universität wien

# MAGISTERARBEIT | MASTER'S THESIS

Titel | Title

Implementation of a Simulation Study to Evaluate the Performance Characteristics of Statistical Methods for Analysis of Time to Event Data under Non-Proportional Hazards


verfasst von | submitted by

Tobias Fellinger BSc


angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Magister der Sozial- und Wirtschaftswissenschaften (Mag.rer.soc.oec.)

Wien | Vienna, 2024

| | |
|---|---|
| Studienkennzahl lt. Studienblatt \| Degree programme code as it appears on the student record sheet: | UA 066 951 |
| Studienrichtung lt. Studienblatt \| Degree programme as it appears on the student record sheet: | Magisterstudium Statistik |
| Betreut von \| Supervisor: | ao. Univ.-Prof. Dr. Martin Posch |

# Acknowledgements

Thanks to my supervisor Martin Posch and to Florian Klinglmüller who supervised my work at AGES. Thanks to Kathrin for proof-reading my English spelling and grammar despite not knowing any statistics. Thanks to all collaborators of the CONFIRMS consortium for valuable input. And thanks to all colleagues at AGES and Med Uni Wien for teaching me invaluable theoretical and practical knowledge about medical statistics. And thanks to the R community for everything R.

# Abstract

In real life applications the proportional hazards assumption commonly used in statistical methods for time to event data often does not hold. This might affect the validity and efficiency of widely used methods.

In this thesis I am discussing some examples of clinical studies where non-proportional hazards were an issue. I am presenting the two R packages I developed in the course of writing this thesis. The R packages facilitate simulation studies to evaluate statistical methods for time to event data under various scenarios of non-proportional hazards. The software was used in a simulation study in an EMA funded project on "Statistical methods for time-to-event endpoints with non-proportional hazards in clinical trials pivotal for benefit risk decision making". As part of this thesis I conducted another simulation study investigating these methods in scenarios for a wider range of parameters. I developed a shiny app to visualize the survival distributions for the scenarios and to browse, filter and plot simulation results.

All commonly used methods maintained the nominal type I error rate in the investigated scenarios. In general methods specifically designed to be used under non-proportional hazards exceeded the power of standard methods. The most widely used estimators targeting summary statistics of the time-to-event distributions performed well in terms of bias and coverage. The software is published on CRAN.

# Kurzfassung

In der Forschungspraxis ist die Proportional-Hazards-Annahme, die bei statistischen Methoden häufig getroffen wird, oft verletzt. Das kann die Validität und Effizienz von häufig eingesetzten statistischen Methoden beeinträchtigen.

In dieser Masterarbeit diskutiere ich einige Beispiele aus klinischen Studien, in denen die Proportional-Hazards-Annahme verletzt war. Ich präsentiere zwei R Pakete, die ich im Rahmen dieser Masterarbeit entwickelt habe. Die R Pakete ermöglichen die Durchführung von Simulationsstudien, um statistische Methoden in Szenarien, in denen die Proportional-Hazards-Annahme verletzt ist, zu evaluieren. Die Software wurde für eine Simulationsstudie im Rahmen des EMA Projektes "Statistical methods for time-to-event endpoints with non-proportional hazards in clinical trials pivotal for benefit risk decision making" eingesetzt. Im Rahmen dieser Masterarbeit habe ich eine weitere Simulationsstudie durchgeführt und Methoden in Szenarien mit unterschiedlichen Parametern untersucht. Weiters habe ich eine shiny app entwickelt um die Überlebenszeitverteilungen der verschiedenen Szenarien visuell darzustellen sowie die Simulationsergebnisse auszuwählen, zu filtern und zu plotten.

Alle betrachteten Methoden haben in allen untersuchten Szenarien die nominelle Typ I Fehlerrate eingehalten. Methoden, die speziell für die Anwendung unter non-proportional hazards entwickelt wurden, weisen im Allgemeinen eine höhere Power als Standardmethoden auf. Die meisten verbreiteten Schätzer für Statistiken der Überlebenszeitverteilung zeigten geringe Verzerrung und nur geringfügige Abweichung von der nominellen Überdeckungswahrscheinlichkeit. Die entwickelte Software ist im CRAN publiziert.

# Contents

# List of Tables

# List of Figures

# 1. Introduction

## 1.1. Non-Proportional Hazards

Analysis of time to event data deals with random variables of times to events that take values in the positive reals. Due to a finite observation time or random drop-out, the data are typically censored. That is for some observations one does not know the time of the event but only that an event did not occur until a certain time. A useful way of expressing time to event distributions is via the hazard function. The hazard function is the event rate in the infinitesimal interval following $t$, given that there has not been an event up until $t$. Here only distributions that have a density will be used. In this case the hazard can be expressed as $\lambda(t) = f(t)/S(t)$ with the density $f(t)$ and the survior function $S(t)$. The most important feature of the hazard function is that there are, under rather weak assumptions, consistent estimators to estimate the hazard function that do not rely on the knowledge of the censoring distribution (See [1] Chapter 2.1. for a discussion of this feature of the hazard function).

One of the most widely used methods in survival analysis is the Cox proportional hazards model that not only allows to test for differences in survival distributions, but also allows to adjust for covariates. The Cox model assumes that the survival distribution depends on an unknown baseline hazard function via covariates that act on the baseline hazard multiplicatively – hence the hazards are proportional to each other. Since the model estimates the log of the hazard ratios, the model parameters are no longer interpretable under non-proportional hazards. Another widely used method is the logrank test. The logrank test is used to test for differences between survival distributions. While the validity of the test does not depend on the proportional hazards assumption the test can only be shown to be the most powerful test under the proportional hazards assumption. Chapter 3.2 gives a description of the analysis methods investigated here.

While both methods make no assumption on the baseline hazard, which makes them rather flexible, the proportional hazards assumption is often not fulfilled. There are many possible data generating mechanisms that can cause different deviations from the proportional hazards assumption impacting the performance of inference methods but also the causal interpretability of estimates. The data generating models investigated here as well as examples from studies from European marketing authorisation applications are described in Chapter 2.

## 1.2. Simulation Studies

When evaluating statistical methods most properties depend on the data generating model the methods are applied to. In some cases there are analytic solutions to calculating the quantities but for more complicated data generating models and for more complicated statistical methods analytic solutions are not always attainable. Important properties also cannot be estimated by applying the methods to data, since the true data generating process has to be known. For example to estimate the bias, the true value of the target of estimation has to be known. And to estimate the type I error rate one has to know whether the data are generated under the null-hypothesis. Therefore simulation studies where a large number of data sets are generated from a known data generating model are an important tool to "measure" the performance of the statistical methods.

While this makes it easy to calculate properties of the statistical methods under a variety of scenarios, care has to be taken on how to choose relevant scenarios and what conclusions the results of the simulation study support. For example type I error rate cannot be claimed to be controlled for all scenarios in the null-hypothesis but just for the scenarios investigated in the simulation study. Additional arguments would be necessary to support the conclusion that this also holds for all other parameter values in the null hypothesis.

The design of the simulation study conducted in this thesis follows the *Clinical Scenario Evaluation* (CSE) framework [2] and the *aims, data-generating mechanisms, estimands, methods, and performance measures* (ADEMP) approach [3].

The CSE Framework focuses specifically on the evaluation of strategies and methods for drug development programs. The framework introduces the concepts of "Assumptions", "Options" and "Metrics". Broadly speaking "Assumptions" are all assumptions about the data generating model that are not under control of the trialist like variability, treatment effect, etc. "Options" refer to everything under the control of the trialist like design of the trials in a development program, decision rules and choice of statistical methods. Finally "Metrics" are the measures by which the different Options are to be compared, like success probability, total running time or sample size or statistical power [2].

The ADEMP approach is more broad in the sense that it does not focus on clinical research programs but biostatistics in general, but more narrow in that it does not consider all the Options for a development program but only a comparison of statistical methods. Morris et al. [3] also put more emphasis on conduct of the simulation study and implementation. According to the ADEMP framework the research question should be made explicit in the "Aim" step. This includes decisions whether the study should be a proof of concept of one method, a comprehensive proof of validity of one method, or a comparison of multiple methods in relevant scenarios. The "Data generating mechanisms" correspond to the "Assumptions" in the CSE Framework. The ADEMP framework also makes the choice of estimands explicit – typically parameters or summary statistics of the data generating mechanism – to be targeted by the methods. The "Methods" item describes the methods – analysis methods, study designs, etc. – to be compared. Finally "Performance Measures" describes the operating characteristics of the methods that should

be compared, this corresponds to the ”Metrics“ item of the CSE framework.

In this thesis the ADEMP approach is followed but for the Assumptions a further distinction between Assumptions and Options is made in line with the CSE framework.

# 2. Data Generating Models

## 2.1. Common Aspects of the Data Generating Processes

All scenarios assumed parallel group randomized clinical trials with a 1:1 randomization to the treatment arms. Enrollment of participants was assumed to either take place uniformly distributed over a pre-defined time or for all patients at the start of the study. An equal number of patients was assigned to each treatment arm, no blocking or stratification was used in the randomization.

A fixed design as well as a group sequential design with one interim analysis were investigated. The timing of the analyses was determined by the number of observed events in both designs. The final analysis was performed when three quarters of the enrolled patients experienced an event. In the group sequential designs the interim analysis was performed when half the events needed for the final analysis were observed. Adjustment for multiple testing was performed using an alpha-spending approach. Those options together with the assumptions on the distribution on the survival times determine the distribution of administrative censoring. See section 4.2.3 for a detailed description of the calculations.

In addition different distributions of random censoring due to withdrawal or loss to follow up were investigated. Exponential distributions with different rate parameters were assumed for time from enrollment into the study to time of random censoring.

## 2.2. Delayed Onset of Treatment Effects and Crossing Hazard Functions

A delayed onset of treatment effect, such that the investigational drug results in a survival benefit only later during the trial has been described in immuno-oncology agents [4][5]. Even more extreme cases exist in which the survival curves show a detrimental effect of the investigational treatment early and a survival benefit later on, such that the survival curves (and therefore the hazards) cross.

### 2.2.1. Example

An example for a trial with a delayed onset of treatment effect is the KEYNOTE-048 trial [6] [7]. The trial compared three treatments against recurrent or metastatic head and neck squamous cell carcinoma: a combination of the humanized monoclonal antibody Prembrolizumab and chemotherapy, Prembrolizumab monotherapy and standard treatment with the monoclonal antibody Cetuximab and chemotherapy. 882 patients

were randomized to the three treatment arms (301 to Prembrolizumab monotherapy, 281 to Prembrolizumab on top of chemotherapy, 287 to standard treatment). The therapies were treatments were compared with respect to overall survival (OS) and progression free survival (PFS). The main comparison was a difference in overall survival between Prembrolizumab and chemotherapy versus standard treatement. The Kaplan-Meier curves of the this comparison are reproduced in Figure 2.1.

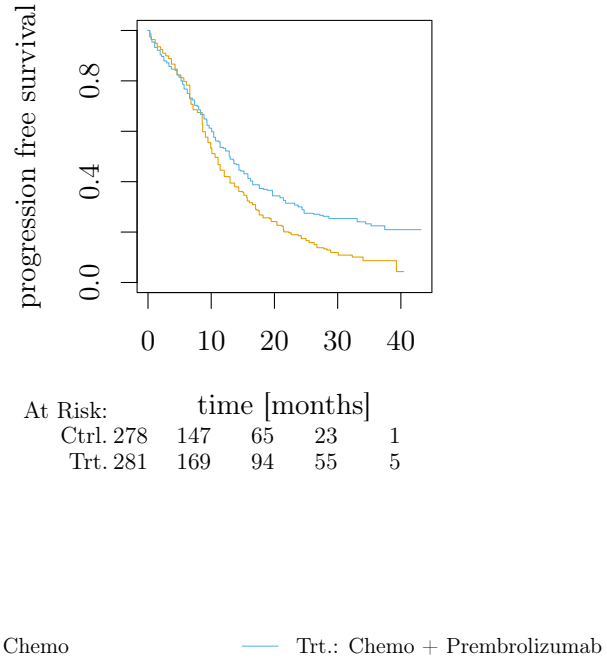



Figure 2.1.: Kaplan-Meier curves for data reconstructed from the Kaplan-Meier curves published in [6] (Figure 4 ITT population).

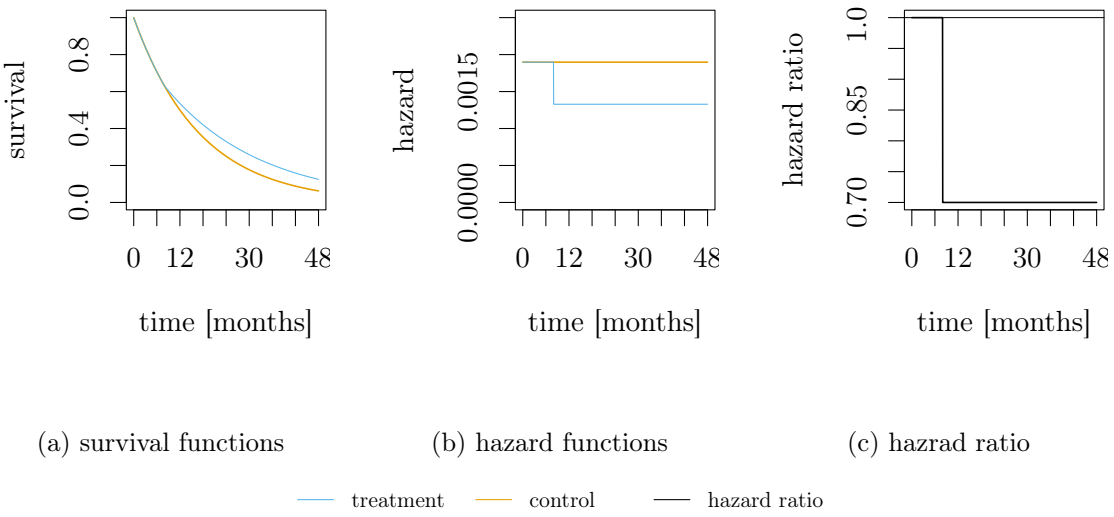
The curves align in the first time interval and separate after a delay. While standard methods control the type I error rate with delayed onset of treatment effects[1], the power of a trial under this setting may be decreased, if design and analysis of the trial are conducted with standard methods.

### 2.2.2. Modelling with Piecewise Constant Hazards

For the scenarios with delayed onset of treatment effect and crossing hazards the survival distributions were modelled using piecewise constant hazards in both arms. For the control arm a constant hazard was assumed in all scenarios. For the treatment arm

[1]The null-hypotehsis of equal survival curves is the same with proportional hazards and delayed onset of treatement effects.

different hazard rates were assumed in two time intervals. In the scenarios with delayed onset of treatment effect, the treatment arm hazard was equal to the control arm hazard in the first time interval and lower in the second. In the scenarios with crossing hazards the hazard in the treatment arm was higher than the hazard in the control arm in the first time interval and lower than the control arm hazard in the second time interval.




(a) survival functions (b) hazard functions (c) hazrad ratio

Figure 2.2.: Survival function, hazard function and hazard ratio for a scenario with delayed effect. Median survival in the control arm 12 months, delay 8 months, hazard ratio after onset of treatment effect 0.7.

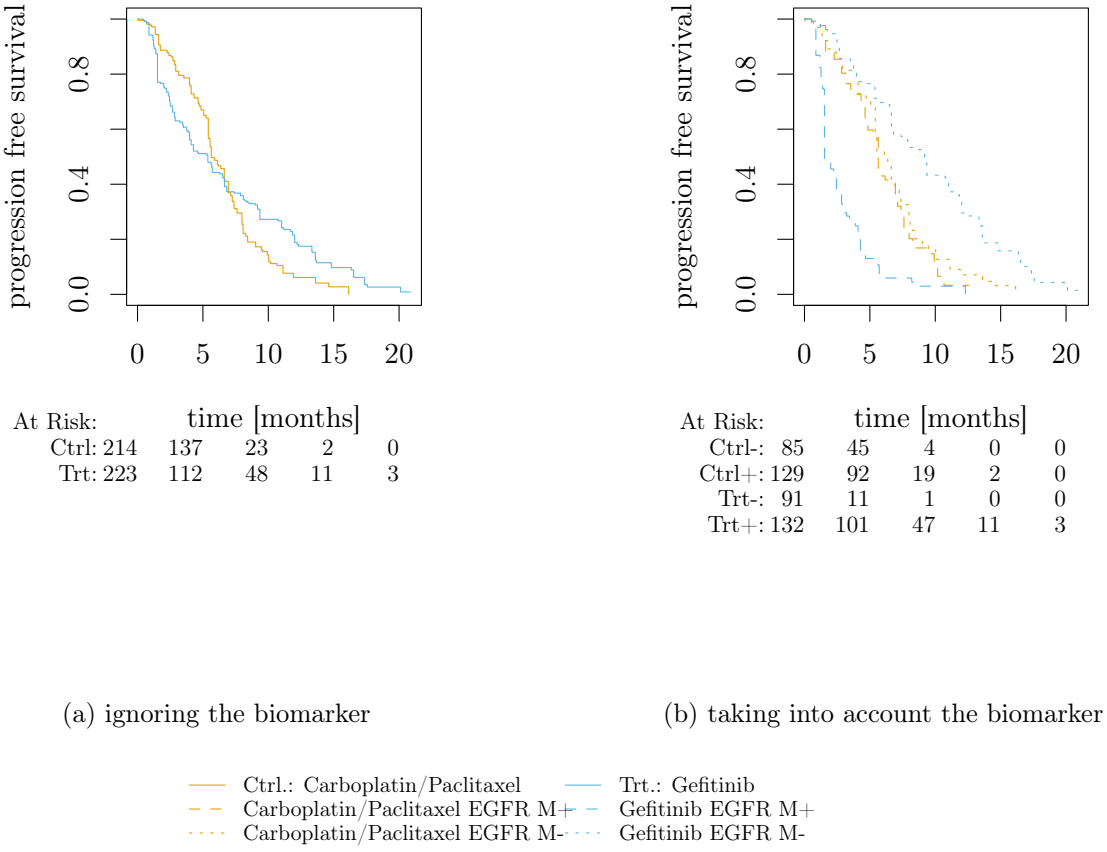
## 2.3. Non-proportional Hazards due to Subgroup Effects

Different susceptibility in a subgroup – for example in carriers of a biomarker – can result in non-propotional hazards if the subgroup is not known or not accounted for in the analysis. Refer to [8] and [9] for a detailed description of the mechanisms.

### 2.3.1. Example

An example is the IPASS trial [10] [11] that compared the EGFR inhibitor Gefitinib to a combination chemotherapy for the treatment of patients with non–small-cell lung cancer. Of the 1329 patients 609 were randomized to receive the experimental treatment Gefitinib and 608 were randomized to receive chemotherapy. The pre-specified main comparison was progression free survival among all randomized patients. For an exploratory biomarker analysis samples were taken to assess a mutation of the epidermal growth factor receptor (EGFR). Evaluable samples were available for 223 participants.

Non-proportional hazards were observed in the main comparison. The survival curves crossed late, with a substantial area between the curves both before and after the crossing (Figure 2.3a). In an exploratory analysis that accounted for the status of a biomarker, the violation of the proportional hazards assumption was less pronounced and the curves did not cross in either group of biomarker positive and biomarker negative participants (Figure 2.3b). The exploratory analysis also sheds light on why the proportional hazards assumption was violated in the primary analysis: in the group of biomarker positive participants the participants on the treatment under investigation fared better than the participants in the comparator arm but the effect was reversed in the group of biomarker negative patients.



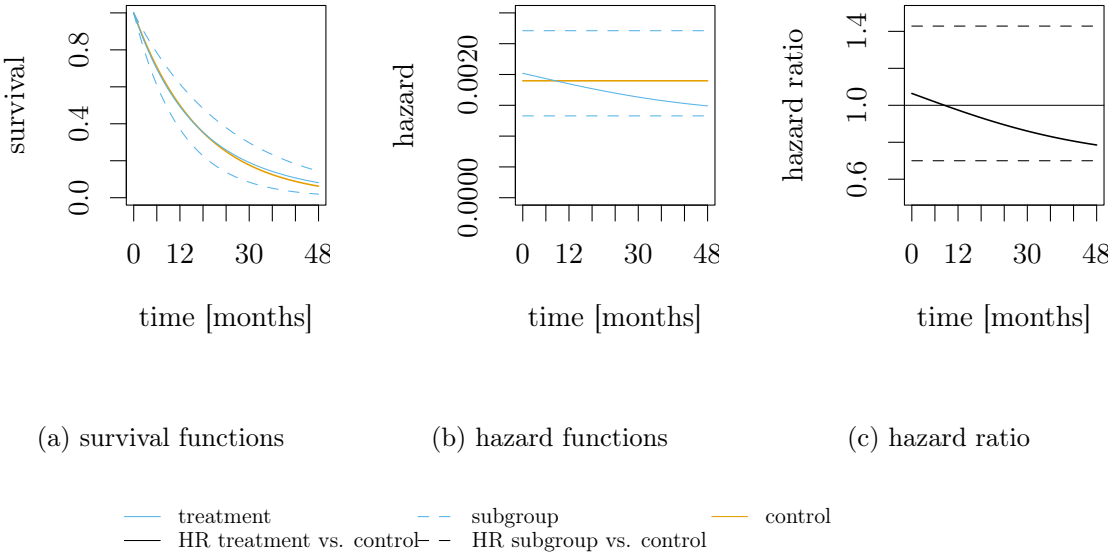

(a) ignoring the biomarker

(b) taking into account the biomarker

Figure 2.3.: Kaplan-Meier curves for data reconstructed from the Kaplan-Meier curves published in [11] (Figure 2 (b) and 2 (c)). Curves for patients with known biomarker status, with and without taking into account the biomarker status of the study participants.

### 2.3.2. Modelling with Mixtures of Distributions with Constant Hazards

For the scenarios with a differential treatment effect in an unknown subgroup again a constant hazard was assumed in the control arm. In the treatment arm, group membership of the subgroup assumed as a binomial random variable with a predefined probability. The time to event distribution was then assumed to have one of two pre-specified constant hazard rates depending on the group membership.




(a) survival functions (b) hazard functions (c) hazard ratio

Figure 2.4.: Survival function, hazard function and hazard ratio for a scenario with non-proportional hazards due to a differential treatment effect in a subgroup. Median survival in the control arm 12 months, hazard ratios for the subgroups 0.7 and 1/0.7 relative to control, prevalence 50%.

## 2.4. Changing Hazards after an Intercurrent Event

If the hazards change after an event that occurs at a fixed or random time, this can lead to non-proportional hazards. The event in this case could be many things, for example disease progression, switching to rescue medication or switching to another treatment within a pre-defined sequence of treatments. The example in this subsection contains a study where participants could switch from placebo to the active comparator after an intercurrent event. In the simulation study a scenario was assumed where participants in both arms could switch to a third rescue medication after disease progression.

### 2.4.1. Example

An example for a similar setting is the TERIKIDS trial [12] [13] that compared terifluomide to placebo in children with relapsing remitting multiple sclerosis. 166 children were included in the study and were randomized two to one to the treatment and control arm. The study consisted of a double blind period in which each participant was given the treatment according to the randomization, followed by an open label period in which patients from the control arm could switch to the active treatment. Patients in the control arm could switch when one of three conditions were met: end of the double blind period, relapse (which was counted as an event) or high MRI activity. The main analysis compared time to relapse in the double-blind period. Since observations of participants who switched to the open label period due to high MRI activity were censored the assumption of non-informative censoring is clearly not fulfilled. (Figure 2.5a)

A sensitivity analysis explored the time to relapse in both periods, comparing the treatment policies of administering placebo in the first and terifluomide in the second period versus treatment with terifluomide in both periods (Figure 2.5b). Another post-hoc sensitivity analysis investigated the composite endpoint of relapse or high MRI activity.

Besides the potential for non-proportional hazards due to change of the hazard rate after treatment switching, this study illustrates the importance of well defined estimands and strategies to handle intercurrent events. In the main analysis, in line with a hypothetical strategy, assumptions made for the analysis methods could not be justified. The sensitivity analyses, corresponding to the treatment policy- and composite variable strategies, can justify the assumptions on the censoring process better but answer different questions (See also [14]).

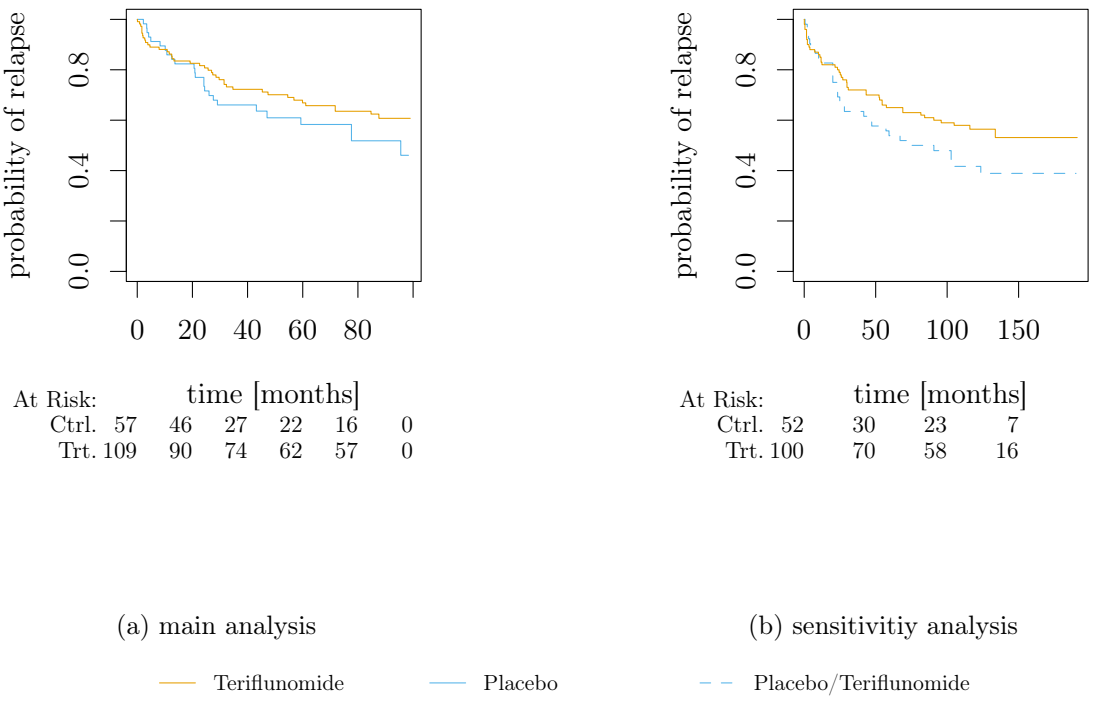



Figure 2.5.: Example with changing hazards after an intercurrent event

### 2.4.2. Modelling with Multi State Models

For the scenarios with disease progression, three states were assumed: non-progressed, progressed disease and death. Participants start in the state of non-progressed disease and can then either enter the state of progressed disease or die before disease progression is observed. Whether disease progression was observed was modeled as the event of the time to death or the time to disease progression being shorter. Both times were assumed to be exponentially distributed with different rates in each treatment arm. The time to death was modeled depending on the event of disease progression either as the time from entering the study to death before disease progression if no progression was observed or as the sum of time from entering the study to disease progression and the time from disease progression to death. The same hazard rate for time from disease progression to death was assumed for the treatment and control arm. (This could for example occur if after disease progression participants from both arms are switched to the same rescue-medication.)

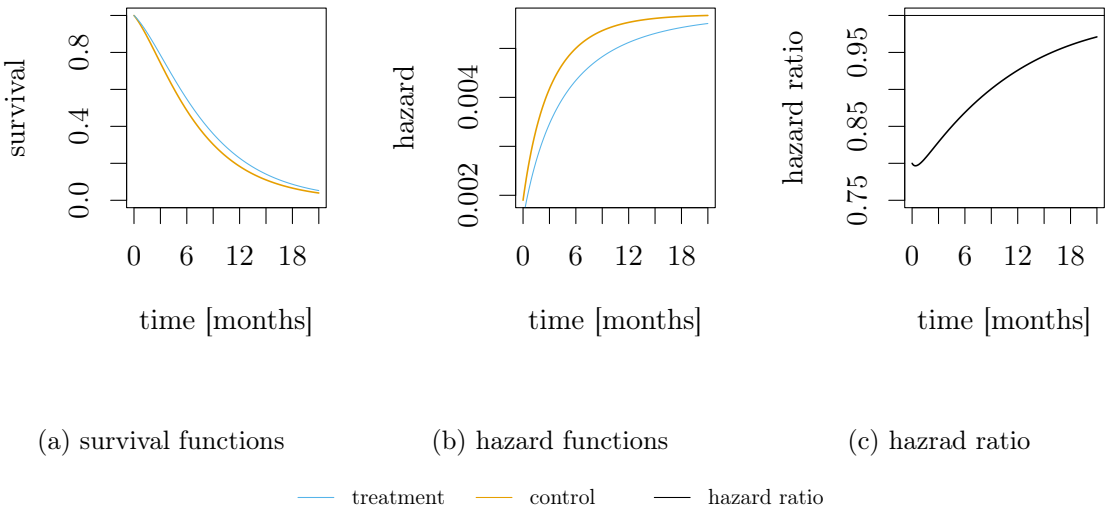



Figure 2.6.: Survival function, hazard function and hazard ratio for a scenario with non-proportional hazards due to changing hazards after disease progression. Median survival in the control arm, before progression 12 months, hazard ratio for death before progression treatment vs. control 0.8, median time to progression in the control arm 2 months, hazard ratio for progression treatment vs. control 0.7, hazard ratio for death after progression (both arms) vs. hazard for death before progression in the control arm 3.

# 3. Estimands and Analysis Methods

## 3.1. Estimands and Population Level Summaries

In line with the ICH E9 addendum on estimands an estimand should consider treatment, population, the variable (endpoint), intercurrent events and a population-level summary for the variable. This population-level summary is then used as the target of estimation [14].

For estimands in which the endpoint is a difference in the distribution of survival times, many population-level summaries have been proposed and estimators have been constructed that target those summaries.

### 3.1.1. Commonly Used Population-Level Summaries

**Milestone Survival Probabilities** Milestone survival probabilities are the probabilities of surviving up to a certain time point. Differences of survival distributions can be quantified as differences and ratios of the survival probabilities at certain time points.

**Quantiles of the Survival Distribution** Quantiles of the survival distribution can also be used as a summary statistic of the survival distribtuion. The median survival time as well as the difference of median survival times are commonly reported.

**Mean Survival Times** The mean survival time cannot be estimated, since this would require a long – under some assumptions infinite – observation time. What can however be estimated from a finite time frame of observation is the restricted mean survival time (RMST). The RMST is defined as $\mathbb{E}[\min(t, T)]$ with the survival time $t$ and a cutoff time $T$ choosen by the user. The RMST equals the area under the survival curve from 0 to $T$ [15].

**Average Hazard Ratios** While under proportional hazards the hazard ratio is a useful summary statistic of the difference of survival distributions, this fails under non-proportional hazards. Since under non-proportional hazards the hazard ratio is a function of time it is intuitive to try to average the hazard ratio over time in a meaningful way. Different approaches exist to define average hazard ratios. The definitions for the average hazard ratio and geometric average hazard ratio given in [16] are described here.

Both definitions use a weighting function $w(t)$, typical choices are $w(t) = 1$ and $w(t) = S(t)$, where $S(t)$ is an estimate of the survival function for the pooled data of both arms. The former corresponding to the Cox proportional hazards model and the

latter corresponding to the test statistic of the Wilcoxon test under proportional hazards. Let further $h(t) = h_1(t) + h_2(t)$ be the sum of the hazard functions of the compared distributions and $f(t)$ the density of the distribution of event times for all groups and $S_0$ and $S_1$ the survival functions of the respective groups.

The average hazard ratio is defined as ([16], equation (3)):

$$\text{AHR} = \frac{\int (h_1(t)/h(t)) w(t) f(t) dt}{\int (h_0(t)/h(t)) w(t) f(t) dt}$$

The geometric average hazard ratio is defined as ([16], equation (2)):

$$\text{gAHR} = \exp\left( \int \log\left( \frac{h_1(t)}{h_0(t)} \right) w(t) f(t) dt \right)$$

### 3.1.2. Null Hypotheses for Different Tests

Under proportional hazards the null hypotheses of stochastic dominance of the survival times, a hazard ratio unequal to 1, differences in the distribution of survival times, etc. are all equivalent. When dealing with non-proportional hazards the situation is more difficult. It is important to take care for which hypotheses one assesses the properties of the tests – most importantly the error rates.

Magirr et. al. call the null hypothesis of equal survival distributions the *weak null hypothesis* and the null hypothesis that the survival times of the treatment arm are stochastically less than or equal to the survival times in the control arm the *strong null hypothesis*. A rejection of the weak null hypothesis gives no information on the character of the differnce between the survival curves and control of the type I error rate with respect to the strong null hypothesis is not guaranteed for methods that control the type I error rate for the weak null hypothesis [17].

## 3.2. Inference Methods

### 3.2.1. The Cox Proportional Hazards Regression

Proportional hazards models assumes the hazard to be of the form:

$$\lambda(t, Z) = \lambda_0(t) \exp(Z^T \beta)$$

Where $Z$ is the matrix of covariates, $\beta$ is the vector of regression coefficients and $\lambda_0$ is the baseline hazard as a function of time $t$. Cox regression exploits the fact, that in the likelihood for $\beta$ the contribution of $\lambda_0(t)$ cancels out yielding a partial likelihood. The partial likelihood can be maximized giving estimates for $\beta$ without estimating the baseline hazard $\lambda_0$ [18].

Under proportional hazards this estimator, per construction, corresponds to the log of the hazard ratio. The partial maximum likelihood equations[1] can also be solved if the model assumptions do not hold, including the case of non-proportional hazards. But under non-proportional the solution does no longer correspond to any simple population-level summary statistic of the survival distributions.

### 3.2.2. Accelerated Failure Time Models

Accelerated failure time (AFT) models assume the hazard to be of the form:

$$\lambda(t, Z) = \exp(-Z^T\beta)\lambda_0(t\exp(-Z^T\beta))$$

Where as in the proportional hazards model, $Z$ is the matrix of covariates, $\beta$ is the vector of regression coefficients and $\lambda_0$ is the baseline hazard as a function of time $t$ [18].

Plugging this into the definition of the hazard function, one can see, that the hazard under $Z$ belongs to the distribution corresponding to $\lambda_0$ by scaling the time axis by the factor of $\exp(-Z^T\beta)$ – hence the name accelerated failure time models.

While semi-parametric estimators exist for the acceleration factor, parametric AFT models based on the exponential and Weibull distributions are commonly used as an alternative to proportional hazards models. The data distribution of accelerated failure time models does in general not satisfy the proportional hazards assumption. Like for the Cox model the maximum likelihood estimate can also be computed if the data generating model does not follow the respective accelerated failure time model. Again the parameters do no longer have a straightforward interpretation if estimated for data that does not follow the assumed data generating mechanism.

### 3.2.3. Estimators Based on Summary Statistics of the Survival Curve

Many of the estimands described above can be expressed as summary statistics of the survival curve. There are various ways of calculating the statistics. For example one can just plug-in an estimate for the survival curve, like the Kaplan-Meier or Nelsen-Aalen estimator into the definition and directly calculate the point estimates.

A versatile approach to compare estimates for summary measures of the survival curve between groups has been developed by Ristl et. al. [19]. This approach allows to construct estimators for the differences of functionals of the survival distribution in the groups and also yields the asymptotic joint distribution. Based on the joint distribution multiple testing procedures and simultaneous confidence intervals can be constructed. Estimators for AHR, gAHR, the difference of quantiles of the survival functions and differences and quotients of milestone survival probabilities derived by this method are implemented in the nph package [4].

[1] The partial likelihood here refers to the partial likelihood of the proportional hazards model, not the partial likelihood of the data generating model, which violates the proportional hazards assumption.

### 3.2.4. (Weighted) Logrank Test

The logrank test was first proposed by Mantel in 1966. The logrank tests compares the number of expected events to the number of observed events in the control arm at each event time. Formulas for the test statistic can be found for example in [20] or [1]. Here the notation from [4] is used.

Let $\mathcal{D}$ be the set of unique event times and for each $t \in \mathcal{D}$ let $n_{t,ctrl}$ and $n_{t,trt}$ be the number of participants at risk immediately prior to $t$, $d_{t,ctrl}$, $d_{t,trt}$ the events at $t$ in the respective groups, the expected number of events in the control group at time $t$ under the null hypothesis $e_{t,ctrl}$. Plugging in and using the variance for the hypergeometric distribution yields:

$$e_{t,ctrl} = (d_{t,ctrl} + d_{t,trt})\frac{n_{t,ctrl}}{n_{t,ctrl} + n_{t,trt}}$$

$$Var(d_{t,ctrl}) = \frac{n_{t,ctrl}n_{t,trt}(d_{t,ctrl} + d_{t,trt})(n_{t,ctrl} + n_{t,trt} - d_{t,ctrl} - d_{t,trt})}{(n_{t,ctrl} + n_{t,trt})^2(n_{t,ctrl} + n_{t,trt} - 1)}$$

With the weighting function $w(t)$ we get the test statistic:

$$z = \frac{\sum_{t\in\mathcal{D}} w(t)(d_{t,ctrl} - e_{t,ctrl})}{\sqrt{\sum_{t\in\mathcal{D}} w(t)^2 Var(d_{t,ctrl})}}$$

With the weight function $w(t) = 1$ one gets the standard logrank test. Another common family of weight functions is the Fleming-Harrington $\rho$-$\gamma$ family of weight functions $w(t) = \hat{S}(t)^\rho(1 - \hat{S}(t))^\gamma$. The Fleming-Harrington family also contains the standard logrank test ($\rho = 0$, $\gamma = 0$) and the Peto-Peto variant of the Gehan-Wilcoxon test ($\rho = 1$, $\gamma = 0$).

### 3.2.5. Max-Combo Test

The max-combo test combines multiple weighted logrank tests, typically of the Fleming-Harrington $\rho$-$\gamma$ family, to achieve higher power for multiple different deviations from proportional hazards. The test exploits that the test-statistics of the combined tests are correlated, so a more efficient correction for multiple testing is possible. Under the null hypothesis the joint distribution of the test-statistics is asymptotically multivariate normal and the variance-covariance matrix can be estimated from the data and the weights. With this the multiplicity adjusted p-value can be computed with the distribution function of the multivariate normal distribution. [21]

### 3.2.6. Modestly Weighted Logrank Test

While the logrank test controls the type I error under the strong null hypothesis, for some weights the weighted logrank test fails to do so. Moreover, the direction of the deviation from the weak null hypothesis is not clear from the test statistic. To still be able to gain

power but still control the type I error rate under the strong null hypothesis, Magirr and Burman proposed the *modestly weighted logrank test* [17].

The construction is based on the idea, that when changing the event times a longer survival in the experimental arm and a shorter survival in the control arm the p-value should always decrease. For the score test this directly translates to non-increasing scores. Leton and Zuluaga [22] showed equivalence between weighted logrank tests and score tests and derived formulae to convert weights to scores and vice versa. Based on this Magirr and Burman derive weights corresponding to scores that are non-increasing in time[2]. The modestly weighted logrank test has weights increasing from 0 to a user-defined parameter $t^*$ and constant weights after $t^*$ [17].

[2]Or equivalently non-increasing with respect to the rank of the event times.

# 4. Simulation Study

This chapter describes the simulation study conducted for this thesis as well as some aspects of the simulation study conducted by the CONFIRMS consortium in EMA research project on non-proportional hazards [23]. The simulation studies were conducted according to the *aims, data-generating mechanisms, estimands methods and performance-measures* (ADEMP) framework introduced in [3]. This chapter describes the details of the conduct of the simulation study and implementation of the necessary computer programs.

For the *Data-Generating Mechanisms* part of the ADEMP framework the choice of parameter values in the models as well as the implementation of the data generation are described. This consists of the aspects relating to the survival distribution, referred to as *Assumptions* in the *clinical scenario evaluation* (CSE) framework introduced in [2] as well as to aspects of trial design, referred to as *Options* in the CSE framework. For the *Estimands* [1] part, the calculation of the true values of the summary statistics for each scenario is described, since there is not always a one to one correspondence to the parameters in the parametrization used in the implementation. The *Methods* investigated are described in section 3.2. The implementations used in the simulations are listed in section 4.3. Finally the used *Performance Measures* and their graphical, tabular and interactive presentation are described.

## 4.1. Simulation Parameters

### 4.1.1. Extracting Realistic Values from Results of Clinical Studies

Most clinical studies reports contain similar summary statistics of the results, yet many properties of the underlying hazards can not directly be reconstructed from those summary statistics, even under very strong assumptions. Almost all studies report the median times of survival in both groups and a plot of the Kaplan-Meier estimates of the survival curves. Most studies where a delay of onset of treatment effect or crossing survival curves were observed contain a rough estimate of the time of onset of treatment effect or the time of crossing of the survival curves. Refer to [24] for a complete listing and discussion of values extracted from European Public Assessment Reports (EPARs).

[1] Note that the term Estimand is used differently in the ADEMP framework and in the ICH E9 addendum R1. In this chapter estimands refers to the term from the ADEMP framework unless explicitly stated otherwise.

### 4.1.2. Parameter Values used for Simulation

**CONFIRMS Simulation Study** In the simulation study conducted by the CONFIRMS consortium [25] the chosen parameters were the control arm hazard, the maximum number of recruited patients, the recruitment time and the proportion of censored patients and the effect size in terms of power of the log-rank test. For the different scenario classes time to onset of treatment effect, time to crossing of the hazard curves, hazard ratio before crossing of the hazard curves, the prevalence of the biomarker subgroup, the hazard ratio between subgroup and complement, the proportion of patients who progress in the control and treatment arm respectively and the hazard ratio between before and after progression in the control arm were fixed.

The hazards not explicitly specified were then set as described in 4.1.3. This resulted in scenarios for which the power of the log-rank test is in a range that would be expected for adequately powered clinical trials. But in some cases this also lead to scenarios with unrealistically large or small hazard rates. This approach also lead to an interdependency between effect size, number of patients and the respective hazards, which made the simulation results harder to interpret and to communicate.

**Simulation Study for this Thesis** In the simulation study conducted in this thesis, simulation parameters were chosen to more directly correspond to the parameters of the data generating process. This should make interpretation of the results as well as comparison between scenarios easier. The numbers extracted in [24] and calculated for [25] helped inform the choice of parameters.

### 4.1.3. Choice of Parameters in the CONFIRMS Simulation Study

When assuming a constant hazard (exponential distribution) for the control arm, the hazard rate $\lambda$ can be directly calculated from the median survival time $t_{1/2}$ using

$$\lambda = \frac{\log(2)}{t_{1/2}} \tag{4.1}$$

Besides the treatment arm hazard, different free parameters remain to be fixed for the different scenarios. In [25] the following approach was used:

For the scenario with a delayed onset of treatment effect, the time of onset of treatment was chosen in the ranges reported in the EPARs reviewed in [24]. In the scenarios with crossing hazard curves the hazards before crossing of the hazards were set to fixed values. The time of crossing was set to the same times as the times of onset of treatment effect. This yielded survival curves with a similar time of crossing of the survival curves as observed in [24].

In the scenarios with a subgroup effect, the values of the prevalence were chosen to cover good part of the range of possible values between 0 and 1. The hazard rates in the subgroup and the compliment were set by fixing the hazard ratios between the compliment and the subgroup to give slight to strong benefits for the subgroup compared to the remainder of the treatment arm.

In the scenarios with disease progression, the hazard rate for death after disease progression in was set to be a fixed factor times the hazard rate for death before progression in the control arm. The same hazard for death after disease progression was assumed for both arms. The hazard rate for disease progression was set by fixing the proportion of participants in the respective trial arms who experience disease progression before death. (The number of patients who experienced disease progression is typically reported in clinical studies, especially when both overall survival and progression free survival are reported.) The hazard rate for disease progression was then set based on the control arm hazard for death before progression for both arms, since the hazard for death before progression in the treatment was not yet fixed. (See 4.1.3 for the calculations.)

### Setting the Treatment Arm Hazard

To choose parameter values that produce realistic values with respect to study size and power, the remaining free parameters of the treatment arm hazards (treatment arm hazard after onset of treatment effect in the scenarios with delayed onset of treatment effect, treatment arm hazard after crossing of the hazards in the scenarios with crossing hazards, hazard in the treatment arm without the subgroup in the scenarios with a differential treatment effect in a subgroup, and hazard in the treatment arm before disease progression in the scenarios with disease progression) were set to approximately give a certain power to detect a treatment benefit with a one-sided logrank test. This heuristic does not guarantee, that the log-rank test does indeed have the nominal power in the resulting scenarios.

Since there is no closed form of the power of the log-rank test under non-proportional hazards the following heuristic was used:

- calculate the hazard ratio required to archive the pre-specified power with the pre-specified number of events by re-arranging the Schönfeld sample size formula [26] to solve for the hazard ratio.

$$n = \frac{(Z_{1-\alpha} + Z_\beta)^2}{\log(\lambda_1/\lambda_0)^2 P_1 P_0 d}$$
$$(\lambda_1/\lambda_0) = \exp\left(\frac{Z_{1-\alpha} + Z_\beta}{\sqrt{P_1 P_0 d n}}\right)$$

- from the hazard ratio calculated in the first step, calculate the treatment arm hazard rate, that would give the target power under proportional hazards and the associated median survival time $m$.

- find a treatment arm hazard rate such that the median survival of the treatment arm matches the one calculated in the previous step by finding a root of either of the functions $\lambda \mapsto Q_\lambda(0.5) - m$ or $\lambda \mapsto F_\lambda(m) - 0.5$. With the median $m$ computed in the second step and $Q_\lambda$ and $F_\lambda$ the quantile and cumulative distribution function with the treatment arm hazard $\lambda$. Calculation of $Q_\lambda$ and $F_\lambda$ is described in section 4.2 for the different used data distributions.

This produced hazard functions for which the log-rank test had a rejection rate within an acceptable range of the pre-specified power.

#### Setting the Censoring Rate and Hazard for Progression

After calibration of the survival distribution, the withdrawal rate was set to match the pre-specified proportion of randomly censored patients. A censoring distribution with a constant hazard was used. The proportion of subjects who are censored before a set time and before experiencing an event before a set time can be calculated as:

$$p_{\text{censored}}(t) = \frac{\Lambda_{\text{censored}}(t)}{\Lambda_{\text{censored}}(t) + \Lambda_{\text{event}}(t)}$$

This was calculated for a time point well after the end of the study and the hazard rate for the censoring distribution was set to yield the given proportion. The same approach was used to set the rate of disease progression, replacing the cumulative hazard for censoring with the cumulative hazard for disease progression and the cumulative hazard for an event with the cumulative hazard for an event before disease progression.

#### Numerical Issues

The numeric root finding algorithm used in the calculation of the treatment arm hazard and the progression- and censoring- rates is ill conditioned if the function has a very small derivative around the root. This becomes an issue when dealing with very small hazards (For example when dealing with daily hazard rates in studies that run for months or years). This obstacle can however easily be avoided by scaling hazard rates with $\lambda_1^{-1}$ and times with $\lambda_1$ and back-transforming after calculating the root.

### 4.1.4. Reconstruction of IPD

For selected studies individual patient data (IPD) was reconstructed from digitized survival curves using an algorithm proposed in [27] and implemented in the R-package IPDfromKM [28]. From this reconstructed datasets are simulated by sampling with replacement. To assess the type I error rate survival times for the treatment and control group are both sampled from the control group in the reconstructed IPD. To compare rejection rates under the alternative [2] survival times for the simulated data are sampled from the respective treatment and control groups in the reconstructed data.

### 4.1.5. Choice of Parameter Values for the Simulation Study for this Thesis

The simulation parameters were set in a way that more directly corresponds to the parameters of the data generating process.

[2] Assuming the positive results from the original trials were true positives and no spurious findings.

The parameters needed for all scenarios are the control arm hazard (the hazard for death before to progression in the progression scenario), the rate of random withdrawal, the (maximum) number of patients recruited and the time span of recruitment.

For the scenarios with delayed effect the additional parameters needed are time to onset of treatment effect and the treatment arm hazard after onset of treatment effect. In the crossing hazards scenarios time of crossing of the hazard curves and the treatment arm hazard before and after crossing of the hazard need to be set. For the subgroup scenarios the hazards for the subgroups of the treatment arm and the prevalence of the subgroups have to be defined. For the scenario with disease progression the parameters that need to be fixed are: the hazards for death before progression in the treatment arm, the hazard for death after progression in both arm, and the rates of progression in both arms.

Hazard rates were either expressed in terms of medians and converted using formula 4.1 or defined via a hazard ratio relative to other pre-defined hazard rates. All combinations of parameter values were simulated in a fully factorial design. See table 4.1 for list of parameter values.

Table 4.1.: Parameter values used for simulations

| Scenario | Parameter | Values |
|---|---|---|
| All | Median Survival in the Control Arm | 12, 24, 36 months |
| | Median Time to Random withdrawal [a] | $\infty$, 120 months |
| | Number of Patients [b] | 300, 600, 1000 |
| | Recruitment Time [c] | 0, 6 months |
| Delayed Effect | Delay | 0, 2, 4, 6, 8 months |
| | HR treatment after onset vs. control | 0.7, 0.8, 0.9, 1 |
| Crossing Hazards | Crossing | 0, 2, 4, 6, 8 months |
| | HR treatment before crossing vs. control | 1/0.7, 1/0.8, 1/0.9 |
| | HR treatment after crossing vs. control | 0.7, 0.8, 0.9 |
| Subgroup | HR complement vs. control | 0.7, 0.8, 0.9 |
| | HR subgroup vs. control [d] | 0.7, 0.8, 0.9, 1/0.7, 1/0.8, 1/0.9 |
| | Subgroup prevalence | 0.1, 0.3, 0.5 |
| Disease Progression | HR death before progression treatment vs. control | 0.7, 0.8, 0.9, 1 |
| | HR death after progression vs. control before progression | 2, 3 |
| | Median time to progression control | 12 months |
| | HR time to progression treatment vs. control | 0.7, 0.8. 0.9, 1 |

a) $\infty$ denotes no random censoring.
b) Participants were randomized 1:1 into treatment and control arm. Interim and final analysis were performed after 3/8 and 3/4 of the participants experienced an event.
c) 0 denotes instantaneous recruiting, all patients enter the trial at the start of the trial.
d) Scenarios with equal hazard in the subgroup and the complement were excluded.

### 4.1.6. Samples Size and Power to detect violation of nominal Type I error rate

The number of replications for each scenario was chosen to give a Monte Carlo standard error of one percentage point in the worst case of 50% coverage of the confidence interval. This gives a number of 2500 replications for each scenario. (see equation 4.2 compare also [25])

$$n_{sim} = \frac{0.5 \cdot (1 - 0.5)}{0.01^2} = 2500 \tag{4.2}$$

This sample size yields a power of about 83% to detect a deviation to 3.5% from the nominal type I error rate of 2.5% for an exact test or equivalently using the Clopper-Pearson [29] confidence interval.

## 4.2. Implementation of the Data Generating Mechanisms

Functions to calibrate simulation parameters, generate simulated data, calculate true values of summary statistics and to aggregate simulation results were implemented in the R package SimNPH [23]. Estimation and test procedures from published packages are used wherever possible and wrappers are implemented to account for different data structure used in the packages. Calculations and random number generators for time to event data were implemeted in the miniPCH package [30].

### 4.2.1. The miniPCH Package

The miniPHC package [30] contains functions to calculate density, distribution function, hazard, cumulative hazard, and survival function for survival distributions for piecewise constant hazards as well as multi-state models used in the scenarios with disease progression. For piecewise constant hazards the quantile function and a random number generator based on the quantile function are also implemented. The miniPCH package is used in random number generation and calculations in the SimNPH package which in turn is used for the simulation studies. This is used to generate data (see section 4.2.2), to calculate the true values of summary statistics of the distribtuions (see section 4.2.4) and for other calculations during the simulation studies.

#### Piecewise constant hazards

For piecewise constant hazards the hazard, cumulative hazard, cumulative density, probability density, survival and quantile functions can be calculated analytically. Random numbers can then be generated by inverse transform sampling (See [31, Section 3.2]) facilitating the uniform random number generator implemented in R and the quantile function implemented in the miniPCH package.

All calculations use the left interval borders $t_i$ of the time intervals $[t_i, t_{i+1})$, the hazard rates in the time intervals $\lambda_i$, and the point at which to compute the function value $x$ as input. $t_i$ and $\lambda_i$ are assumed to be ordered in chronological order.

**Hazard function** The hazard function is piecewise defined, the value at $x$ is therefore:

$$\lambda(x) = \lambda_{\max_i\{i:t_i<=x\}}$$

**Cumulative hazard function** The cumulative hazard function is the integral over the hazard function from 0 to $x$. For a piecewise constant function this is just the sum of the values multiplied with the interval lengths for all intervals left of $x$, 0 for all intervals right of $x$ and the piecewise constant value multiplied with the time from the left interval border to $x$ for the interval containing $x$:

$$\Lambda(x) = \sum w_i \lambda_i$$

with

$$w_i := \begin{cases} t_{i+1} - t_i, & x > t_{i+1} \\ x - ti, & x \in [t_i, t_{i+1}) \\ 0, & x >= t_{i+1} \end{cases}$$

**Survival function, cumulative density function** The survival function can be calculated as $S(x) = \exp(-\Lambda(x))$ and the cumulative density function as $F(x) = 1 - S(x)$.

**Probability density function** The probability density function can be calculated from the survival function and the hazard function by re-arranging the definition of the hazard function:

$$\lambda(x) = \frac{f(x)}{S(x)}$$
$$f(x) = \lambda(x)S(x)$$

**Quantile function** The quantile function is the inversion of the cumulative density function. The $p$-quantile can therefore be calculated by solving the equations defining the cumulative density function for the time $t$. Let $i^*$ be defined as the index such that $t_{i^*}$ is the largest interval border smaller than $t$. Then the cumulative hazard at $t$ can be written as $\Lambda(t) = \Lambda(t_{i^*}) + (t - t_{i^*})\lambda_{t_{i^*}}$.

$$\begin{aligned} p &= F(t) \\ &= 1 - \exp(-\Lambda(t)) \\ &= 1 - \exp(-\Lambda(t_{i^*}) - (t - t_{i^*})\lambda_{t_{i^*}}) \\ &= 1 - \frac{\exp(-\Lambda(t_{i^*}))}{\exp((t - t_{i^*})\lambda_{t_{i^*}})} \\ &= 1 - \frac{1 - F(t_{i^*})}{\exp((t - t_{i^*})\lambda_{t_{i^*}})} \end{aligned}$$

Note that $i^*$ can easily be computed by computing $F(t_i)$ for all $i$ and then finding the largest $F(t_i)$ larger than $p$ since $F(t)$ is monotonous. It remains to solve the equation:

$$\begin{aligned} p &= 1 - \frac{1 - F(t_{i^*})}{\exp((t - t_{i^*})\lambda_{i^*})} \\ \frac{1 - F(t_{i^*})}{1 - p} &= \exp((t - t_{i^*})\lambda_{t_{i^*}}) \\ \log\left(\frac{1 - F(t_{i^*})}{1 - p}\right) &= (t - t_{i^*})\lambda_{t_{i^*}} \\ t_{i^*} + \log\left(\frac{1 - F(t_{i^*})}{1 - p}\right)\lambda_{t_{i^*}}^{-1} &= t \\ t_{i^*} - \log\left(1 - \frac{p - F(t_{i^*})}{1 - F(t_{i^*})}\right)\lambda_{t_{i^*}}^{-1} &= t \end{aligned}$$

### Multistate Models

The scenarios with disease progression can be described as time continuous Markov chains with a finite state space. The time spent in each state (the holding times) is exponentially distributed and the jump process is a discrete Markov chain. Part three of the book by Beyersmann, Allignol and Schumacher [1] gives a good introduction to multistate models. This section uses notation and results about continuous time Markov chains from part two of the book by Norris [32].

Let $N$ be the number of states. An $n \times n$ Q-matrix is defined as a matrix where all off-diagonal entries are positive and the diagonal entries are minus the sum of the off-diagonal entries of the same row. (Such that the row-sums are zero for all rows.) The off diagonal entries $Q_{i,j} \quad i \neq j$ can then be interpreted as the hazards of transitioning from state $i$ to state $j$. And the transition probabilities from state $i$ to state $j$ within a time interval of length $t$ can be calculated as $(\exp(tQ))_{i,j}$ where exp denotes the matrix exponential function. With the distribution $\pi$ at time $t = 0$ the probability of being in state $j$ at time $t$ is then $(\pi \cdot \exp(tQ))_j$. This can be re-written as $\pi \cdot \exp(tQ) \cdot \mathbb{I}_{\{j\}}$ where $\mathbb{I}_{\{A\}}$, $A \subseteq \{1 \dots N\}$ is the $N \times 1$ vector that is one in the components $i \in A$ and zero everywhere else.

As in the section of piecewise constant hazards, $Q$ is assumed to be piecewise defined (as a function from $t > 0$ into the space or Q-matrices) with $t_i$ denoting the left interval

boundaries and $Q_i$, $i \in \{1 \dots n\}$ the Q-matrix describing the state transitions in the $i$-th time interval.

**Distribution function, survival function and cumulative hazard** The probability of being in one of the states in $A \subseteq \{1 \dots N\}$ at time $t$ is then:

$$p_A(x) = \pi \cdot \underbrace{\exp((t_1 - t_0)Q_0) \cdot \exp((t_2 - t_1)Q_1) \cdot \ldots \cdot \exp((x - t_{n^*})Q_{n^*})}_{=:P} \cdot \mathbb{I}_A = F(x)$$

Where $n^*$ is again the largest $n$ such that $t > t_n$. If $A$ is an absorbing set, this is exactly the distribution function of the hitting time of $A$. (Since once the process enters $A$ it cannot leave it, the probability of having hit the set by time $t$ is equal to the probability of being in $A$ by time $t$.) Note that $A$ has to be absorbing in the sense, that a transition out of $A$ is not possible in any time interval.

The survival function can then be calculated as $S(x) = 1 - F(x)$ The cumulative hazard function can subsequently be calculated as $\Lambda(x) = -\ln(S(x))$

**Density and hazard** Recall the differential equation $\frac{\partial}{\partial t}\exp(tQ) = \exp(tQ)Q$ and the fact that $\exp((a+b)Q) = \exp(aQ) \cdot \exp(bQ)$ for scalar $t$, $a$, $b$ and matrices $Q$. The density can then be calculated as the derivative of the distribution function with respect to $x$. In the product of matrices all but the last factor are constants with respect to $x$, yielding:

$$\begin{aligned} d(x) &= \frac{\partial}{\partial x}\pi \cdot \overbrace{\exp((t_1 - t_0)Q_0) \cdot \exp((t_2 - t_1)Q_1) \cdot \ldots \cdot \exp((x - t_{n^*})Q_{n^*})}^{=P} \cdot \mathbb{I}_A \\ &= \pi \cdot \exp((t_1 - t_0)Q_0) \cdot \exp((t_2 - t_1)Q_1) \cdot \ldots \cdot \exp(-t_{n^*}Q_{n^*}) \left(\frac{\partial}{\partial x}\exp(xQ_{n^*})\right) \cdot \mathbb{I}_A \\ &= \pi \cdot P \cdot Q_{n^*} \cdot \mathbb{I}_A \end{aligned}$$

The hazard can then be calculated as $\lambda(x) = \frac{d(x)}{S(x)}$. Note that this can be numerically unstable at the tail of the survival function where $d$ and $S$ go to zero. This however was not an issue for the computations in this simulation study.

### 4.2.2. Data Generation: Assumptions

Data generation can be split into two parts. In alignment with the CSE framework the part of the data generation not under control of the experimenter is called *assumptions*.

#### Piecewise Constant Hazards

The scenarios with piecewise constant hazards can be directly simulated with the functions implemented in the miniPCH package [30].

#### Senarios with Subgroups

For the scenarios with differential treatment effect in a subgroup, the survival time distribution is a mixture of two distributions with piecewise constant hazards. For a mixture the probability density function, the cumulative density function and the survival function can be directly calculated as a weighted sum of the probability density, cumulative density and survival functions of the mixture components. The hazard can then be calculated as the pointwise quotient of probability density function over the survival function. The cumulative hazard can be calculated as minus the logarithm of the survival function. For the quantile function no closed form exists, but quantiles can easily be computed numerically by finding a solution to $F(x) - p = 0$ with the cumulative density function $F(x)$ and the probablity $p$. Random numbers can be generated without explicitly calculating the quantile function by sampling from a multinomial distribution and then sampling from the mixture components according to the multinomial sample.

#### Scenarios with Disease Progression

Simulation of survival times in multi-state models could be done by directly sampling from the all-cause hazard and then running a multinomial experiment to determine the event type [1, chapter 3.2] (in this case death before disease progression and death after disease progression). In the SimNPH package a different approach was used. (Potentially unobserved) event times for death before progression, time to progression and time from progression to death were generated for each individual participant. The final event times were calculated from intermediate results. If time to death before progression was smaller than time to progression, time to death was set to time to death before progression and time to progression was set to be censored at time of death before progression. If time to death before progression was larger than time to progression, time of death was set to time to progression plus time from progression to death and time to progression was recorded. This has two advantages: one does not rely on the time-consuming approximation of the survival function and numerically solving for the quantiles in each simulation; and the time of the intercurrent event can be recorded in the simulation to enable evaluating not only overall survival but also progression-free survival.

#### Random Censoring

Random censoring was implemented by modifying the generated dataset of survival times. Censoring times are generated from an exponential distribution. Whenever the censoring time is before an event, the respecitve event indicator is set to censored and the respective event time is set to the censoring time. This is done for death as well as disease progression.

### 4.2.3. Data Generation: Options

Another important part of the data generating mechanism lies within the choice of the experimenter, in line with the CSE framework they are referred to as *options*.

### Administrative Censoring

To handle administrative censoring a recruitment time was generated. In the simplest case it was set to 0 for all patients, in the case of recruitment over a time span it sampled from a pre-specified distribution. In the simulations a uniform distribution over a pre-specified time period was used.

At time of analysis, which is determined by an observed number of events or a pre-specified time, participants for whom the sum of recruitment time $r_i$ and survival time $t_i$ is after the time of study end $T$ are censored at the time of study end. The event indicator $\delta_i$ is set to 0 (censored) if $r_i + t_i > T$ and is left unchanged if $r_i + t_i \leq T$.

$$\delta_{i,\text{new}} = \begin{cases} \delta_i, & t_i + r_i \leq T \\ 0, & t_i + r_i > T \end{cases} \qquad t_{i,\text{new}} = \begin{cases} t_i, & t_i + r_i \leq T \\ T - r_i, & t_i + r_i > T \end{cases}$$

Or equivalently $t_{i,\text{new}} = \min\{t_i + r_i, T\} - r_i$. If the study ends before all patients are recruited, patients not yet recruited are excluded from the dataset. If the study is to be stopped after a number of events, but less events can be observed, the study is stopped after the last observed event.

### Group sequential tests

Group sequential tests were implemented as follows: times or numbers of events and the alpha levels to use at each step are specified. At the first interim analysis administrative censoring, as described above is applied to the dataset. Any of the implemented test procedures is applied to the administratively censored dataset, if the p-value is lower than the alpha level given for the respective stage of the testing sequence, the test rejects the null. The function then returns whether the null hypothesis was rejected and if so at which stage, the number of recruited patients, the number of observed events, the time from start of recruitment to study termination, and all details the used test procedures at each stage return.

Note that this implementation does not correspond one to one to the data generating mechanism of a true group sequential trial, since data that would not be generated in a group sequential trial are generated and later discarded. This approach however allows to compare trials with and without the possibility to stop early for efficacy using the same simulated dataset.

Interim analyses were performed after half of the maximum number of events were observed. The nominal alpha at each stage was set using an O'Brien Fleming alpha spending approach [33] as implemented in the ldbounds package [34]. An information fraction of 0.5 and 1 respectively and an overall one sided alpha of 2.5% giving a nominal alpha of 0.153% at the interim analysis and a nominal alpha of 2.45% at the final analysis.

## 4.2.4. True Values of Summary Statistics

To assess bias and mean squared error of estimators the true values of summary statistics are needed. The calculations of the true value of most summary statistics from the

underlying distributions is simple but some care is required when choosing the weights for the average hazard ratios. The *RMST* is just the area under the survival curve up to the cutoff time and was computed by numerical integration of the survival function. *Milestone survival* can be computed using the survival functions derived above.

Quantiles can be computed directly for the scenarios with piecewise constant hazards. For the other scenarios quantiles can be calculated by numerically solving $F(t) - q = 0$ for $t$.

#### Average Hazard Ratios

The geometric average hazard ratio (gAHR), the average hazard ratio (AHR) and the average hazard ratio corresponding to the odds of concordance (AHRoc) where computed according to formula (2), formula (3) and formula (7) in Schemper et al. [16]. For the gAHR and AHR a weight function of $w(t) = 1$ was used. The integration was performed with R's built in `integrate` function.

## 4.3. Packages and Methods used for Estimation

For methods already implemented in other R packages the existing implementations were used. Wrapper functions were written so that all methods could readily be applied to the simulated data despite differing interfaces. The implementations uses for the different procedures are listed in Table 4.2.

**Difference in median survival from Weibull models** As suggested by members of the CONFIRMS consortium, an additional parametric estimate for the difference of median survival was calculated by extracting the median survival times from two Weibull models for the two treatment arms and constructing confidence intervals using the delta-method [23]. The implementations used were the `survreg` function from the survival package to get estimates for the parameters and covariance matrices for the Weibull model and the `deltaMethod` function from the car package [37] to calculate the confidence intervals from the estimates. The survival package uses a parametrization of the Weibull distribution also used in Kalbfleisch and Prentice 2011 [18]. Therefore the median can be computed from the fitted intercept ($\lambda$) and scale ($\gamma$) parameters as:

$$\exp(\lambda) \cdot \log(2)^{\log(\gamma)}$$

To apply the delta-method, the covariance of the model parameters and the derivatives of the summary statistics (difference in median survival) with respect to the original parameters ($\lambda$ and $\gamma$, respectively for both treatment arms) are needed. The covariance of the two parameters for both arms is estimated by the `survreg` function. No correlation between the parameters for the two models is assumed, giving a block-diagonal covariance structure. The derivatives are automatically computed by symbolic differentiation in the `deltaMethod` function.

Table 4.2.: Implementations of analysis procedures used

| Procedure | Function | Package |
|---|---|---|
| Weibull accelerated failure time model | survreg | survival [35] |
| Cox proportional hazards model | coxph | |
| Gehan-Wilcoxon test | survdiff [a] | |
| AHR<br>gAHR<br>difference of median survival<br>ratio of milestone survival | nphparams | nph [4] |
| logrank test<br>Fleming Harrington weighted logrank test | logrank.test | |
| max-combo test | logrank.maxtest | |
| modestly weighted logrank test | wlrt | nphRCT [36] |
| difference of median survival (Weibull models) | survreg, deltaMethod [b] | survival, car |

a) The test was implemented by comparing the test statistic from survdiff to the $\chi^2$ quantiles with the appropriate degrees of freedom.
b) see 4.3 for implementation details

## 4.4. Performance measures

The performance measures used for estimators were bias and mean squared error. If confidence intervals were available, average confidence interval length and coverage were reported.

For all fixed sample and group sequential hypothesis test power and type I error rate were reported. Since the design for both, fixed sample tests and group sequential tests, was event driven (the analyses are performed after a pre-defined number of events is observed) both the running time and the number of recruited patients is subject to chance, average numbers of patients and running times are reported for all test-procedures.

All performance measures used for estimators, fixed sample tests and group sequential tests as well as their definitions are listed in Table 4.3.

Table 4.3.: Performance Measures

| Methods | Performance Measure | Definition |
|---|---|---|
| Estimators | Bias | $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta)$ |
| | MSE | $\frac{1}{n_{sim}} \sum_{i=1}^{n} (\hat{\theta}_i - \theta)^2$ |
| | CI-Coverage | $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \{\text{CI covers true value}\}$ |
| | Average CI-Width | $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (u_i - l_i)$ |
| | One Sided CI-Based Rejection | $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \{\text{CI lies over/under value under } \mathcal{H}_0\}$ |
| (Group-Sequential) Tests | Power | under $\mathcal{H}_1$: $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{I}_{\{\text{procedure rejects}\}}(i)$ |
| | Type-I Error Rate | under $\mathcal{H}_0$: $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{I}_{\{\text{procedure rejects}\}}(i)$ |
| | Average Sample Size | $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} n_{pat,i}$ |
| | Average Running Time | $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (t_{end,i})$ |

$\theta$ the quantity targeted for estimation
$\theta_i$ the estimate in the $i$th simulation
$n_{sim}$ the number of simulations
$u_i$, $l_i$ the upper and lower bounds of the confidence interval in the $i$th simulation
$n_{pat,i}$ the number of recruited patients in the $i$th simulation
$t_{end,i}$ the time from start of recruitment to end of study in the $i$th simulation
$\mathbb{I}_{\{\text{procedure rejects}\}}(i)$ the indicator function of the event, that the procedure rejects the null-hypothesis for the $i$-th simulated dataset.

## 4.5. Summarization and Presentation

For estimators of continuous metrics the mean and Monte Carlo standard error were reported. For coverage and rejection probabilities the proportion of simulations in which the CI covered the true value and the rejection probabilities were reported.

For confidence intervals the proportion of simulations in which the lower confidence bound was above the value of the corresponding test statistic under the null hypothesis (or the upper bound was below the value under the null) was reported to evaluate the rejection probability of the corresponding one sided test. Care has to be taken in the interpretation, since depending on the estimand a higher or lower value corresponds to survival benefit for the treated subjects. For continuous metrics as well as proportions asymptotic confidence intervals based on a normal approximation were computed.

### 4.5.1. Graphical Presentation: Hybrid Nested Loop Plots

As the main method for graphical representation hybrid nested loop plots [38] were used. Hybrid nested loop plots are a combination of Trellis plots and nested loop plots. In Trellis plots one performance metric is plotted against three dimensions of the parameter space as facet-rows, facet-columns and the x-axis. In nested loop plots all combinations of parameters are ordered lexicographically by all parameters. The order of parameters is arbitrary but often a natural ordering of parameters arises (compare [38]). Hybrid nested loop plots combine the two approaches such that for one performance metric one or two dimensions of the parameter space are plotted in facets and the remaining dimensions are presented in nested loop plots in each facet. Additionally parameter values can be selected to further constrain the parameter space shown.

### 4.5.2. Shiny App

In addition to the tabular and graphical presentation a shiny [39] app was developed. The app contains a section to visualise the simulation scenarios and a section to visualise the simulation results.

In the section to visualise the simulation scenarios the user can select parameters for a simulation scenario. Survival functions, hazard functions and hazard ratio of the selected scenario are plotted as a function of time and the values of summary statistics are listed in a table.

In the section to visualise the simulation results, a scenario class and performance metric can be selected. The scenarios in the scenario class can be filtered by parameter values. A hybrid nested loop plot [38] is drawn for the selected scenarios. Parameters can be assigned to be used to create facets and reordered to change the nesting of the loops on the x-axis. Horizontal lines at the nominal alpha of 2.5% and at 3.1% can be added to the plot. Under the null hypothesis and if the method maintains the nominal alpha level of 2.5%, 95% of the rejection rates should lie under 3.1% at 2500 replications (refer to 4.1.6). In a similar manner horizontal lines at the nominal coverage of 95% of the CIs and

a ribbon from from 95.84% to 94.12% can be added to the plot. If the method maintains the nominal coverage, 95% of the coverage proportions should lie within the ribbon.

The source code for the app [40] as well as the app [41] can be accessed online. Screenshot of the app are provided in Appendix C.

# 5. Results

For each of the 5760 scenarios (720 with delayed effect, 1620 with crossing hazards, 1152 with disease progression and 2268 with a subgroup effect) 2500 replications were simulated. The performance of 9 estimation methods and 4 testing procedures was investigated. For methods with user specified parameters (cutoff time for geometric average hazard ratio, average hazard ratio and RMST, time for milestone survival, $\rho$ and $\gamma$ parameters for the Fleming-Harrington weighted logrank tests and $t^*$ parameter for modestly weighted logrank test), multiple parameter values were explored such that 13 estimators and 7 tests were investigated. The performance of the tests was also investigated when used in a group sequential design. Of the 5760 scenarios investigated 648 were under the strong null-hypothesis (180 with delayed effect, 0 with crossing hazards, 72 with disease progression, 396 with subgroup effect). Abbreviations were used for all methods in the following figures, see table 5.1 for reference.

Table 5.1.: Abbreviations for Methods Used in Plots

| Label | Method |
|---|---|
| `logrank` | logrank test |
| `fh_0_1` | Fleming Harrington weighted logrank test $\rho = 0$, $\gamma = 1$ |
| `fh_1_0` | Fleming Harrington weighted logrank test $\rho = 1$, $\gamma = 0$ |
| `fh_1_1` | Fleming Harrington weighted logrank test $\rho = 1$, $\gamma = 1$ |
| `max_combo` | max-combo test |
| `modest_6` | modestly weighted logrank test $t^* = 6$months |
| `modest_8` | modestly weighted logrank test $t^* = 8$months |
| `fh_gs_0_1` | group sequential Fleming Harrington weighted logrank test $\rho = 0$, $\gamma = 1$ |
| `fh_gs_1_0` | group sequential Fleming Harrington weighted logrank test $\rho = 1$, $\gamma = 0$ |
| `fh_gs_1_1` | group sequential Fleming Harrington weighted logrank test $\rho = 1$, $\gamma = 1$ |
| `logrank_gs` | group sequential logrank test |
| `max_combo_gs` | group sequential max-combo test |
| `modest_gs_6` | group sequential modestly weighted logrank test $t^* = 6$months |
| `modest_gs_8` | group sequential modestly weighted logrank test $t^* = 8$months |
| `cox` | Cox proportional hazards model |
| `median_surv` | median survival |
| `ahr_6m` | average hazard ratio 0 to 6 months |
| `ahr_12m` | average hazard ratio 0 to 12 months |
| `gahr_6m` | geometric average hazard ratio 0 to 6 months |
| `gahr_12m` | geometric average hazard ratio 0 to 12 months |
| `milestone_6m` | milestone survival at 6 months |
| `milestone_12m` | milestone survival at 12 months |
| `rmst_diff_6m` | difference in RMST up to 6 months |
| `rmst_diff_12m` | difference in RMST up to 12 months |
| `aft_weibull` | Weibull accelerated failure time model |
| `aft_lognormal` | log-normal accelerated failure time model |
| `diff_med_weibull` | difference in median survival estimated from Weibull models |

Refer to Table 4.2 and Section 4.3 for used implementations

## 5.1. Estimators

### 5.1.1. Tests based on the CI

All one sided tests constructed from confidence intervals for the investigated estimators controlled the nominal type I error rate for the strong null-hypothesis. From the investigated estimators the Weibull-AFT model and the Cox model performed best with respect to power, which is not surprising since those tests are closely related to the logrank test. From the tests targeting an interpretable summary measure, the estimator for the difference in median survival showed the highest power across a wide range of scenarios. See figures 5.1, 5.2, 5.3 and 5.4. While the CI based test for a difference of median survival estimated from a Weibull model showed a good power, under the alternative the estimator exhibits strong bias and does not attain the nominal coverage. The difference in median survival estimated from a Weibull model can therefore not be considered a reliable estimator in the considered scenarios (See figures 5.5, 5.6, 5.7 and 5.8).

### 5.1.2. Bias

Bias was assessed separately for methods targeting an estimand on the time scale (RMST and median survival) and an the methods targeting an estimand on a risk-ratio or hazard ratio scale (average hazard ratio, geometric average hazard ratio and ratio of milestone survival). The estimator for the RMST showed a consistently low bias across all scenarios. The estimator for difference in median survival showed large variance and depending on the scenario. (see figures 5.5, 5.6, 5.7 and 5.8). The estimator for milestone survival ratio at different time points showed the lowest bias among the methods on a ratio scale. All methods on the ratio scale had a bias lower than 5 percent points (see sigure A.5, A.6, A.7 and A.8).

### 5.1.3. Coverage of Confidence Intervals

The only method with a strong deviation from the nominal coverage was the estimate of the difference in median survival based on Weibull models and the delta method which had a significant deviation from the nominal coverage in 449 of the 5760 scenarios. The estimators for the average hazard ratio in the first 12 months and for the median survival had a significant deviation in 106 and 29 scenarios. All methods that had at least one deviation from the nominal coverage that was significant at the 0.05 level after adjustment for false discovery rate (FDR) are listed in table 5.2. A nested loop plot of the coverage of the confidence intervals is listed in figures A.1, A.2, A.3 and A.4.

Coverage of the estimator for milestone survival could not be evaluated due to a coding error in the simulation code, where coverage of the reciprocal value was calculated instead of the correct value.

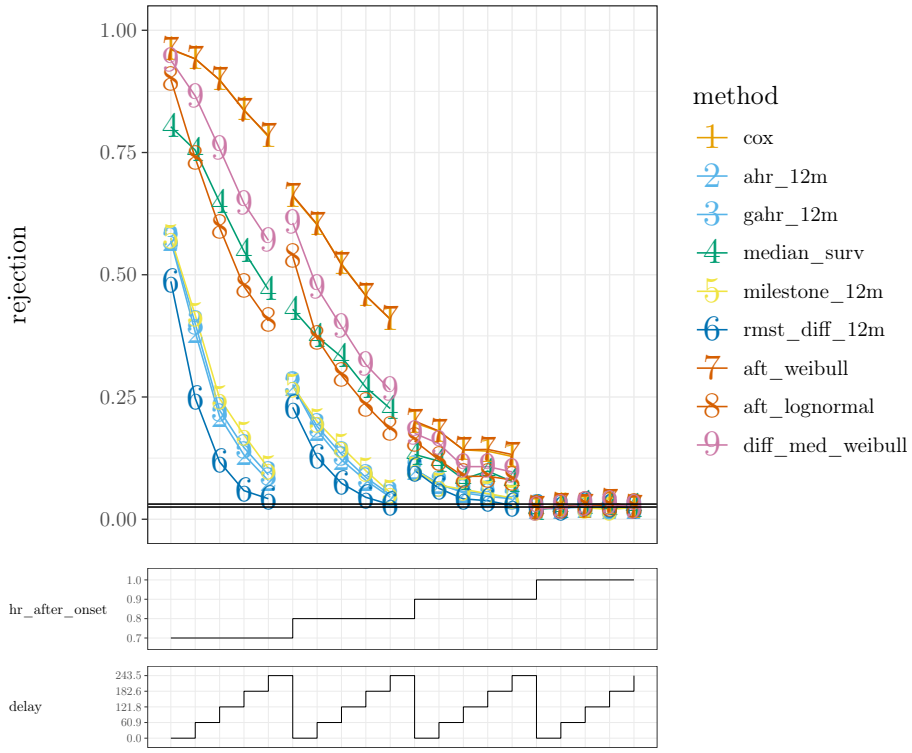



Figure 5.1.: Rejection rates of the tests based on the confidence intervals in scenarios with delayed onset of treatment effect. Cox model, accelerated failure time models, average hazard ratio, geometric average hazard ratio, difference in median survival, difference of median survival from a Weibull model, difference in RMST and ratio of milestone survival. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

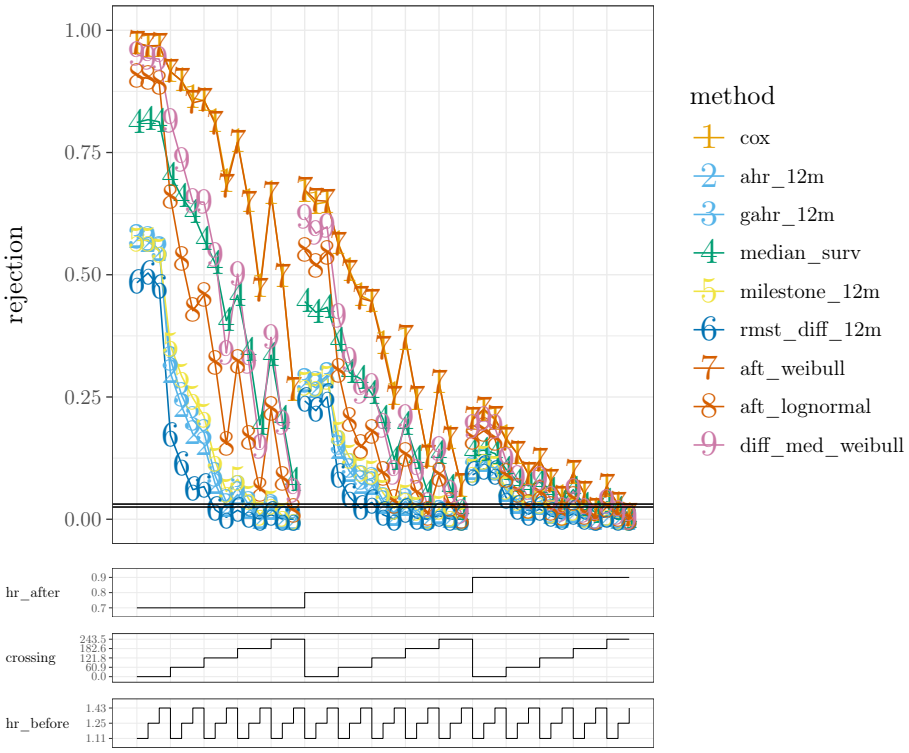



Figure 5.2.: Rejection rates of the tests based on the confidence intervals in scenarios with crossing hazard functions. Cox model, accelerated failure time models, average hazard ratio, geometric average hazard ratio, difference in median survival, difference of median survival from a Weibull model, difference in RMST and ratio of milestone survival. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

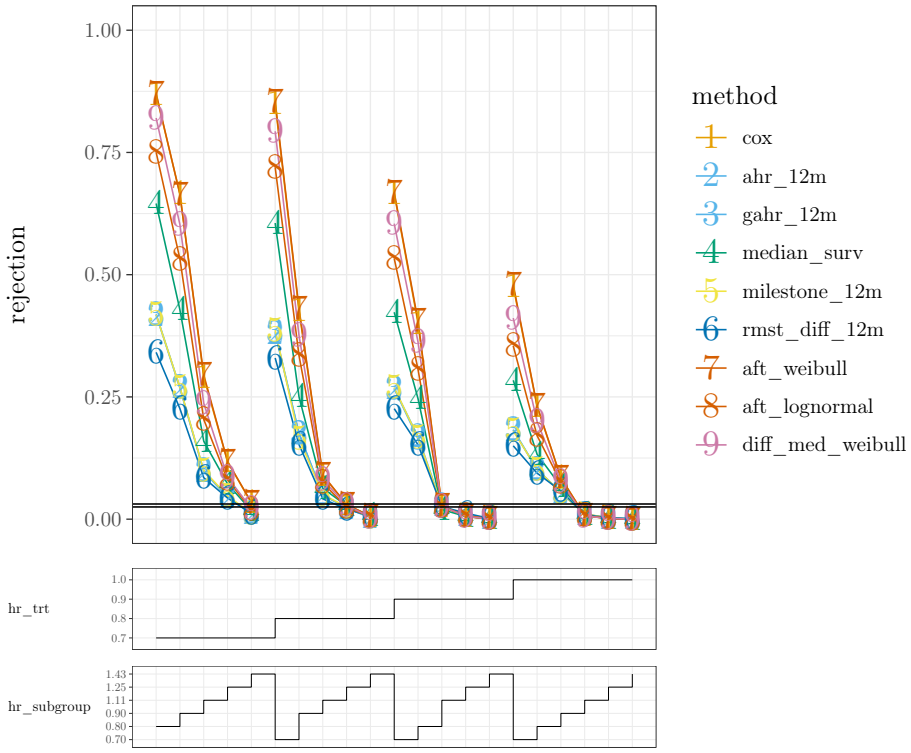



Figure 5.3.: Rejection rates of the tests based on the confidence intervals in scenarios with subgroup effect. Cox model, accelerated failure time models, average hazard ratio, geometric average hazard ratio, difference in median survival, difference of median survival from a Weibull model, difference in RMST and ratio of milestone survival. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months, prevalence of the subgroup 50%. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

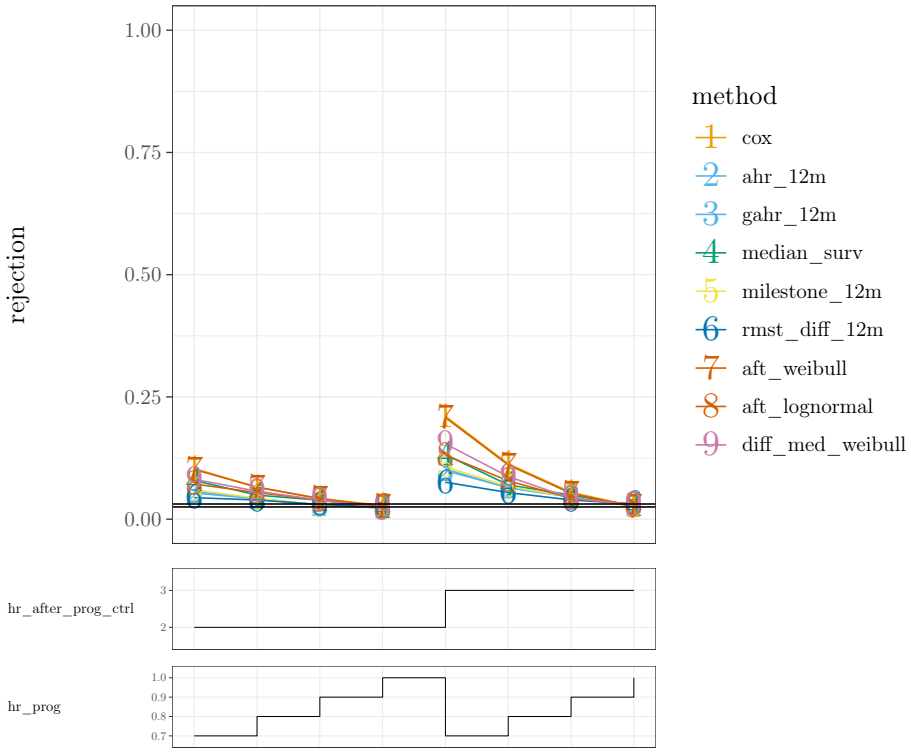



Figure 5.4.: Rejection rates of the tests based on the confidence intervals in scenarios with changing hazards after disease progression. Cox model, accelerated failure time models, average hazard ratio, geometric average hazard ratio, difference in median survival, difference of median survival from a Weibull model, difference in RMST and ratio of milestone survival. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months and equal hazards for death before disease progression in both arms. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

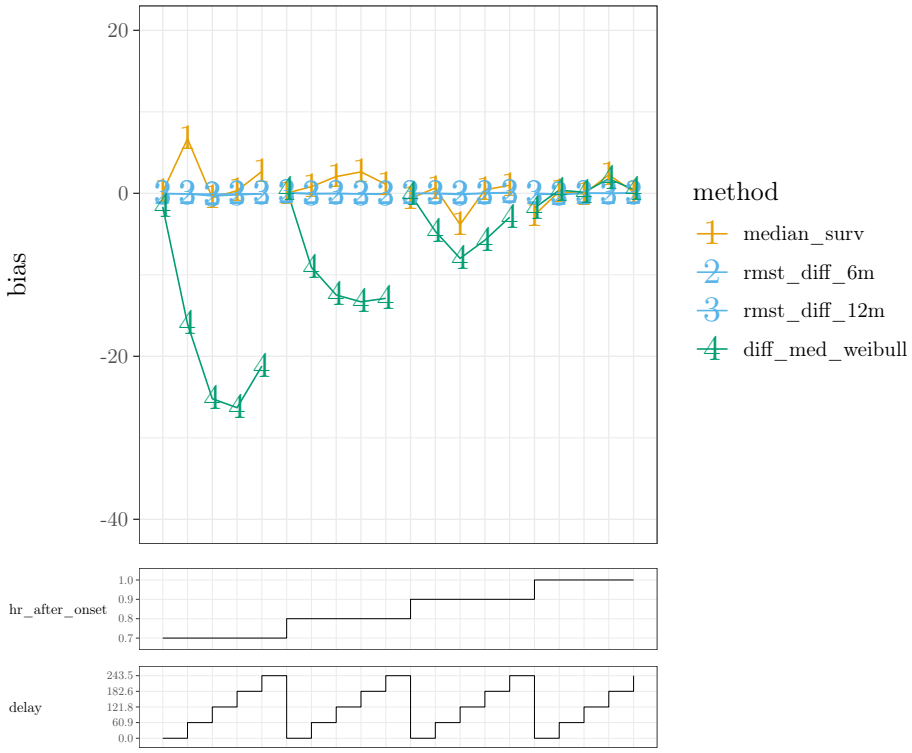



Figure 5.5.: Bias of the estimators targeting summary statistics on the time scale in scenarios with delayed onset of treatment effect. Difference in median survival, RMST difference and difference in median survival estimated from a Weibull model. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

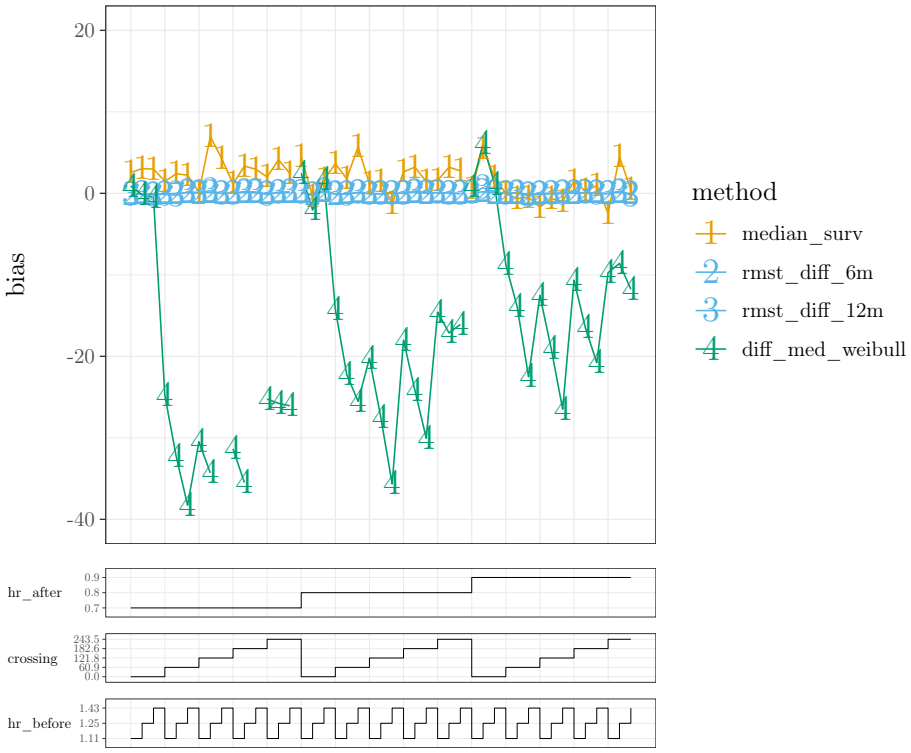



Figure 5.6.: Bias of the estimators targeting summary statistics on the time scale in scenarios with crossing hazard functions. Difference in median survival, RMST difference and difference in median survival estimated from a Weibull model. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

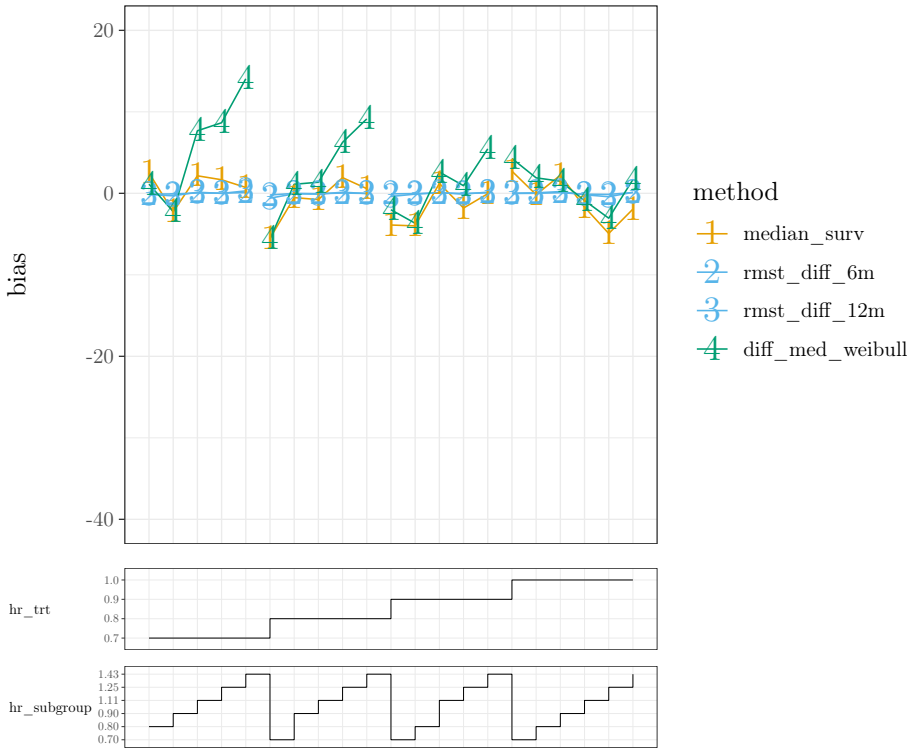



Figure 5.7.: Bias of the estimators targeting summary statistics on the time scale in scenarios with subgroup effect. Difference in median survival, RMST difference and difference in median survival estimated from a Weibull model. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months, prevalence of the subgroup 50%. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

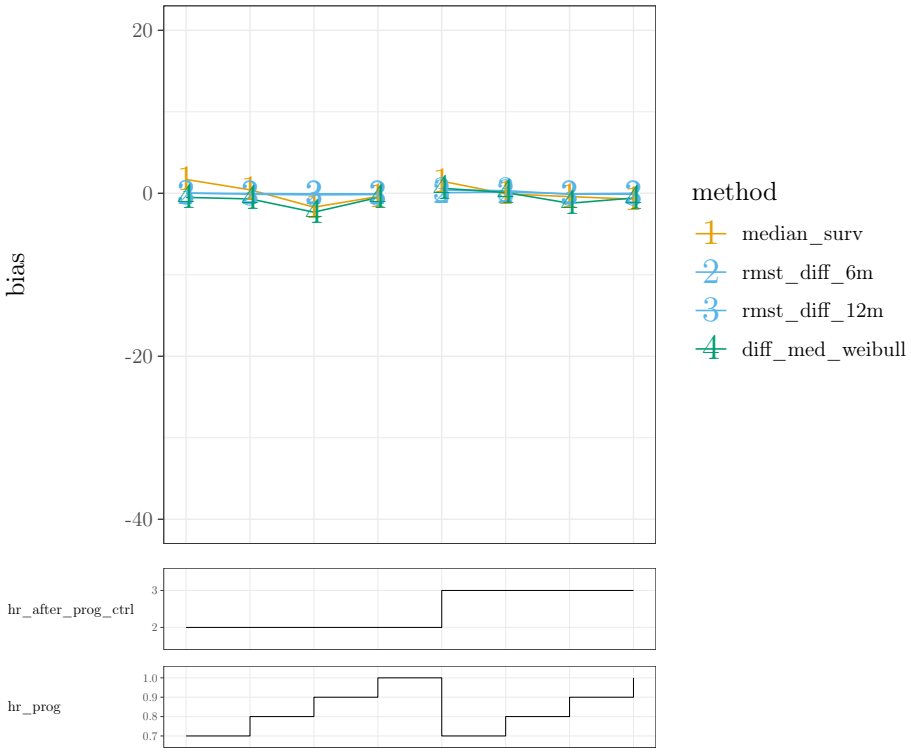



Figure 5.8.: Bias of the estimators targeting summary statistics on the time scale in scenarios with changing hazards after disease progression. Difference in median survival, RMST difference and difference in median survival estimated from a Weibull model. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months and equal hazards for death before disease progression in both arms. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

Table 5.2.: Estimators with deviation from the nominal-coverage

| Method | #Scenarios [a] | FDR adj. p-value [b] | FWER adj. p-value [c] |
|---|---|---|---|
| difference in median survival based on Weibull glm | 449 | 0.00e+00 | 0.00e+00 |
| average hazard ratio 12m | 106 | 5.69e-25 | 3.75e-23 |
| difference in median survival | 29 | 2.13e-07 | 4.22e-05 |
| average hazard ratio 6m | 5 | 2.83e-03 | 1.00e+00 |
| RMST (6m) difference | 4 | 7.27e-03 | 1.00e+00 |
| RMST (12m) difference | 5 | 9.62e-03 | 1.00e+00 |
| geometric average hazard ratio 6m | 2 | 1.29e-02 | 1.00e+00 |
| geometric average hazard ratio 12m | 1 | 4.84e-02 | 1.00e+00 |

a) Each method that showed a significant deviation from the nominal coverage of 95% at the 5% level after Benjamini-Hochberg adjustment for FDR in at least one scenario are shown. The column #Scenarios indicates in how many scenarios this occurred. p-values are based on a one-sided exact binomial test.
b) Column FDR-adjusted p-value shows the minimum of the FDR adjusted p-values of all scenarios in which the method showed significant under-coverage.
c) Column FWER-adjusted p-value shows the minimum of the p-values with the Holm adjustment for the family wise error rate (FWER).
Refer to Table 4.2 and Section 4.3 for used implementations

## 5.2. Tests

All investigated tests controlled the type I error rate for all explored scenarios under the strong null-hypothesis. Appropriately weighted logrank tests were able to gain some power in comparison to other tests. The modestly weighted test showed similar performance to the max-combo test, only in the scenarios with a late crossing of the hazard curves the max-combo test outperformed the modestly weighted test. Figures 5.9, 5.10, 5.11 and 5.12 shows the results for a selection of scenarios.

### 5.2.1. Group Sequential Tests

The group sequential tests were able to reject the null-hypothesis earlier in terms of observed events in scenarios with a more pronounced deviation from the null-hypothesis (See Figures 5.13, 5.14, 5.15 and 5.16). Depending on the hazards in the treatment and control arm this translated to an earlier rejection in terms of running time of the study. The group sequential tests had no benefit regarding the number of recruited patients in the investigated scenarios, this is because due to the quick recruitment the interim analysis was always scheduled after all patients were recruited. An interim analysis being scheduled after all patients have been recruited is not uncommon in trials with a survival outcome and the interim analysis might still have benefits for example that the participants can be switched to the more effective treatment earlier or that a marketing authorization can be submitted earlier.

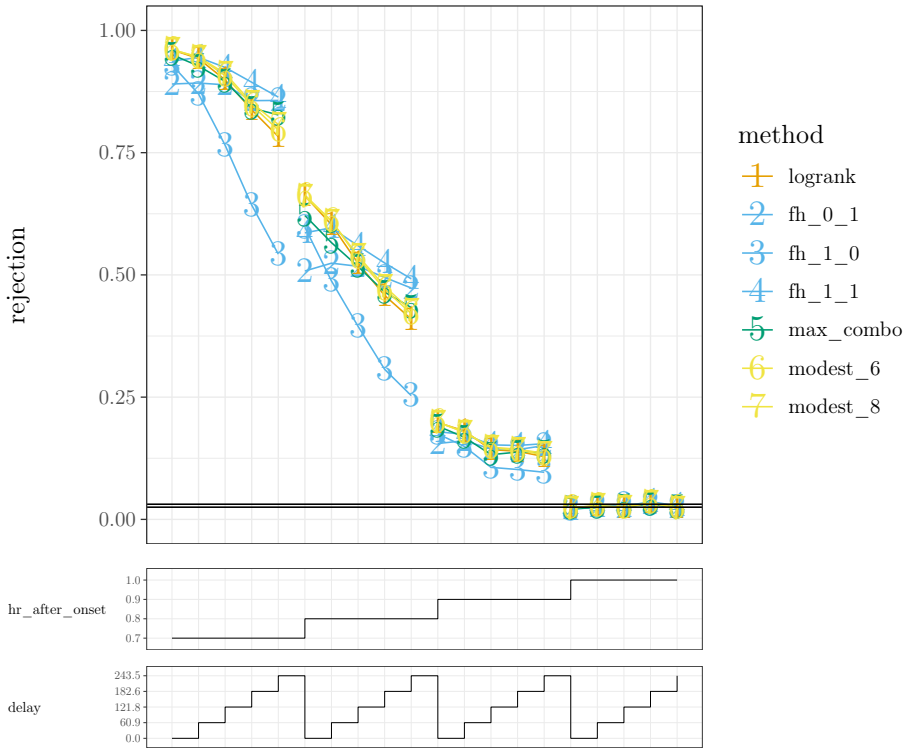



Figure 5.9.: Rejection Rates of logrank test, weighted logrank test, max-combo test and modestly weighted logrank test in scenarios with delayed onset of treatment effect. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

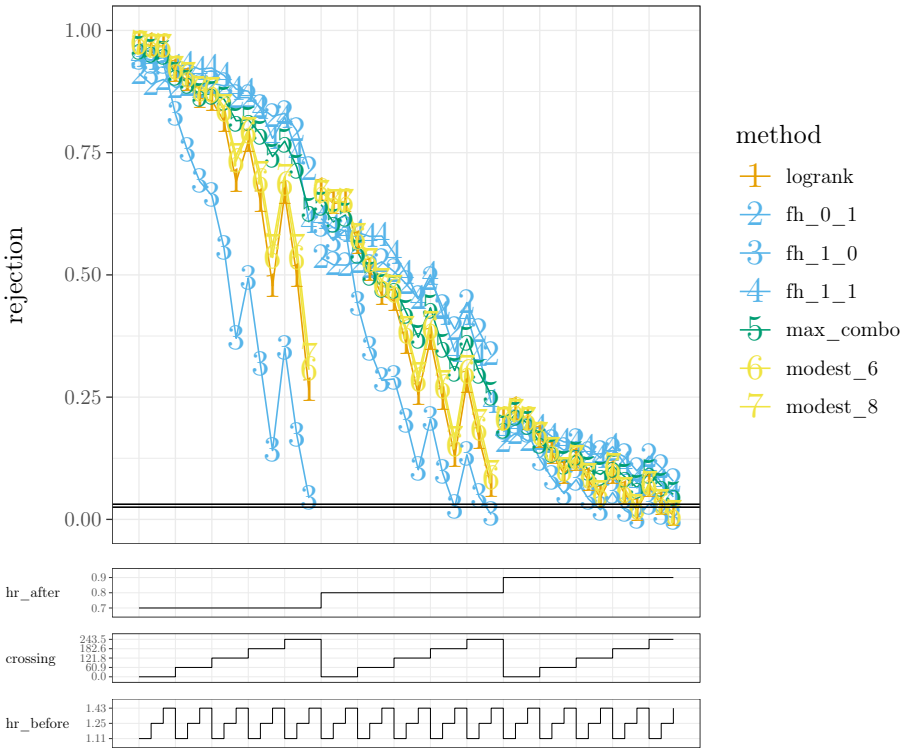



Figure 5.10.: Rejection Rates of logrank test, weighted logrank test, max-combo test and modestly weighted logrank test in scenarios with crossing hazard functions. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

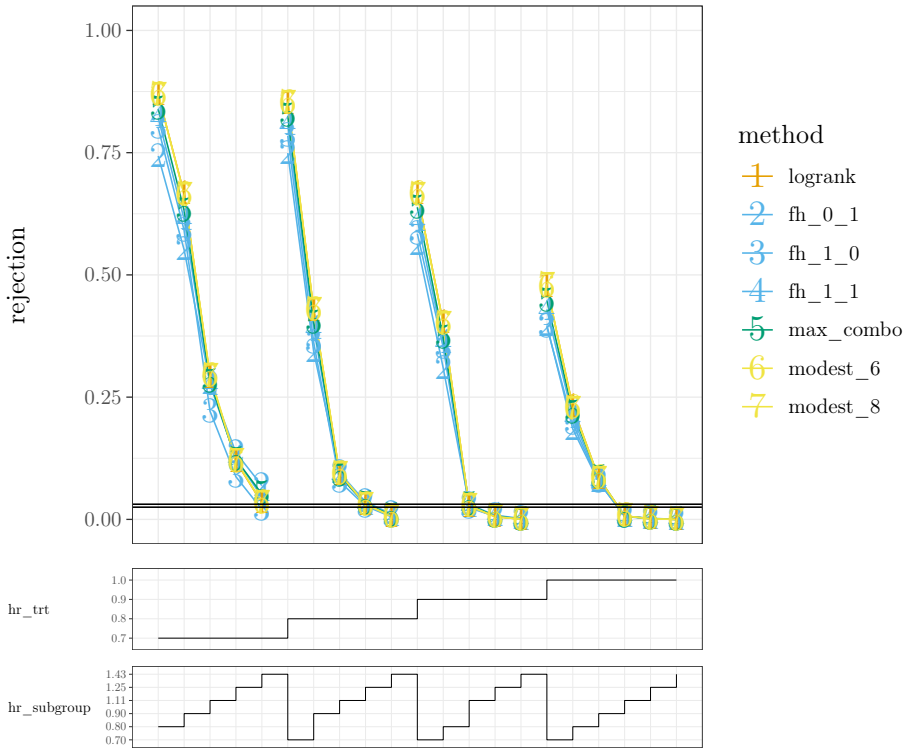



Figure 5.11.: Rejection Rates of logrank test, weighted logrank test, max-combo test and modestly weighted logrank test in scenarios with a subgroup effect. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months, prevalence of the subgroup 50%. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

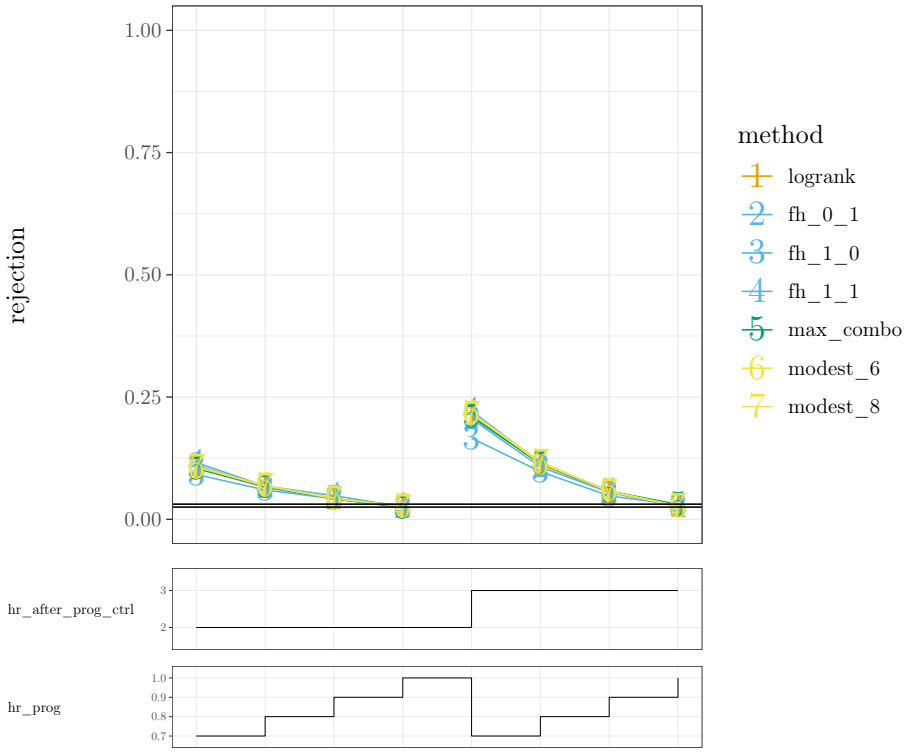



Figure 5.12.: Rejection Rates of logrank test, weighted logrank test, max-combo test and modestly weighted logrank test in scenarios with changing hazards after disease progression. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months and equal hazards for death before disease progression in both arms. The horizontal line at 0.025 indicates the nominal type I error rate under the null hypothesis. With 2500 replications 95% of the tests that maintain the nominal type I error rate should have a rejection rate under 0.031 – indicated by the second horizontal line.

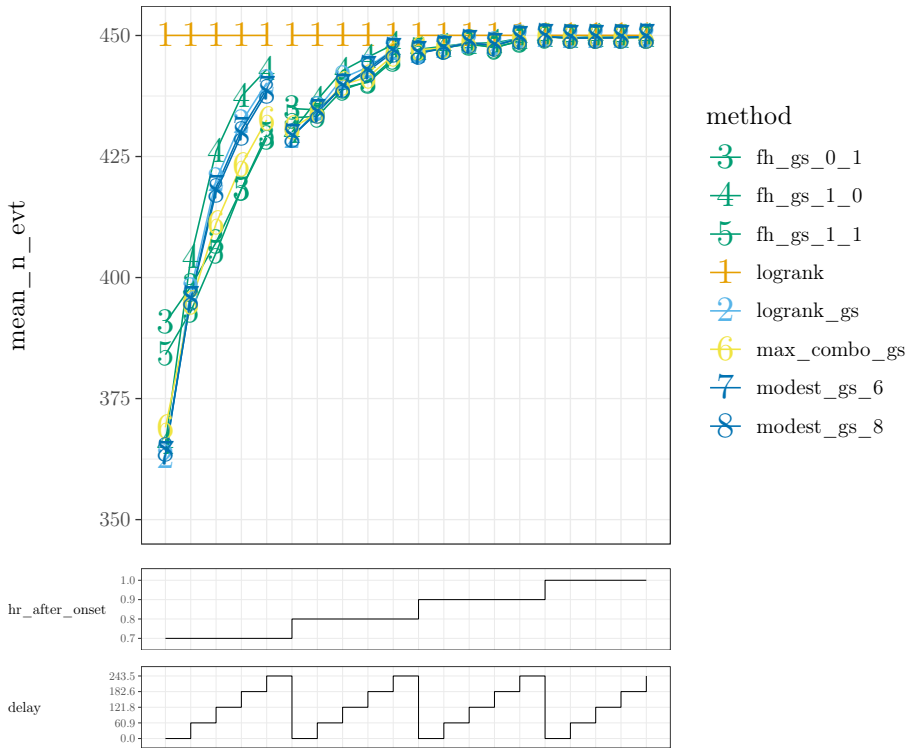



Figure 5.13.: Average number of events, scenarios with delayed onset of treatment effect. Logrank test and group sequential logrank test, weighted logrank tests, max-combo test and modestly weighted logrank test. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months.

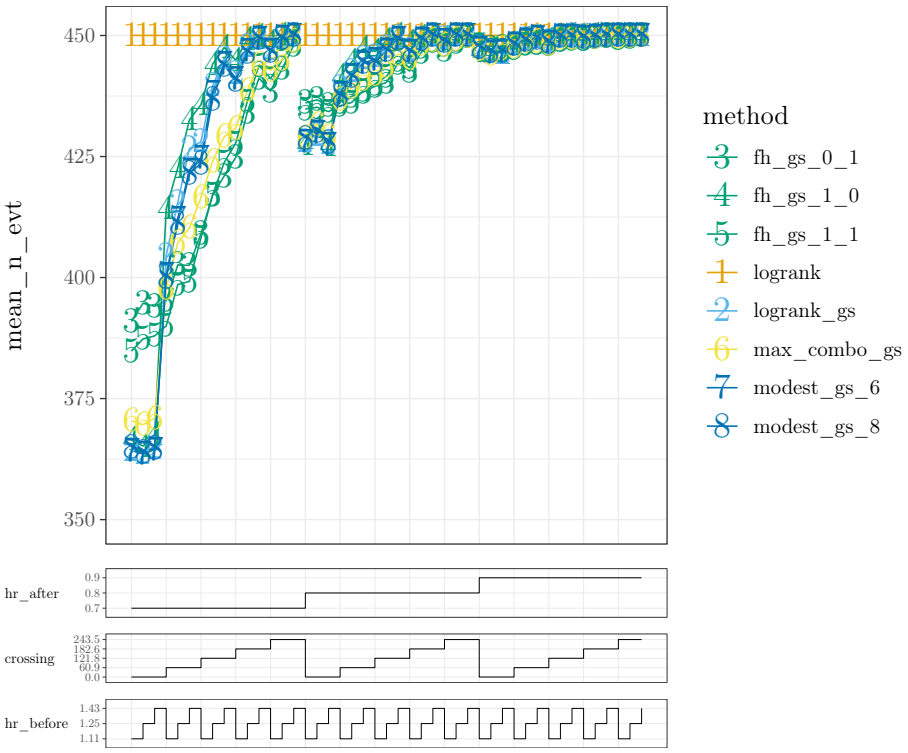



Figure 5.14.: Average number of events, Scenarios with crossing hazard functions. Logrank test and group sequential logrank test, weighted logrank tests, max-combo test and modestly weighted logrank test. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months.

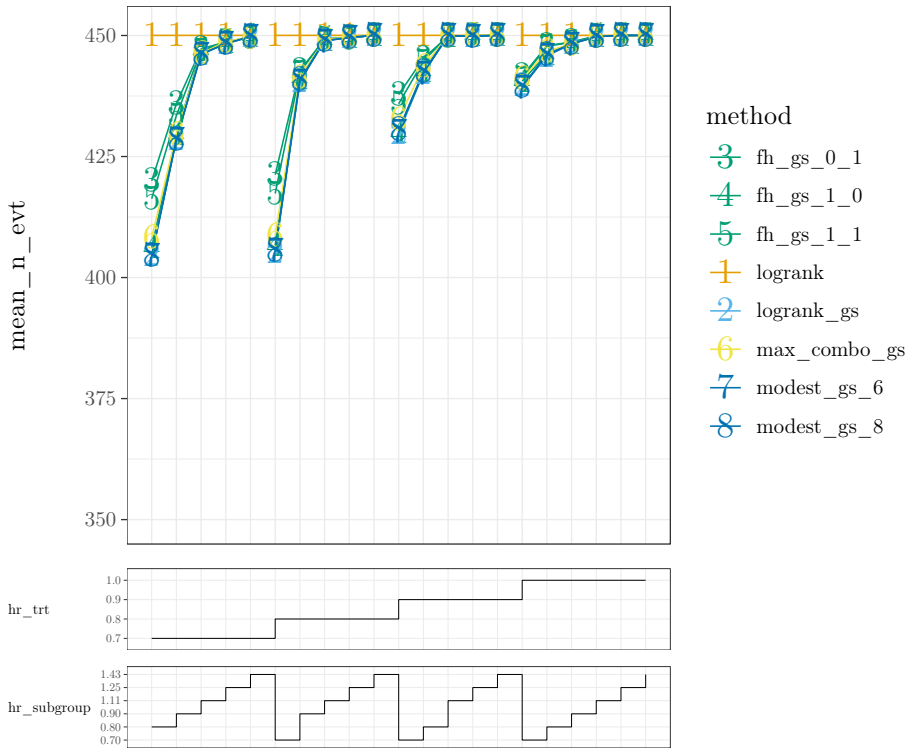



Figure 5.15.: Average number of events, scenarios with subgroup effect. Logrank test and group sequential logrank test, weighted logrank tests, max-combo test and modestly weighted logrank test. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months, prevalence of the subgroup 50%.

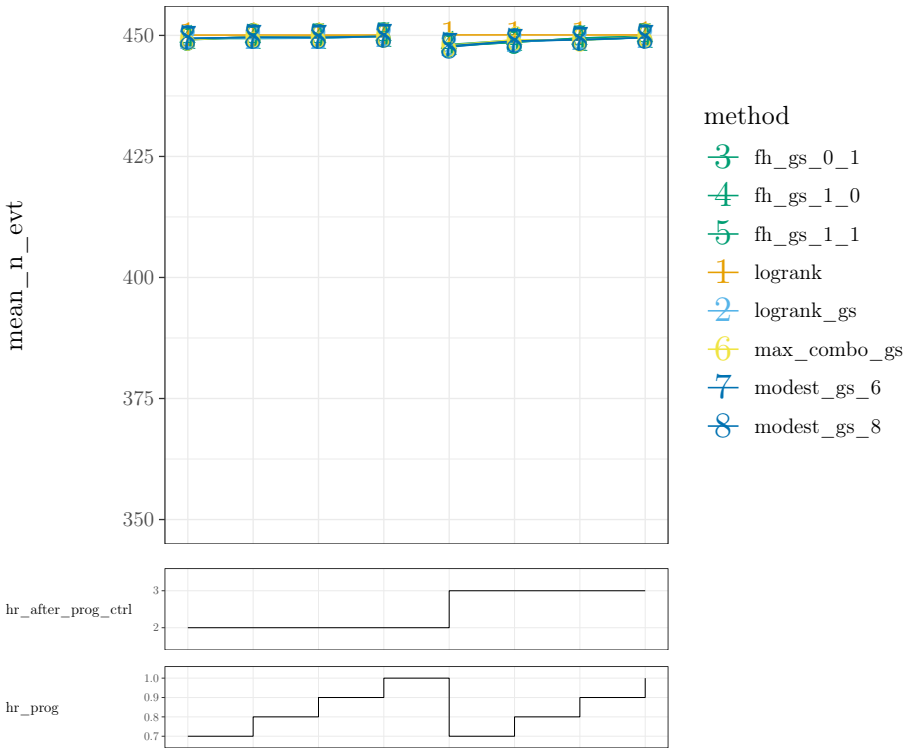



Figure 5.16.: Average number of events, scenarios with changing hazards after disease progression. Logrank test and group sequential logrank test, weighted logrank tests, max-combo test and modestly weighted logrank test. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months and equal hazards for death before disease progression in both arms.

# 6. Discussion

## 6.1. Implementation of the Simulation Study

**Parametrization of the Scenarios** While the methods to set parameter values described in 4.1.3 enable to directly translate summary measures found in the literature to simulation parameters, they also introduce a complex interdependence between the simulation parameters. For example, in the setting with delayed onset of treatment effect when pre-specifying: sample size, control arm hazard, delay and power of the logrank test – the treatment arm hazard after onset of treatment as well as the hazard for random censoring depend on the combination of all those pre-specified parameters. This can hinder the interpretation of the results, for instance the power of the weighted logrank test with weights at the end of the study period increases with a longer delay. When just changing the hazard ratio after onset of treatment the power of the weighted test stays similar. But since the logrank test looses power, the hazard ratio after onset of treatment effect is calibrated to more extreme values.

While calibrating the parameters in such a way ensures, that the scenarios are realistic in terms of effect size, setting the values for each parameter on a separate grid aids interpretation and visualization fo the results.

**Implementation** It is preferable to investigate existing implementations of statistical methods instead of re-implementing them. Implementations are often published and used in practice and there is a good chance that bugs are already found and fixed. Most implementations however are intended to be used to analyze few datasets instead of being applied to thousands of datasets. Some implementations are therefore not optimized for speed which impacts the computation time for the simulation study. The inclusion of some methods in the simulation study was not feasible at all, for example the weighted Cox regression from the coxphw package [42].

When the data generating model is simple enough or can be expressed in terms of simple building blocks for which an efficient implementation already exists, choosing the efficient methods can improve the performance. Lastly (re-)implementing parts of the code that are run very often during simulations in native C++ code improves runtime considerably.

While shorter in runtime than the actual simulations the calculation of true values of summary statistics also took substantial time when done naively. Especially when an inefficient R implementation is called inside routines for numerical integration (when calculating the RMST or AHRs) or numerical root-finding (when calculating quantiles from the survival function) a lot of speed can be gained by an efficient implementation.

## 6.2. Performance of the Statistical Methods

**Type I error rate** All investigated statistical methods kept the nominal type I error rate in the simulated scenarios contained in the null-hypotheses. The method to estimate the difference in median survival based on the Weibull model and the delta-method showed large bias in many scenarios and did not maintain the nominal coverage probability. All other estimators estimating a summary statistic on the time scale ha a bias of less than 5 days. All estimators targeting summary statistics on a ratio scale had a bias of less than 5 percentage points. Some estimators had a coverage proportion that was significantly lower than would be expected for the nominal coverage probability. It remains to investigate whether this is due to an underestimated variance or due to bias. This could be done with the "bias-eliminated coverage" introduced in [3].

**Specialized, weighted tests versus general methods** The logrank test is the most powerful test under proportional hazards and maintains good power for slight deviations from proportional hazards. Weighted tests have the largest power in the scenarios with the largest difference between the hazards in the times where the test puts most weight. But they loose power in other scenarios. The max-combo test has good power over a large range of scenario but is always outperformed by the weighted test that is most powerful in the respective scenario. This is clear since the max-combo test used has exactly the weighted tests also investigated as components in the combination. The max-combo test therefore uses the test statistic of the most powerful component but has to pay a penalty to account for multiple testing. The modestly weighted logrank test had a similar power to the max-combo test in all but the scenarios with a late crossing of the hazards. The modestly weighted logrank test seems to be a powerful and flexible option that also provides a certain safety due to its desirable theoretical properties.

**Estimators versus Tests** Methods targeting an interpretable summary measure of the survival distribution were in general less powerful than pure hypothesis tests that test a difference in survival distributions. This is to be expected. RMST and average hazard ratios do per definition only take into account information up to a certain time. Quantiles of the survival distribution only take into account the respective part of the information. The loss in power is lower for longer cutoff times of the summary statistics. While tests based on a summary statistic might be impractical as the main endpoint due to their lower power and therefore requirement for a larger sample size, it is useful to report a selection of meaningful summary statistics.

**Limitations** Summary statistics that would require to do calculations across replications which are more involved than calculating summary statistics were not included in the simulation study. This includes the data to construct zip plots and bias-eliminated coverage. While the SimDesign package provides functionality to save the estimates for all methods for each replication saving the necessary data for the high number of methods and high number of replications would have impacted the running time of the simulations.

Except in the scenarios with disease progression, the control arm hazard was always assumed to be constant. This is not an issue for methods based only on ranks, but scenarios relevant for other methods might not have been thoroughly investigated. Among the scenarios simulated there were few that were contained in the strong null hypothesis but not in the weak null hypothesis. A comparison between the modestly weighted logrank test and the max-combo test would have been interesting in those scenarios.

## 6.3. Conclusions

**Understanding the data and choosing an appropriate estimand** The examples in sections 2.3.1 and 2.4.1 illustrate the importance of understanding the data generating mechanism and carefully choosing an appropriate estimand when expecting non-proportional hazards. In the IPASS study having to account for non-proportional hazards could have been entirely avoided, if the analysis of the main outcome would have accounted for the effect of the biomarker. Besides avoiding the technical difficulties of dealing with non-proportional hazards, the data are better understood with regards to the causal relations of the treatment effect. The TERIKIDS trial had a complicated design with the possibility for crossover to the arm of the investigational treatment. Besides potentially introducing non-proportional hazards the intercurrent event has to be taken into account when defining the estimand [14].

**Investigating rejection of the strong null hypothesis** Besides the aim to better describe the data, another reason not to rely on test decisions alone is, that tests that target the weak null hypothesis might reject even under the strong null hypothesis. Inspection of Kaplan-Meier curves remains an important part of assessing the efficacy in survival settings, especially when weighted logrank tests are used that put weight on the later observations (compare to the example given in [17]).

**Recommendations** Tests for the difference in survival distributions have larger power to detect differences than methods testing for the difference in a summary statistic or estimating summary statistics. Easily interpretable summary statistics should nevertheless be reported, since they often translate to outcomes of interest for patients and decision makers.

Reporting other measures can also aide the causal interpretation of the results for which the hazard ratios pose challenges [9]. Other measures are also more useful for meta-analyses, for example the RMST does not depend on the proportional hazards assumption, that has to be assumed to aggregate hazard ratios across trials [43].

An inspection of the Kaplan-Meier curves remains an essential part of the interpretation of the results of studies with time to event outcomes. The Kaplan-Meier estimator can indicate issues of non-proportional hazards and Kaplan-Meier curves for subgroups can shed light on possible causes for non-proportional hazards in the case of a selection effect.

Overall the design and interpretation of studies of time to event endpoints remains an interdisciplinary exercise and statisticians and subject-matter experts need to collaborate

on choosing the appropriate methodology and study design to assess meaningful outcomes under the expected circumstances.

# Bibliography


[1] J. Beyersmann, A. Allignol, and M. Schumacher, *Competing Risks and Multistate Models with R.* Springer New York, 2012.

[2] N. Benda, M. Branson, W. Maurer, and T. Friede, "Aspects of modernizing drug development using clinical scenario planning and evaluation," *Drug Inf. J.*, vol. 44, no. 3, pp. 299–315, May 2010.

[3] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods," *Stat. Med.*, vol. 38, no. 11, pp. 2074–2102, May 2019.

[4] R. Ristl, N. Ballarini, H. Götte, A. Schüler, M. Posch, and F. König, "Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze rcts in oncology," *Pharmaceutical statistics*, vol. 20, no. 1, pp. 129–145, 2021.

[5] V. Anagnostou, M. Yarchoan, A. R. Hansen, H. Wang, F. Verde, E. Sharon, D. Collyar, L. Q. Chow, and P. M. Forde, "Immuno-oncology trial endpoints: Capturing clinically meaningful activity," *Clinical Cancer Research*, vol. 23, no. 17, pp. 4959–4969, Aug. 2017.

[6] European Medicines Agency, "KEYTRUDA (prembrolizumab) - Public Assessment Report," European Medicines Agency, Tech. Rep., 2019, accessed on 2024-02-11. [Online]. Available: https://www.ema.europa.eu/en/documents/variation-report/keytruda-h-c-3820-ii-0065-epar-assessment-report-variation_en.pdf

[7] B. Burtness, K. J. Harrington, R. Greil, D. Soulières, M. Tahara, G. de Castro, A. Psyrri, N. Basté, P. Neupane, Å. Bratland *et al.*, "Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (keynote-048): a randomised, open-label, phase 3 study," *The Lancet*, vol. 394, no. 10212, pp. 1915–1928, 2019.

[8] O. O. Aalen, "Effects of frailty in survival analysis," *Statistical methods in medical research*, vol. 3, no. 3, pp. 227–243, 1994. [Online]. Available: https://doi.org/10.1177/096228029400300303

[9] M. A. Hernán, "The hazards of hazard ratios," *Epidemiology*, vol. 21, no. 1, pp. 13–15, jan 2010.

[10] European Medicines Agency, "IRESSA (gefitinib) - Public Assessment Report," European Medicines Agency, Tech. Rep., 2009, accessed on 2023-11-20. [Online].

Available: https://www.ema.europa.eu/en/documents/assessment-report/iressa-epar-public-assessment-report_en.pdf

[11] T. S. Mok, Y.-L. Wu, S. Thongprasert, C.-H. Yang, D.-T. Chu, N. Saijo, P. Sunpaweravong, B. Han, B. Margono, Y. Ichinose, Y. Nishiwaki, Y. Ohe, J.-J. Yang, B. Chewaskulyong, H. Jiang, E. L. Duffield, C. L. Watkins, A. A. Armour, and M. Fukuoka, "Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma," *New England Journal of Medicine*, vol. 361, no. 10, p. 947–957, Sep. 2009. [Online]. Available: http://dx.doi.org/10.1056/NEJMoa0810699

[12] T. Chitnis, B. Banwell, L. Kappos, D. L. Arnold, K. Gücüyener, K. Deiva, N. Skripchenko, L.-Y. Cui, S. Saubadu, W. Hu *et al.*, "Safety and efficacy of teriflunomide in paediatric multiple sclerosis (terikids): a multicentre, double-blind, phase 3, randomised, placebo-controlled trial," *The Lancet Neurology*, vol. 20, no. 12, pp. 1001–1011, 2021.

[13] European Medicines Agency, "Aubagio (teriflunomide) - Public Assessment Report," European Medicines Agency, Tech. Rep., 2021, accessed on 2023-11-29. [Online]. Available: https://www.ema.europa.eu/documents/variation-report/aubagio-h-c-002514-x-0031-g-epar-assessment-report-variation_en.pdf

[14] ICH (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use), "ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials," 2019. [Online]. Available: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf

[15] P. Royston and 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 Medical Research Methodology*, vol. 13, no. 1, dec 2013.

[16] M. Schemper, S. Wakounig, and G. Heinze, "The estimation of average hazard ratios by weighted cox regression," *Statistics in medicine*, vol. 28, no. 19, pp. 2473–2489, 2009.

[17] D. Magirr and C. Burman, "Modestly weighted logrank tests," *Statistics in Medicine*, vol. 38, no. 20, pp. 3782–3790, May 2019.

[18] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

[19] R. Ristl, H. Götte, A. Schüler, M. Posch, and F. König, "Simultaneous inference procedures for the comparison of multiple characteristics of two survival functions," *Statistical Methods in Medical Research*, vol. 33, no. 4, pp. 589–610, 2024.

[20] T. R. Fleming and D. P. Harrington, *Counting processes and survival analysis*. John Wiley & Sons, 2013, vol. 625.

[21] S.-H. Lee, "On the versatility of the combination of the weighted log-rank statistics," *Computational statistics & data analysis*, vol. 51, no. 12, pp. 6557–6564, 2007.

[22] E. Letón and P. Zuluaga, "Equivalence between score and weighted tests for survival curves," *Communications in Statistics - Theory and Methods*, vol. 30, no. 4, pp. 591–608, Mar. 2001.

[23] F. Klinglmüller, T. Fellinger, F. König, T. Friede, A. C. Hooker, H. Heinzl, M. Mittlböck, J. Brugger, M. Bardo, C. Huber, N. Benda, M. Posch, and R. Ristl, "A neutral comparison of statistical methods for time-to-event analyses under non-proportional hazards," 2023.

[24] CONFIRMS, "Review of ema epars where nonproportional hazards were identified. deliverable 2," 2022.

[25] CONFIRMS, "Protocol: A simulation study to evaluate the performance characteristics of statistical methods for the analysis of time- to- event data under non-proportional hazard," 2022. [Online]. Available: https://www.encepp.eu/encepp/viewResource.htm?id=49976

[26] D. SCHOENFELD, "The asymptotic properties of nonparametric tests for comparing survival distributions," *Biometrika*, vol. 68, no. 1, pp. 316–319, 1981, sample size formula for the log-rank test.

[27] P. Guyot, A. Ades, M. J. Ouwens, and N. J. Welton, "Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves," *BMC medical research methodology*, vol. 12, no. 1, pp. 1–13, 2012.

[28] N. Liu and J. Lee, *IPDfromKM: Map Digitized Survival Curves Back to Individual Patient Data*, 2020, r package version 0.1.10. [Online]. Available: https://CRAN.R-project.org/package=IPDfromKM

[29] C. J. CLOPPER and E. S. PEARSON, "THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[30] T. Fellinger and F. Klinglmueller, *miniPCH: Survival Distributions with Piece-Wise Constant Hazards*, 2023, r package version 0.2.0. [Online]. Available: https://CRAN.R-project.org/package=miniPCH

[31] B. Ripley, *Stochastic Simulation*. John Wiley & Sons, Inc., 1987.

[32] J. R. Norris, *Markov chains*. Cambridge university press, 1998, no. 2.

[33] T. R. Fleming, D. P. Harrington, and P. C. O'Brien, "Designs for group sequential tests," *Controlled Clinical Trials*, vol. 5, no. 4, pp. 348–361, dec 1984.

[34] C. Casper, T. Cook, and O. A. P. B. on FORTRAN program ld98., *ldbounds: Lan-DeMets Method for Group Sequential Boundaries*, 2023, r package version 2.0.1. [Online]. Available: https://CRAN.R-project.org/package=ldbounds

[35] T. M. Therneau, *A Package for Survival Analysis in R*, 2022, r package version 3.3-1. [Online]. Available: https://CRAN.R-project.org/package=survival

[36] D. Magirr and I. Barrott, *nphRCT: Non-Proportional Hazards in Randomized Controlled Trials*, 2022, r package version 0.1.0. [Online]. Available: https://CRAN.R-project.org/package=nphRCT

[37] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks CA: Sage, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[38] G. Rücker and G. Schwarzer, "Presenting simulation results in a nested loop plot," *BMC Medical Research Methodology*, vol. 14, no. 1, dec 2014.

[39] W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges, *shiny: Web Application Framework for R*, 2023, r package version 1.7.4.1. [Online]. Available: https://CRAN.R-project.org/package=shiny

[40] T. Fellinger, "Shinylive app to present the simulated scenarios and simulation results of my masters thesis." [Online]. Available: https://github.com/TobiasFellinger/ShinyDiplomarbeit

[41] ——, "Visualisation of simulation results." [Online]. Available: https://tobiasfellinger.github.io/ShinyDiplomarbeit/

[42] D. Dunkler, M. Ploner, M. Schemper, and G. Heinze, "Weighted cox regression using the R package coxphw," *Journal of Statistical Software*, vol. 84, no. 2, pp. 1–26, 2018.

[43] Y. Wei, P. Royston, J. F. Tierney, and M. K. Parmar, "Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data," *Statistics in Medicine*, vol. 34, no. 21, pp. 2881–2898, 2015.

# Appendices

# A. Results, Additional Figures

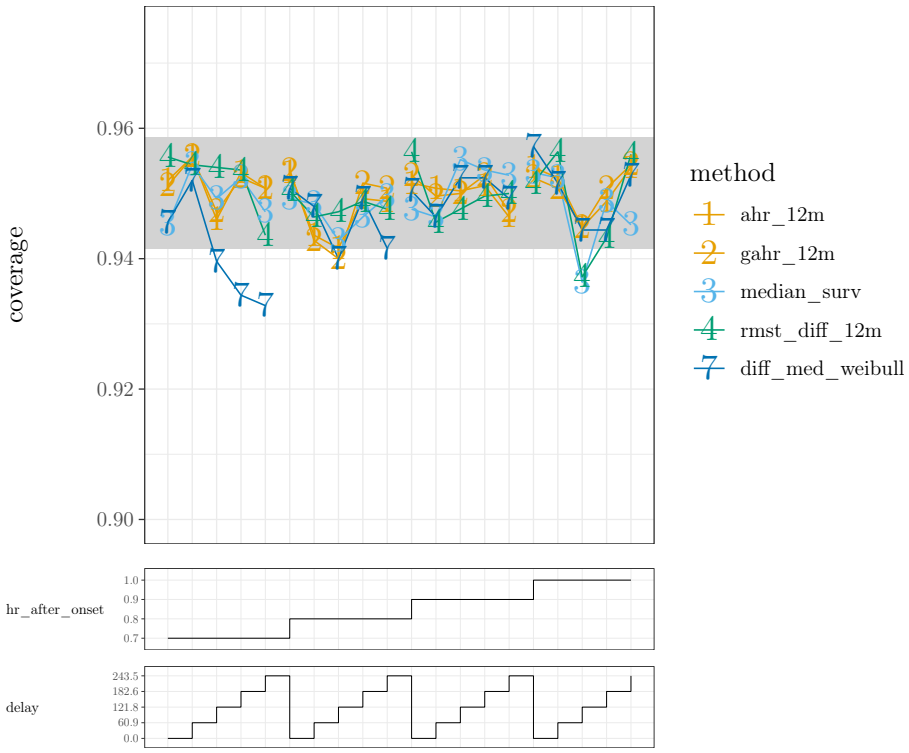



Figure A.1.: Coverage of the confidence intervals of the estimators where the true value of a summary statistic was available, scenarios with delayed onset of treatment effect. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months. The gray area indicated the range in which 95% of coverage proportions would fall if estimated with 2500 replications for methods that exactly maintain the 95% coverage probability.

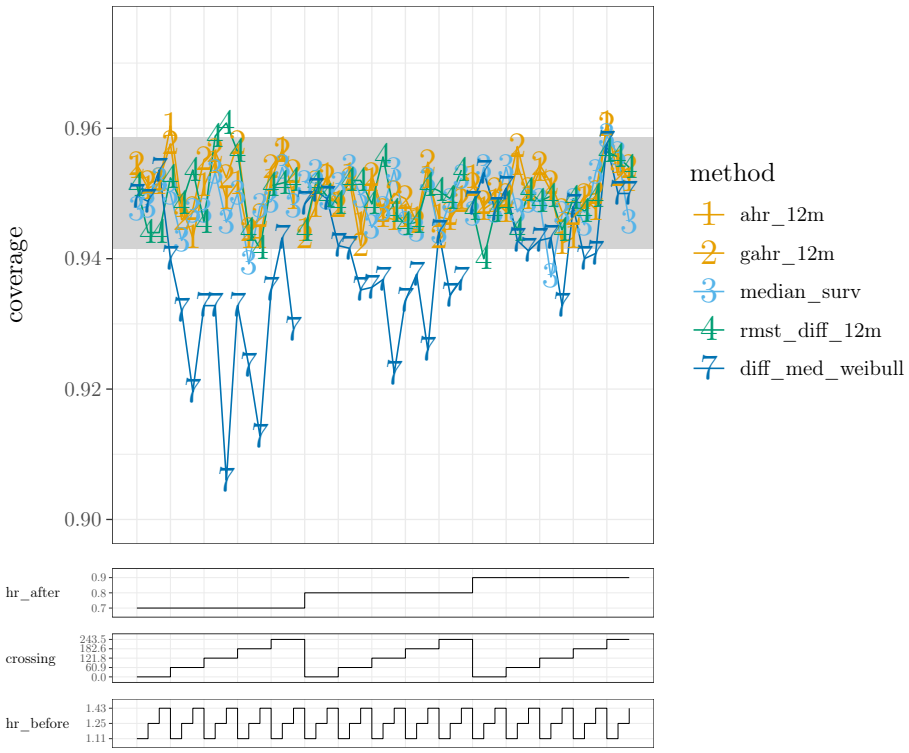



Figure A.2.: Coverage of the confidence intervals of the estimators where the true value of a summary statistic was available, scenarios with crossing hazard functions. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months. The gray area indicated the range in which 95% of coverage proportions would fall if estimated with 2500 replications for methods that exactly maintain the 95% coverage probability.

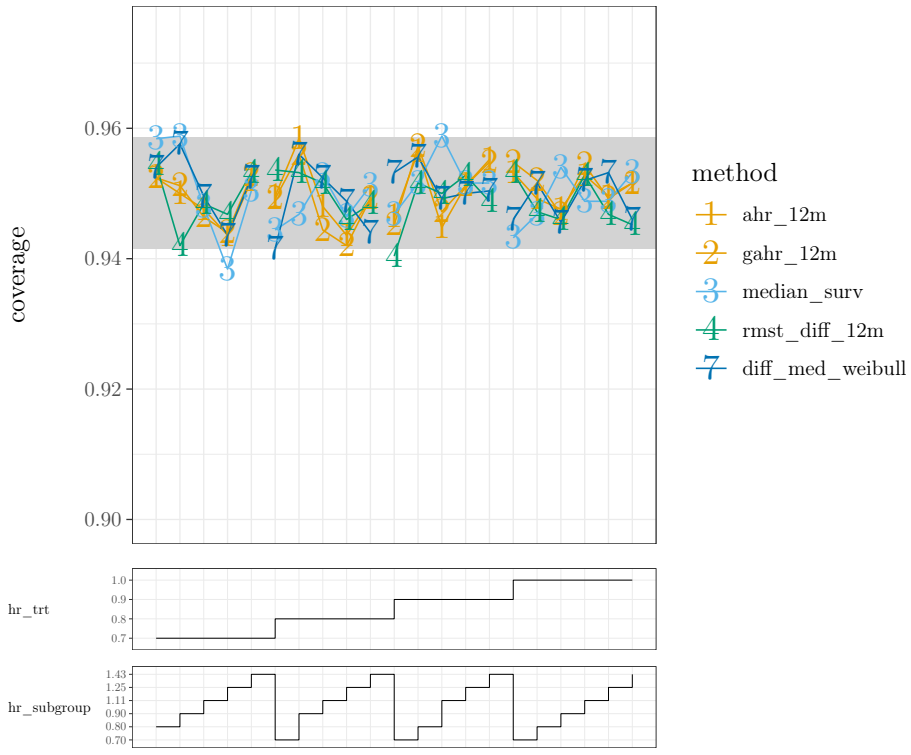



Figure A.3.: Coverage of the confidence intervals of the estimators where the true value of a summary statistic was available scenarios with subgroup effect. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months, prevalence of the subgroup 50%. The gray area indicated the range in which 95% of coverage proportions would fall if estimated with 2500 replications for methods that exactly maintain the 95% coverage probability.

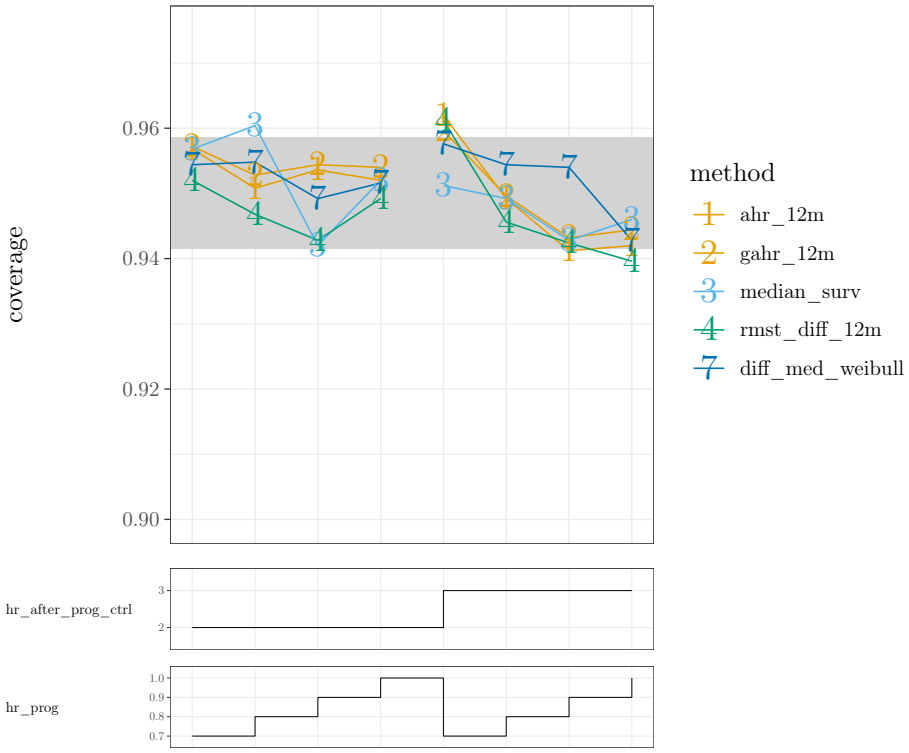



Figure A.4.: Coverage of the confidence intervals of the estimators where the true value of a summary statistic was available, scenarios with changing hazards after disease progression. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months and equal hazards for death before disease progression in both arms. The gray area indicated the range in which 95% of coverage proportions would fall if estimated with 2500 replications for methods that exactly maintain the 95% coverage probability.

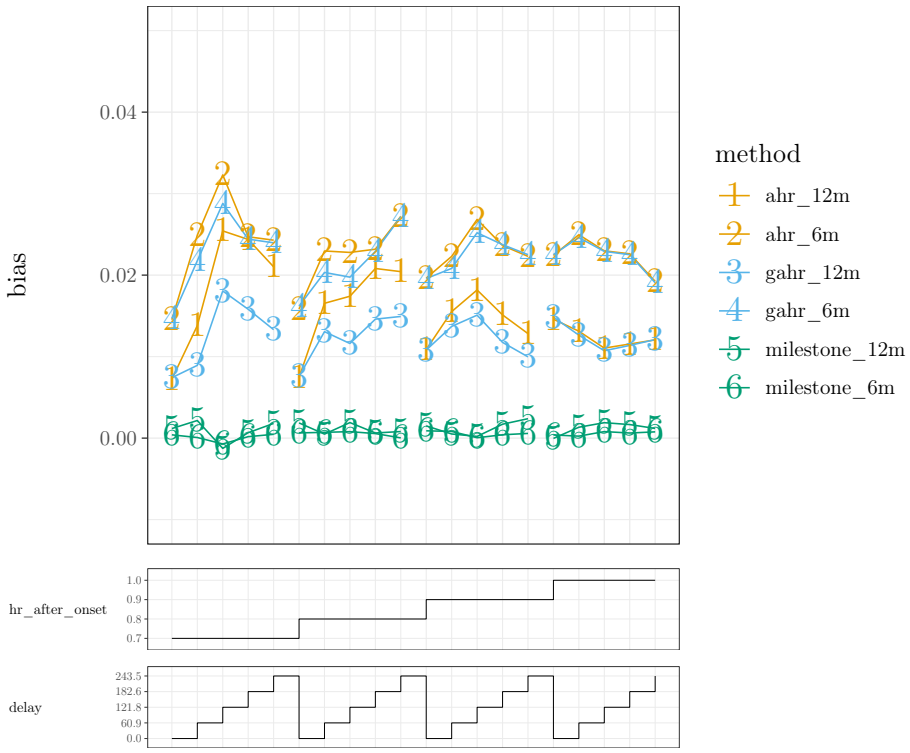



Figure A.5.: Bias of the estimators targeting summary statistics on the scales of ratios, scenarios with delayed onset of treatment effect. Average hazard ratio, geometric average hazard ratio and ratio of milestone survival. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months.

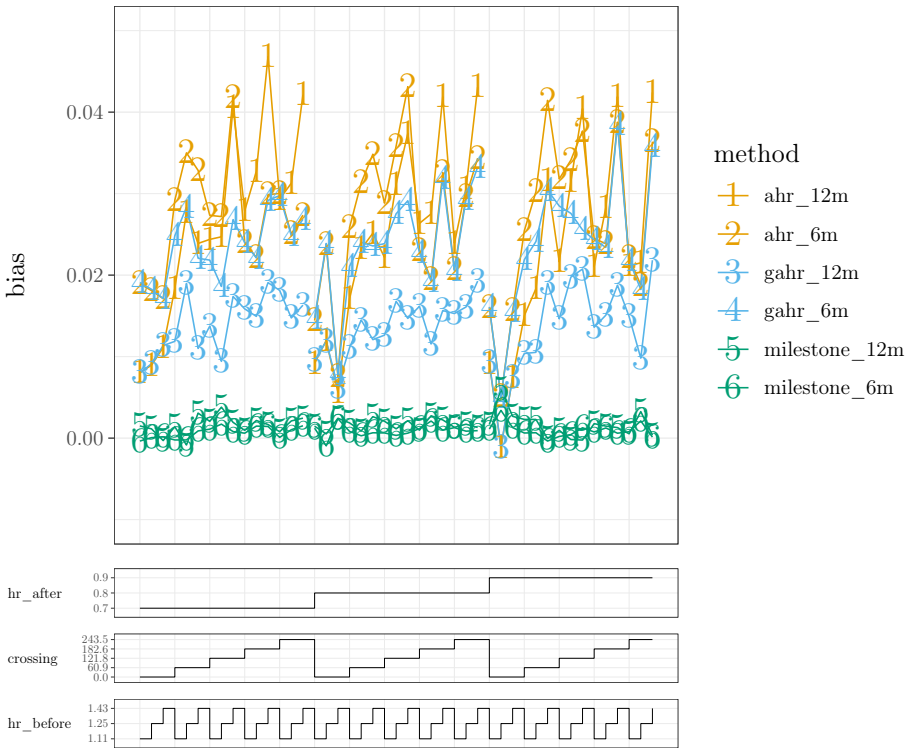



Figure A.6.: Bias of the estimators targeting summary statistics on the scales of ratios, scenarios with crossing hazard functions. Average hazard ratio, geometric average hazard ratio and ratio of milestone survival. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months.

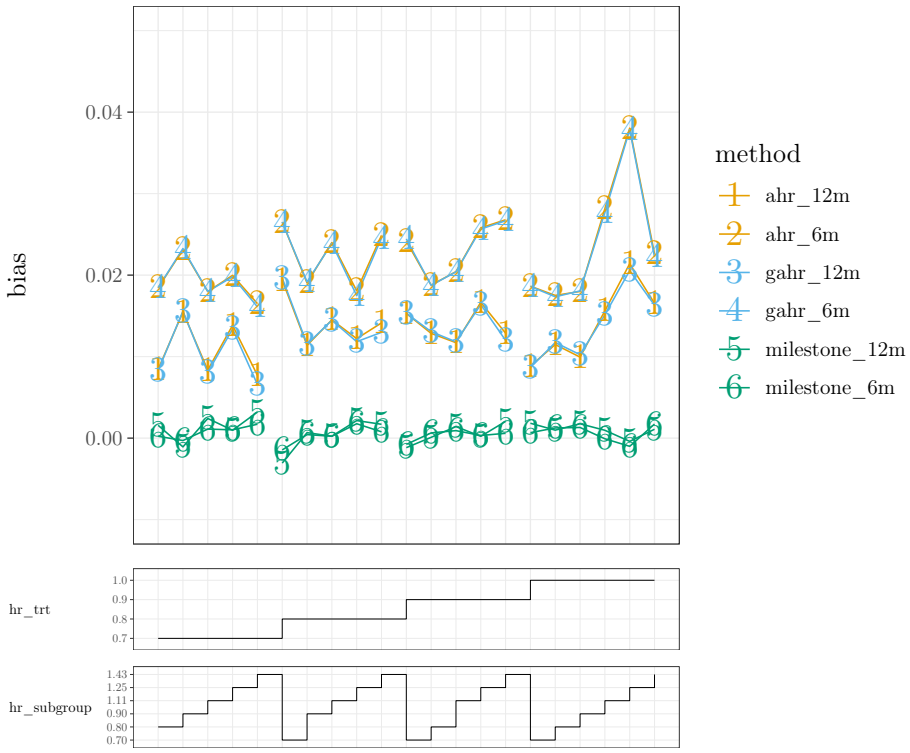



Figure A.7.: Bias of the estimators targeting summary statistics on the scales of ratios, scenarios with subgroup effect. Average hazard ratio, geometric average hazard ratio and ratio of milestone survival. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months, prevalence of the subgroup 50%.

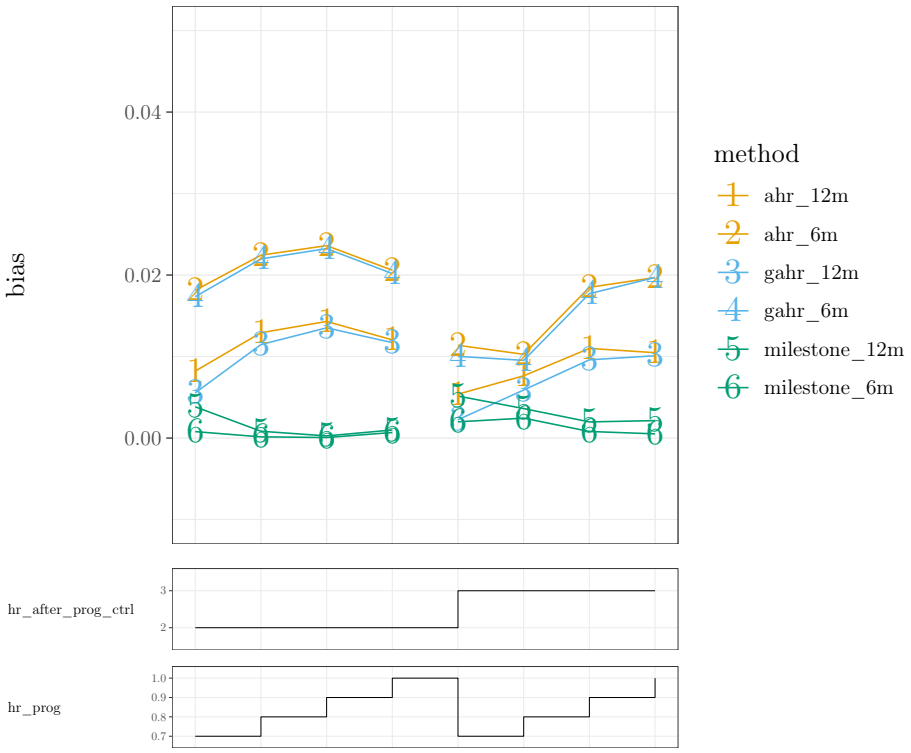



Figure A.8.: Bias of the estimators targeting summary statistics on the scales of ratios, scenarios with changing hazards after disease progression. Average hazard ratio, geometric average hazard ratio and ratio of milestone survival. Scenarios were restricted to the following parameter values: number of patients: 600, recruitment: 6 months, median time to random withdrawal: 120 months, median survival in the control arm: 24 months and equal hazards for death before disease progression in both arms.

# B. Scripts used for Simulation Study

Scripts used to execute the simulation study:

- parameters.R: set the simulation parameters for all scenarios
- run_simulations_common.R: define the analysis methods and the performance measures, reused in the following simulation scripts
- run_simulations_delayed_effect.R: run simulations for scenarios with delayed effect
- run_simulations_crossing_hazards.R: run simulations for scenarios with crossing hazards
- run_simulations_subgroup.R: run simulations for scenarios with subgroup effect
- run_simulations_disease_progression.R: run simulations for scenarios with disease progression

Listing B.1: parameters.R

```
1
2  library(SimNPH)
3
4  # assumptions for all scenarios
       ------------------------------------------
5
6  assumptions_all <- expand.grid(
7    hazard_ctrl       = m2r(c(12, 24, 36)),
8    random_withdrawal = c(0, m2r(120))
9  )
10
11 design_all <- expand.grid(
12   n_trt       = c(150, 300, 500),
13   recruitment = m2d(c(0,6))
14 ) |>
15   within({
16     n_ctrl = n_trt
17     n_pat  = n_trt + n_ctrl
18     interim_events = ceiling(n_pat * 0.75 * 0.5)
19     final_events   = ceiling(n_pat * 0.75)
20   })
21
22 parameters_all <- merge(
23   assumptions_all,
```

```
  design_all,
  by=NULL
)

hr_s <- c(0.7, 0.8, 0.9)
cutoffs <- m2d(c("6m"=6, "12m"=12))
milestones <- m2d(c("6m"=6, "12m"=12))

# delayed effect
    ------------------------------------------------------------

parameters_delayed_effect <- expand.grid(
  delay          = m2d(seq(0, 8, by=2)),
  hr_after_onset = c(hr_s, 1)
) |>
  merge(parameters_all, by=NULL) |>
  within({
    hazard_trt = hazard_ctrl * hr_after_onset
  }) |>
  true_summary_statistics_delayed_effect(
    cutoff_stats = cutoffs,
    milestones = milestones
  )

# crossing hazards
    ------------------------------------------------------------

parameters_crossing_hazards <- expand.grid(
  crossing  = m2d(seq(0, 8, by=2)),
  hr_before = 1/hr_s,
  hr_after  = hr_s
)|>
  merge(parameters_all, by=NULL) |>
  within({
    hazard_trt_before = hazard_ctrl * hr_before
    hazard_trt_after  = hazard_ctrl * hr_after
  }) |>
  true_summary_statistics_crossing_hazards(
    cutoff_stats = cutoffs,
    milestones = milestones
  )

# disease progression
    ------------------------------------------------------

parameters_progression <- expand.grid(
  hr_death_before_prog = c(1, hr_s),
  hr_after_prog_ctrl   = c(2, 3),
  prog_rate_ctrl = m2r(12),
  hr_prog        = c(1, hr_s)
) |>
  merge(parameters_all, by=NULL) |>
  within({
```

```
    hazard_trt = hazard_ctrl * hr_death_before_prog
    hazard_after_prog = hazard_ctrl * hr_after_prog_ctrl
    prog_rate_trt = prog_rate_ctrl * hr_prog
  }) |>
  true_summary_statistics_progression(
    cutoff_stats = cutoffs,
    milestones = milestones
  )

# subset
  ------------------------------------------------------------------

parameters_subset <- expand.grid(
  hr_trt       = c(hr_s, 1),
  hr_subgroup  = c(hr_s, 1/hr_s),
  prevalence   = seq(0.1, 0.5, by=0.2)
) |>
  subset(hr_trt != hr_subgroup) |>
  merge(parameters_all, by=NULL) |>
  within({
    hazard_trt = hazard_ctrl * hr_trt
    hazard_subgroup = hazard_ctrl * hr_subgroup
  }) |>
  true_summary_statistics_subgroup(
    cutoff_stats = cutoffs,
    milestones = milestones
  )

# save
  ------------------------------------------------------------------

session_info <- sessionInfo()
sys_info <- Sys.info()
today <- Sys.time()

save(
  parameters_delayed_effect,
  parameters_crossing_hazards,
  parameters_progression,
  parameters_subset,
  session_info,
  sys_info,
  today,
  file="parameters_simulations.Rdata"
  )
```

Listing B.2: run_simulations_common.R

```
# define analyse functions
  ------------------------------------------------

my_analyse <- list(
  # estimation
  ahr_6m   = analyse_ahr(type="AHR", max_time=m2d(6), alternative = "one.
```

```
    sided"),
  ahr_12m  = analyse_ahr(type="AHR", max_time=m2d(12), alternative = "one
    .sided"),
  gahr_6m  = analyse_ahr(type="gAHR", max_time=m2d(6), alternative = "one
    .sided"),
  gahr_12m = analyse_ahr(type="gAHR", max_time=m2d(12), alternative = "
    one.sided"),
  median_surv = analyse_diff_median_survival(alternative = "one.sided"),
  milestone_6m = analyse_milestone_survival(times=m2d(6), alternative = "
    one.sided"),
  milestone_12m = analyse_milestone_survival(times=m2d(12), alternative =
     "one.sided"),
  rmst_diff_6m  = analyse_rmst_diff(max_time = m2d(6), alternative = "one
    .sided"),
  rmst_diff_12m = analyse_rmst_diff(max_time = m2d(12), alternative = "
    one.sided"),
  cox = analyse_coxph(alternative = "one.sided"),
  aft_weibull = analyse_aft(dist="weibull", alternative = "one.sided"),
  aft_lognormal = analyse_aft(dist="lognormal", alternative = "one.sided"
    ),
  diff_med_weibull = analyse_weibull(alternative = "one.sided"),
  # tests
  fh_0_0 = analyse_logrank_fh_weights(rho=0, gamma=0, alternative = "one.
    sided"),
  fh_0_1 = analyse_logrank_fh_weights(rho=0, gamma=1, alternative = "one.
    sided"),
  fh_1_0 = analyse_logrank_fh_weights(rho=1, gamma=0, alternative = "one.
    sided"),
  fh_1_1 = analyse_logrank_fh_weights(rho=1, gamma=1, alternative = "one.
    sided"),
  logrank = analyse_logrank(alternative = "one.sided"),
  max_combo = analyse_maxcombo(alternative = "one.sided"),
  modest_6 = analyse_modelstly_weighted(t_star=m2d(6)),
  modest_8 = analyse_modelstly_weighted(t_star=m2d(8)),
  # tests group sequential
  fh_gs_0_0 = analyse_group_sequential(
    followup = c(condition$interim_events, condition$final_events),
    followup_type = c("event", "event"),
    alpha = nominal_alpha,
    analyse_functions = analyse_logrank_fh_weights(rho=0, gamma=0,
    alternative = "one.sided")
  ),
  fh_gs_0_1 = analyse_group_sequential(
    followup = c(condition$interim_events, condition$final_events),
    followup_type = c("event", "event"),
    alpha = nominal_alpha,
    analyse_functions = analyse_logrank_fh_weights(rho=0, gamma=1,
    alternative = "one.sided")
  ),
  fh_gs_1_0 = analyse_group_sequential(
    followup = c(condition$interim_events, condition$final_events),
    followup_type = c("event", "event"),
    alpha = nominal_alpha,
```

```
    analyse_functions = analyse_logrank_fh_weights(rho=1, gamma=0,
    alternative = "one.sided")
  ),
  fh_gs_1_1 = analyse_group_sequential(
    followup = c(condition$interim_events, condition$final_events),
    followup_type = c("event", "event"),
    alpha = nominal_alpha,
    analyse_functions = analyse_logrank_fh_weights(rho=1, gamma=1,
    alternative = "one.sided")
  ),
  logrank_gs = analyse_group_sequential(
    followup = c(condition$interim_events, condition$final_events),
    followup_type = c("event", "event"),
    alpha = nominal_alpha,
    analyse_functions = analyse_logrank(alternative = "one.sided")
  ),
  max_combo_gs = analyse_group_sequential(
    followup = c(condition$interim_events, condition$final_events),
    followup_type = c("event", "event"),
    alpha = nominal_alpha,
    analyse_functions = analyse_maxcombo(alternative = "one.sided")
  ),
  modest_gs_6 = analyse_group_sequential(
    followup = c(condition$interim_events, condition$final_events),
    followup_type = c("event", "event"),
    alpha = nominal_alpha,
    analyse_functions = analyse_modelstly_weighted(t_star=m2d(6))
  ),
  modest_gs_8 = analyse_group_sequential(
    followup = c(condition$interim_events, condition$final_events),
    followup_type = c("event", "event"),
    alpha = nominal_alpha,
    analyse_functions = analyse_modelstly_weighted(t_star=m2d(8))
  ),
  descriptive = analyse_describe()
)

my_analyse <- wrap_all_in_trycatch(my_analyse)

# define summaries
    ------------------------------------------------------------

my_summarise <- create_summarise_function(
  # estimation
  ahr_6m  = summarise_estimator(est=AHR, real=AHR_6m, lower=AHR_lower,
    upper=AHR_upper, null=1),
  ahr_12m = summarise_estimator(est=AHR, real=AHR_12m, lower=AHR_lower,
    upper=AHR_upper, null=1),
  gahr_6m  = summarise_estimator(est=gAHR, real=gAHR_6m, lower=gAHR_lower
    , upper=gAHR_upper, null=1),
  gahr_12m = summarise_estimator(est=gAHR, real=gAHR_12m, lower=gAHR_
    lower, upper=gAHR_upper, null=1),
  median_surv = summarise_estimator(est=diff_Q, real=median_survival_trt-
```

```
    median_survival_ctrl, lower=diff_Q_lower, upper=diff_Q_upper, null=0),
  milestone_6m = summarise_estimator(est=milestone_surv_ratio, real=
    milestone_survival_ctrl_6m/ milestone_survival_trt_6m, lower=milestone
    _surv_ratio_lower, upper=milestone_surv_ratio_upper, null=1),
  milestone_12m = summarise_estimator(est=milestone_surv_ratio, real=
    milestone_survival_ctrl_12m/milestone_survival_trt_12m, lower=
    milestone_surv_ratio_lower, upper=milestone_surv_ratio_upper, null=1),
  rmst_diff_6m  = summarise_estimator(est=rmst_diff, real=rmst_trt_6m -
    rmst_ctrl_6m, lower=rmst_diff_lower, upper=rmst_diff_upper, null=0),
  rmst_diff_12m = summarise_estimator(est=rmst_diff, real=rmst_trt_12m-
    rmst_ctrl_12m, lower=rmst_diff_lower, upper=rmst_diff_upper, null=0),
  cox = summarise_estimator(est=hr, real=gAHR_12m, lower=hr_lower, upper=
    hr_upper, null=1),
  aft_weibull = summarise_estimator(est=coef, real=NA_real_, lower=lower,
     upper=upper, null=0),
  aft_lognormal = summarise_estimator(est=coef, real=NA_real_, lower=
    lower, upper=upper, null=0),
  diff_med_weibull = summarise_estimator(est=diff_med_est, real=median_
    survival_trt-median_survival_ctrl,lower=diff_med_lower, upper=diff_med
    _upper, null=0),
  # tests
  fh_0_0 = summarise_test(alpha),
  fh_0_1 = summarise_test(alpha),
  fh_1_0 = summarise_test(alpha),
  fh_1_1 = summarise_test(alpha),
  logrank = summarise_test(alpha),
  max_combo = summarise_test(alpha),
  modest_6 = summarise_test(alpha),
  modest_8 = summarise_test(alpha),
  # tests group sequential
  fh_gs_0_0 = summarise_group_sequential(),
  fh_gs_0_1 = summarise_group_sequential(),
  fh_gs_1_0 = summarise_group_sequential(),
  fh_gs_1_1 = summarise_group_sequential(),
  logrank_gs = summarise_group_sequential(),
  max_combo_gs = summarise_group_sequential(),
  modest_gs_6 = summarise_group_sequential(),
  modest_gs_8 = summarise_group_sequential(),
  descriptive = summarise_describe()
)
```

Listing B.3: run_simulations_delayed_effect.R

```

library(SimNPH)
library(SimDesign)
library(parallel)

if(packageVersion("SimNPH") != "0.5.3"){
  stop("Please run the simulations with the correct vesion of the SimNPH
    package for reproducability.")
}

N_sim <- 2500
```

```
11
12 # setup cluster
    ----------------------------------------------------------
13
14 n_cores <- parallel::detectCores()
15 cl <- makeCluster(n_cores-1)
16
17 clusterEvalQ(cl, {
18   library(SimDesign)
19   library(SimNPH)
20   library(parallel)
21 })
22
23 # load parameters
    ----------------------------------------------------------
24
25 load("parameters_simulations.Rds")
26
27 # setup data generation
    ----------------------------------------------------
28
29 # load parameters
30 design <- parameters_delayed_effect
31
32 # define generator
33 my_generator <- function(condition, fixed_objects=NULL){
34   generate_delayed_effect(condition, fixed_objects) |>
35     recruitment_uniform(condition$recruitment) |>
36     random_censoring_exp(condition$random_withdrawal) |>
37     admin_censoring_events(condition$final_events)
38 }
39
40 alpha <- 0.025
41 nominal_alpha <- ldbounds::ldBounds(c(0.5,1), sides=1, alpha = 0.025)$nom
    .alpha
42
43 clusterExport(cl, "nominal_alpha")
44
45
46 # define analysis and summarise functions
    ---------------------------------
47 # this is the same for all scenarios
48
49 source("scripts/run_simulations_common.R")
50
51 # run
    ------------------------------------------------------------------------
52
53 save_folder <- paste0(paste0("data/simulation_delayed_effect_", Sys.info
    ()["nodename"], "_", strftime(Sys.time(), "%Y-%m-%d_%H%M%S")))
54 dir.create(save_folder, recursive = TRUE)
55
56 results <- runSimulation(
```

```
  design,
  replications = N_sim,
  generate = my_generator,
  analyse = my_analyse,
  summarise = my_summarise,
  cl = cl,
  parallel = TRUE,
  save_details = list(
    out_rootdir = save_folder
  ),
  extra_options = list(
    store_warning_seeds = TRUE,
    allow_na = TRUE
  )
)

saveRDS(results, paste0(save_folder, "/results.Rds"))
```

Listing B.4: run_simulations_crossing_hazards.R

```

library(SimNPH)
library(SimDesign)
library(parallel)

if(packageVersion("SimNPH") != "0.5.3"){
  stop("Please run the simulations with the correct vesion of the SimNPH
    package for reproducability.")
}

N_sim <- 2500

# setup cluster
    ------------------------------------------------------------

n_cores <- parallel::detectCores()
cl <- makeCluster(n_cores-1)

clusterEvalQ(cl, {
  library(SimDesign)
  library(SimNPH)
  library(parallel)
})

# load parameters
    ------------------------------------------------------------

load("parameters_simulations.Rds")

# setup data generation
    ----------------------------------------------------

# load parameters
design <- parameters_crossing_hazards
```

```r

# define generator
my_generator <- function(condition, fixed_objects=NULL){
  generate_crossing_hazards(condition, fixed_objects) |>
    recruitment_uniform(condition$recruitment) |>
    random_censoring_exp(condition$random_withdrawal) |>
    admin_censoring_events(condition$final_events)
}

alpha <- 0.025
nominal_alpha <- ldbounds::ldBounds(c(0.5,1), sides=1, alpha = 0.025)$nom
    .alpha

clusterExport(cl, "nominal_alpha")


# define analysis and summarise functions
    ---------------------------------
# this is the same for all scenarios

source("scripts/run_simulations_common.R")

# run
    ------------------------------------------------------------------------

save_folder <- paste0(paste0("data/simulation_crossing_harzards_", Sys.
    info()["nodename"], "_", strftime(Sys.time(), "%Y-%m-%d_%H%M%S")))
dir.create(save_folder, recursive = TRUE)

results <- runSimulation(
  design,
  replications = N_sim,
  generate = my_generator,
  analyse = my_analyse,
  summarise = my_summarise,
  cl = cl,
  parallel = TRUE,
  save_details = list(
    out_rootdir = save_folder
  ),
  extra_options = list(
    store_warning_seeds = TRUE,
    allow_na = TRUE
  )
)

saveRDS(results, paste0(save_folder, "/results.Rds"))
```

Listing B.5: run_simulations_subgroup.R

```r

library(SimNPH)
library(SimDesign)
library(parallel)
```

```
5
6  if(packageVersion("SimNPH") != "0.5.3"){
7    stop("Please run the simulations with the correct vesion of the SimNPH
       package for reproducability.")
8  }
9
10 N_sim <- 2500
11
12 # setup cluster
       ------------------------------------------------------------
13
14 n_cores <- parallel::detectCores()
15 cl <- makeCluster(n_cores-1)
16
17 clusterEvalQ(cl, {
18   library(SimDesign)
19   library(SimNPH)
20   library(parallel)
21 })
22
23 # load parameters
       ------------------------------------------------------------
24
25 load("parameters_simulations.Rds")
26
27 # setup data generation
       ------------------------------------------------
28
29 # load parameters
30 design <- parameters_subset
31
32 # define generator
33 my_generator <- function(condition, fixed_objects=NULL){
34   generate_subgroup(condition, fixed_objects) |>
35     recruitment_uniform(condition$recruitment) |>
36     random_censoring_exp(condition$random_withdrawal) |>
37     admin_censoring_events(condition$final_events)
38 }
39
40 alpha <- 0.025
41 nominal_alpha <- ldbounds::ldBounds(c(0.5,1), sides=1, alpha = 0.025)$nom
       .alpha
42
43 clusterExport(cl, "nominal_alpha")
44
45
46 # define analysis and summarise functions
       ---------------------------------
47 # this is the same for all scenarios
48
49 source("run_simulations_common.R")
50
51 # run
```

```
    ------------------------------------------------------------------

save_folder <- paste0(paste0("data/simulation_subgroup_", Sys.info()["
    nodename"], "_", strftime(Sys.time(), "%Y-%m-%d_%H%M%S")))
dir.create(save_folder, recursive = TRUE)

results <- runSimulation(
  design,
  replications = N_sim,
  generate = my_generator,
  analyse = my_analyse,
  summarise = my_summarise,
  cl = cl,
  parallel = TRUE,
  save_details = list(
    out_rootdir = save_folder
  ),
  extra_options = list(
    store_warning_seeds = TRUE,
    allow_na = TRUE
  )
)

saveRDS(results, paste0(save_folder, "/results.Rds"))
```

Listing B.6: run_simulations_disease_progression.R

```

library(SimNPH)
library(SimDesign)
library(parallel)

if(packageVersion("SimNPH") != "0.5.3"){
  stop("Please run the simulations with the correct vesion of the SimNPH
    package for reproducability.")
}

N_sim <- 2500

# setup cluster
    ------------------------------------------------------------

n_cores <- parallel::detectCores()
cl <- makeCluster(n_cores-1)

clusterEvalQ(cl, {
  library(SimDesign)
  library(SimNPH)
  library(parallel)
})

# load parameters
    ------------------------------------------------------------

```

```r
load("parameters_simulations.Rds")

# setup data generation
    ------------------------------------------------

# load parameters
design <- parameters_progression

# define generator
my_generator <- function(condition, fixed_objects=NULL){
  generate_progression(condition, fixed_objects) |>
    recruitment_uniform(condition$recruitment) |>
    random_censoring_exp(condition$random_withdrawal) |>
    admin_censoring_events(condition$final_events)
}

alpha <- 0.025
nominal_alpha <- ldbounds::ldBounds(c(0.5,1), sides=1, alpha = 0.025)$nom
    .alpha

clusterExport(cl, "nominal_alpha")


# define analysis and summarise functions
    ---------------------------------
# this is the same for all scenarios

source("scripts/run_simulations_common.R")

# run
    ------------------------------------------------------------------------

save_folder <- paste0(paste0("data/simulation_disease_progression_", Sys.
    info()["nodename"], "_", strftime(Sys.time(), "%Y-%m-%d_%H%M%S")))
dir.create(save_folder, recursive = TRUE)

results <- runSimulation(
  design,
  replications = N_sim,
  generate = my_generator,
  analyse = my_analyse,
  summarise = my_summarise,
  cl = cl,
  parallel = TRUE,
  save_details = list(
    out_rootdir = save_folder
  ),
  extra_options = list(
    store_warning_seeds = TRUE,
    allow_na = TRUE
  )
)

```

```
73 saveRDS(results, paste0(save_folder, "/results.Rds"))
```

# C. Screenshots of the Shiny App

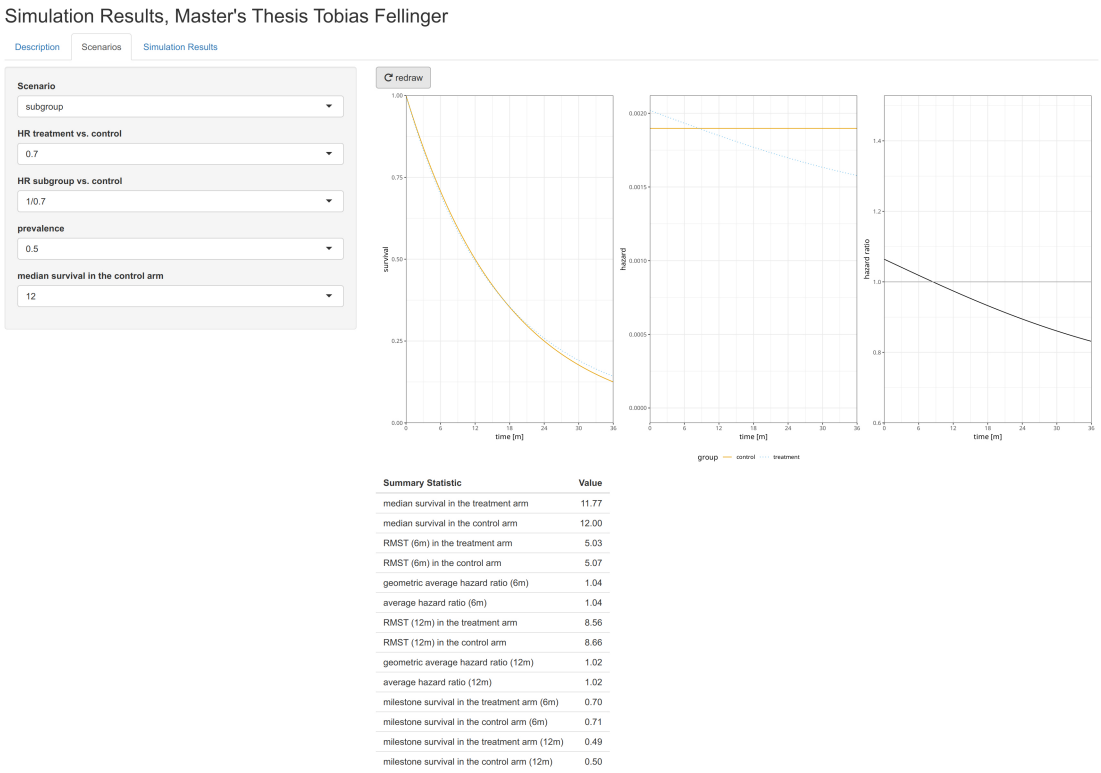



Figure C.1.: Screenshot of the scenarios section of the shiny app. The user has selected a scenario with subgroup effect with the parameters HR treatment vs. control: 0.7, HR subgroup vs. control: 1/0.7, prevalence 0.5 and median survival in the control arm 12. Plots for the survival function, hazard function and hazard rate are shown. Values of summary statistics are listed in a table.

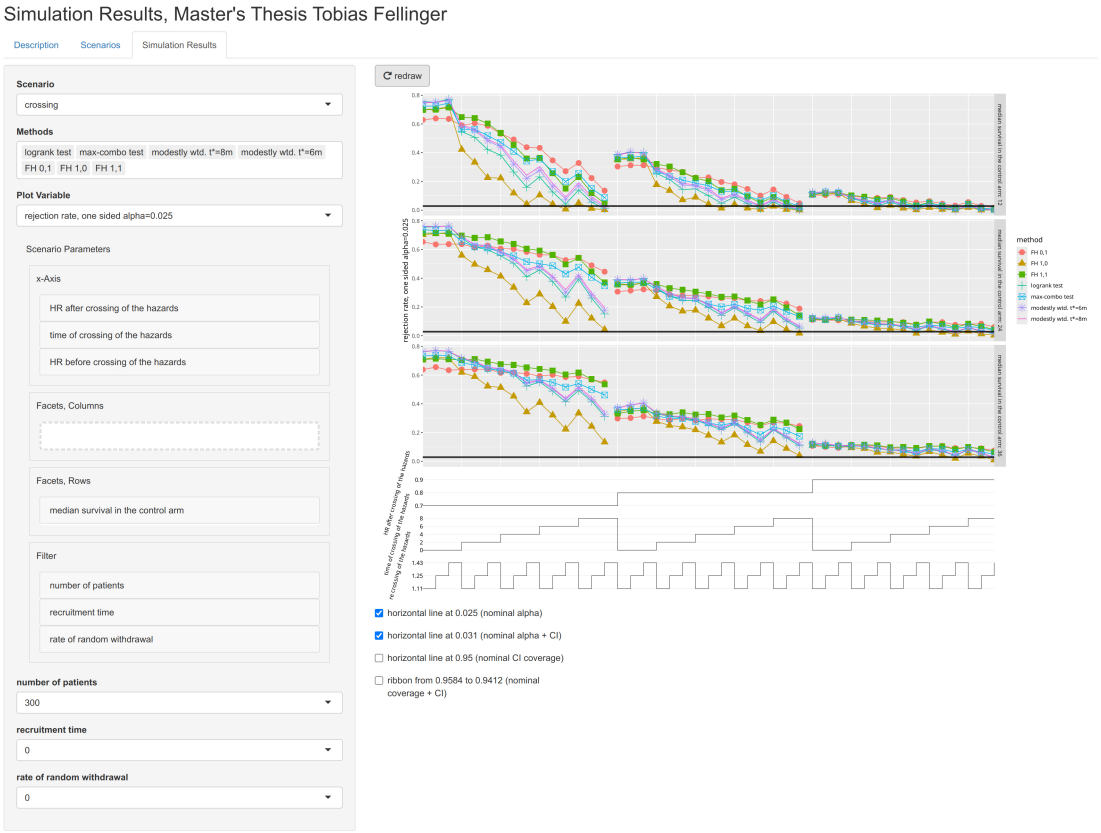



Figure C.2.: Screenshot of the results section of the shiny app. The user has selected a crossing hazards scenario. A number of Methods is selected, the selected performance measure is the rejection rate with a one sided alpha of 0.025. Parameters on the x-Axis are HR after crossing of the hazards, time of crossing of the hazards, HR before crossing of the hazards and HR before crossing of the hazards. Facets, Rows contains the variable median survival in the control arm. The data are filtered by number of patients, recruitment time and rate of random withdrawal. Values for the filter variables are selected with dropdown menus. A plot with 3 vertical facets is shown on the right, below are the levels of the x-axis variables at the corresponding x-value. Checkboxes allow to add lines and areas to the plot.