best power performance. Although slightly less powerful, the new methods provided more consistent results across different scenarios, highlighting their robustness.

The results suggest potential areas for further methodological improvements. While the Bonferroni correction is widely used for controlling type I error rates, its conservative nature may result in lower power, particularly in settings with many comparisons. More sophisticated adjustment techniques, like the Holm procedure, could better balance error rate control and power, as discussed in previous studies.

Additionally, evaluating the performance of these methods in unbalanced designs could provide a more comprehensive understanding of their behavior in practical applications. This could be particularly interesting since [12] showed that such conditions could boost the power of specific local designs.

In the illustrative data example involving patients with multiple myeloma, the new methods produced consistent results with those obtained from the adjusted log-rank tests. Although the novel approaches identified less significant differences than the traditional methods, their findings were largely aligned, underscoring their reliability in practical scenarios. This consistency and the conservative behavior in terms of FWER control suggest that the new methods still offer a robust alternative for analyzing time-to-event data in clinical studies. Future research would include more efficient exploitation of the FWER for the new approaches, e.g., by incorporating closed testing approaches. In general, it is essential to critically assess whether a higher number of statistically significant results truly reflects a superior testing approach, as statistical significance does not inherently equate to clinical relevance.

Acknowledgements

This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within https://uaruhr.de. The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359. Moreover, the work of Marc Ditzhaus and Markus Pauly was supported by the joined DFG Sachbeihilfe-project (project number 352692197). These data were generated as part of the Multiple Myeloma Research Foundation CoMMpass [SM] (Relating Clinical Outcomes in MM to Personal Assessment of Genetic Profile) study (www.themmrf.org).

In memory of our esteemed colleague, Marc Ditzhaus, who passed away in September 2024. His invaluable input will always be remembered, and he will be deeply missed.

References

[1] Maximilian Bardo, Cynthia Huber, Norbert Benda, Jonas Brugger, Tobias Fellinger, Vaidotas Galaune, Judith Heinz, Harald Heinzl, Andrew C Hooker, Florian Klinglmüller, et al. Methods for non-proportional hazards in clinical trials: A systematic review. *arXiv preprint arXiv:2306.16858*, 2023.

[2] Frank Konietschke, Sandra Bösiger, Edgar Brunner, and Ludwig A Hothorn. Are multiple contrast tests superior to the anova? *The International Journal of Biostatistics*, 9(1):63–73, 2013.

[3] Marc Ditzhaus, Jon Genuneit, Arnold Janssen, and Markus Pauly. CASANOVA: Permutation inference in factorial survival designs. *Biometrics*, pages 1–13, October 2021.

[4] Frank Konietschke, Ludwig A. Hothorn, and Edgar Brunner. Rank-based multiple test procedures and simultaneous confidence intervals. *Electronic Journal of Statistics*, 6, 2012.

[5] Brent R Logan, Hong Wang, and Mei-Jie Zhang. Pairwise multiple comparison adjustment in survival analysis. *Statistics in medicine*, 24(16):2509–2523, 2005.

[6] Xin Gao, Mayer Alvo, Jie Chen, and Gang Li. Nonparametric multiple comparison procedures for unbalanced one-way factorial designs. *Journal of Statistical Planning and Inference*, 138(6):2574–2591, 2008.

[7] Xin Gao and Mayer Alvo. Nonparametric multiple comparison procedures for unbalanced two-way layouts. *Journal of Statistical Planning and Inference*, 138(12):3674–3686, 2008.

[8] Frank Bretz, Alan Genz, and Ludwig A. Hothorn. On the numerical availability of multiple comparison procedures. *Biometrical Journal*, 43(5):645–656, 2001.

[9] Frank Schaarschmidt, Egbert Biesheuvel, and Ludwig A Hothorn. Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials. *Journal of Biopharmaceutical Statistics*, 19(2):292–310, 2009.

[10] Mario Hasler and Ludwig A Hothorn. Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(5):793–800, 2008.

[11] Paul Blanche, Jean-François Dartigues, and Jérémie Riou. A closed max-t test for multiple comparisons of areas under the roc curve. *Biometrics*, 78(1):352–363, 2022.

[12] Merle Munko, Marc Ditzhaus, Dennis Dobler, and Jon Genuneit. RMST-based multiple contrast tests in general factorial designs. *Statistics in Medicine*, 43(10):1849–1866, 2024.

[13] Ina Dormuth, Tiantian Liu, Jin Xu, Menggang Yu, Markus Pauly, and Marc Ditzhaus. Which test for crossing survival curves? a user's guideline. *BMC medical research methodology*, 22(1):1–7, 2022.

[14] Ina Dormuth, Tiantian Liu, Jin Xu, Markus Pauly, and Marc Ditzhaus. A comparative study to alternatives to the log-rank test. *Contemporary Clinical Trials*, 128:107165, 2023.

[15] Ivar Sonbo Kristiansen. Prm39 survival curve convergences and crossing: a threat to validity of meta-analysis? *Value in health*, 15(7):A652, 2012.

[16] Ludovic Trinquart, Justine Jacot, Sarah C Conner, and Raphaël Porcher. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, 34(15):1813–1819, 2016.

[17] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[18] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 625. John Wiley & Sons, 2013.

[19] Michael Brendel, Arnold Janssen, Claus-Dieter Mayer, and Markus Pauly. Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, 41(3):742–761, September 2014.

[20] Marc Ditzhaus and Sarah Friedrich. More powerful logrank permutation tests for two-sample survival data. *Journal of Statistical Computation and Simulation*, 90(12):2209–2227, 2020.

[21] Terry M Therneau. *A Package for Survival Analysis in R*, 2023. R package version 3.5-5.

[22] Marc Ditzhaus and Sarah Friedrich. *mdir.logrank: Multiple-Direction Logrank Test*, 2018. R package version 0.0.4.

[23] Frank Konietschke, Sandra Bösiger, Edgar Brunner, and Ludwig A Hothorn. Are Multiple Contrast Tests Superior to the ANOVA? *The International Journal of Biostatistics*, 9(1), January 2013.

[24] Marc Ditzhaus, Dennis Dobler, Markus Pauly, Philipp Steinhauer, and Merle Munko. *GFDsurv: Tests for Survival Data in General Factorial Designs*, 2022. R package version 0.1.1.

[25] Tobias Bluhmki, Dennis Dobler, Jan Beyersmann, and Markus Pauly. The wild bootstrap for multivariate nelson–aalen estimators. *Lifetime data analysis*, 25:97–127, 2019.

[26] Tobias Bluhmki, Claudia Schmoor, Dennis Dobler, Markus Pauly, Juergen Finke, Martin Schumacher, and Jan Beyersmann. A wild bootstrap approach for the aalen–johansen estimator. *Biometrics*, 74(3):977–985, 2018.

[27] Regina Y Liu. Bootstrap procedures under some non-iid models. *The annals of statistics*, 16(4):1696–1708, 1988.

[28] Enno Mammen. *When does bootstrap work?: asymptotic results and simulations*, volume 77. Springer Science & Business Media, 2012.

[29] R Core Team. R: A language and environment for statistical computing, 2021.

[30] Guus JJE Heynen, Francis Baumgartner, Michael Heider, Upayan Patra, Maximilian Holz, Jan Braune, Melanie Kaiser, Isabell Schäffer, Stefanos A Bamopoulos, Evelyn Ramberger, et al. Sumoylation inhibition overcomes proteasome inhibitor resistance in multiple myeloma. *Blood Advances*, 7(4):469–481, 2023.

Single CASANOVA? Not in multiple comparisons

Supporting Information

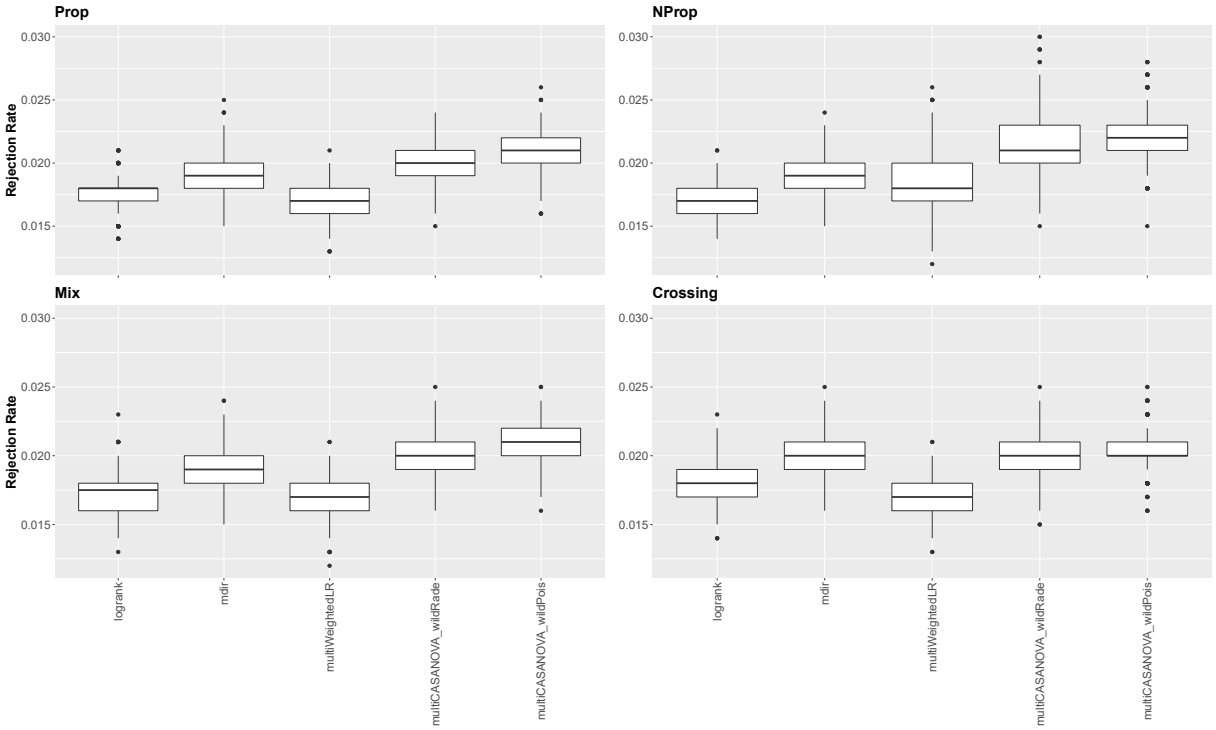


Figure 5: Rejection rate of the local tests with no difference in survival for the Dunnett-type matrix.

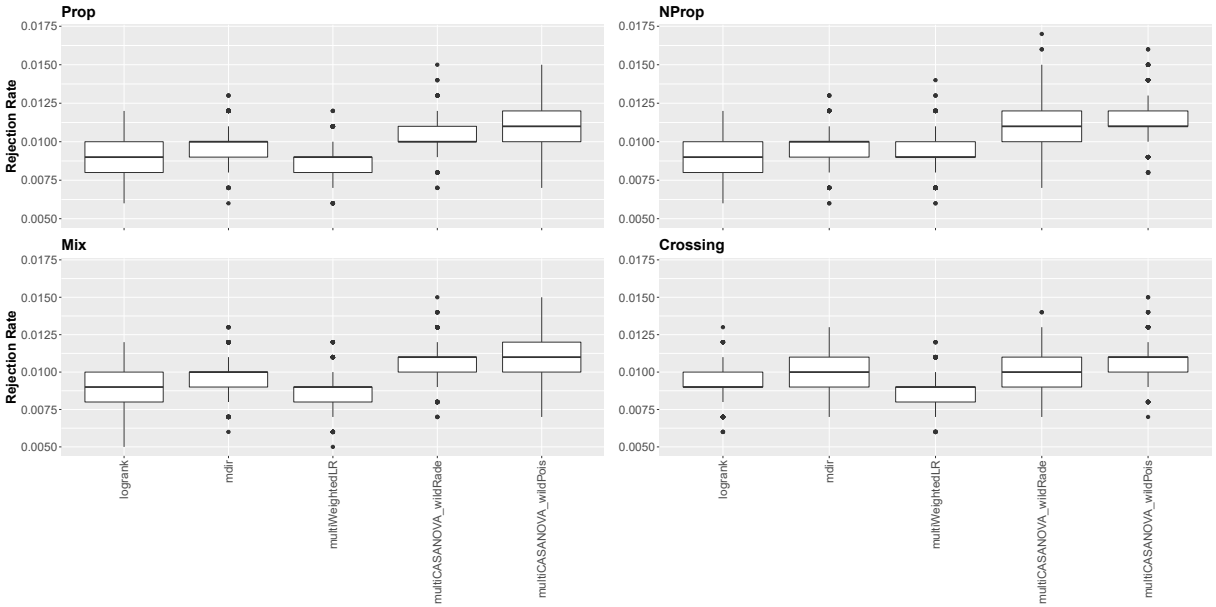


Figure 6: Rejection rate of the local tests with no difference in survival for the Tukey-type matrix.

Table 3: p-values for the local comparisons in the data example (Section 4). Significant tests to level α or $\alpha_{\text{Bonferroni}}$ are highlighted in bold font.

	log-rank	mdir	multiWeightedLR	multiCASANOVA_wildRade	multiCASANOVA_wildPois
2 - 1	0.196	0.411	0.130	0.301	0.304
3 - 1	0.100	0.152	0.415	0.699	0.727
4 - 1	**0.001**	**0.003**	**0.013**	**0.030**	**0.023**
5 - 1	**0.010**	**0.028**	0.135	0.203	0.210
6 - 1	0.502	0.515	0.981	1.000	1.000
7 - 1	0.040	0.136	0.366	0.571	0.566
3 - 2	**<0.001**	**0.001**	**0.006**	**0.012**	**0.005**
4 - 2	**<0.001**	**<0.001**	**0.001**	**0.003**	**<0.001**
5 - 2	**<0.001**	**<0.001**	**0.003**	**0.007**	**0.001**
6 - 2	**0.002**	**0.001**	**0.040**	0.092	**0.091**
7 - 2	**<0.001**	**<0.001**	**0.006**	**0.011**	**0.003**
4 - 3	0.406	0.217	0.821	0.980	0.991
5 - 3	0.612	0.759	0.999	1.000	1.000
6 - 3	0.242	0.534	0.929	0.985	0.995
7 - 3	0.889	0.943	1.000	1.000	1.000
5 - 4	0.885	0.990	0.956	0.981	0.993
6 - 4	0.034	0.046	0.175	0.415	0.414
7 - 4	0.777	0.768	0.760	0.922	0.942
6 - 5	0.195	0.329	0.649	0.776	0.789
7 - 5	0.959	1.000	0.999	1.000	1.000
7 - 6	0.424	0.248	0.907	0.941	0.959

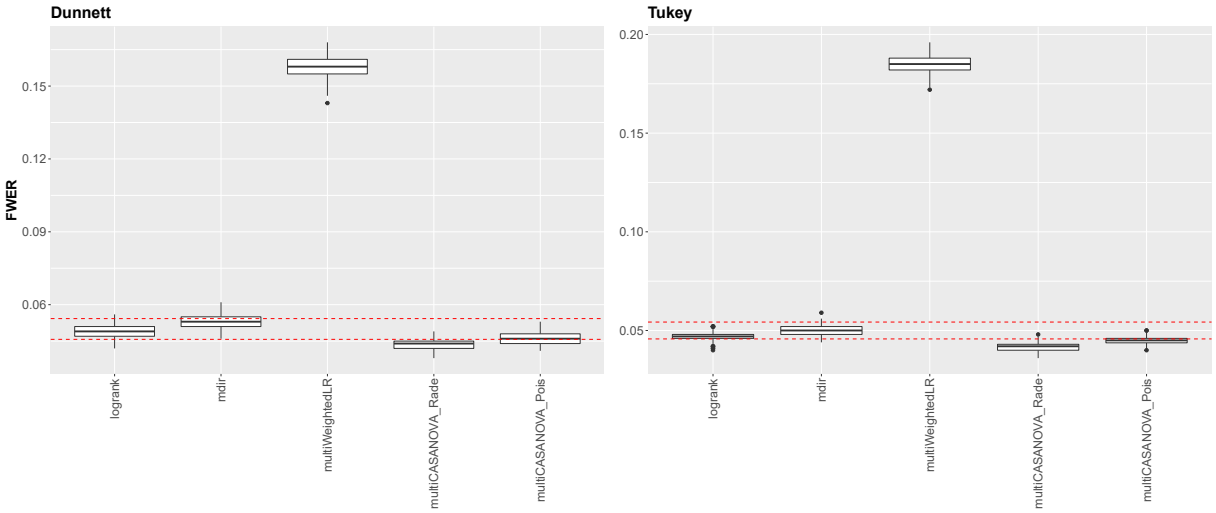


Figure 7: FWER under $\mathcal{H}_0$ for all settings for the Dunnett-type (left) and Tukey-type (right) contrast matrices for $n = 50$. The dashed lines represent the borders of the binomial confidence interval $[4.57\%, 5.43\%]$

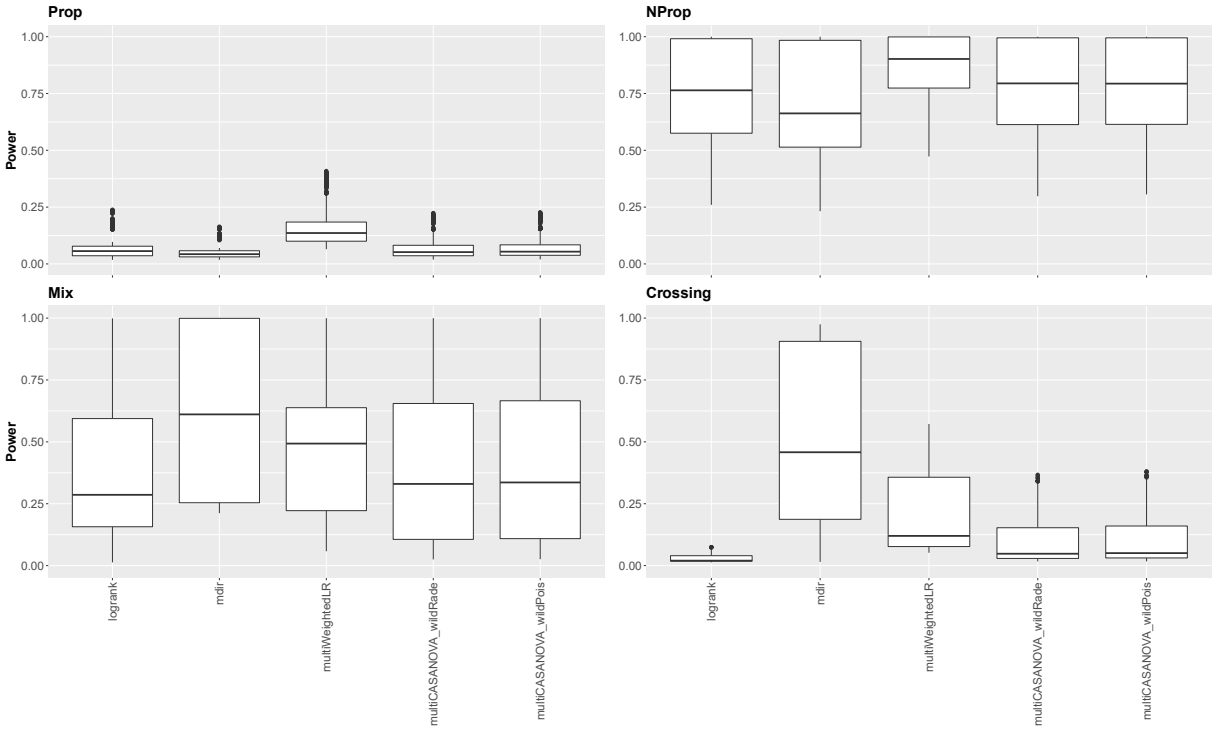


Figure 8: Local power over all tests under the alternative for $n = 50$ for Dunnett-type contrasts for all four scenarios (each boxplot contains 1136 data points).

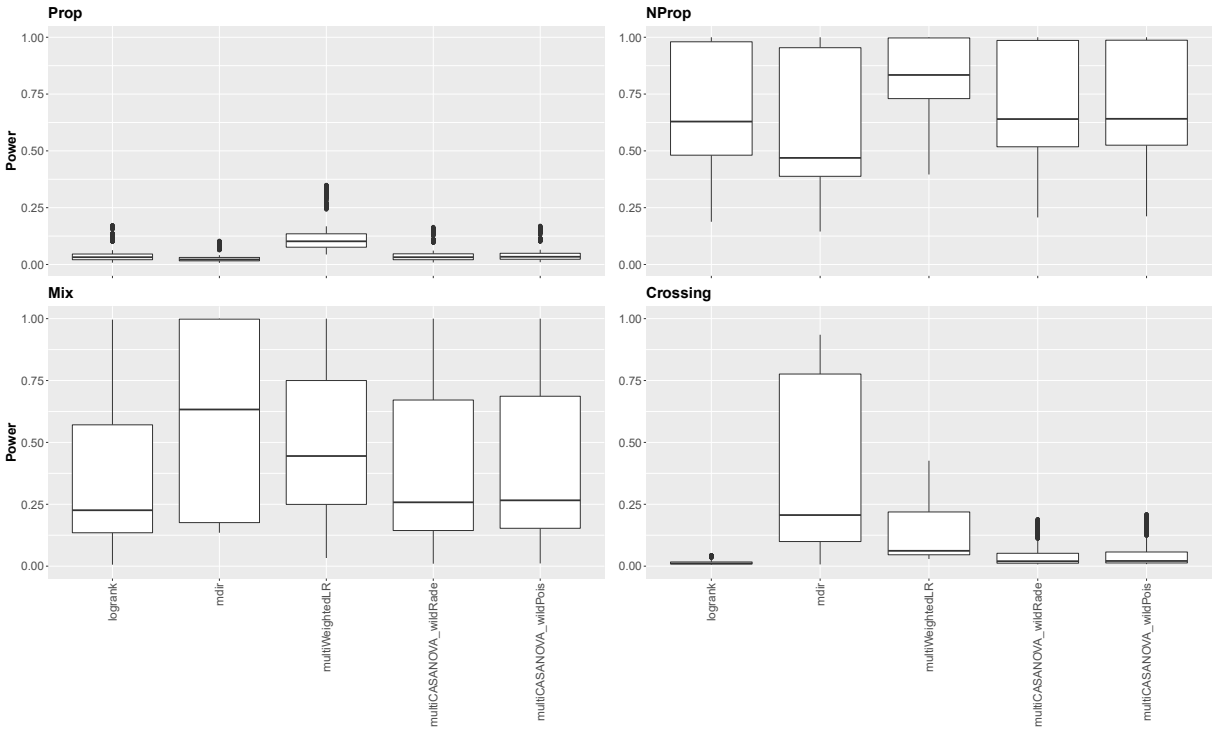


Figure 9: Local power over all tests under the alternative for $n = 50$ for Tukey-type contrasts for all four scenarios (each boxplot contains 2016 data points).

Article 5

Danzer M.F. & Dormuth, I. (2024). Adaptive weight selection for time-to-event data under non-proportional hazards. arXiv preprint arXiv:2409.15145.

Adaptive weight selection for time-to-event data under non-proportional hazards

Moritz Fabian Danzer[1] and Ina Dormuth[2]

[1]Institute of Biostatistics and Clinical Research, University of Münster, 48149 Münster, Germany
[2]Department of Statistics, TU Dortmund University, 44227 Dortmund, Germany

September 2024

Abstract

When planning a clinical trial for a time-to-event endpoint, we require an estimated effect size and need to consider the type of effect. Usually, an effect of proportional hazards is assumed with the hazard ratio as the corresponding effect measure. Thus, the standard procedure for survival data is generally based on a single-stage log-rank test. Knowing that the assumption of proportional hazards is often violated and sufficient knowledge to derive reasonable effect sizes is usually unavailable, such an approach is relatively rigid. We introduce a more flexible procedure by combining two methods designed to be more robust in case we have little to no prior knowledge. First, we employ a more flexible adaptive multi-stage design instead of a single-stage design. Second, we apply combination-type tests in the first stage of our suggested procedure to benefit from their robustness under uncertainty about the deviation pattern. We can then use the data collected during this period to choose a more specific single-weighted log-rank test for the subsequent stages. In this step, we employ Royston-Parmar spline models to extrapolate the survival curves to make a reasonable decision. Based on a real-world data example, we show that our approach can save a trial that would otherwise end with an inconclusive result. Additionally, our simulation studies demonstrate a sufficient power performance while maintaining more flexibility.

KEY WORDS: survival data, adaptive designs, conditional power, interim analysis, weighted log-rank tests, combination-type tests

1 Introduction

There are two common sources of uncertainty in the planning phase of a clinical trial with a survival endpoint. On the one hand, we need to identify the effect size on which case number planning is based. On the other hand, we must make assumptions about the type of effect. Traditionally, an effect of proportional hazards is considered, whereby the effect size is referred to as the hazard ratio. However, the assumption of proportional hazards must often be questioned. This is particularly the case if therapies with different mechanisms of action are to be compared.
Adaptive designs are widely accepted as a possible solution for the uncertainty regarding the effect size [1;2]. Such designs allow for interim analyses at multiple time points. In these interim analyses, the study can be terminated early, either with or without rejection of the null hypothesis, or the further course of the study can be adapted. This concerns, for example, the planning of future interim analyses or an adjustment of the sample size. For this purpose, the data collected to date on the endpoint to be investigated can be used as a basis for planning. This data can be used to calculate the conditional power of the design. The design is often adapted so that the conditional power reaches a specific target value. In particular, such an adaptive procedure allows us to revisit the assumptions regarding the effect size made at the beginning of the study and correct them if needed.

Stage-wise log-rank statistics are often employed in standard adaptive designs for survival data. Then, the initial sample size planning and the sample size recalculation are commonly based on the hazard

ratio that can also be re-estimated during the trial. This is, e.g., demonstrated in Wassmer (2006)[3]. The technical foundation is given in Tsiatis (1982)[4].
As in the adaptive (multi-stage) approach described above, the standard log-rank test is commonly applied in single-stage designs, with one single analysis at the end of the study. This is because, in proportional hazard situations, the log-rank test is the optimal test in terms of power. When this assumption is violated, it might, however, lose power and thus lead to poor test decisions. One way to increase the power for non-proportional hazard alternatives is by implementing a weight function. Multiple weight functions have been introduced over time[5;6]. More recent approaches facilitate the weight-choosing procedure and increase robustness against various hazard patterns by introducing combination-type tests[7–9]. One combination approach is the multi-directional log-rank (*mdir*) test, originally proposed by Brendel et al.[7] and revisited by Ditzhaus and Friedrich[8]. This Wald-type test statistic consists of multiple weighted log-rank tests covering several alternatives and their linear combinations. Dormuth et al.[10] illustrated the robust power behavior of the two-sided version of the *mdir*test. Ditzhaus et al.[11] extended the test for a one-sided testing problem.
Employing more robust testing procedures is also crucial in multi-stage designs[12]. Group sequential designs for weighted log-rank tests[13] and the *max combo* combination test[14] have already been proposed. One straightforward extension of such procedures is to select testing procedures for the forthcoming stages based on the available information. Of course, this only makes sense if the test statistics in each stage test the same hypothesis. Selecting the test depending on a proportional hazards check has already been considered[15]. However, it must be taken into account here that such two-stage procedures generally increase the significance level, and a corresponding adjustment must be made. An adaptation of the testing procedure for survival endpoints has only been considered in Lawrence (2002)[16]. We will go beyond this work in several respects:
Using a slightly more formal approach, we obtain second-stage test statistics explicitly as increments of the corresponding stochastic processes. An additional calculation of covariances is therefore not necessary due to the use of prominent asymptotic results[4;17]. Furthermore, this approach also allows combination tests in individual design stages. In particular, we would like to use these in the first stage to apply a robust test under uncertainty about the type of deviation. For the following stages, we use the information available to select the best weighted log-rank test regarding conditional power. Therefore, we present flexible calculations for this quantity. The full first-stage primary endpoint data can be used to guide this decision. In addition, the basic design still allows for adjustments to the sample size or changes to the analysis schedule.
The manuscript is organized as follows. In Section 2, we give a brief overview of the methods and techniques we will combine to construct our adaptive testing procedure, which will be presented in Section 3. We illustrate its application on a reconstructed data set in Section 4 and study its characteristics via simulation in Section 5. Finally, in section 6, we summarize our findings, discuss them, and give outlooks for further research. Further material on our application example, the simulation study, and technical foundations are provided in the Supplementary Material.

2 Methods

This section will present various methods and techniques we require to formulate our proposed method. We will start by introducing some notation.
Any patient $i \in \{1, \dots, n\}$ enters the trial at the random time $R_i \geq 0$ in calendar time. That patient is assigned to treatment group $Z_i \in \{0, 1\}$. The patient experiences the event of interest at the random time $T_i \geq 0$ after recruitment. However, the observation of this event may be censored. This may either happen because of a random dropout or administrative censoring. The former occurs at $C_i^\star$. The latter depends on the time of analysis. If this analysis is performed at calendar time t, the latter censoring is given by $(t - R_i)_+$. Overall, censoring at this calendar date is thus given by $C_i(t) \coloneqq C_i^\star \wedge (t - R_i)_+$ where $\wedge$ indicates the minimum of two real numbers. The observation that can hence be made at that time is the probably censored event date $X_i(t) \coloneqq T_i \wedge C_i(t)$ and the corresponding indicator function $\delta_i(t) \coloneqq \mathbb{1}_{\{T_i \leq C_i(t)\}}$.
We assume that the tuples $(R_i, Z_i, C_i^\star, T_i)$ are independent and identically distributed for all $i \in \{1, \dots, n\}$. In particular, they shall be independent replicates of some tuple $(R, Z, C^\star, T)$. Furthermore, the censoring through $C(t)$ shall constitute an independent censoring mechanism and we assume that R and $C^\star$ are independent random variables with cumulative distribution functions F_R and $F_{C^\star}$, respectively.
Based on this, we can now define counting processes and at-risk indicators for the event of interest. For

any $i \in \{1, \dots, n\}$, the multivariate process $(N_i(t,s))_{t\geq 0, s\geq 0}$ defined by

$$N_i(t,s) \coloneqq \mathbb{1}_{\{T_i \leq s \wedge C_i(t)\}}$$

indicates whether the event of interest was observed before calendar time t and before patient i spent s time units in the study. These processes can be aggregated over the complete study sample to obtain the overall number of events $N(t,s) \coloneqq \sum_{i=1}^{n} N_i(t,s)$ observed before calendar time t and trial time s. Additionally, we define those processes in the subgroups of patients that are assigned to the same treatment, i.e.

$$N^{Z=k}(t,s) \coloneqq \sum_{i=1}^{n} N_i(t,s) \cdot \mathbb{1}_{\{Z_i=k\}}$$

for treatment group $k \in \{0,1\}$.
Similarly, the multivariate process $(Y_i(t,s))_{t\geq 0, s\geq 0}$ indicates whether it is known at calendar time t that patient $i \in \{1, \dots, n\}$ remained event-free in the trial just before s time units after its enrollment, i.e.

$$Y_i(t,s) \coloneqq \mathbb{1}_{\{N_i(t,s-)=0\}} \cdot \mathbb{1}_{\{s \leq C_i(t)\}}$$

The term $N_i(t,s-)$ denotes the left-hand limit in the second argument, i.e.

$$N_i(t,s-) \coloneqq \lim_{u \nearrow s} N_i(t,u).$$

As above, we can aggregate these quantities over the complete study sample or the two treatment groups to obtain the processes $(Y(t,s))_{t\geq 0, s\geq 0}$ respectively $(Y^{Z=k}(t,s))_{t\geq 0, s\geq 0}$ for $k \in \{0,1\}$.
Given these quantities, we can estimate the pooled and group-specific survival functions $S, S_k \colon [0,\infty) \to [0,1]$ of T which are defined by $S(s) \coloneqq \mathbb{P}[T \geq s]$ and $S_k(s) \coloneqq \mathbb{P}[T \geq s | Z = k]$, respectively, using the information collected up to calendar time t by

$$\hat{S}(t,s) \coloneqq \prod_{i \colon X_i(t) \leq s} \left(1 - \frac{\delta_i(t)}{Y(t, X_i(t))}\right)$$

resp.

$$\hat{S}_k(t,s) \coloneqq \prod_{i \colon X_i(t) \leq s, Z_i = k} \left(1 - \frac{\delta_i(t)}{Y^{Z=k}(t, X_i(t))}\right)$$

for $k \in \{0,1\}$. Corresponding cumulative distribution functions F, F_k and estimators $\hat{F}, \hat{F}_k$ are given by the probabilities of the respective complementary events.
The corresponding cumulative hazard functions $A, A_k \colon [0,\infty) \to [0,\infty)$ defined by $A(t) \coloneqq -\log(S(t))$ and $A_k(t) \coloneqq -\log(S_k(t))$, respectively, can be estimated at calendar time t by the Nelson-Aalen estimates

$$\hat{A}(t,s) \coloneqq \int_{[0,s]} \frac{\mathbb{1}_{\{Y(t,u)>0\}}}{Y(t,u)} dN(t,u) = \sum_{i \colon X_i(t) \leq s} \frac{\delta_i(t)}{Y(t, X_i(t))}$$

resp.

$$\hat{A}_k(t,s) \coloneqq \int_{[0,s]} \frac{\mathbb{1}_{\{Y^{Z=k}(t,u)>0\}}}{Y^{Z=k}(t,u)} dN^{Z=k}(t,u) = \sum_{i \colon X_i(t) \leq s, Z_i = k} \frac{\delta_i(t)}{Y^{Z=k}(t, X_i(t))}$$

for $k \in \{0,1\}$.
In what follows, we present testing procedures to investigate the null hypothesis of equal survival distributions

$$H_0 \colon \{A_0 \equiv A_1\} = \{S_0 \equiv S_1\}. \tag{1}$$

It will be tested against the one-sided alternative of superiority

$$H_{\geq} \colon \{A_0 \geq A_1, A_0 \not\equiv A_1\} = \{S_0 \leq S_1, S_0 \not\equiv S_1\}. \tag{2}$$

In particular, weighted log-rank tests and combination tests can be applied to address this issue. These will be presented in the following subsections.

2.1 Weighted log-rank tests

Weighted log-rank tests allow giving more emphasis to different parts of the survival function depending on the selection of appropriate weights[18]. The non-standardized test statistic at calendar time t is defined as

$$T_{\hat{Q}}(t) \coloneqq n^{-\frac{1}{2}} \sum_{i:\, X_i(t)\leq s, \delta_i(t)=1} \hat{Q}(t, X_i(t)) \underbrace{\left(Z_i - \frac{Y^{Z=1}(t, X_i(t))}{Y(t, X_i(t))}\right)}_{\text{observed event}-\text{expected event}}, \tag{3}$$

with some weight function $\hat{Q}\colon [0,\infty)^2 \to \mathbb{R}$ that fulfills the standard assumptions for repeated testing with log-rank type tests[4;5] (see also Supplementary Material, Section A). In particular, its large sample limit Q shall not depend on calendar time t and hence be only a function of s. Additionally, we require the weight functions to be positive such that the test is valid to test the null hypothesis (1) against the one-sided alternative from (2). A set of weight functions that satisfies these conditions are referred to as the Fleming-Harrington weights. The weight functions are based on pooled Kaplan-Meier estimates and parameters $\rho, \gamma \geq 0$:

$$\hat{Q}(t,s) = w^{(\rho,\gamma)}(\hat{F}(t,s-)) \coloneqq \hat{F}(t,s-)^{\rho} \cdot \hat{S}(t,s-)^{\gamma}. \tag{4}$$

The test statistic can be standardized using the corresponding variance. For big enough sample sizes, the standardized test statistic is approximately standard normally distributed under the null. The p-values are then obtained employing the associated tables.

We obtain the classical log-rank test when setting $\rho = \gamma = 0$. The resulting test is optimal regarding power under proportional hazards[18]. Higher values of ρ result in a weight function that gives higher weights to events that occur later in time relative to earlier events. On the other side, higher values of γ emphasize events that occur earlier in time. When both values are high, events that occur in the middle are given more weight than early and late events. The challenge in utilizing weighted log-rank tests often lies in selecting meaningful weights.

2.2 mdir

One way to simplify the weight selection procedure is combining multiple weighted log-rank tests. Such methods allow using several weighted log-rank tests in one testing approach[7–9]. One way of combining multiple weighted tests is in a Wald-type test statistic[7;8]. In the following, we define the multi-directional log-rank test (*mdir*) in terms of the multivariate process

$$\mathbf{T}_{\hat{\mathcal{Q}}}(t) \coloneqq (T_{\hat{Q}_1}(t), \ldots, T_{\hat{Q}_m}(t))$$

where

$$T_{\hat{Q}_\ell}(t) \coloneqq n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_{[0,t]} \hat{Q}_\ell(t,s) \left(Z_i - \frac{Y^{Z=1}(t,s)}{Y(t,s)}\right) dN_i(t,s),$$

for a set of weights $\hat{\mathcal{Q}} \coloneqq \{\hat{Q}_1, \ldots, \hat{Q}_m\}$. This process is asymptotically equivalent to a multivariate martingale and has asymptotically independent and jointly normally distributed increments (expanding on Tsiatis, 1982[4]). These properties are essential for the adaptive approach introduced in Subsection 3. In adaptive designs, we are only interested in one-sided testing. This allows us to avoid situations where we obtain a final significant result but with effects pointing in different directions in subsequent stages. Thus, in the following, we focus on the one-sided test statistic of the *mdir* test.
Ditzhaus and Pauly (2019)[11] derived the one-sided test statistic from the initial set of weights. The main idea is to restrict the space in which the test statistic is spanned to only positive values. This results in the test statistic

$$W(t) \coloneqq \max\{0, \mathbf{T}_{\hat{\mathcal{L}}}(t)^T \hat{\mathbf{\Sigma}}^{-}_{\hat{\mathcal{L}}}(t) \mathbf{T}_{\hat{\mathcal{L}}}(t) \colon \emptyset \neq \hat{\mathcal{L}} \subseteq \hat{\mathcal{Q}}; \hat{\mathbf{\Sigma}}^{-}_{\hat{\mathcal{L}}}(t) \mathbf{T}_{\hat{\mathcal{L}}}(t) \geq 0\},$$

with $\mathbf{T}_{\hat{\mathcal{L}}}(t) = (T_{\hat{Q}}(t))_{\hat{Q}\in\hat{\mathcal{L}}}$ for all $t \geq 0$. The covariance matrix can be consistently estimated by $\hat{\mathbf{\Sigma}}(t)$. Its entries are given by

$$(\hat{\mathbf{\Sigma}}(t))_{\ell\ell^\star} = n^{-1} \sum_{i=1}^{n} \int_{[0,t]} \hat{Q}_\ell(t, X_i(t)) \hat{Q}_{\ell^\star}(t, X_i(t)) \frac{Y^{Z=1}(t,s)}{Y(t,s)} \left(1 - \frac{Y^{Z=1}(t,s)}{Y(t,s)}\right) dN_i(t,s).$$

Furthermore, $\hat{\mathbf{\Sigma}}(t)^{-}$ denotes the Moore-Penrose inverse of $\hat{\mathbf{\Sigma}}(t)$.
The authors state that an asymptotic limit distribution was not derived and propose a wild bootstrap approach with Rademacher weights instead. For more details, see Ditzhaus and Pauly (2019)[11]. The test is implemented in the R package `mdir.logrank`.[19]

2.3 Royston-Parmar splines

As described at the beginning of this section, the pooled and group-specific survival functions S, S_k can be estimated non-parametrically by Kaplan-Meier estimators. Such a non-parametric estimation can potentially be inefficient compared to a parametric estimation approach if the distribution of T resp. $T|Z = k$ lies within the parametric family assumed for the estimation process. Additionally, a parametric approach allows extrapolation of the survival curve beyond the time horizon that is present in the data. This is because the estimated parameters directly specify a distribution on $[0, \infty)$.
Despite the potential efficiency, such parametric approaches can be criticized for being too restrictive in terms of the shape of the distribution. Hence, important characteristics of the survival mechanism could potentially not be captured[20]. Consequently, it is difficult to quantify differences among treatment groups. For example, estimation within the family of exponential distributions in several treatment groups directly leads to the assumption of proportional hazards.
As a more flexible alternative, Royston & Parmar introduced an estimation procedure based on natural cubic splines[21]. A transformation of the survival function S, given by a link function $g : [0, 1] \to \mathbb{R}$, is modeled by

$$g(S(t; \mathbf{z})) = g(S_{\text{baseline}}(t)) + \boldsymbol{\beta}^T \mathbf{z} = s(x; \phi) + \boldsymbol{\beta}^T \mathbf{z}$$

where $S(t; \mathbf{z})$ denotes the survival distribution given covariates $\mathbf{z}$ and $x = \log(t)$. The function $s\colon \mathbb{R} \times \mathbb{R}^{p+2} \to \mathbb{R}$ is a natural cubic spline which is parameterized by $\phi \in \mathbb{R}^{p+2}$. The number of internal knots p determines the number of polynomials employed in the cubic spline.
When fitting such a model, it is suggested to place the boundary knots at the smallest and the largest uncensored logarithmized survival time and the internal knots evenly at the centiles of the uncensored logarithmized survival times[21]. For example, for $p = 3$, one would place the internal knots at the two quartiles and the median of the uncensored logarithmized survival times.
For the scale on which the function should be modeled, three prominent suggestions have been proposed in the literature[21;22] that are also implemented in software[23]. They are shown in Table 1.
The choice of p and the scale can be guided by criteria such as the AIC (Akaike information criterion)

scale	link function $g(u)$
hazard	$\log(-\log(u))$
odds	$\log(1/u - 1)$
normal	$-\Phi^{-1}(u)$

Table 1: Popular link functions for Royston-Parmar splines; naming according to the function `flexsurvspline` in the R package `flexsurv`[23]

or BIC (Bayesian information criterion). However, it is warned against using these criteria mechanically, and an informal choice based on the appearance of the fitted survival functions is also suggested[21;22]. Simulation studies have shown that the correct choice of knots is not mandatory because the splines are flexible enough if a sufficient number of knots is used[24].
Like simple parametric models, Royston-Parmar splines admit an extrapolation of the survival curve beyond the available time horizon. It should be mentioned here that the transformed survival function beyond the upper boundary knot is linear.

2.4 Adaptive designs

We briefly outline some cornerstones for constructing group-selective adaptive trial designs with unblinded interim analyses. For the sake of simplicity, we restrict ourselves to two-stage designs with one interim and one final analysis. In these two stages, we define a test statistic and a corresponding p-value to test a null hypothesis H_0. These p-values will be denoted by p_1 and p_2, respectively. At best, these are independent and uniformly distributed on $[0, 1]$. However, it is possible to relax this assumption to what is known as the *p-clud* property[25]. This property is fulfilled if

$$\mathbb{P}_{H_0}[p_1 \leq u] \leq u \quad \text{and} \quad \mathbb{P}_{H_0}[p_2 \leq u | p_1 = v] \leq u \qquad \forall 0 \leq u, v \leq 1.$$

If this is warranted, an adaptive design can be defined by a combination function $C\colon [0, 1]^2 \to [0, 1]$ that is non-decreasing in p_1 and p_2 and continuous in p_2, bounds α_0 and α_1 for stopping of the trial in the interim analysis and a critical value c for the combined p-value in the final analysis. In such a design, the trial will stop with the rejection of H_0 in the interim analysis if $p_1 \leq \alpha_1$, and it stops for futility, i.e., with

acceptance of H_0 if $p_1 \geq \alpha_0$. The null hypothesis can be rejected at the final analysis if $C(p_1, p_2) \leq c$ and otherwise it stops without rejection of H_0. It adheres to the nominal type I error level α if

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbb{1}_{\{C(p_1,p_2)\leq c\}} dp_1\, dp_2. \tag{5}$$

Popular choices for C are the combination function arising from Fisher's product test $C(p_1, p_2) = p_1 p_2$ [26;27] and the inverse normal combination function [28]

$$C(p_1, p_2) = 1 - \Phi\left(w_1 \cdot \Phi^{-1}(1 - p_1) + w_2 \cdot \Phi^{-1}(1 - p_2)\right) \tag{6}$$

where Φ and Φ^{-1} denote the cumulative distribution function and the quantile function of the standard normal distribution, respectively, and the weights $w_1, w_2 \geq 0$ underly the constraint $w_1^2 + w_2^2 = 1$.
The choice of the sequential decision bounds will determine the values of α_1 and c. One can use standard bounds according to the sequential plans of Pocock [29] or O'Brien and Fleming [30] or some α-spending approach [31].
All the methods mentioned here are implemented in the comprehensive R package `rpact` [32]. More details on group-sequential adaptive designs going far beyond what we sketched here can be found in Wassmer and Brannath (2016) [2].
The feature of adaptive designs that needs to be emphasized again for our purposes is the capability to redesign the second stage using the data collected on the primary endpoint up to the interim analysis.

3 Adaptive testing procedure

In this section, we present our adaptive testing procedure. While focusing on technical aspects here, it will also be presented in an example based on real data in Section 4.
To put it briefly, we conduct a multi-stage adaptive design that addresses the uncertainty about the type of effect by application of a combination testing procedure in the first stage and uses the information from this early stage to choose a well-suited weighted log-rank test for later stages. For the sake of simplicity, we will restrict ourselves to considering two-stage designs. However, an extension to more than two stages follows straightforward.
At the beginning of the trial, we fix two sets of weight functions $\hat{\mathcal{Q}}_{\text{mdir}} \coloneqq \{\hat{Q}_{\text{mdir},1}, \ldots, \hat{Q}_{\text{mdir},m_1}\}$ and $\hat{\mathcal{Q}}_{\text{cand}} \coloneqq \{\hat{Q}_{\text{cand},1}, \ldots, \hat{Q}_{\text{cand},m_2}\}$ and its union $\hat{\mathcal{Q}}_{\text{all}} \coloneqq \hat{\mathcal{Q}}_{\text{mdir}} \cup \hat{\mathcal{Q}}_{\text{cand}}$. The only things that need to be ensured are that all weights meet standard conditions required for repeated significance testing with log-rank type tests [5] (see also Supplementary Material, Section A) and linear independence of the weights in $\hat{\mathcal{Q}}_{\text{mdir}}$ [11].
Additionally, we set an interim analysis date t_1. Based on the results shown in Section A of the Supplementary Material, the calendar time process

$$(\mathbf{T}_{\hat{\mathcal{Q}}_{\text{all}}}(t))_{t\geq 0}, \text{ with } \mathbf{T}_{\hat{\mathcal{Q}}_{\text{all}}}(t) = (T_{\hat{Q}}(t))_{\hat{Q}\in\hat{\mathcal{Q}}_{\text{all}}}\ \forall t \geq 0.$$

is asymptotically equivalent to a multi-dimensional Gaussian process. In particular, its asymptotically independent increments asymptotically follow a joint normal distribution.
Additionally, we fix decision bounds for the first stage α_0 (futility) and α_1 (efficacy), a combination function C and a decision bound c for the combined p-value. These quantities are chosen such that (5) is maintained for the prefixed type 1 error rate α. The analyses shall take place at calendar times t_1 (interim analysis) and t_2 (final analysis).
At the time of the interim analysis, the weighted testing procedure for the next stage is determined. Therefore, let D be a random variable that assumes values in $\{1, \ldots, m_2\}$ and is measurable w.r.t. the information about the primary endpoint collected up to the interim analysis.
In the first stage, we obtain the test statistic and p-value

$$\begin{aligned} S_1 &\coloneqq \max\{0, \mathbf{T}_{\hat{\mathcal{L}}}(t_1)^T \hat{\mathbf{\Sigma}}^-_{\hat{\mathcal{L}}}(t_1)\mathbf{T}_{\hat{\mathcal{L}}}(t_1) \colon \emptyset \neq \hat{\mathcal{L}} \subseteq \hat{\mathcal{Q}}_{\text{mdir}}; \hat{\mathbf{\Sigma}}^-_{\hat{\mathcal{L}}}(t_1)\mathbf{T}_{\hat{\mathcal{L}}}(t_1) \geq 0\} \quad \text{resp.} \\ p_1 &\coloneqq \inf\{p \in [0,1] \colon q^G_{1-p} \geq S_1\} \end{aligned}$$

where q^G_{1-p} is the quantile of the distribution resulting from the wild bootstrap procedure. Together with the second stage test statistic and p-value

$$\begin{aligned} S_{2,\hat{Q}_{\text{cand},D}} &\coloneqq (T_{\hat{Q}_{\text{cand},D}}(t_2) - T_{\hat{Q}_{\text{cand},D}}(t_1))/\sqrt{\hat{\Sigma}_{\hat{Q}_{\text{cand},D}}(t_2) - \hat{\Sigma}_{\hat{Q}_{\text{cand},D}}(t_1)} \quad \text{resp.} \\ p_2 &\coloneqq 1 - \Phi(S_2), \end{aligned} \tag{7}$$

we obtain an adaptive testing approach that asymptotically keeps the type I error level α when applied as described in 2.4. I.e., we stop the trial for futility at the interim analysis if $p_1 > \alpha_0$, we reject H_0 at the interim analysis if $p_1 \leq \alpha_1$ and we can reject the trial at the final analysis if neither of those is the case and $C(p_1, p_2) \leq c$.
Although the design would also permit sample size recalculation (e.g., by extending the recruitment period) and concomitantly postpone the final analysis date, we only consider adapting the weight in the testing procedure.

3.1 Determination of the second stage test statistic

The critical question here concerns the choice of test statistics in the second stage. As mentioned above, we can use all the information about the primary endpoint collected up to the interim analysis. Of course, this consideration is only necessary if we proceed to a second stage, i.e., if $\alpha_1 < p_1 \leq \alpha_0$. Then, we suggest the following procedure:
Based on the p-value p_1 of the first stage, we can compute the conditional error probability

$$\tilde{\alpha}_2 \coloneqq \mathbb{P}_{H_0}[C(p_1, p_2) \leq c | p_1] = \int_0^1 \mathbb{1}_{\{C(p_1,u)\leq c\}} du = \sup\{u \in [0,1] \colon C(p_1, u) \leq c\}.$$

If $p_2 < \tilde{\alpha}_2$, we will reject (1) in favour of (2) at the final analysis. Our aim is, therefore, to estimate the probability of this event. The associated considerations, which we now present, are based on the calculations shown in Yung and Liu (2020)[33].
We fit a Royston-Parmar spline model for each group as presented in Section 2.3. Please note that we do not fit one joint model for the two groups. This would mean that the difference between the groups would follow a fixed pattern, e.g., a pattern of proportional hazards if the hazard scale was chosen. The hyperparameters of the spline model (number of knots, scale) are chosen based on information criteria or based on visual inspection (see Section 4 for exemplary applications). To highlight quantities based on estimates or assumptions made at the interim analysis, we will add a tilde to all of them. Hence, we denote the resulting pooled and group-specific survival, density and hazard functions by $\tilde{S}$ and $\tilde{S}_k$, $\tilde{f}$ and $\tilde{f}_k$, and $\tilde{\lambda}$ and $\tilde{\lambda}_k$, respectively. We assume we can identify the large sample limit Q of $\hat{Q}$ under knowledge of the true distribution of T and $T|Z = k$. This is, e.g., the case for the Fleming-Harrington weights presented in (8). We do not know the true distribution, so we insert $\tilde{S}_k$ instead and denote the resulting weight functions by $\tilde{Q}$.
The same could be done to assess the distribution of the recruitment date R and random dropout $C^\star$. We denote our planning assumptions about the distribution functions of these two random variables at the time of the interim analysis by $\tilde{F}_R$ and $\tilde{F}_{C^\star}$, respectively. Based on those assumptions, the probability that some individual is allocated to treatment group k and has spent at least s time units at risk (i.e., without any censoring and without experiencing the event of interest) at calendar time t is given by

$$\tilde{\pi}_k(t, s) \coloneqq \mathbb{P}[Z = k]\tilde{F}_R((t - s)_+)(1 - \tilde{F}_{C^\star}(s))\tilde{S}_k(s).$$

Now, we estimate the drift of the weighted log-rank test (3) by

$$\tilde{\xi}_{\tilde{Q}}(t) \coloneqq \int_0^t \tilde{Q}(s) \cdot \frac{\tilde{\pi}_0(t, s)\tilde{\pi}_1(t, s)}{\tilde{\pi}_0(t, s) + \tilde{\pi}_1(t, s)} \cdot (\tilde{\lambda}_0(s) - \tilde{\lambda}_1(s))\, ds.$$

Its asymptotic variance is estimated by

$$\tilde{\sigma}^2_{\tilde{Q}}(t) \coloneqq \int_0^t \tilde{Q}(s)^2 \cdot \left(\frac{\tilde{\pi}_0(t, s)\tilde{\pi}_1(t, s)}{\tilde{\pi}_0(t, s) + \tilde{\pi}_1(t, s)}\right)^2 \cdot \tilde{F}_R((t - s)_+)(1 - \tilde{F}_{C^\star}(s)) \cdot ((1 - r)\tilde{f}_0(s) + r\tilde{f}_1(s))\, ds.$$

where r denotes the proportion of patients allocated to treatment group $k = 1$. Based on the property of asymptotically independent and normally distributed increments of the process $(T_{\tilde{Q}}(t))_{t\geq 0}$, the standardized increment (7) would then follow the distribution

$$\tilde{S}_{2,\tilde{Q}} \sim \mathcal{N}\left(\frac{\tilde{\xi}_{\tilde{Q}}(t_2) - \tilde{\xi}_{\tilde{Q}}(t_1)}{\sqrt{\tilde{\sigma}^2_{\tilde{Q}}(t_2) - \tilde{\sigma}^2_{\tilde{Q}}(t_1)}}, 1\right)$$

To maximize the power of our procedure, one would choose the weight for which the largest conditional power is assumed based on our planning assumptions, i.e.

$$D \coloneqq \underset{d\in\{1,\ldots,m_2\}}{\arg\max}\; 1 - \Phi\left(\Phi^{-1}(1 - \tilde{\alpha}_2) - \frac{\tilde{\xi}_{\tilde{Q}_{\text{cand},d}}(t_2) - \tilde{\xi}_{\tilde{Q}_{\text{cand},d}}(t_1)}{\sqrt{\tilde{\sigma}^2_{\tilde{Q}_{\text{cand},d}}(t_2) - \tilde{\sigma}^2_{\tilde{Q}_{\text{cand},d}}(t_1)}}\right).$$

This selection process is demonstrated in Section 4.

4 Real data example

We illustrate the proposed procedure based on reconstructed data from the FAKTION trial[34] (NCT01992952). In this multicentre, randomized, placebo-controlled, phase 2 trial, the addition of capivasertib to fulvestrant was investigated in patients with advanced breast cancer. It was found that the experimental therapy extended progression-free survival, which was the primary endpoint. However, no effect could be proven for overall survival, possibly the primary endpoint in a consecutive phase III trial. Therefore, we reanalyze overall survival data using the method presented above. For further details, we refer to the published results of this trial[34].

We obtained the individual patient data using the algorithm by Guyot et al. (2012)[35]. The obtained data includes the observed survival time and a group and censoring indicator. However, the recruitment dates could not be reconstructed. Hence, we processed the data to receive the required data structure, including enrollment. The observed censoring suggests that recruitment took place evenly over the entire duration of the study (see Supplementary Material, Section B). Hence, we assume that new patients were recruited over the entire duration of the trial.

For patients with censored event time C_i, the recruitment date was set to $t_2 - C_i$ where $t_2 = 35.98$ is the calendar date of the final analysis. For patients with uncensored event time T_i a recruitment date that is uniformly distributed on the interval $[0, t_2 - T_i]$ was simulated. We considered an adaptive design with one interim analysis after 24 months, i.e., approximately one year before the final analysis. Stage-wise p-values are combined with the inverse normal combination function from (6) with equal weights $w_1 = w_2 = 1/\sqrt{2}$. Decision bounds were calculated according to the design of O'Brien and Fleming[30] without any futility bound. Accordingly, the null hypothesis (1) will be rejected in favor of the alternative (2) if $p_1 \leq 0.002583$ or $C(p_1, p_2) \leq 0.023996$.

When applying the adaptive testing procedure proposed in Section 3, we consider the set of candidate weights $\hat{\mathcal{Q}}_{\text{cand}} = \{w^{(0,0)} \circ \hat{F}, w^{(1,0)} \circ \hat{F}, w^{(0,1)} \circ \hat{F}, w^{(1,1)} \circ \hat{F}\}$ from the Fleming-Harrington family from (8). Here, $\circ$ denotes function composition. For the first stage, we consider different choices of $\hat{\mathcal{Q}}_{\text{mdir}}$. To this end, we look at all elements of the power set of $\hat{\mathcal{Q}}_{\text{cand}}$ that contain the standard log-rank weight $w^{(0,0)} \circ \hat{F}$. Resulting p-values can be found in the second column of Table 2.

For the interim analysis data, we fit Royston-Parmar splines to the interim data as presented in Section 2.3. This is done for the number of internal knots $p \in \{0, 1, 2\}$ and all three scales presented in Table 1. The final model used for conditional power calculation is chosen based on the AIC. The model with the lowest AIC has 0 internal knots and is computed on the normal scale. It is displayed in Figure 1. Fitted curves for other parameter configurations can be found in Supplementary Figure S8. The extrapolation performance can be judged based on Supplementary Figure S9. AIC values for all models are displayed in Supplementary Table S5. Of course, we are convinced that a medical expert should also be consulted at this stage to assess the plausibility of the extrapolated curves. In this particular case, the extrapolation of the model selected on the basis of the AIC proves to be plausible.

For this model, conditional power calculations as presented in Section 3.1 were executed. The conditional power computations favor the $(1, 1)$-weighted Fleming-Harrington test statistic for the second stage. However, the conditional power values differ greatly between the models. For the model that is chosen based on the AIC (see Figure 1), the conditional power ranges between 73.45% and 83.34%, depending on the chosen test statistic for the first stage. The results of the conditional power calculations for all modeling parameter choices and first-stage test statistics choices can be found in S6.

Based on the decision rule lined out above, the combined p-values in Table 2 indicate that the null hypothesis can be rejected whenever the $(1, 1)$-weighted Fleming-Harrington test is chosen as the test statistic for the second stage. Interestingly, this is the test statistic suggested by our approach, as lined out above. Whenever, the $(1, 1)$-weighted Fleming-Harrington test is included in the combination test of the first stage, a rejection can also be achieved if the standard log-rank test or the $(1, 0)$-weighted Fleming-Harrington test is chosen for the second stage. Additionally, rejection also occurs if the weights $(0, 0)$ and $(1, 0)$ are combined in the first stage and $(1, 0)$ is chosen for the second stage. The trial would always end with the acceptance of the null hypothesis if the $(0, 1)$-weighted Fleming-Harrington test is selected for the second stage.

In summary, we could show that an adaptation of the weight could have rejected the hypothesis of equal distributions of overall survival in the two treatment groups. In particular, applying our proposed procedure would have yielded such a result.

We emphasize that these results rely on simulated recruitment dates for uncensored patients. Neverthe-

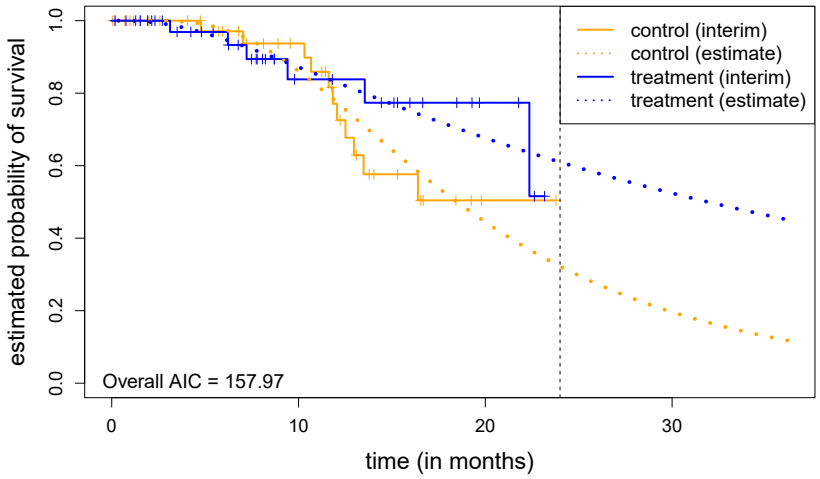


Figure 1: Interim data with fitted Royston-Parmar spline models with 0 interior knots on the normal scale. This model has the lowest AIC among all models considered. Vertical line at 24 months indicates the calendar date of the interim analysis.

$\hat{\mathcal{Q}}_{\text{mdir}}$	p_1	$C(p_1, p_2)$			
		$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,0),(1,0),(0,1),(1,1)$	0.133	0.021	0.034	0.021	0.009
$(0,0),(1,0),(0,1)$	0.168	0.027	0.043	0.027	0.012
$(0,0),(1,0),(1,1)$	0.116	0.018	0.030	0.018	0.008
$(0,0),(0,1),(1,1)$	0.125	0.020	0.032	0.019	0.009
$(0,0),(1,0)$	0.152	0.024	0.039	0.024	0.011
$(0,0),(0,1)$	0.188	0.031	0.048	0.030	0.014
$(0,0),(1,1)$	0.110	0.017	0.028	0.017	0.008
$(0,0)$	0.172	0.028	0.044	0.027	0.013

Table 2: p-values from first stage data for different choices of $\hat{\mathcal{Q}}_{\text{mdir}}$ and combined p-value for different choices of the single weighted log-rank test for the second stage test

less, this is necessary for demonstration purposes. In Section B.4 of the Supplementary Material, the influence of this simulation is assessed.

5 Simulations

In this simulation study, we want to examine compliance with our adaptive selection procedure's nominal type I error rate. Furthermore, we compare power curves for various testing procedures to illustrate the merits of an adaptive selection procedure.

In both parts, the calendar time schedule of our fictional clinical trial will be the same. We consider a design with one interim analysis after $t_1 = 5$ and a final analysis after $t_2 = 8$ years. We will not consider any sample size adaptation with an accompanying shift of t_2 as we only want to focus on the advantages and disadvantages of the selection procedure. Nevertheless, this possibility exists and represents an undisputed advantage of our design. The participating individuals enter the trial uniformly up until $a = 6$ years have passed. We only consider administrative censoring and no additional loss to follow-up. The time-to-event variable in the control group is exponentially distributed with parameter $-\log(1-0.3)$, i.e., an annual event rate of 30%. For the flexible estimation and extrapolation of the survival curves, the same nine Royston-Parmar spline models as in Section 4 will be considered. For each run, the best among these models will be determined based on the AICs of the models. The test for the second stage is then chosen based on conditional power calculations based on the extrapolated curves from this particular model.

For the two-stage designs examined here, combinations of stagewise p-values and sequential decision

bounds are chosen as in the previous example from Section 4.

5.1 Empirical type I error rates

For the first stage, we consider 7 different combination tests and 8 differently weighted log-rank tests based on Fleming-Harrington weights for the second stage. In particular, we set $\hat{\mathcal{Q}}_{\text{cand}} = \{w^{(0,0)} \circ \hat{F}, w^{(1,0)} \circ \hat{F}, w^{(2,0)} \circ \hat{F}, w^{(3,0)} \circ \hat{F}, w^{(1,1)} \circ \hat{F}, w^{(0,1)} \circ \hat{F}, w^{(0,2)} \circ \hat{F}, w^{(0,3)} \circ \hat{F}\}$. The sample sizes for each group vary in the set $\{50, 100, 200, 500\}$. We only considered balanced group sizes. The results are based on 10,000 simulation runs. Hence, for a true underlying rate of 0.025, the empirical rate lies within the interval $[0.0219, 0.0281]$ with a probability of 95%.
The empirical rejection rates for any pre-fixed combination of test statistics in the two stages (i.e., not determined by a selection procedure at the interim analysis but already predefined at the start of the trial) can be found in Supplementary Tables S10 - S13.

$\hat{\mathcal{Q}}_{\text{mdir}}$	n			
	100	200	400	1000
$(0,0),(1,0),(0,1),(1,1)$	0.0238	0.0256	0.0248	0.0260
$(0,0),(1,0),(0,1)$	0.0241	0.0269	0.0237	0.0260
$(0,0),(1,0),(1,1)$	0.0238	0.0259	0.0240	0.0264
$(0,0),(0,1),(1,1)$	0.0242	0.0277	0.0244	0.0259
$(0,0),(1,0)$	0.0233	0.0259	0.0235	0.0279
$(0,0),(0,1)$	0.0251	0.0273	0.0230	0.0269
$(0,0),(1,1)$	0.0258	0.0267	0.0231	0.0265
$(0,0)$	0.0269	0.0285	0.0235	0.0265

Table 3: Empirical type I error rates

All rates shown here lie in the confidence interval mentioned above. We can observe a slight inflation of the empirical type I error level for small sample sizes. We can assume that this is due to the lack of agreement between the actual distribution of the weighted log-rank test statistics and its asymptotical approximation by a normal distribution for small sample sizes. This is supported by the fact that we can see similar inflations for the fixed combinations displayed in the Supplementary Tables S10 - S13. However, this fact is well-known. For small sample sizes or high censoring percentages, this problem could easily be solved by applying a permutation version of the log-rank test [36]. Slight inflations of the type I error levels in Table 3 do not exceed the reported rates. As those rates refer to group sequential designs without any adaptation, we can claim that the adaptive weight choice does not introduce any further complications regarding type I error levels.

5.2 Power comparisons

We examine 7 different types of deviations of the distribution in the experimental group from the distribution in the control group. For each type, the strength of the deviation will be given by a parameter θ. These types of deviations are chosen in such a way that one particular test based on Fleming-Harrington weight $w^{(\rho^\star,\gamma^\star)} \circ \hat{F}$ will be optimal. We consider combinations $(\rho^\star, \gamma^\star) \in \{(0,0),(1,0),(2,0),(3,0),(0,1),(0,2),(0,3)\}$. For $\rho^\star = \gamma^\star = 0$, the type of deviation is just given by proportional hazards, and the standard log-rank test is optimal. For $\rho^\star > \gamma^\star = 0$ and $\gamma^\star > \rho^\star = 0$, the construction of the corresponding mechanisms is described in Garès et al. (2017) [37] and Chapter 7.4 of Fleming and Harrington (2011) [18], respectively. Thus, $\theta = 0$ yields no difference between the survival curves in the two groups, and the advantage of the experimental group increases as θ decreases.
All simulation runs' sample size is 500 patients per group. We consider 7 different values of θ for each type of deviation. At first, we determine some value θ_0 s.t. the two-stage test with the weighted log-rank test, that would be optimal in this case (i.e., $w^{(\rho^\star,\gamma^\star)} \circ \hat{F}$) achieves an overall power of 50%. Some analytical calculations can accomplish this. In the simulations, we then consider values of θ in the set $\{0.4 \cdot \theta_0, 0.6 \cdot \theta_0, 0.8 \cdot \theta_0, \theta_0, 1.2 \cdot \theta_0, 1.4 \cdot \theta_0, 1.6 \cdot \theta_0\}$ in order to cover a broad power range.
For the sake of brevity, we only show the results for deviations with $(\rho^\star, \gamma^\star) \in \{(0,0),(2,0),(0,2)\}$ in the main manuscript, see Figures 3-5. Survival curves and results for the other types can be found in Section C.2.1 of the Supplementary Material. The selection comprises scenarios with proportional hazards and late and early effects. For these three types, the corresponding survival curves can be found in Figure 2.
For each type, we show two graphs. Subfigure **(A)** shows the power curves of six different testing

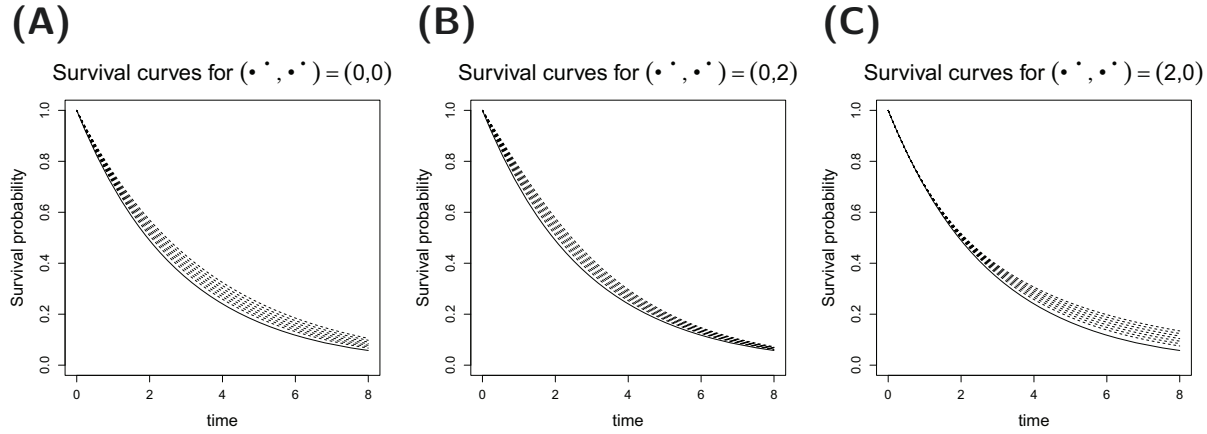


Figure 2: Survival curves for three types of deviation of the distribution in the experimental group from the distribution in the control group. The solid line gives the survival curve in the control group. Dashed lines are survival curves in the experimental group for the seven effect sizes $\{0.4 \cdot \theta_0, 0.6 \cdot \theta_0, 0.8 \cdot \theta_0, \theta_0, 1.2 \cdot \theta_0, 1.4 \cdot \theta_0, 1.6 \cdot \theta_0\}$. **(A)** Survival curves in the proportional hazards case ($(\rho^\star, \gamma^\star) = (0, 0)$) **(B)** Survival curves in the early effect case ($(\rho^\star, \gamma^\star) = (0, 2)$) **(C)** Survival curves in the late separation case ($(\rho^\star, \gamma^\star) = (2, 0)$)

procedures, which are abbreviated as follows:

Abbreviation	Description
OS-MDIR	One-stage testing procedure with the *mdir* combination test based on the weights $w^{(0,0)} \circ \hat{F}$, $w^{(1,0)} \circ \hat{F}$ and $w^{(0,1)} \circ \hat{F}$
OS-restrMDIR	One-stage testing procedure with an *mdir* combination test with a restricted set of weights in some cases ($w^{(0,0)} \circ \hat{F}$, $w^{(1,0)} \circ \hat{F}$ if $\rho^\star > \gamma^\star = 0$ and $w^{(0,0)} \circ \hat{F}$, $w^{(0,1)} \circ \hat{F}$ if $\gamma^\star > \rho^\star = 0$)
TS-AD	Two-stage adaptive design with *mdir* combination test as for OS-MDIR in the first stage and a selection of the test for the second stage among the weights in $\hat{\mathcal{Q}}_{\text{cand}}$ as defined above
TS-LR	Two-stage standard log-rank test
TS-optFH	Two-stage weighted log-rank test with the optimal weighting by $w^{(\rho^\star, \gamma^\star)} \circ \hat{F}$
TS-restrAD	Two-stage adaptive design with *mdir* combination test as for the design OS-restrMDIR in the first stage and a selection of the test for the second stage among a restricted subset of $\hat{\mathcal{Q}}_{\text{cand}}$ that only includes weights with $\rho \gtrless \gamma$ and the standard weight if $\rho^\star \gtrless \gamma^\star$ and weights with $\lvert\rho - \gamma\rvert \leq 1$ if $\rho^\star = \gamma^\star = 0$.

Table 4: Overview of the different testing procedures.

TS-optFH serves as a benchmark because the optimal test is used here. In comparison, applying TS-LR shall illustrate the disadvantages of not applying techniques that can guard against deviations from proportional hazards. To quantify the advantages of using a robust *mdir* combination test, we consider a one-stage version of the *mdir* test (OS-MDIR). Suppose it can be anticipated that a particular deviation from proportional hazards is more likely. In that case, choosing the weights for the *mdir* combination test according to this prior knowledge can be beneficial (OS-restrMDIR). Analogously, for the two-stage design, one can use the robust *mdir* and choose the second stage test out of a wide variety of single-weighted tests (TS-AD) or make pre-selections according to the anticipated possible effects (TS-restrAD).
Please note that the overall power curves do not consider any additional advantages of the adaptive

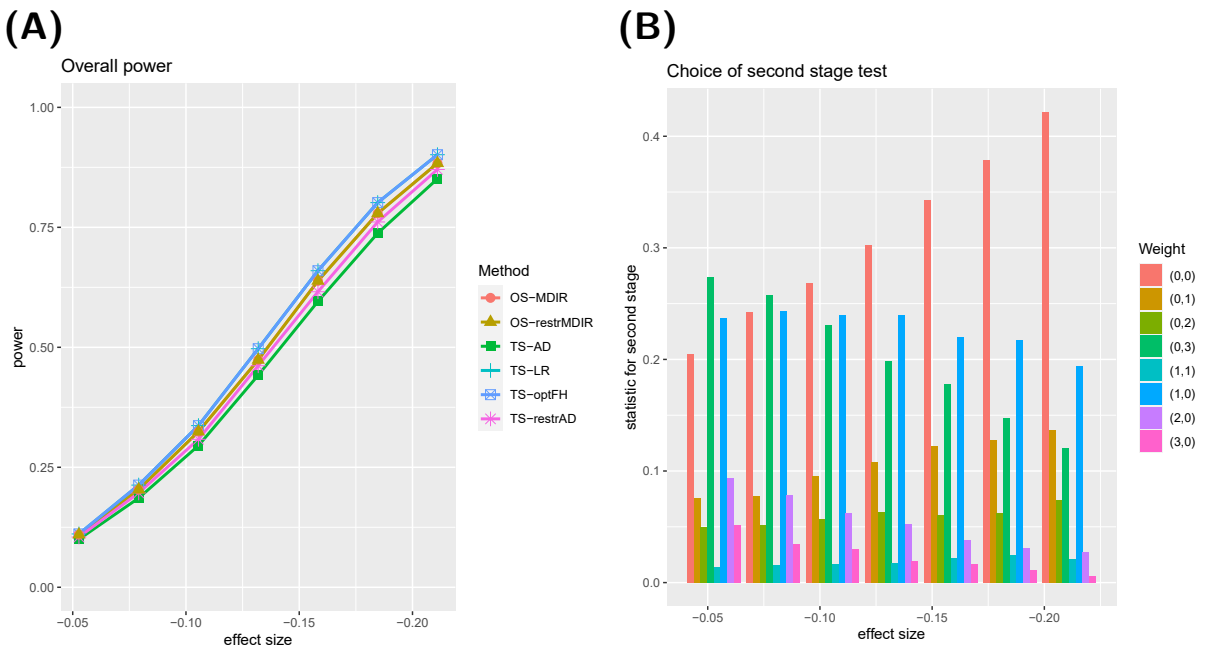


Figure 3: **(A)** Power curve for six testing procedures in case of proportional hazards ($(\rho^\star, \gamma^\star) = (0, 0)$). Please note that the two procedures, TS-AD and TS-optFH, and the two procedures, OS-MDIR and OS-restrMDIR, coincide in this case. **(B)** Relative frequencies of the choice of single-weighted tests for the second stage of the TS-AD testing procedure.

designs, e.g., early rejection or sample size recalculations. For comparisons between one-stage *mdir* combination tests and one-stage weighted log-rank tests we refer to Dormuth et al. (2023) [10].
The second graph **(B)** always illustrates the choice of the test statistic for the second stage of the testing procedure TS-AD. It is shown how often each of the second stage tests from $\hat{\mathcal{Q}}_{\text{cand}}$ is chosen (using the selection procedure described above) in those simulated trials which proceeded to the second stage.

There is not much difference between the six approaches concerning proportional hazard scenarios. As expected, the two-stage standard log-rank test performs best. The two-stage designs with a combination test in the first and a weight selection in the second stage perform very similarly to the one-stage *mdir* combination tests. Of course, some efficiency of the adaptive selection designs is lost because of a non-optimal test selection for the second stage. However, the optimal choice is made quite often. Additionally, the procedures TS-AD and TS-restrAD exhibit early rejection rates of about 9% for the effect size θ_0 and even more than 30% for the effect size $1.6 \cdot \theta_0$.

For the scenarios of early separation, TS-LR now performs worst among the six competitors. The correctly weighted log-rank test performs best. The two procedures OS-MDIR and TS-AD perform very similarly. The weight selection graph shows that suitable tests are often chosen for the design TS-AD. However, the test based on the weights $w^{(0,3)} \circ \hat{F}$ is chosen more often than the optimal weighting scheme $w^{(0,2)} \circ \hat{F}$. We assume this behavior is rooted in the spline models, but this speculation has to be investigated further in future research. The two power curves for the restricted procedures are also almost equal. As the pre-selection of weights introduces useful information for the procedure, these perform slightly better (about five percentage points) than the unrestricted procedures OS-MDIR and TS-AD. Similarly to the case of proportional hazards, the procedures TS-AD and TS-restrAD exhibit early rejection rates of about 10% for the effect size θ_0 and even more than 35% for the effect size $1.6 \cdot \theta_0$.

In the late separation scenario, TS-LR performs worst, and TS-optFH performs best again. The adaptive designs TS-AD and TS-restrAD have less overall power than the one-stage counterparts OS-MDIR and OS-restrMDIR. This may be because it is more difficult to identify a well-fitting test here since, at the time of the interim analysis, less information is available about the period in which the hazards in the two groups differ greatly. The selection procedure seems to prefer the test based on the weight $w^{(1,0)} \circ \hat{F}$ to the optimal test. Again, this behavior might result from the spline extrapolation. Additionally, we could observe that the model selection based on the AIC often prefers models with 0 interior knots (see Supplementary Material, Section C.2.2). Unfortunately, these models often fail to

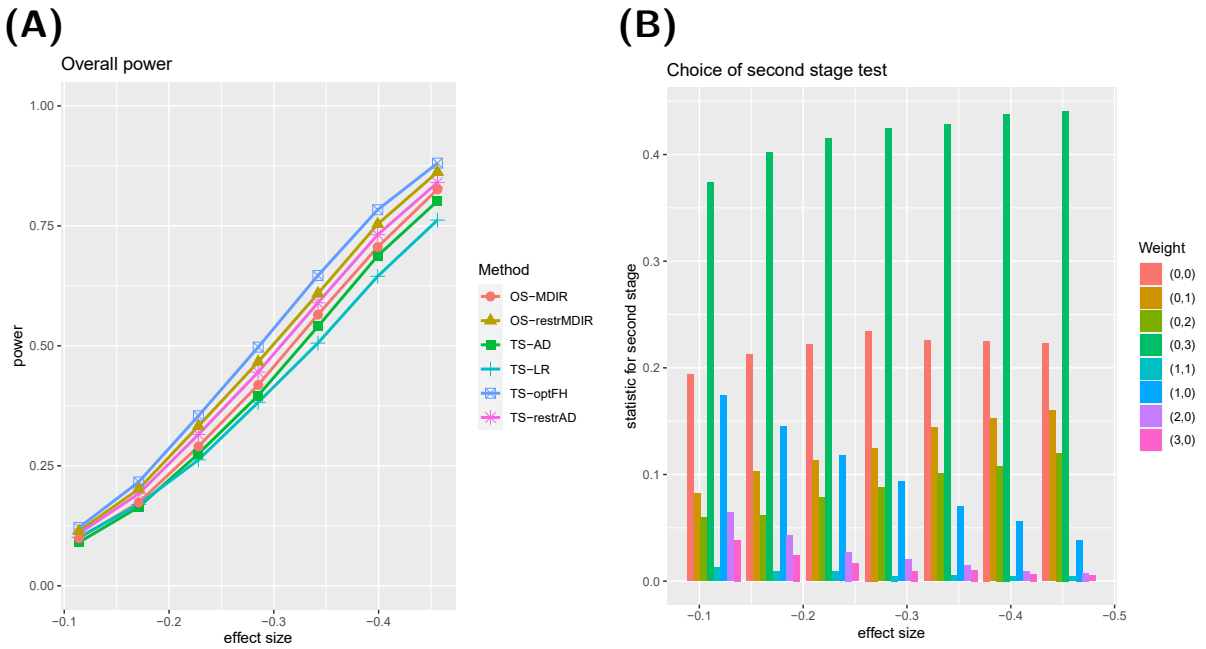


Figure 4: **(A)** Power curve for six testing procedures in case of an early effect ($(\rho^\star, \gamma^\star) = (0, 2)$). **(B)** Relative frequencies of the choice of single-weighted tests for the second stage of the TS-AD testing procedure.

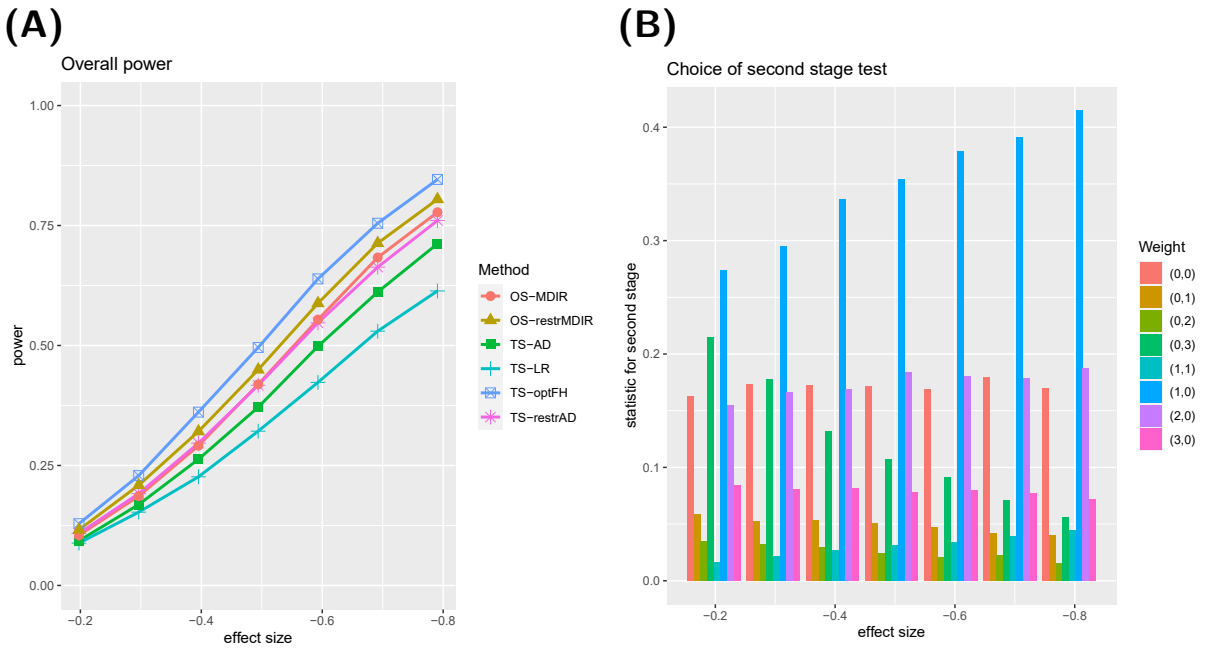


Figure 5: **(A)** Power curve for six testing procedures in case of a late separation ($(\rho^\star, \gamma^\star) = (2, 0)$). **(B)** Relative frequencies of the choice of single-weighted tests for the second stage of the TS-AD testing procedure.

detect late separation of the survival curves (see Supplementary Material, Section C.2.3). However, all tests perform remarkably better than TS-LR. The early rejection rates are smaller than in the previous scenarios, with about 3% for the effect size θ_0 and even more than 7% for the effect size $1.6 \cdot \theta_0$. This is again due to the sparse information about late events at the date of the interim analysis.
In all scenarios, we could see that the adaptive selection procedure can close the power gap between one- and two-stage designs that is caused by the known inefficiencies of adaptive designs [38]. However, as seen in Figures 4 and 5, our selection procedure often selects well-suited but not optimal tests. This shows there is room for improvement in the selection procedure.

6 Discussion

In the previous sections, an adaptive design for a survival time endpoint was presented and examined, allowing for adjustment of test statistics at the time of interim analyses and the use of combination tests. Such a combination test is particularly beneficial in the first stage. In contrast, in the second stage, the information already collected can be used to select a suitable test for the further course. Our application example demonstrated that adapting the weight can save a trial that would otherwise end with an inconclusive result. Our simulation studies demonstrated that our two-stage procedures are superior to the two-stage log-rank test in non-proportional hazard settings. At the same time, they do not lose much power in a proportional hazard setting. Compared to one-stage combination tests, they provide much more flexibility as other adaptations are still allowed [2].
However, we could observe that pre-selection of the involved weights based on some initial assumptions can markedly improve the performance. In this work, we only consider Fleming-Harrington weights in the combination tests and our selection procedure due to their popularity. The Fleming-Harrington weights were subject to increased criticism when applied to prove superiority due to their inconsistency regarding the scoring in late effect situations [39]. However, our procedure can also be used for differently weighted log-rank tests that take this issue into account [40] as shown in Section B.5 of the Supplementary Material. This concerns the combination tests as well as the weight selection procedure.
The extrapolation of the survival curve of the variable under investigation beyond the time horizon observed so far is crucial for the decision to be made during the interim analysis. We applied the model of Royston and Parmar [21] as it is flexible, provides extrapolation, and incorporates some standard parametric distributions. We did not tailor the spline approach to our specific data set up to obtain a fair comparison between the two-stage and the one-stage procedures. However, we assume that this might lead to the selection of a weighted log-rank test that is not the optimal test by design. To tackle this issue, further investigation of the proposed selection procedure of Royston-Parmar splines or alternative extrapolation methods is necessary. In principle, our proposed procedure allows for any other extrapolation approach. For example, the Kaplan-Meier estimator could be combined with a parametric tail for extrapolation [41], or a penalized version of the chosen approach [42] can be employed. Of course, it would also be desirable to incorporate expert knowledge or prior knowledge from other data sets using Bayesian methods [43]. We assume that such methods can further improve the extrapolation and, thus, the entire procedure. However, we need to make sure that information beyond the primary endpoint should not be incorporated into the decision process about the adaptation as it might compromise the type I error rate [44] if nor precautions are taken against this [45].
The analysis schedule also plays a major role here. Suppose the survival curves of the two groups separate so late that no corresponding observations can be made until the interim analysis. In that case, no reasonably informed decision can be made about how to continue the study.
While we have limited ourselves to calculating conditional power as an instrument for determining adaptation, other tools are also possible. It is well-known that the conditional power tends to assume quite extreme values [46]. Similar concepts, such as the predictive power [47], are less prone to this problem. Such an approach could also be excellently combined with a Bayesian approach to the survival extrapolation as mentioned above [43].
Finally, we would like to point out that the main aim of the proposed procedure is to be as flexible and robust as possible. This inevitably leads to a loss of efficiency. First, this applies to adaptive designs in general since the fixed weights selected for the combination function do not necessarily correspond to the amount of information that comes in from the individual stages [38]. This challenge is even more accentuated here, as the information processes of differently weighted log-rank tests are not proportional to each other and generally do not correspond to the number of events observed [48]. We would also like to mention again that combination tests have a good sensitivity to a wide range of alternatives [10] but are never optimal.

We regard our contribution as methodological phase I/II in the sense of Heinze et al. (2022) [49]. From our point of view, the adaptation of test statistics has been woefully neglected, although discussions about the choice of an appropriate test in situations of non-proportional hazards continue. We believe that combination tests and adaptive designs complement each other here naturally. Hence, we would like to present this fundamental possibility here. Nevertheless, we are aware that further research is still required. This includes the incorporation of different weight functions [39], consideration of other extrapolation approaches [41–43] as well as fine-tuning the selection of the final extrapolation model and the investigation of alternative quantities for decision making regarding design adaptations [47]. Furthermore, we want to investigate different approaches to combine the stagewise p-values and derive recommendations on executing the sample size reestimation within the framework.

Acknowledgments

The work of Moritz Fabian Danzer was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 413730122.
Parts of the calculations for this publication were performed on the HPC cluster PALMA II of the University of Münster, subsidised by the DFG (INST 211/667-1).
In memory of our esteemed colleague, Marc Ditzhaus, who passed away in September 2024. His invaluable input will always be remembered, and he will be deeply missed.

Conflict of Interest

The authors have declared no conflict of interest.

References

1. Bauer P, Bretz F, Dragalin V, König F, Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine.* 2016;35(3):325-347.

2. Wassmer G, Brannath W. *Group sequential and confirmatory adaptive designs in clinical trials*;301. Springer 2016.

3. Wassmer G. Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal.* 2006;48(4):714–729.

4. Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association.* 1982;77(380):855–861.

5. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika.* 1982;69(3):553–566.

6. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika.* 1977;64(1):156–160.

7. Brendel M, Janssen A, Mayer CD, Pauly M. Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics.* 2014;41(3):742–761.

8. Ditzhaus M, Friedrich S. More powerful logrank permutation tests for two-sample survival data. *Journal of Statistical Computation and Simulation.* 2020;90:2209–2227.

9. Lin RS, Lin J, Roychoudhury S, *et al.* Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards: A Comparative Analysis. *Statistics in Biopharmaceutical Research.* 2020;12(2):187–198.

10. Dormuth I, Liu T, Xu J, Pauly M, Ditzhaus M. A comparative study to alternatives to the log-rank test. *Contemporary Clinical Trials.* 2023;128:107165.

11. Ditzhaus M, Pauly M. Wild bootstrap logrank tests with broader power functions for testing superiority. *Computational statistics & data analysis.* 2019;136:1–11.

12. Jiménez JL. Quantifying treatment differences in confirmatory trials under non-proportional hazards. *Journal of Applied Statistics.* 2022;49(2):466–484.

13. Hasegawa T. Group sequential monitoring based on the weighted log-rank test statistic with the Fleming–Harrington class of weights in cancer vaccine studies. *Pharmaceutical Statistics.* 2016;15(5):412-419.

14. Ghosh P, Ristl R, König F, *et al.* Robust group sequential designs for trials with survival endpoints and delayed response. *Biometrical Journal.* 2022;64(2):343-360.

15. Campbell H, Dean C. The consequences of proportional hazards based model selection. *Statistics in Medicine.* 2014;33(6):1042-1056.

16. Lawrence J. Strategies for changing the test statistic during a clinical trial. *Journal of biopharmaceutical statistics.* 2002;12(2):193–205.

17. Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association.* 1997;92(440):1342–1350.

18. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis.* John Wiley & Sons 2011.

19. Ditzhaus M, Friedrich S. *mdir.logrank: Multiple-Direction Logrank Test* 2018. R package version 0.0.4.

20. Latimer NR, Adler AI. Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. *BMJ Medicine.* 2022;1(1).

21. Royston P, Parmar MK. 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.* 2002;21(15):2175–2197.

22. Royston P, Lambert PC, others . *Flexible parametric survival analysis using Stata: beyond the Cox model*;347. Stata press College Station, TX 2011.

23. Jackson C. flexsurv: A Platform for Parametric Survival Modeling in R. *Journal of Statistical Software.* 2016;70(8):1–33.

24. Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation.* 2015;85(4):777–793.

25. Brannath W, Gutjahr G, Bauer P. Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association.* 2012;107(498):824–832.

26. Fisher RA. Statistical methods for research workers. in *Breakthroughs in statistics: Methodology and distribution*:66–70Springer 1970.

27. Bauer P. Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie.* 1989;20(4):130–148.

28. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics.* 1999;55(4):1286–1290.

29. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika.* 1977;64(2):191–199.

30. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics.* 1979:549–556.

31. Lan K, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika.* 1983;70(3):659–663.

32. Wassmer G, Pahlke F. *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis* 2023. R package version 3.3.4.

33. Yung G, Liu Y. Sample size and power for the weighted log-rank test and Kaplan-Meier based tests with allowance for nonproportional hazards. *Biometrics.* 2020;76(3):939–950.

34. Jones RH, Casbard A, Carucci M, *et al.* Fulvestrant plus capivasertib versus placebo after relapse or progression on an aromatase inhibitor in metastatic, oestrogen receptor-positive breast cancer (FAKTION): a multicentre, randomised, controlled, phase 2 trial. *The Lancet Oncology.* 2020;21(3):345–357.

35. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced Secondary Analysis of Survival Data: Reconstructing the Data from Published Kaplan-Meier Survival Curves. *BMC Medical Research Methodology.* 2012;12(1):9.

36. Neuhaus G. Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics.* 1993:1760–1779.

37. Garès V, Andrieu S, Dupuy JF, Savy N. On the Fleming—Harrington test for late effects in prevention randomized controlled trials. *Journal of Statistical Theory and Practice.* 2017;11:418–435.

38. Tsiatis AA, Mehta C. On the Inefficiency of the Adaptive Design for Monitoring Clinical Trials. *Biometrika.* 2003;90(2):367–378.

39. Magirr D, Burman CF. The MaxCombo Test Severely Violates the Type I Error Rate. *JAMA Oncology.* 2023;9(4):571-572.

40. Magirr D, Burman CF. Modestly weighted logrank tests. *Statistics in medicine.* 2019;38(20):3782–3790.

41. Gelber RD, Goldhirsch A, Cole BF, Group IBCS, others . Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. *Controlled Clinical Trials.* 1993;14(6):485–499.

42. Liu XR, Pawitan Y, Clements M. Parametric and penalized generalized survival models. *Statistical methods in medical research.* 2018;27(5):1531–1546.

43. Jackson CH. survextrap: a package for flexible and transparent survival extrapolation. *BMC Medical Research Methodology.* 2023;23(1):282.

44. Bauer P, Posch M. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections by H. Schäfer and H.-H. Müller, Statistics in Medicine 2001; 20: 3741–3751. *Statistics in Medicine.* 2004;23(8):1333-1334.

45. Danzer MF, Faldum A, Simon T, Hero B, Schmidt R. Confirmatory adaptive group sequential designs for clinical trials with multiple time-to-event outcomes in Markov models. 2023.

46. Bauer P, Koenig F. The reassessment of trial perspectives from interim data—a critical view. *Statistics in medicine.* 2006;25(1):23–36.

47. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power?. *Controlled clinical trials.* 1986;7(1):8–17.

48. Kundu MG, Sarkar J. On information fraction for Fleming-Harrington type weighted log-rank tests in a group-sequential clinical trial design. *Statistics in Medicine.* 2021;40(10):2321–2338.

49. Heinze G, Boulesteix AL, Kammer M, Morris TP, White IR, STRATOS initiative SP. Phases of methodological research in biostatistics—Building the evidence base for new methods. *Biometrical Journal.* 2022:2200222.

Adaptive weight selection for time-to-event data under non-proportional hazards - Supplementary Material

A Technical appendix

In this section of the Supplementary Material, we state some technical results that justify the validity of our adaptive weight selection procedure. We adopt the notation from the main manuscript.

Lemma 1. *Let $(\mathbf{X}^{(n)})_{n\geq 0}$ be a sequence of $\mathbb{R}^d$-valued random vectors s.t. $X_c^{(n)} \xrightarrow{\mathbb{P}} X_c$ for each $c \in \{1,\dots,d\}$ as $n \to \infty$. Then it also holds*

$$\mathbf{X}^{(n)} \xrightarrow{\mathbb{P}} \mathbf{X} =: (X_1, \dots X_d)$$

as $n \to \infty$ in $\mathbb{R}^d$.

Proof. As all norms are equivalent on $\mathbf{R}^d$, it is enough to show it for the 1-norm, i.e.

$$\mathbb{P}\left[\sum_{c=1}^{d} |X_c^{(n)} - X_c| > \varepsilon\right] \to 0$$

for any $\varepsilon > 0$. Because $\sum_{c=1}^{d} |X_c^{(n)} - X_c| > \varepsilon$ implies that there is at least one c s.t. $|X_c^{(n)} - X_c| > \varepsilon/d$, we get

$$\begin{aligned}
&\mathbb{P}\left[\sum_{c=1}^{d} |X_c^{(n)} - X_c| > \varepsilon\right] \\
\leq &\mathbb{P}\left[\cup_{c=1}^{d} |X_c^{(n)} - X_c| > \varepsilon/d\right] \\
\leq &\sum_{c=1}^{d} \mathbb{P}[|X_c^{(n)} - X_c| > \varepsilon/d]
\end{aligned}$$

As all of the summands in the last sum converge to 0, the sum becomes arbitrary small for increasing n. □

Extending the notation from the manuscript, we define the bivariate $\mathbb{R}$-valued stochastic process $(T_{\hat{Q}}(t,s))_{t,s\geq 0}$ by

$$T_{\hat{Q}}(t,s) \coloneqq n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_{[0,s]} \hat{Q}(t,u) \left(Z_i - \frac{Y^{Z=1}(t,u)}{Y(t,u)}\right) dN_i(t,u),$$

that sums up the information available at calendar time t about the time-to-event endpoint until trial time s. This process has to be adapted to the bivariate filtration $(\mathcal{F}(t,s))_{t,s\geq 0}$ with

$$\mathcal{F}(t,s) = \sigma\left(\cup_{i=1}^{n} \mathcal{F}_i(t,s)\right),$$

i.e. these σ-algebras are generated by patient-specific σ-algebras. These $\mathcal{F}_i(t,s)$ are in turn generated by the random variables

$$\begin{aligned}
\mathbb{1}_{\{R_i \leq t\}}, R_i \cdot \mathbb{1}_{\{R_i \leq t\}}, \mathbb{1}_{C_i^\star \leq s \wedge (t-R_i)_+}, C_i^\star \cdot \mathbb{1}_{C_i^\star \leq s \wedge (t-R_i)_+}, \\
\mathbb{1}_{T_i \leq s \wedge C_i(t)}, T_i \mathbb{1}_{T_i \leq s \wedge C_i(t)}.
\end{aligned}$$

The following results are valid under the null hypothesis of equal distributions of the time-to-event variable T

Theorem 1. *If for all $t \geq 0$, the assumptions*

A1 For any $\tau < t$

$$\sup_{0\leq s\leq \tau} |\hat{Q}(t,s) - Q(t,s)| \xrightarrow{\mathbb{P}} 0$$

A2 In its second argument, $\hat{Q}(t,s)$ is bounded over $[0,t]$, is left-continuous and has right hand limits

A3 For any $\tau_1 > 0$ and $\tau_2 < t$

$$\sup_{\tau_1 \leq s \leq \tau_2} \left| \frac{Y^{Z=1}(t,s)}{Y(t,s)} - \frac{y^{Z=1}(t,s)}{y(t,s)} \right| \xrightarrow{\mathbb{P}}$$

where $y(t,s) \coloneqq \mathbb{E}[Y(t,s)]$ and $y^{Z=1}(t,s) \coloneqq \mathbb{E}[Y^{Z=1}(t,s)]$

are fulfilled, then the process $(T_{\hat{Q}}(t))_{t\geq 0}$ is asymptotically equivalent to the process $(\tilde{T}_Q(t))_{t\geq 0}$, i.e.

$$(T_{\hat{Q}}(t) - \tilde{T}_Q(t)) \xrightarrow{\mathbb{P}} 0 \qquad \forall t \geq 0$$

This process is defined by

$$\tilde{T}_Q(t,s) \coloneqq n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_{[0,s]} Q(t,s) \left(Z_i - \frac{Y^{Z=1}(t,s)}{Y(t,s)} \right) dM_i(t,s),$$

where $(M_i(t,s))_{t,s\geq 0}$ is the counting process martingale based on the counting process $(N_i(t,s))_{t,s\geq 0}$.

For a proof, we refer to Theorem 1 of the Supplementary Material of[45]. In particular, it should be noted that the random quantities $\hat{Q}$, Y and $Y^{Z=1}$ have been replaced by deterministic quantities. Applying Lemma 1, we obtain the following Corollary.

Corollary 1. *The multivariate processes $(\mathbf{T}_{\hat{\mathcal{Q}}}(t))_{t\geq 0}$ as in Section 2.2 and the process $(\tilde{\mathbf{T}}_{\mathcal{Q}}(t))_{t\geq 0}$, where $\mathcal{Q}$ denotes the set of deterministic limit functions of those functions in $\hat{\mathcal{Q}}$ and $\tilde{\mathbf{T}}_{\mathcal{Q}}(t) \coloneqq (\tilde{T}_Q)_{Q\in\mathcal{Q}}$, are asymptotically equivalent.*

Lemma 2. *If all functions $Q \in \mathcal{Q}$ and $y^{Z=1}(t,s)/y(t,s)$ are independent of their first arguments, the multivariate processes $(\tilde{\mathbf{T}}_{\mathcal{Q}}(t))_{t\geq 0}$ is a martingale w.r.t. the filtration $(\mathcal{F}(t))_{t\geq 0}$ that comprises all available information in calendar time, i.e. $\mathcal{F}(t) \coloneqq \mathcal{F}(t,t)$.*

The proof follows analogously to the proof of Lemma 2 in Danzer et al. (2023)[45].
Please note that the assumptions made in this Lemma are naturally fulfilled in our applications. For Fleming-Harrington weights, the limit functions are given by $F(t)^\rho \cdot S(t)^\rho$ and under the null hypothesis $y^{Z=1}(t,s)/y(t,s)$ reduces to the (constant) probability that an individual is assigned to the treatment group.

Theorem 2. *As $n \to \infty$, $(\tilde{\mathbf{T}}_{\mathcal{Q}}(t))_{t\geq 0}$ converges in distribution to a Gaussian mean-zero vector martingale on some interval $[0, t_{mas}]$ with the $|\mathcal{Q}| \times |\mathcal{Q}|$-matrix-valued covariance function $\mathbf{\Sigma}_{\mathcal{Q}} \colon [0, t_{max}] \to \mathbb{R}^{|\mathcal{Q}|\times|\mathcal{Q}|}$ given by*

$$(\mathbf{\Sigma}(\boldsymbol{t}))_{kl} \coloneqq \int_{[0,t]} Q_k(s) Q_l(s) \cdot \mathbb{P}[s \leq C(t) \wedge T] \cdot \mathbb{P}[Z=1](1-\mathbb{P}[Z=1]) dA(s).$$

The proof follows analogously to the proof of Theorem 2 in Danzer et al. (2023)[45]. The covariance function can be consistently estimated as stated in Section 2.2.

Corollary 2. *For a sequence of analysis dates $0 =: t_0 < t_1 < \cdots < t_m$ in calendar time, the test multivariate test statistics $\mathbf{T}_{\hat{\mathcal{Q}}}$ are asymptotically jointly normally distributed with asymptotically independent increments, i.e.*

$$(\mathbf{T}_{\hat{\mathcal{Q}}}(t_1), \ldots, \mathbf{T}_{\hat{\mathcal{Q}}}(t_m)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{\Sigma}_{\mathcal{Q},acc})$$
$$(\mathbf{T}_{\hat{\mathcal{Q}}}(t_1) - \mathbf{T}_{\hat{\mathcal{Q}}}(t_0), \ldots, \mathbf{T}_{\hat{\mathcal{Q}}}(t_m) - \mathbf{T}_{\hat{\mathcal{Q}}}(t_{m-1})) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{\Sigma}_{\mathcal{Q},inc})$$

where both $\mathbf{\Sigma}_{\mathcal{Q},acc}$ and $\mathbf{\Sigma}_{\mathcal{Q},inc}$ are $m|\mathcal{Q}| \times m|\mathcal{Q}|$ matrices consisting of m^2 blocks of size $|\mathcal{Q}| \times |\mathcal{Q}|$. The block in row r_1 and column r_2 of $\mathbf{\Sigma}_{\mathcal{Q},acc}$ is given by $\mathbf{\Sigma}_{\mathcal{Q}}(t_{r_1} \wedge t_{r_2})$ and $\mathbf{\Sigma}_{\mathcal{Q},inc}$ is a block diagonal matrix with $\mathbf{\Sigma}_{\mathcal{Q},inc} = diag(\mathbf{\Sigma}_{\mathcal{Q}}(t_1) - \mathbf{\Sigma}_{\mathcal{Q}}(t_0), \ldots, \mathbf{\Sigma}_{\mathcal{Q}}(t_m) - \mathbf{\Sigma}_{\mathcal{Q}}(t_{m-1}))$
Accordingly,

$$(\Psi_1(\mathbf{T}_{\hat{\mathcal{Q}}}(t_1) - \mathbf{T}_{\hat{\mathcal{Q}}}(t_0)), \ldots, \Psi_m(\mathbf{T}_{\hat{\mathcal{Q}}}(t_m) - \mathbf{T}_{\hat{\mathcal{Q}}}(t_{m-1})))$$

forms a set of asymptotically independent random variables for any set of Borel-measurable functions $\{\Psi_1, \ldots, \Psi_m\}$ with $\Psi_j \colon \mathbb{R}^{|\mathcal{Q}|} \to \mathbb{R}^{k_j}$ for any $k_j \in \mathbb{N}$ for all $j \in \{1, \ldots, m\}$.

As in Corollary 3 of Danzer et al. (2023), this is a consequence of the previous results.
The last statement allows us to apply the results and techniques of Brendel et al. (2014) and Ditzhaus & Friedrich (2019)[7;11] also to the increments of the multivariate test statistics.

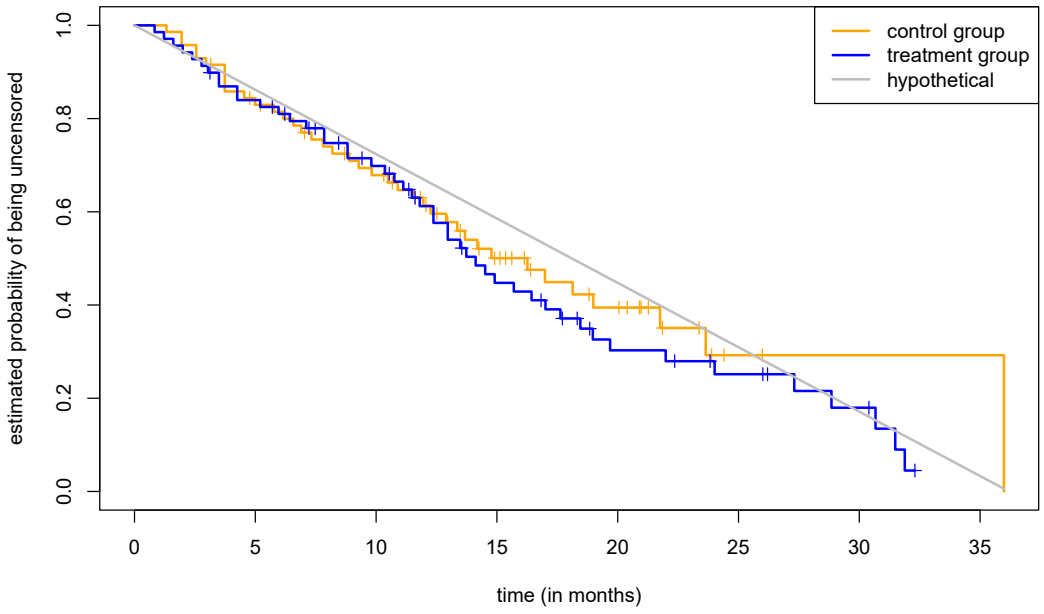
B Additional details to the real data example in Section 4 of the main manuscript

In Section 4 of the main manuscript, we reanalysed reconstructed data from the FAKTION trial[34] (NCT number NCT01992952). Further details, that are also mentioned in the main manuscript, will be given here.

B.1 Recruitment and censoring mechanism

Here, we give some additional details on how and on what basis we reconstructed/simulated recruitment dates. Recruitment dates are required to apply administrative censoring at the interim analysis date to obtain hypothetical interim data.
According to the reconstructed data, the last observation has been censored at $t_2 := 35.98$ months, which is approximately 36 months. This corresponds to the total recruitment duration specified in the published manuscript[34]. Additionally, the hypothetical censoring distribution at the final analysis that emerges from a uniform recruitment over the whole trial duration is very similar to the estimated distribution from reconstructed data (see Figure S6). Hence, we assume that $C^\star = \infty$ with probability 1 and $R \sim \text{Unif}[0, t_2]$ according to the notation introduced in Section 2.

Therefore, we set $R_i = t_2 - X_i(t_2)$ if $\delta_i(t_2) = 0$ and simulate $R_i \sim [0, t_2 - X_i(t_2)]$ if $\delta_i(t_2) = 1$. This

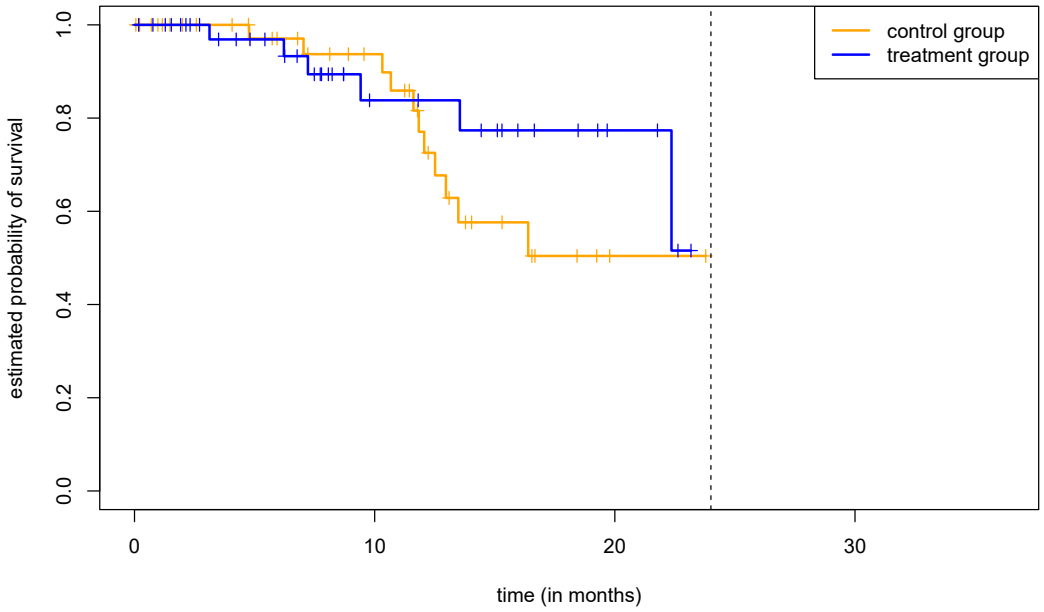


Supplementary Figure S6: Kaplan-Meier estimates of the survival function of the group-wise censoring distribution from reconstructed data. Grey line indicates hypothetical function if recruitment occurs uniformly over the whole trial duration and no additional loss to follow-up occurs.

simulation is in accordance with our preceding comments as $R|R \leq r \sim \text{Unif}[0, r]$ for any uniformly distributed random variable R on $[0, r_{\max}]$ with $r \leq r_{\max}$.
Obviously, the final results will depend on the simulated recruitment dates for uncensored observations. This dependence will be investigated further in Section B.4.

B.2 Royston-Parmar spline fits for interim data

After administrative censoring at calendar time $t_1 = 24$ was applied to the reconstructed and simulated data, the Kaplan-Meier estimates in Figure S7 can be obtained. Royston-Parmar splines were fitted to this data for each group separately using the function `flexsurvspline` from the R package `flexsurv`[23]. The Nelder-Mead optimization method has been chosen. Models with the number of interior knots

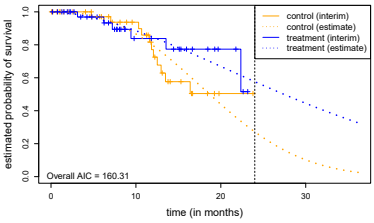


Supplementary Figure S7: Group-wise Kaplan-Meier estimates from interim data.

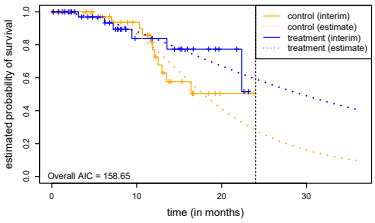
$p \in \{0, 1, 2\}$ and on all three available scales (hazard, odds, normal) were fitted. For each fit, the combined AIC was computed. The values can be found in Table S5. The lowest AIC is achieved for $p = 0$ on the normal scale. This results in a log-normal distribution[21]. It would also be possible to choose a different p and scale for the two groups. However, we restricted ourselves to the application of the same modeling parameters for both groups. On the next page, in Figure S8, the fitted Royston-Parmar spline models for all considered configurations are shown. On the page following afterwards (Figure S9), the Kaplan-Meier estimate based on the complete data is also shown in order to display the extrapolation performance.

		scale		
		hazard	odds	normal
p	0	160.31	158.65	157.97
	1	162.03	161.82	162.03
	2	160.70	161.04	161.51

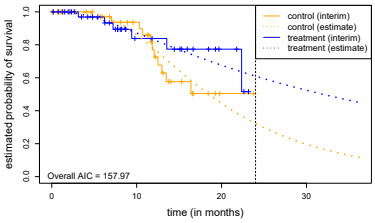
Supplementary Table S5: AIC values for various Royston-Parmar spline models when fitted to the interim data



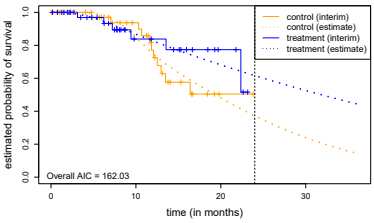
(a) $p = 0$, hazard scale



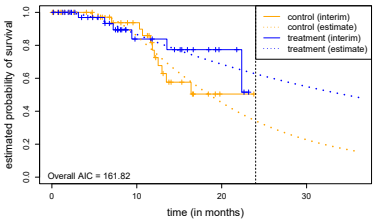
(b) $p = 0$, odds scale



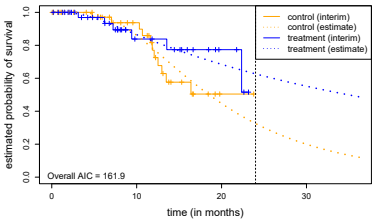
(c) $p = 0$, normal scale



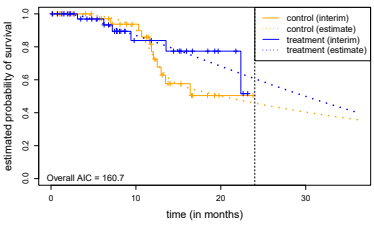
(d) $p = 1$, hazard scale



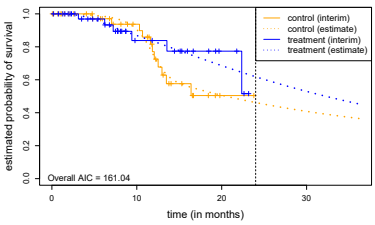
(e) $p = 1$, odds scale



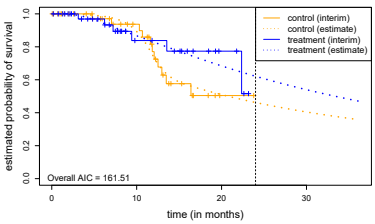
(f) $p = 1$, normal scale



(g) $p = 2$, hazard scale

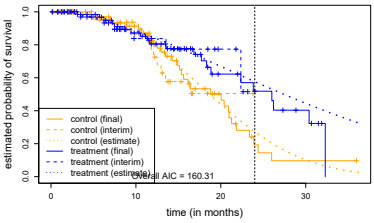


(h) $p = 2$, odds scale

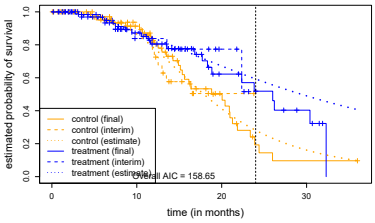


(i) $p = 2$, normal scale

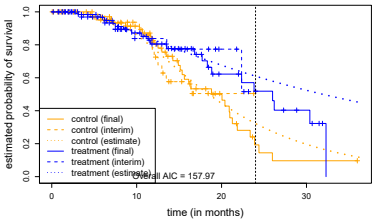
Supplementary Figure S8: Fits of Royston-Parmar spline models to interim data.



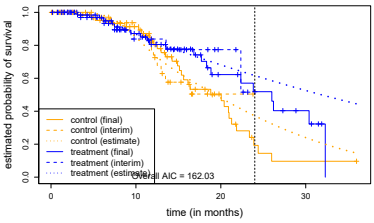
(a) $p = 0$, hazard scale



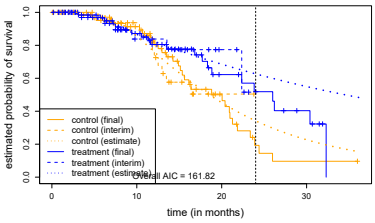
(b) $p = 0$, odds scale



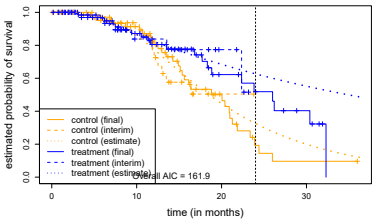
(c) $p = 0$, normal scale



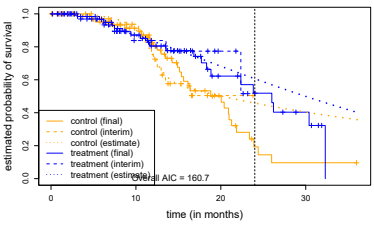
(d) $p = 1$, hazard scale



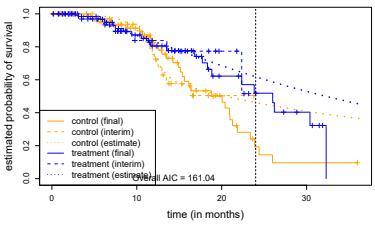
(e) $p = 1$, odds scale



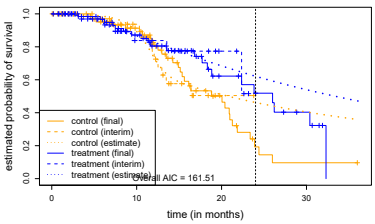
(f) $p = 1$, normal scale



(g) $p = 2$, hazard scale



(h) $p = 2$, odds scale



(i) $p = 2$, normal scale

Supplementary Figure S9: Fits of Royston-Parmar spline models to interim data with additional display of the final Kaplan-Meier estimates.

B.3 Conditional power calculations

Here, we present the results of the conditional power calculations. We consider

- 8 different mdir combination tests for the first stage (as presented in Table 2 in the main manuscript); the sets of weights used in these tests will be indexed as follows

  $\mathcal{Q}_{\mathbf{mdir},1} = \{(0,0),(1,0),(0,1),(1,1)\}$
  $\mathcal{Q}_{\mathbf{mdir},2} = \{(0,0),(1,0),(0,1)\}$
  $\mathcal{Q}_{\mathbf{mdir},3} = \{(0,0),(1,0),(1,1)\}$
  $\mathcal{Q}_{\mathbf{mdir},4} = \{(0,0),(0,1),(1,1)\}$
  $\mathcal{Q}_{\mathbf{mdir},5} = \{(0,0),(1,0)\}$
  $\mathcal{Q}_{\mathbf{mdir},6} = \{(0,0),(0,1)\}$
  $\mathcal{Q}_{\mathbf{mdir},7} = \{(0,0),(1,1)\}$
  $\mathcal{Q}_{\mathbf{mdir},8} = \{(0,0)\}$

- 9 different parameter constellations to fit Royston-Parmar splines to the interim data (number of knots $p \in \{0,1,2\}$, hazard, odds or normal scale)

- 4 different single-weighted tests for the second stage

For each combination of the mdir combination tests in the first stage and parameter constellation of the Royston-Parmar spline mode, the weighted test achieving the highest conditional power is marked in bold. Obviously, the choice does not depend on the first stage test statistic, as it is a monotone function of the standardized drift that is computed for each of the weighted tests in the second stage.
It is remarkable that for most of the choices of parameters for the Royston-Parmar splines (6 out of 9), the $(1,1)$-weighted test is favoured. This test actually has the best performance (see main manuscript). However, the test with the worst performance (weight $(1,0)$) is also chosen once. In the remainig two cases, the $(0,1)$-weighted test is favoured.

		hazard scale				odds scale				normal scale			
2nd stage weight		$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$p=0$	$\mathcal{Q}_{\text{mdir},1}$	0.8089	0.6396	**0.8868**	0.87	0.7537	0.6233	0.8105	**0.8295**	0.7052	0.5649	0.7955	**0.8033**
	$\mathcal{Q}_{\text{mdir},2}$	0.7653	0.5821	**0.8553**	0.8356	0.704	0.5651	0.7671	**0.7888**	0.6514	0.5052	0.7503	**0.7591**
	$\mathcal{Q}_{\text{mdir},3}$	0.8307	0.6702	**0.9019**	0.8868	0.7791	0.6543	0.8321	**0.8497**	0.7331	0.5972	0.8182	**0.8255**
	$\mathcal{Q}_{\text{mdir},4}$	0.8191	0.6538	**0.8939**	0.8779	0.7655	0.6376	0.8206	**0.839**	0.7182	0.5798	0.8061	**0.8137**
	$\mathcal{Q}_{\text{mdir},5}$	0.7851	0.6076	**0.8698**	0.8513	0.7263	0.5908	0.7867	**0.8073**	0.6754	0.5314	0.7707	**0.7791**
	$\mathcal{Q}_{\text{mdir},6}$	0.7411	0.5519	**0.8371**	0.8158	0.6769	0.5347	0.743	**0.7659**	0.6226	0.4746	0.7253	**0.7345**
	$\mathcal{Q}_{\text{mdir},7}$	0.8384	0.6814	**0.9072**	0.8927	0.7882	0.6658	0.8398	**0.8569**	0.7433	0.6093	0.8264	**0.8334**
	$\mathcal{Q}_{\text{mdir},8}$	0.7605	0.5759	**0.8517**	0.8316	0.6985	0.5588	0.7622	**0.7842**	0.6455	0.4989	0.7453	**0.7541**
$p=1$	$\mathcal{Q}_{\text{mdir},1}$	0.598	0.4835	0.6715	**0.6932**	0.679	0.5493	0.7718	**0.7878**	0.74	0.5875	**0.8501**	0.85
	$\mathcal{Q}_{\text{mdir},2}$	0.539	0.4241	0.6156	**0.6386**	0.6235	0.4895	0.7239	**0.7417**	0.6891	0.5282	**0.8123**	0.8123
	$\mathcal{Q}_{\text{mdir},3}$	0.6297	0.5166	0.7009	**0.7217**	0.7081	0.5819	0.7961	**0.811**	0.7662	0.6194	**0.8686**	0.8685
	$\mathcal{Q}_{\text{mdir},4}$	0.6126	0.4987	0.6852	**0.7065**	0.6925	0.5643	0.7831	**0.7987**	0.7522	0.6022	**0.8588**	0.8587
	$\mathcal{Q}_{\text{mdir},5}$	0.565	0.45	0.6405	**0.663**	0.6482	0.5157	0.7455	**0.7625**	0.7119	0.5543	**0.8295**	0.8295
	$\mathcal{Q}_{\text{mdir},6}$	0.5084	0.3942	0.5859	**0.6095**	0.594	0.4589	0.6977	**0.7163**	0.6615	0.4976	**0.791**	0.7909
	$\mathcal{Q}_{\text{mdir},7}$	0.6415	0.5291	0.7117	**0.7321**	0.7188	0.5941	0.8048	**0.8194**	0.7757	0.6313	**0.8751**	0.8751
	$\mathcal{Q}_{\text{mdir},8}$	0.5327	0.4179	0.6095	**0.6327**	0.6175	0.4832	0.7186	**0.7366**	0.6835	0.5219	**0.8081**	0.808
$p=2$	$\mathcal{Q}_{\text{mdir},1}$	0.1992	**0.222**	0.1167	0.191	0.2345	0.2411	0.1687	**0.2442**	0.2532	0.2496	0.2033	**0.2743**
	$\mathcal{Q}_{\text{mdir},2}$	0.1599	**0.1799**	0.0898	0.1528	0.1909	0.1968	0.1336	**0.1996**	0.2076	0.2044	0.1636	**0.2266**
	$\mathcal{Q}_{\text{mdir},3}$	0.2231	**0.2474**	0.1338	0.2143	0.2607	0.2677	0.1904	**0.271**	0.2805	0.2767	0.2276	**0.3026**
	$\mathcal{Q}_{\text{mdir},4}$	0.21	**0.2334**	0.1243	0.2015	0.2463	0.2531	0.1785	**0.2563**	0.2655	0.2618	0.2143	**0.2872**
	$\mathcal{Q}_{\text{mdir},5}$	0.1764	**0.1977**	0.101	0.1688	0.2094	0.2156	0.1483	**0.2185**	0.227	0.2236	0.1803	**0.2469**
	$\mathcal{Q}_{\text{mdir},6}$	0.1419	**0.1604**	0.078	0.1354	0.1707	0.1762	0.1177	**0.1788**	0.1864	0.1833	0.1453	**0.2042**
	$\mathcal{Q}_{\text{mdir},7}$	0.2326	**0.2574**	0.1407	0.2236	0.2709	0.2781	0.199	**0.2814**	0.2911	0.2872	0.2371	**0.3136**
	$\mathcal{Q}_{\text{mdir},8}$	0.1561	**0.1758**	0.0873	0.1491	0.1867	0.1925	0.1302	**0.1952**	0.2032	0.2	0.1597	**0.2219**

Supplementary Table S6: Conditional power for different choices of $\mathcal{Q}_{\text{mdir}}$ single weighted log-rank test for the second stage test for different parameter configurations of the Royston-Parmar spline model

B.4 Dependence from simulated recruitment data

The results in Section 4 and in the previous subsections of course depend on the simulated recruitment data for patients with uncensored event time data.
Here, we want to assess the dependence from this simulation. Therefore, we repeat the procedure 10,000 times. Each time, interim data is analysed with the 8 different mdir combination tests mentioned above and 9 different Royston-Parmar spline models are fitted to the data. The second stage data is analysed with one of the for different single-weighted log-rank tests listed above. In Supplementary Table **??**, the empirical power of the respective combinations of first- and second-stage tests is shown.

first stage test	second stage test			
	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\mathcal{Q}_{\text{mdir},1}$	0.5492	0.3827	0.3596	0.8997
$\mathcal{Q}_{\text{mdir},2}$	0.5297	0.3782	0.3478	0.8859
$\mathcal{Q}_{\text{mdir},3}$	0.5994	0.4228	0.4032	0.9147
$\mathcal{Q}_{\text{mdir},4}$	0.5380	0.3594	0.3355	0.8942
$\mathcal{Q}_{\text{mdir},5}$	0.5768	0.4169	0.3924	0.9044
$\mathcal{Q}_{\text{mdir},6}$	0.0769	0.0042	0.1570	0.6228
$\mathcal{Q}_{\text{mdir},7}$	0.5915	0.4040	0.3800	0.9108
$\mathcal{Q}_{\text{mdir},8}$	0.0890	0.0081	0.1770	0.7065

Supplementary Table S7: Empirical power of the simulation study for all combinations of first stage combination testing procedures and single-weighted log-rank testing procedures in the second stage.

Please note that the combination of $\mathcal{Q}_{\text{mdir},8}$ in the first stage and the $(0,0)$-weighted log-rank test in the second stage basically constitutes a two-stage standard log-rank test as in[3]. In about 9% of all simulation runs, this leads to a rejection although the originally applied simple one-stage standard log-rank leads to a rejection.
Power can strongly be increased if a different test is chosen in the first stage and the $(1,1)$-weighted log-rank test is chosen for the second stage. However, this is a retrospective assessment and one cannot guarantee a better choice for the first-stage test. Nevertheless, if a deviation from proportional hazards is anticipated, it is reasonable to consider a different test for the first stage.
For the second stage, we consider a choice based on an extrapolation with Royston-Parmar splines, a model selection based on AIC values and a consecutive conditional power calculation. In Supplementary Table S8, it is shown how often the 9 Royston-Parmar spline models used in this example so far are chosen based on the AIC values. Based on these choices, the second stage tests considered here, are

		scale		
		hazard	odds	normal
p	0	0.1869	0.1741	0.4637
	1	0.0258	0.0032	0.0541
	2	0.0824	0.0011	0.0087

Supplementary Table S8: Relative frequency with which the various Royston-Parmar spline models are selected based on the AIC

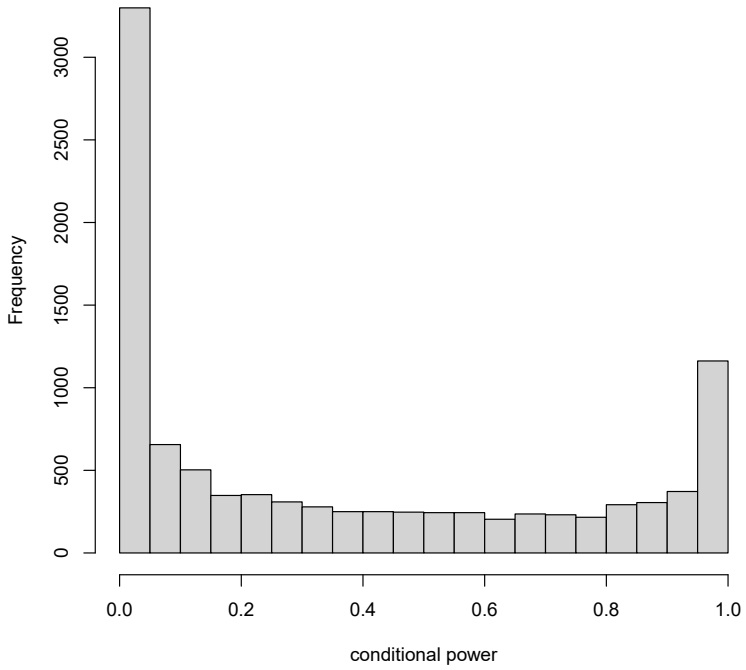
chosen with the following frequencies if $\mathcal{Q}_{\text{mdir},8}$ is applied in the first stage: $(0,0)$: 0.1755; $(1,0)$: 0.4293; $(0,1)$: 0.1258; $(1,1)$: 0.2694. It seems like the chosen model prefers the $(1,0)$-weighted test although the performance for the weight $(1,1)$ is much better according to Supplementary Table S7.
Finally, we can evaluate the performance of this procedure in terms of the power. This needs to be done separately for each first stage test statistic. The results can be found in Supplementary Table S9

first stage test	$\mathcal{Q}_{\text{mdir},1}$	$\mathcal{Q}_{\text{mdir},2}$	$\mathcal{Q}_{\text{mdir},3}$	$\mathcal{Q}_{\text{mdir},4}$	$\mathcal{Q}_{\text{mdir},5}$	$\mathcal{Q}_{\text{mdir},6}$	$\mathcal{Q}_{\text{mdir},7}$	$\mathcal{Q}_{\text{mdir},8}$
empirical power	0.4802	0.4699	0.5103	0.4613	0.5027	0.2836	0.4915	0.3057

Supplementary Table S9: Relative frequency with which the various Royston-Parmar spline models are selected based on the AIC

One can see that the power can be increased by about 6 percentage points even if the standard log-rank test is chosen in the first stage. If the first stage test statistic is chosen appropriately, it can even

raise to about 50%.
In Supplementary Figure S10, we show the empirical distribution of the maximal conditional power among the four tests for the second stage, computed based on the Royston-Parmar spline model with the lowest AIC. We restrict ourselves to the case that the weights in $\mathcal{Q}_{\text{mdir},1}$ have been chosen for the first stage. Correpsponding histograms for other choices for the first stage are very similar. We encounter a well-known problem in conditional power calculation in adaptive designs. This is characterised by the fact that the distribution of this variable tends towards extremes, i.e. very large and very small values [46]. It could therefore make sense to consider alternative concepts here.



Supplementary Figure S10: Empirical distribution of the maximal conditional power that can be obtained for the four candidate tests for the second stage. The first stage test is an mdir combination test with weights in $\mathcal{Q}_{\text{mdir},1}$. The Royston-Parmar spline model is chosen based on the AIC.

B.5 Application of modestly weighted log-rank tests

As an additional analysis, we conduct a similar analysis to the one of Section B.4 with a different class of weights. The class of Fleming-Harrington weight has been criticized for its undesirable properties when applied in a one-sided testing procedure [39]. This applies equally when combining them in a combination testing procedure. In this context, Magirr and Burman suggested a different class of weights, which they termed "modest weights" [40].
The class of modest weights is parametrized by some threshold time $s^\star \geq 0$ and given by

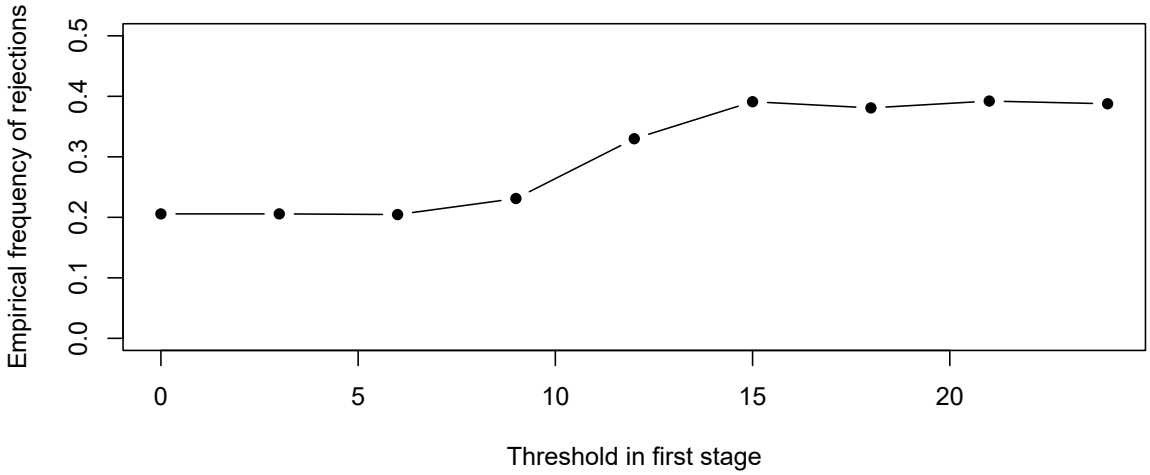
$$\hat{Q}(t,s) = w_{\text{modest},s^\star}(\hat{S}(t,s-)) \coloneqq \frac{1}{\max\left(\hat{S}(t,s-), \hat{S}(t,s^\star-)\right)}. \tag{8}$$

As the weights of Fleming and Harrington with $\rho > 0$ and $\gamma = 0$, this function is increasing in the second argument. Differently from these functions, it is bounded from below and it stays constant after the second argument exceeds the threshold $s^\star$. As previously shown, are an appropriate tool to detect late effects [40]. Notably, for $s^\star = 0$ the modestly weighted log-rank test is the same as the standard log-rank test as it is constantly equal to 1. The modest weights also fulfill the assumptions of Theorem 1 and can hence also be incorporated into our framework.
Here, we consider designs where a single modestly weighted log-rank test is applied in the first stage and

the second test stage is chosen among the modestly weighted tests with the thresholds $s^\star \in \{0, 3, 6, 9, 12, 15, 18, 21, 24\}$. The modest weights could also be used in a combination testing procedure as e.g. the *mdir* test. However, we do not consider this option here.

As previously, 10,000 simulation runs are made where the unreconstructable recruitment dates are simulated. Conditional power calculations for the 9 different modestly weighted tests are made from the same 9 different Royston-Parmar spline models as above. The best spline model is chosen based on the AIC and the second stage test is chosen as the modestly weighted test with the highest conditional power according to this spline model. The results in terms of the empirical power of the procedure in dependence from the (fixed) test in the first stage are displayed in Figure S11.

For comparison, the empirical rejection rate for the two-stage procedure which applies the standard



Supplementary Figure S11: Empirical power for an adpaitve design with a modestly weighted log-rank test in the first stage and a modestly weighted log-rank test in the second stage that is chosen based on conditional power consideration based on Royston-Parmar spline models. 9 different modestly weighted log-rank tests are considered. The threshold for the (fixed) test in the first stage is given by the x-axis.

log-rank test in both stages (i.e. $s^\star = 0$) is 9.31%. It is obvious that the selection procedure increases the rejection rate, even if the standard log-rank test is chosen for the first stage. The increase is even more articulate if a better-suited test is chosen in the first stage.

C Additional simulation results

In this section, we provide additional results to the simulation study in Section 5 of the main manuscript.

C.1 Empirical type I error rates

Here, we present the empirical type I error rates that emerge from fixed (in particular non-adaptive) combinations of testing procedures in the two stages. This shall demonstrate that our selection procedure (see Table 3) induces no additional type I error rate inflation.
For the sake of presentability, we use the following numbering system for the various tests:

1: Standard log-rank test

2: Weighted log-rank test with Fleming-Harrington weight $w^{(0,1)} \circ \hat{F}$

3: Weighted log-rank test with Fleming-Harrington weight $w^{(0,2)} \circ \hat{F}$

4: Weighted log-rank test with Fleming-Harrington weight $w^{(0,3)} \circ \hat{F}$

5: Weighted log-rank test with Fleming-Harrington weight $w^{(1,1)} \circ \hat{F}$

6: Weighted log-rank test with Fleming-Harrington weight $w^{(1,0)} \circ \hat{F}$

7: Weighted log-rank test with Fleming-Harrington weight $w^{(2,0)} \circ \hat{F}$

8: Weighted log-rank test with Fleming-Harrington weight $w^{(3,0)} \circ \hat{F}$

9: *mdir* combination test based on the set of weights $\{w^{(0,0)} \circ \hat{F}, w^{(1,0)} \circ \hat{F}\}$

10: *mdir* combination test based on the set of weights $\{w^{(0,0)} \circ \hat{F}, w^{(1,1)} \circ \hat{F}\}$

11: *mdir* combination test based on the set of weights $\{w^{(0,0)} \circ \hat{F}, w^{(0,1)} \circ \hat{F}\}$

12: *mdir* combination test based on the set of weights $\{w^{(0,0)} \circ \hat{F}, w^{(1,0)} \circ \hat{F}, w^{(0,1)} \circ \hat{F}\}$

13: *mdir* combination test based on the set of weights $\{w^{(0,0)} \circ \hat{F}, w^{(1,0)} \circ \hat{F}, w^{(1,1)} \circ \hat{F}\}$

14: *mdir* combination test based on the set of weights $\{w^{(0,0)} \circ \hat{F}, w^{(1,1)} \circ \hat{F}, w^{(0,1)} \circ \hat{F}\}$

15: *mdir* combination test based on the set of weights $\{w^{(0,0)} \circ \hat{F}, w^{(1,0)} \circ \hat{F}, w^{(1,1)} \circ \hat{F}, w^{(0,1)} \circ \hat{F}\}$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.028	0.029	0.0302	0.0317	0.0274	0.027	0.0263	0.0259	0.0241	0.0255	0.0259	0.0254	0.0255	0.0262	0.0251
2	0.0301	0.0302	0.0316	0.0323	0.0308	0.0292	0.0289	0.0282	0.0279	0.028	0.0275	0.0274	0.0275	0.0273	0.0278
3	0.0313	0.0318	0.032	0.0337	0.0311	0.0306	0.0305	0.0305	0.0279	0.0289	0.0294	0.0278	0.0275	0.0287	0.028
4	0.0315	0.0319	0.0328	0.034	0.0317	0.0297	0.0305	0.0312	0.0286	0.0298	0.0293	0.0287	0.0294	0.0299	0.0284
5	0.0292	0.0292	0.0292	0.0301	0.0286	0.0273	0.0272	0.0278	0.0262	0.0267	0.0272	0.0275	0.0264	0.0268	0.0265
6	0.0269	0.0273	0.0302	0.0313	0.0269	0.0263	0.025	0.025	0.0256	0.0253	0.0259	0.0264	0.0252	0.0263	0.0258
7	0.0267	0.0281	0.0297	0.0315	0.0272	0.0254	0.0245	0.0245	0.0249	0.0255	0.0255	0.0255	0.0252	0.0253	0.0259
8	0.0269	0.0277	0.03	0.0317	0.0269	0.0253	0.0252	0.0244	0.0249	0.0257	0.0257	0.0259	0.0253	0.0261	0.0262
9	0.0258	0.0268	0.0274	0.0283	0.0255	0.025	0.0236	0.0237	0.0227	0.0236	0.0232	0.0234	0.0233	0.0236	0.0234
10	0.0267	0.0275	0.0278	0.0285	0.0261	0.0257	0.0253	0.0256	0.0231	0.0241	0.0248	0.0242	0.0241	0.0251	0.0246
11	0.0253	0.0258	0.0288	0.0304	0.0262	0.0251	0.0241	0.0235	0.0237	0.0239	0.0234	0.0234	0.0234	0.0229	0.0236
12	0.0257	0.0272	0.0282	0.0291	0.0265	0.0252	0.024	0.0243	0.0234	0.0238	0.0237	0.0232	0.0236	0.0246	0.0237
13	0.0263	0.0268	0.0276	0.0275	0.0267	0.0257	0.0245	0.0254	0.0232	0.0241	0.0242	0.0241	0.0238	0.025	0.024
14	0.0256	0.0268	0.0282	0.0299	0.0266	0.0262	0.0245	0.0244	0.0238	0.0248	0.0242	0.0237	0.0241	0.0245	0.0246
15	0.0258	0.027	0.0274	0.0287	0.0266	0.0244	0.0242	0.0249	0.0227	0.0235	0.0239	0.0231	0.0232	0.0252	0.0232

Supplementary Table S10: Empirical type I error rates for fixed combinations of stagewise tests with 50 patients in each group. Rows correspond to first-stage tests and columns correspond to second-stage tests.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.0283	0.0273	0.0278	0.0272	0.0285	0.0281	0.0283	0.0264	0.0259	0.0282	0.0279	0.0279	0.0266	0.0285	0.0275
2	0.0264	0.0274	0.0258	0.0258	0.0269	0.0261	0.0267	0.026	0.0248	0.0257	0.0258	0.0269	0.0265	0.0267	0.027
3	0.0272	0.0258	0.0271	0.028	0.0275	0.0271	0.0254	0.0256	0.0259	0.0274	0.0271	0.0288	0.0277	0.0277	0.029
4	0.0281	0.0264	0.0273	0.0283	0.0284	0.0275	0.0258	0.0261	0.0278	0.0287	0.0286	0.0285	0.029	0.0283	0.0288
5	0.026	0.0275	0.027	0.0273	0.0272	0.0266	0.0276	0.0265	0.0256	0.0262	0.026	0.0274	0.0265	0.0277	0.0267
6	0.0276	0.0266	0.0267	0.0263	0.0283	0.0294	0.0277	0.0277	0.025	0.0277	0.0282	0.0276	0.0262	0.0285	0.027
7	0.0278	0.0262	0.0266	0.0266	0.0283	0.0303	0.0294	0.0278	0.0264	0.0268	0.0294	0.0277	0.0266	0.0289	0.0278
8	0.0276	0.026	0.0259	0.0264	0.0279	0.03	0.0282	0.0287	0.0264	0.0276	0.0287	0.0285	0.0258	0.0279	0.0272
9	0.0261	0.0251	0.026	0.0253	0.0272	0.026	0.0258	0.0244	0.0243	0.0264	0.0265	0.027	0.0263	0.0269	0.0267
10	0.0268	0.0269	0.0261	0.0246	0.0283	0.0265	0.0263	0.0245	0.0245	0.0278	0.0266	0.0265	0.0262	0.0273	0.0258
11	0.0256	0.0263	0.0262	0.0257	0.027	0.0279	0.0272	0.0263	0.0239	0.0261	0.0275	0.0266	0.0248	0.0279	0.0252
12	0.027	0.0249	0.0264	0.0267	0.0276	0.0272	0.0269	0.0247	0.0241	0.0264	0.0272	0.0271	0.0256	0.0276	0.027
13	0.0264	0.0261	0.026	0.0252	0.0271	0.0263	0.0262	0.0256	0.0244	0.0272	0.0262	0.0272	0.0264	0.0269	0.0267
14	0.0273	0.0263	0.0266	0.0258	0.0291	0.0284	0.0275	0.0256	0.025	0.0281	0.0283	0.0278	0.0264	0.0283	0.0273
15	0.0271	0.0264	0.0258	0.0257	0.0276	0.0277	0.0258	0.0251	0.0237	0.0269	0.0275	0.0272	0.0256	0.0277	0.0261

Supplementary Table S11: Empirical type I error rates for fixed combinations of stagewise tests with 100 patients in each group. Rows correspond to first-stage tests and columns correspond to second-stage tests.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.0228	0.0211	0.0214	0.0218	0.0232	0.024	0.0236	0.0238	0.0224	0.0231	0.0233	0.0231	0.023	0.0231	0.0231
2	0.0257	0.0243	0.0234	0.0237	0.0249	0.0254	0.024	0.0261	0.025	0.0255	0.0254	0.0256	0.0246	0.0241	0.0251
3	0.0262	0.0252	0.0251	0.0249	0.026	0.0258	0.0262	0.0274	0.0255	0.0252	0.026	0.0256	0.0247	0.0255	0.0259
4	0.0261	0.026	0.0255	0.026	0.0268	0.0264	0.0264	0.0276	0.0271	0.0262	0.027	0.0272	0.026	0.0266	0.027
5	0.0237	0.0221	0.0222	0.0229	0.0242	0.0248	0.0246	0.0247	0.0225	0.0238	0.0229	0.0235	0.0224	0.0242	0.0239
6	0.0244	0.0216	0.0216	0.0218	0.0237	0.0239	0.0237	0.0237	0.0225	0.0245	0.0245	0.0233	0.0222	0.0237	0.0243
7	0.024	0.0234	0.0228	0.0227	0.0239	0.0236	0.024	0.0223	0.0235	0.0248	0.0241	0.0243	0.023	0.0242	0.0234
8	0.0242	0.0224	0.0225	0.0228	0.0233	0.0236	0.0226	0.0228	0.0232	0.024	0.0239	0.0238	0.0235	0.0239	0.0228
9	0.0237	0.0229	0.0221	0.0226	0.0236	0.0245	0.0244	0.0246	0.0227	0.0242	0.0238	0.024	0.0227	0.024	0.0241
10	0.0219	0.0217	0.021	0.0212	0.0233	0.0237	0.0237	0.0241	0.0225	0.0223	0.023	0.0227	0.0218	0.023	0.0227
11	0.0231	0.0218	0.0221	0.0222	0.0225	0.0248	0.0236	0.0241	0.0218	0.0238	0.0238	0.0237	0.0223	0.0239	0.0237
12	0.0234	0.0213	0.0213	0.0224	0.0237	0.0239	0.0247	0.0241	0.0217	0.0239	0.0239	0.0231	0.022	0.0239	0.0235
13	0.0237	0.0228	0.0216	0.0228	0.0245	0.0235	0.0253	0.025	0.0233	0.024	0.0238	0.0235	0.023	0.0245	0.0238
14	0.023	0.0217	0.021	0.0219	0.0244	0.0251	0.0245	0.0239	0.0228	0.0246	0.0237	0.0228	0.0226	0.0228	0.023
15	0.0239	0.0224	0.0214	0.0223	0.0233	0.0242	0.0255	0.0243	0.0223	0.0239	0.024	0.0235	0.0229	0.0243	0.0236

Supplementary Table S12: Empirical type I error rates for fixed combinations of stagewise tests with 200 patients in each group. Rows correspond to first-stage tests and columns correspond to second-stage tests.

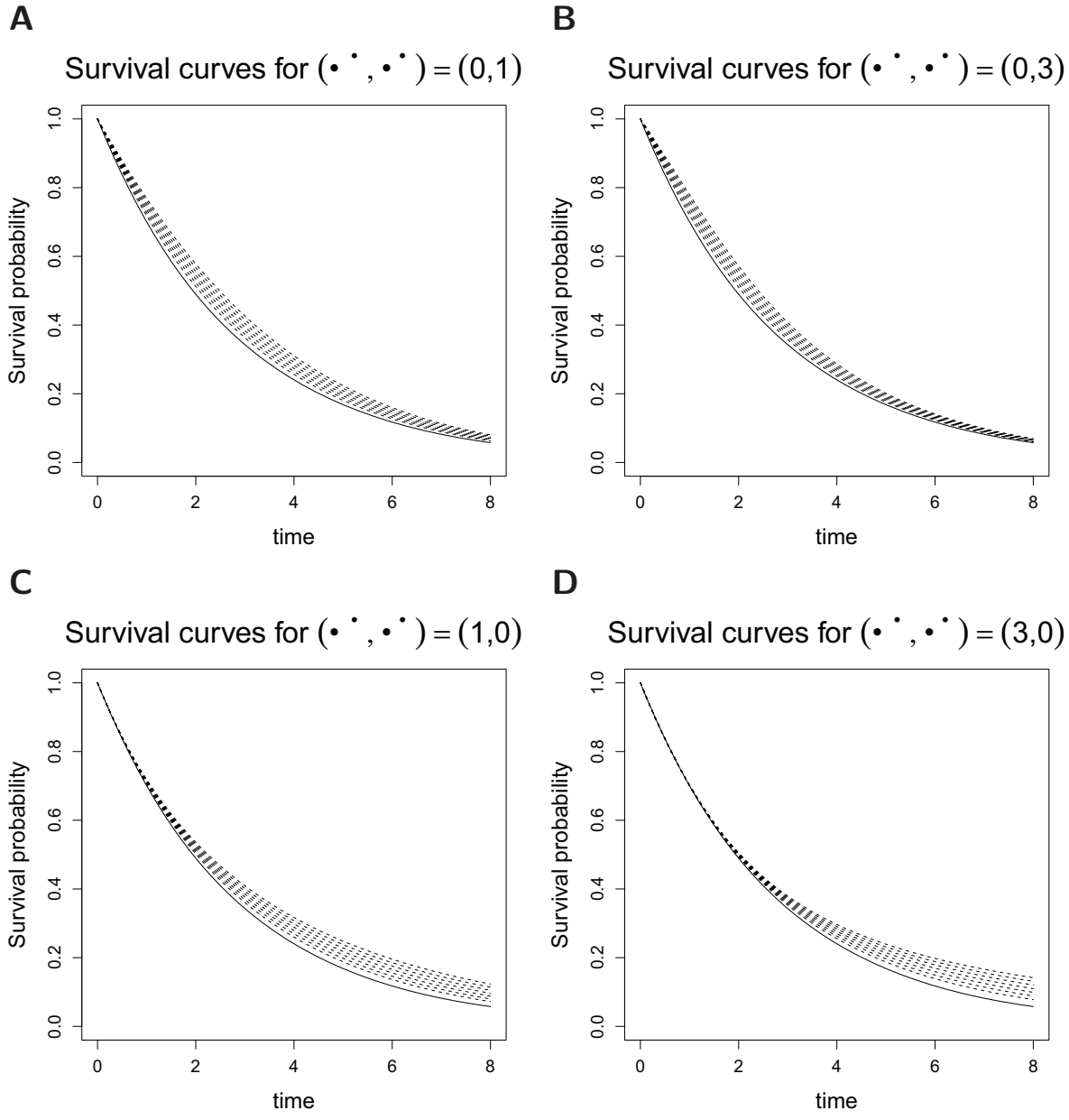
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.0263	0.0286	0.0288	0.029	0.0269	0.026	0.0254	0.0258	0.0277	0.027	0.0269	0.0278	0.0284	0.028	0.0282
2	0.0252	0.0279	0.0272	0.0275	0.0275	0.0257	0.025	0.0247	0.0271	0.0264	0.0243	0.025	0.0269	0.0256	0.0268
3	0.0263	0.0255	0.0259	0.0254	0.0265	0.0248	0.0247	0.0239	0.0261	0.0262	0.025	0.0252	0.0261	0.025	0.0262
4	0.0271	0.0266	0.0266	0.0265	0.0257	0.0235	0.0257	0.0251	0.0256	0.0276	0.0259	0.0259	0.0268	0.0258	0.0269
5	0.0255	0.0277	0.0283	0.0279	0.0273	0.0263	0.0263	0.0254	0.0263	0.0257	0.025	0.0256	0.0271	0.0256	0.0256
6	0.0279	0.0298	0.0286	0.0291	0.0288	0.0263	0.026	0.0256	0.0286	0.0288	0.0272	0.0284	0.0294	0.0284	0.0284
7	0.0276	0.0286	0.0279	0.0281	0.0285	0.0278	0.0269	0.0267	0.0282	0.0285	0.0278	0.028	0.0291	0.0284	0.0287
8	0.0275	0.0277	0.0271	0.0276	0.0284	0.027	0.027	0.0258	0.0282	0.0278	0.0268	0.0269	0.0277	0.0272	0.0276
9	0.0259	0.0285	0.0288	0.0286	0.0278	0.0261	0.0271	0.0261	0.0264	0.0266	0.0256	0.0269	0.0273	0.0262	0.0274
10	0.026	0.0274	0.0276	0.0276	0.0264	0.0258	0.0262	0.0256	0.0267	0.0265	0.0251	0.0269	0.0266	0.0259	0.0266
11	0.0274	0.0287	0.0285	0.0289	0.0286	0.0269	0.0264	0.0258	0.0281	0.0274	0.0267	0.0276	0.0284	0.0269	0.0266
12	0.0253	0.0276	0.0281	0.0272	0.0263	0.0258	0.0264	0.0269	0.0258	0.0267	0.0252	0.0261	0.0267	0.0254	0.0258
13	0.0253	0.0274	0.0275	0.0273	0.0265	0.025	0.0266	0.0264	0.0259	0.0263	0.0252	0.0256	0.0271	0.0261	0.0266
14	0.0265	0.028	0.0278	0.0275	0.0265	0.0258	0.0254	0.0262	0.0276	0.0266	0.0254	0.0265	0.0274	0.0255	0.0267
15	0.0255	0.0278	0.0272	0.027	0.0274	0.0257	0.0261	0.0259	0.0268	0.0265	0.025	0.0262	0.0265	0.0261	0.0258

Supplementary Table S13: Empirical type I error rates for fixed combinations of stagewise tests with 500 patients in each group. Rows correspond to first-stage tests and columns correspond to second-stage tests.

C.2 Power comparisons

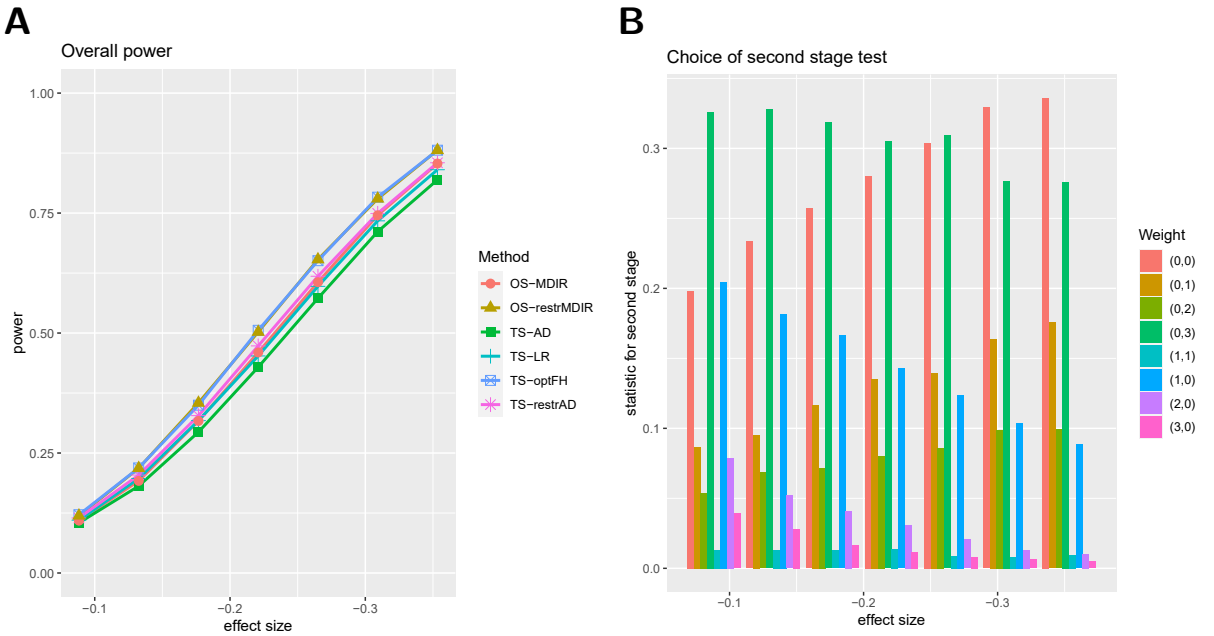
C.2.1 Additional deviation types

In this subsection, we supplement the results from Section 5.2 of the main manuscript by those results for the settings $(\rho^\star, \gamma^\star) \in \{(0,1), (0,3), (1,0), (3,0)\}$. The corresponding survival curves are shown in Figure S12.

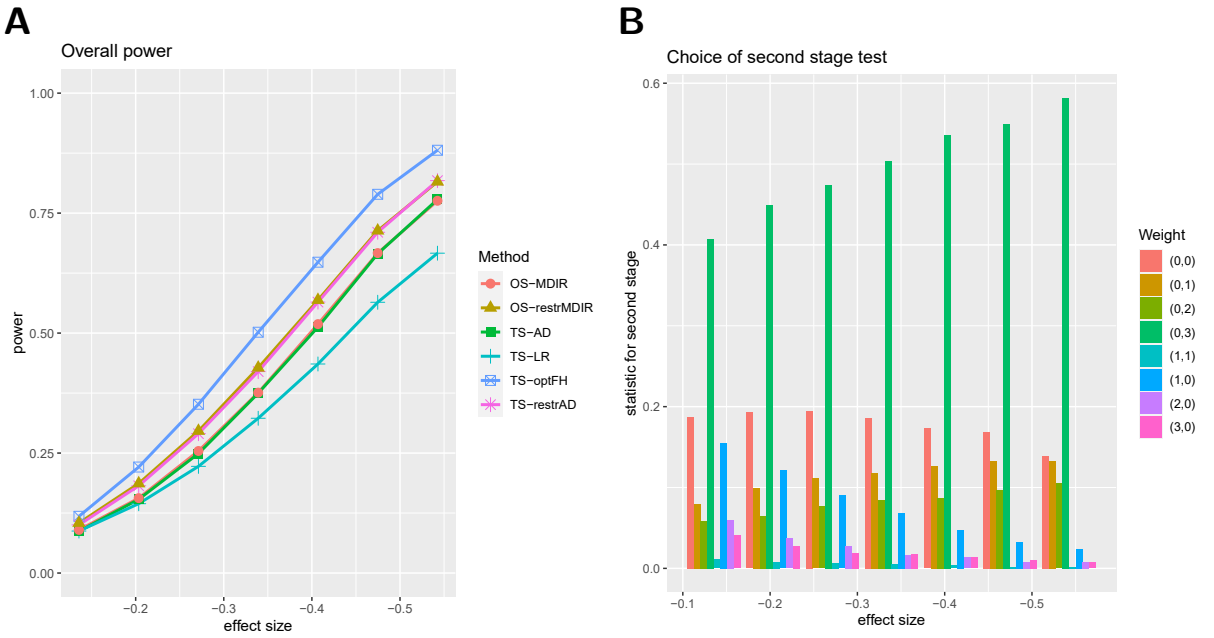


Supplementary Figure S12: Survival curves for three types of deviation of the distribution in the experimental group from the distribution in the control group. The survival curve in the control group is given by the solid line. Dashed lines are survival curves in the experimental group for the seven effect sizes $\{0.4 \cdot \theta_0, 0.6 \cdot \theta_0, 0.8 \cdot \theta_0, \theta_0, 1.2 \cdot \theta_0, 1.4 \cdot \theta_0, 1.6 \cdot \theta_0\}$. A) Survival curves in the slightly early effect case ($(\rho^\star, \gamma^\star) = (0,1)$) B) Survival curves in the very early effect case ($(\rho^\star, \gamma^\star) = (0,3)$) C) Survival curves in the slightly late effect case ($(\rho^\star, \gamma^\star) = (1,0)$) B) Survival curves in the very late effect case ($(\rho^\star, \gamma^\star) = (3,0)$)

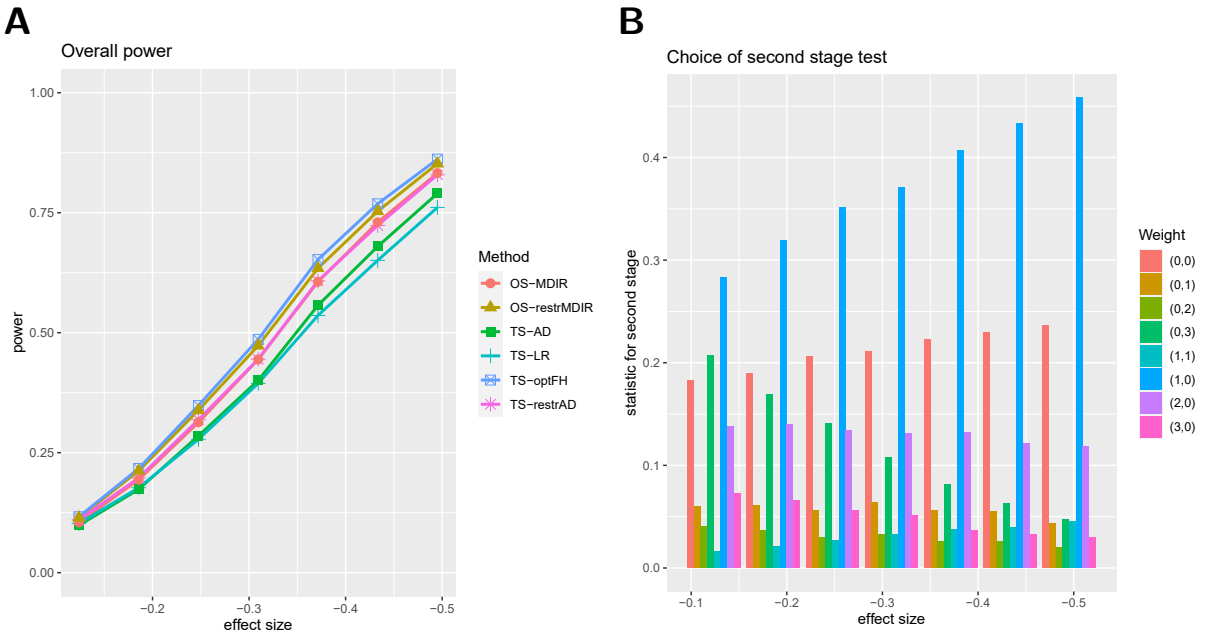
For these scenarios, the power curves and the bar plots showing the choices made by the procedure **TS-AD** for the second stage test can be found in Figures S13-S16. For slightly early and late effects (i.e. $(\rho^\star, \gamma^\star) = (0, 1)$ or $(\rho^\star, \gamma^\star) = (1, 0)$, respectively) we can see that all procedures perform very similarly. For strong early and late effects (i.e. $(\rho^\star, \gamma^\star) = (0, 3)$ or $(\rho^\star, \gamma^\star) = (3, 0)$, respectively), there is a marked gap between the optimal two-stage test and the two-stage standard log-rank test. The one-stage combination testing procedures and the adaptive procedures fill this gap. Once again, we can see that the two restricted procedures with pre-chosen sets of weights perform better than the unrestricted ones. It should be noted that our selection procedure tends to favor the log-rank test with weights $w^{(0,3)} \circ \hat{F}$ in the case of early effects and the log-rank test with weights $w^{(1,0)} \circ \hat{F}$ in the case of late effects. This shows that there is certainly room for improvement in this selection process.



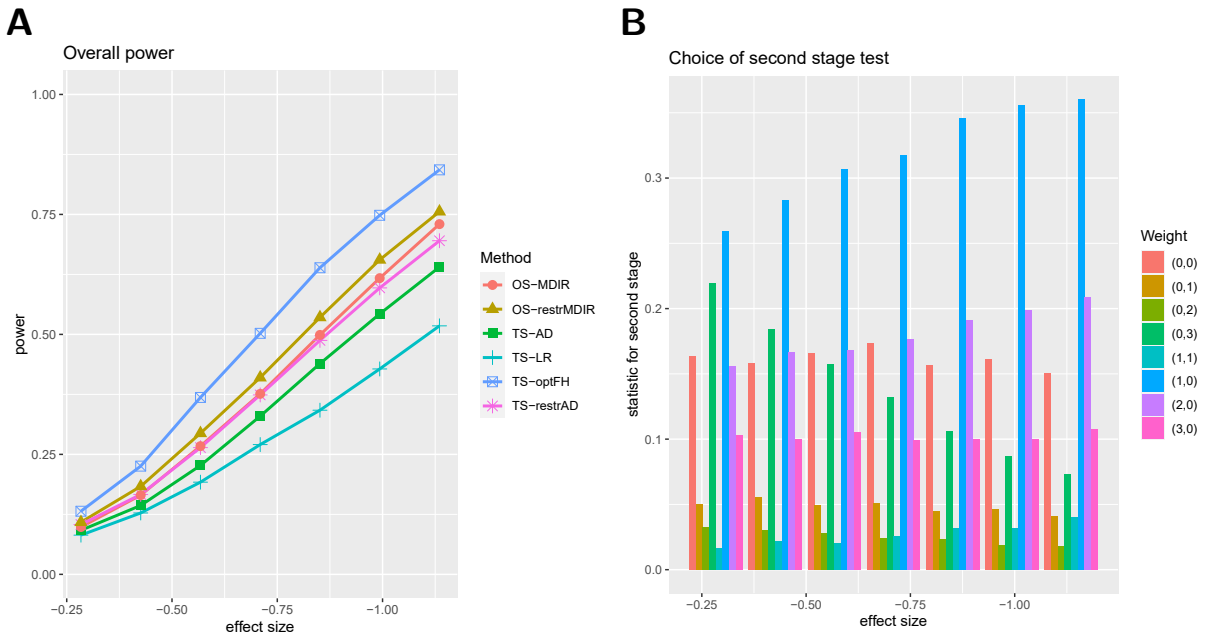
Supplementary Figure S13: A) Power curve for six testing procedures in case of proportional hazards ($(\rho^\star, \gamma^\star) = (0, 1)$). Please note that the two procedures TS-AD and TS-optFH as well as the two procedures OS-MDIR and OS-restrMDIR coincide in this case. B) Relative frequencies of the choice of single weighted tests for the second stage for the testing procedure TS-AD.



Supplementary Figure S14: A) Power curve for six testing procedures in case of proportional hazards ($(\rho^\star, \gamma^\star) = (0, 3)$). Please note that the two procedures TS-AD and TS-optFH as well as the two procedures OS-MDIR and OS-restrMDIR coincide in this case. B) Relative frequencies of the choice of single weighted tests for the second stage for the testing procedure TS-AD.



Supplementary Figure S15: A) Power curve for six testing procedures in case of proportional hazards ($(\rho^\star, \gamma^\star) = (1, 0)$). Please note that the two procedures TS-AD and TS-optFH as well as the two procedures OS-MDIR and OS-restrMDIR coincide in this case. B) Relative frequencies of the choice of single weighted tests for the second stage for the testing procedure TS-AD.



Supplementary Figure S16: A) Power curve for six testing procedures in case of proportional hazards ($(\rho^\star, \gamma^\star) = (3, 0)$). Please note that the two procedures TS-AD and TS-optFH as well as the two procedures OS-MDIR and OS-restrMDIR coincide in this case. B) Relative frequencies of the choice of single weighted tests for the second stage for the testing procedure TS-AD.

C.2.2 Model choice based on AIC

Here we show which of the 9 Royston-Parmar spline models is chosen based on their AIC value. As already mentioned, we can see that the AIC selection mechanism prefers models with 0 interior knots for our simulation scenarios. In particular, the model on the hazard scale is preferred. This spline model induces Weibull distributions for both groups.

Effect size	$p = 0$			$p = 1$			$p = 2$		
	hazard	odds	normal	hazard	odds	normal	hazard	odds	normal
$0.4\theta_0$	0.7892	0.0224	0.0001	0.0799	0.0243	0.0227	0.0233	0.0141	0.0241
$0.6\theta_0$	0.7807	0.0252	0.0001	0.0816	0.0273	0.0254	0.022	0.0143	0.0235
$0.8\theta_0$	0.7728	0.0276	0.0001	0.0806	0.0243	0.0263	0.0255	0.0166	0.0262
θ_0	0.7774	0.0284	0.0001	0.0788	0.0288	0.0253	0.0236	0.0141	0.0234
$1.2\theta_0$	0.777	0.0308	0	0.0755	0.0291	0.0203	0.0247	0.015	0.0276
$1.4\theta_0$	0.7689	0.038	0.0001	0.0754	0.0299	0.0223	0.0235	0.0153	0.0267
$1.6\theta_0$	0.7745	0.0368	0.0001	0.0747	0.0264	0.0244	0.0223	0.0163	0.0245

Supplementary Table S14: For each of the seven effect sizes, this table displays which of the nine Royston-Parmar spline models is rated as the best based on the AIC. As above, p refers to the number of inner knots in the spline model. The empirical rates refer to the total quantity of runs without early stopping of the corresponding simulated trial. This table refers to the scenario $(\rho^\star, \gamma^\star) = (0, 0)$

Effect size	$p = 0$			$p = 1$			$p = 2$		
	hazard	odds	normal	hazard	odds	normal	hazard	odds	normal
$0.4\theta_0$	0.763	0.0314	0.0001	0.0827	0.0262	0.0321	0.0219	0.0169	0.0257
$0.6\theta_0$	0.7614	0.0362	0	0.0799	0.0291	0.0347	0.0231	0.0131	0.0226
$0.8\theta_0$	0.7483	0.0412	0	0.0836	0.0291	0.0338	0.0239	0.0148	0.0253
$1\theta_0$	0.7332	0.0479	0.0001	0.0874	0.032	0.0365	0.0238	0.0149	0.0242
$1.2\theta_0$	0.7192	0.054	0.0005	0.0906	0.0338	0.0373	0.0221	0.0156	0.027
$1.4\theta_0$	0.7104	0.058	0.0003	0.0918	0.0362	0.0409	0.0247	0.0128	0.0248
$1.6\theta_0$	0.6998	0.063	0.0003	0.0936	0.0321	0.0416	0.0249	0.0162	0.0284

Supplementary Table S15: For each of the seven effect sizes, this table displays which of the nine Royston-Parmar spline models is rated as the best based on the AIC. As above, p refers to the number of inner knots in the spline model. The empirical rates refer to the total quantity of runs without early stopping of the corresponding simulated trial. This table refers to the scenario $(\rho^\star, \gamma^\star) = (1, 0)$

Effect size	$p = 0$			$p = 1$			$p = 2$		
	hazard	odds	normal	hazard	odds	normal	hazard	odds	normal
$0.4\theta_0$	0.7595	0.0335	0.0001	0.0828	0.0326	0.035	0.0193	0.014	0.0232
$0.6\theta_0$	0.7429	0.042	0	0.0836	0.0357	0.0368	0.0214	0.0134	0.0242
$0.8\theta_0$	0.7213	0.0518	0.0003	0.0812	0.0395	0.0429	0.0219	0.0157	0.0255
$1\theta_0$	0.6932	0.064	0.0001	0.094	0.0391	0.0489	0.0227	0.0154	0.0226
$1.2\theta_0$	0.6595	0.0711	0	0.0985	0.047	0.0517	0.0267	0.0167	0.0288
$1.4\theta_0$	0.6348	0.0833	0.0002	0.105	0.048	0.0517	0.0268	0.0191	0.031
$1.6\theta_0$	0.6007	0.0943	0.0001	0.1136	0.0533	0.0598	0.0297	0.0174	0.031

Supplementary Table S16: For each of the seven effect sizes, this table displays which of the nine Royston-Parmar spline models is rated as the best based on the AIC. As above, p refers to the number of inner knots in the spline model. The empirical rates refer to the total quantity of runs without early stopping of the corresponding simulated trial. This table refers to the scenario $(\rho^\star, \gamma^\star) = (2, 0)$

Effect size	$p = 0$			$p = 1$			$p = 2$		
	hazard	odds	normal	hazard	odds	normal	hazard	odds	normal
$0.4\theta_0$	0.7586	0.0323	0	0.0765	0.0351	0.0359	0.0223	0.0147	0.0246
$0.6\theta_0$	0.7262	0.0461	0.0001	0.0827	0.0369	0.0435	0.0236	0.0167	0.0243
$0.8\theta_0$	0.7094	0.053	0.0001	0.0851	0.0414	0.0426	0.0234	0.0183	0.0266
$1\theta_0$	0.6718	0.0697	0.0001	0.0954	0.0451	0.0496	0.0233	0.0175	0.0274
$1.2\theta_0$	0.6471	0.074	0.0003	0.0958	0.0522	0.058	0.0267	0.02	0.0259
$1.4\theta_0$	0.6214	0.0853	0.0004	0.1007	0.0557	0.0589	0.0288	0.0216	0.0273
$1.6\theta_0$	0.569	0.1016	0.0002	0.1132	0.0631	0.0644	0.0331	0.0229	0.0324

Supplementary Table S17: For each of the seven effect sizes, this table displays which of the nine Royston-Parmar spline models is rated as the best based on the AIC. As above, p refers to the number of inner knots in the spline model. The empirical rates refer to the total quantity of runs without early stopping of the corresponding simulated trial. This table refers to the scenario $(\rho^\star, \gamma^\star) = (3, 0)$

Effect size	$p = 0$			$p = 1$			$p = 2$		
	hazard	odds	normal	hazard	odds	normal	hazard	odds	normal
$0.4\theta_0$	0.7827	0.0209	0	0.0804	0.0256	0.0226	0.0246	0.0155	0.0276
$0.6\theta_0$	0.7762	0.0189	0	0.0864	0.0259	0.0215	0.0274	0.0163	0.0274
$0.8\theta_0$	0.7787	0.0189	0	0.0861	0.0244	0.0224	0.0259	0.0163	0.0273
$1\theta_0$	0.7788	0.0212	0	0.0903	0.0249	0.0176	0.0252	0.0142	0.0279
$1.2\theta_0$	0.7861	0.0182	0	0.0875	0.0229	0.0183	0.0236	0.0176	0.0258
$1.4\theta_0$	0.7889	0.0158	0	0.0836	0.0249	0.0135	0.0298	0.0132	0.0303
$1.6\theta_0$	0.7798	0.0179	0	0.0913	0.0253	0.0137	0.029	0.0134	0.0297

Supplementary Table S18: For each of the seven effect sizes, this table displays which of the nine Royston-Parmar spline models is rated as the best based on the AIC. As above, p refers to the number of inner knots in the spline model. The empirical rates refer to the total quantity of runs without early stopping of the corresponding simulated trial. This table refers to the scenario $(\rho^\star, \gamma^\star) = (0, 1)$

Effect size	$p = 0$			$p = 1$			$p = 2$		
	hazard	odds	normal	hazard	odds	normal	hazard	odds	normal
$0.4\theta_0$	0.7812	0.0194	0	0.0863	0.0284	0.0201	0.0254	0.0143	0.0248
$0.6\theta_0$	0.7752	0.0172	0	0.0887	0.0294	0.0223	0.0238	0.0159	0.0275
$0.8\theta_0$	0.7659	0.0177	0	0.0969	0.0306	0.0208	0.0238	0.0165	0.0279
$1\theta_0$	0.7668	0.0148	0	0.1012	0.0294	0.0204	0.024	0.0165	0.0269
$1.2\theta_0$	0.7636	0.0145	0	0.1015	0.0294	0.0162	0.0278	0.0192	0.0278
$1.4\theta_0$	0.7588	0.014	0	0.1107	0.0309	0.0148	0.0268	0.0173	0.0267
$1.6\theta_0$	0.7526	0.0123	0	0.1194	0.0271	0.015	0.0306	0.0174	0.0256

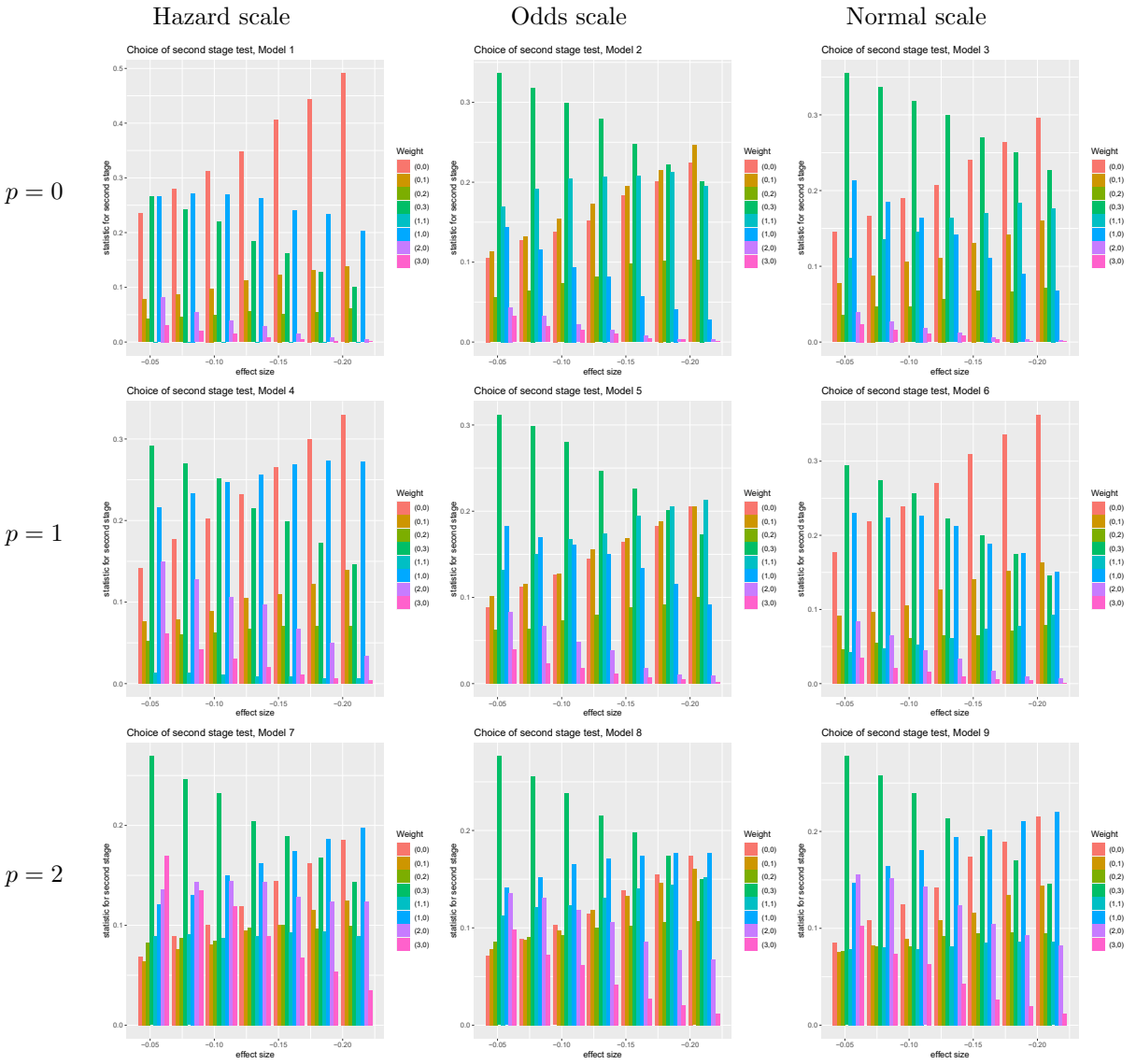
Supplementary Table S19: For each of the seven effect sizes, this table displays which of the nine Royston-Parmar spline models is rated as the best based on the AIC. As above, p refers to the number of inner knots in the spline model. The empirical rates refer to the total quantity of runs without early stopping of the corresponding simulated trial. This table refers to the scenario $(\rho^\star, \gamma^\star) = (0, 2)$

Effect size	$p = 0$			$p = 1$			$p = 2$		
	hazard	odds	normal	hazard	odds	normal	hazard	odds	normal
$0.4\theta_0$	0.7807	0.0142	0	0.0861	0.0289	0.0226	0.0233	0.0168	0.0274
$0.6\theta_0$	0.7706	0.0185	0	0.088	0.0307	0.0254	0.0244	0.0164	0.026
$0.8\theta_0$	0.7672	0.0178	0	0.0964	0.0309	0.0233	0.0229	0.0186	0.0229
$1\theta_0$	0.7573	0.0183	0.0001	0.0984	0.0351	0.0228	0.0259	0.0174	0.0247
$1.2\theta_0$	0.7571	0.0152	0	0.1066	0.0368	0.0181	0.0227	0.0175	0.0261
$1.4\theta_0$	0.7528	0.0159	0	0.1123	0.0383	0.0195	0.022	0.0154	0.0239
$1.6\theta_0$	0.731	0.0151	0	0.1192	0.0413	0.0216	0.0236	0.0228	0.0254

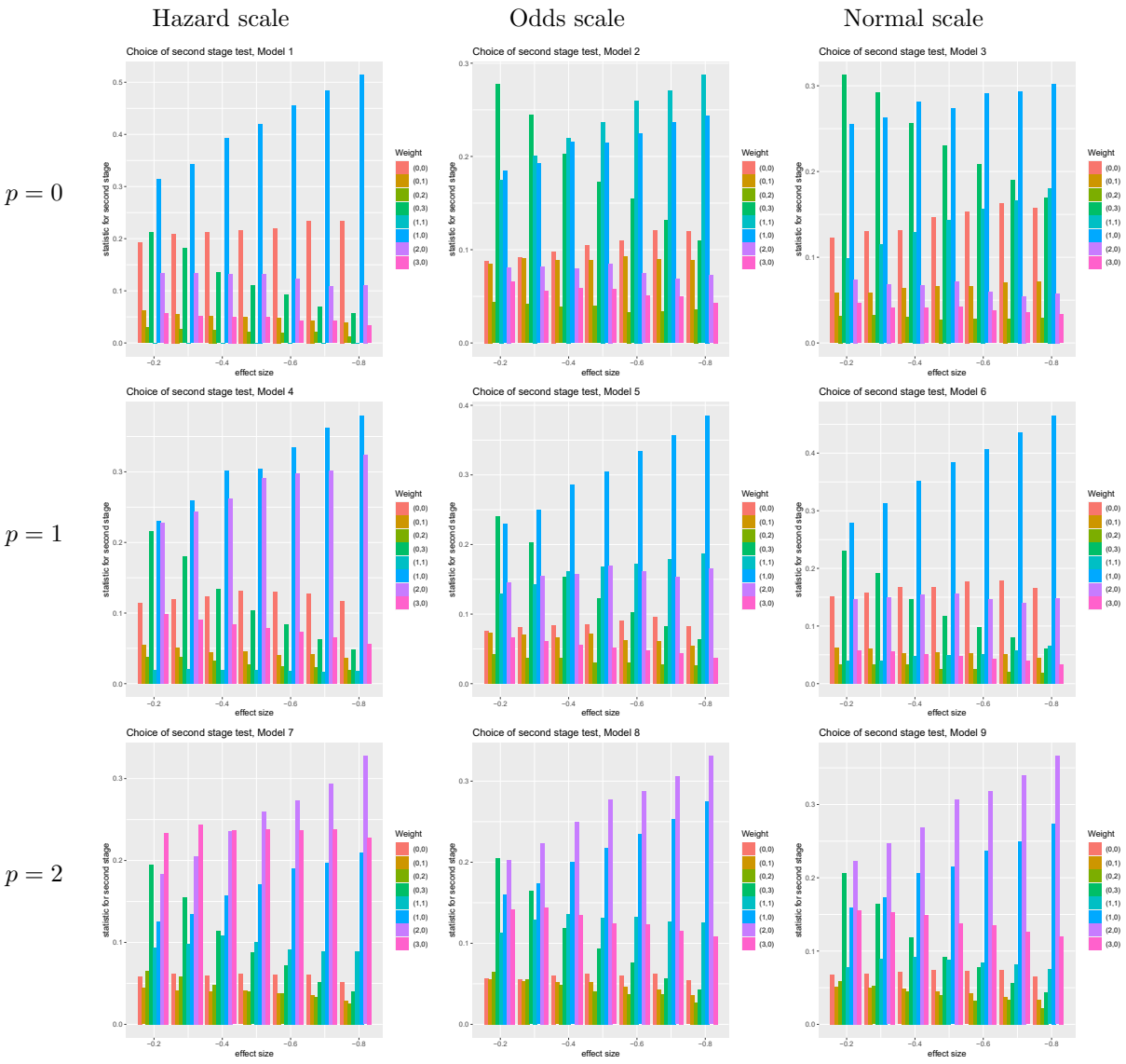
Supplementary Table S20: For each of the seven effect sizes, this table displays which of the nine Royston-Parmar spline models is rated as the best based on the AIC. As above, p refers to the number of inner knots in the spline model. The empirical rates refer to the total quantity of runs without early stopping of the corresponding simulated trial. This table refers to the scenario $(\rho^\star, \gamma^\star) = (0, 3)$

C.2.3 Modelwise choice of test statistics in the second stage

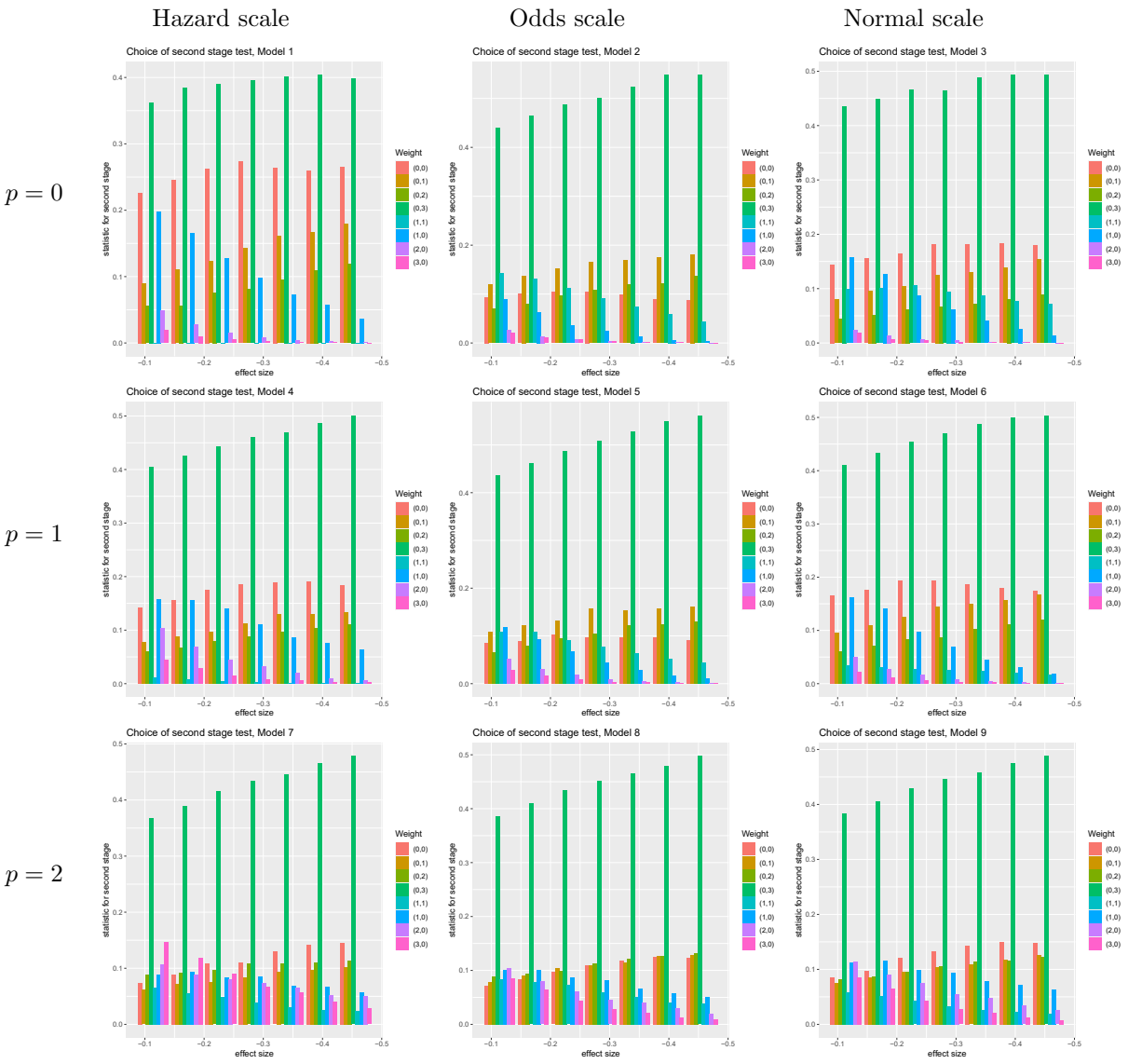
For the three deviation types from the main manuscript, we also show how the choice of the second stage test is distributed for each of the nine spline models. As already mentioned, models with a higher number of inner knots seem to be better suited to choose an appropriate test in late effects settings. Because of that, one should maybe restrict the set of models to a set with a higher number of inner knots if a possible late effect is anticipated.



Supplementary Figure S17: Choices of the log-rank test statistic with Fleming-Harrington weights for our nine different Royston-Parmar spline models. The rates refer to the total quantity of simulation runs in which the corresponding simulated trial proceeded to a second stage (i.e. no early termination). The plots are arranged in a grid where the columns refer to different scales and the rows to different numbers of interior knots p. These figures refer to the deviation type $(\rho^\star, \gamma^\star) = (0, 0)$, i.e. proportional hazards.



Supplementary Figure S18: Choices of the log-rank test statistic with Fleming-Harrington weights for our nine different Royston-Parmar spline models. The rates refer to the total quantity of simulation runs in which the corresponding simulated trial proceeded to a second stage (i.e. no early termination). The plots are arranged in a grid where the columns refer to different scales and the rows to different numbers of interior knots p. These figures refer to the deviation type $(\rho^\star, \gamma^\star) = (2, 0)$, i.e. a late effects scenario.



Supplementary Figure S19: Choices of the log-rank test statistic with Fleming-Harrington weights for our nine different Royston-Parmar spline models. The rates refer to the total quantity of simulation runs in which the corresponding simulated trial proceeded to a second stage (i.e. no early termination). The plots are arranged in a grid where the columns refer to different scales and the rows to different numbers of interior knots p. These figures refer to the deviation type $(\rho^\star, \gamma^\star) = (0, 2)$, i.e. an early effects scenario.

Name: Ina Dormuth Matrikelnummer: 165647

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation mit dem Titel

"Methods for Time-to-Event Data Analysis for Non-Proportional Hazard Settings"

selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe und die Satzung der Technischen Universität Dortmund zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet habe. Ich versichere außerdem, dass ich die beigefügte Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind. Ferner erkläre ich, dass keine Aberkennung eines bereits erworbenen Doktorgrades vorliegt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Dortmund, den .....

Ina Dormuth