Aus dem Institut für Medizinische Biometrie
Universitätsklinikum Heidelberg
Geschäftsführender Direktor: Prof. Dr. sc. hum. Meinhard Kieser

Comparison of methods to analyze time-to-event endpoints when treatment effect is delayed

Inauguraldissertation
zur Erlangung des
„Doctor scientiarum humanarum“

an der Medizinischen Fakultät Heidelberg
der Ruprecht-Karls-Universität

vorgelegt von
Rouven Behnisch
aus
Herrenberg

2024

Dekan: Prof. Dr. Michael Boutros

Doktorvater: Prof. Dr. sc. hum. Meinhard Kieser

Contents

Abbreviations and Symbols

$\mathbb{1}(\cdot)$	indicator function
Aalen	Aalen additive hazard model test
ABC	area between curves
acc	accrual proportion
ADEMP	aim, data generating mechanism, estimand, method and performance measure
AFT	accelerated failure time
AHR	average hazard ratio
AOC	area over the curve
asymLR	asymptotic logrank test
AUC	area under the curve
$B(\cdot)$	Brownian motion
Beta(p, q)	Beta distribution with shape parameters p, q
BEP	Bayesian expected power
CauchyCP	Cauchy changepoint model
CheckPH	checking PH approach
combTest	combined test
Cox	Cox model
CoxTD	Cox model with time-dependent treatment effect
DOT	distance from origin
Exp(λ)	exponential distribution with rate λ
GB	Gehan-Breslow test
GenLin	generalized linear lag test

GLLM	generalized linear lag model
$G(\rho, \gamma)$	Fleming-Harrington family of weighted logrank tests with parameters $\rho, \gamma > 0$
$G(-1, 0)$	Gray-Tsiatis test
H_0	null hypothesis
H_1	alternative hypothesis
HR	hazard ratio
IMBI	institute of medical biometry
jointTest	joint test
k_C	Weibull shape parameter for control arm
KS Cheng	Kolmogorov-Smirnov type tests based on weighted logrank test by Cheng
KS FH	Kolmogorov-Smirnov type tests based on Fleming-Harrington test
KS GB	Kolmogorov-Smirnov type tests based on Gehan-Breslow test
KS LR	Kolmogorov-Smirnov type tests based on logrank test
Landmark	landmark survival model
(m)Lee1/2/3	(modified) versatile tests by Lee
LLRNA	linear combination of logrank and milestone survival based on Nelson-Aalen estimator
$\text{lag}_1, \text{lag}_2$	changepoint and delay proportion
$\lambda(\cdot)$	hazard function
$\Lambda(\cdot)$	cumulative hazard function
LR	logrank test
(m)Logit	(modified) logit model
(m)MaxCombo	(modified) MaxCombo test
MaxRMST	maximum restricted mean survival time
med_C	median survival in control arm
MERT	maximin efficiency robust test
Mile	milestone survival
MileCLL	milestone survival based on complementary log-log transformation
MileNA	milestone survival based on the Nelson-Aalen estimator
MLE	maximum likelihood estimator
mScore	modified score test
MWLR	modestly weighted logrank test

NA	Nelson-Aalen estimator
NPH	non-proportional hazards
NSCLC	non-small cell lung cancer
OR	odds ratio
OS	overall survival
ParGroup	partially grouped logrank test
PFS	progression-free survival
PH	proportional hazards
(m)PP	(modified) Peto-Peto test
ProjTest	projection test
PWExp(Lag)	piecewise exponential (lag) model
QLRNA	quadratic combination of logrank and milestone survival based on Nelson-Aalen estimator
RAT	restrained adaptive test
RCT	randomized clinical trial
RMST	restricted mean survival time
RP.PH	Royston-Parmar model
RP.TD	Royston-Parmar model with time-dependent treatment effect
$S(\cdot)$	survival function
$\hat{S}_{KM}(\cdot)$	Kaplan-Meier estimator of survival function
SCLC	small cell lung cancer
t_1^*, t_2^*	changepoint and delay in generalized linear lag model
τ	overall study duration
θ	maximum treatment effect in generalized linear lag model
Thres	threshold lag test
TLM	threshold lag model
TW	Tarone-Ware test
V0	V0 test
Wei(λ, k)	Weibull distribution with shape k and scale λ
WKM	weighted Kaplan-Meier test
YP	Yang and Prentice model
(m)Zm3	(modified) Zm3 test

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

In recent years the role of immunotherapy with its aim to harness and augment the immune system gained more and more importance in cancer drug development. As outlined by Zhang and Zhang (2020) these oncological immunotherapies can be divided into the following five categories with description of their mechanism of action:

1. **Oncolytic virus therapies:** Possibly genetically modified viruses are used to target and attack tumor cells and stimulate antitumor immune response.

2. **Cancer vaccines:** Tumor-specific markers (antigens) are used as vaccines.

3. **Cytokine therapies:** Cytokines are messenger molecules that regulate communication of the immune system and are mostly used in combination with other immunotherapies due to the poor tolerability as monotherapy.

4. **Adoptive cell transfer:** Own immune cells are isolated, expanded in number and reinfused to eliminate cancer cells.

5. **Immune checkpoint inhibitors:** Immune checkpoints that are frequently manipulated by tumors to inhibit tumor response are blocked so that immune-mediated elimination of cancer cells can be promoted.

Although these therapies have shown to be tremendously successful in clinical trials, their unique mechanism of action is posing challenges to the statistical analysis of randomized clinical trial (RCT) data. This is especially the case in oncological trials where the outcome of interest is very often a time-to-event endpoint. The standard analysis of such endpoints - the logrank test and Cox model - relies on the proportional hazards (PH) assumption, which means that the hazard, i.e. the instantaneous rate at which an event of interest occurs, is proportional between the treatment arms. As the focus throughout this thesis lies on methods to analyze two-arm trials, this assumption implies that the instantaneous event rate in the experimental arm is accelerated or in case of negative events ideally decelerated by a time-independent constant factor compared to the instantaneous rate in the control arm.

For many applications the Cox model has been of great use and the PH assumption can reasonably be considered to hold true, but this is unfortunately not the case when investigating immunotherapeutic agents. One commonly observed feature of these treatments is a delayed onset of treatment effect, as the immune system needs time to respond to this therapy.

This feature has, for example, been observed in the CHECKMATE trials, a number of 21 trials on non-small-cell lung cancer (NSCLC) and 3 trials focusing on small-cell lung cancer (SCLC) evaluating the immune-checkpoint inhibitor Nivolumab developed by Bristol-Myers Squibb. Of these CHECKMATE trials, two trials will be used as an illustrative example in this thesis and presented in the following.

CHECKMATE067 A multicentre, randomised, controlled, double-blind phase 3 trial where 945 patients have been randomly assigned 1:1:1 to receive either nivolumab plus ipilimumab or nivolumab alone or ipilimumab alone. For the final analysis the nivolumab containing groups were compared to the ipilimumab alone group for the co-primary endpoints progression-free survival (PFS) and overall survival (OS) with an α allocation of 0.01 and 0.04. The results at the 4 years follow-up were published by Hodi et al. (2018) and can be seen in Figure 1

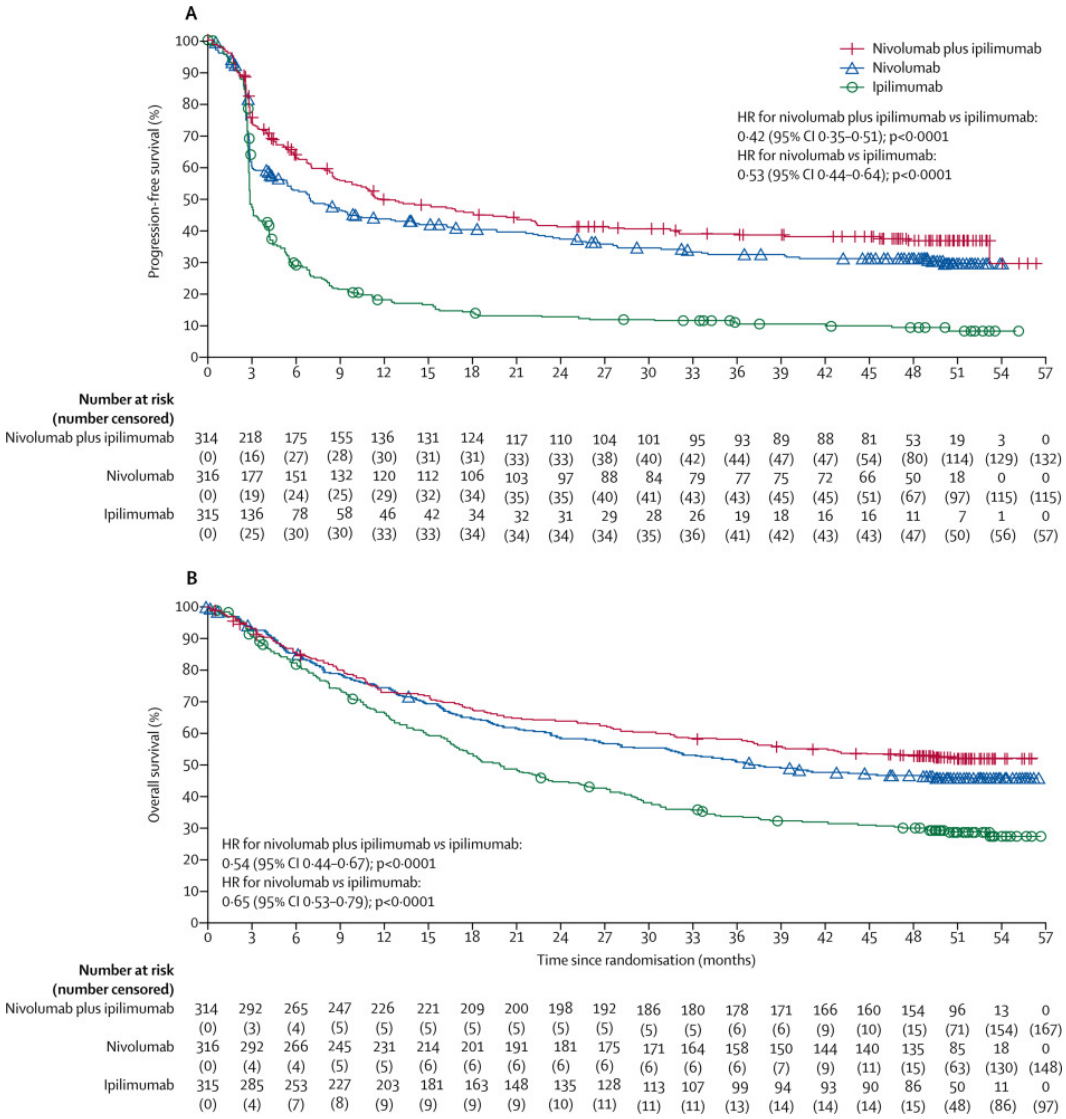

Figure 1: *Results of the CHECKMATE 067 trial for PFS (panel A) and OS (panel B) (Hodi et al., 2018). Figure used with permission of Elsevier.*

As one can see from these plots the curves are identical up to 3 months after start of treatment and then slowly start to diverge, which shows the delayed onset of treatment effect at approximately 3 months. This effect is even more pronounced for PFS (Figure 1, panel A) as the phase of slow divergence is preceded by a drop in PFS of 20%-50%.

CHECKMATE078 A multicentre, randomised, controlled, open-label phase 3 trial in chinese and russian patients with NSCLC that had progressed after chemotherapy. In total 504 patients have been randomly assigned 2:1 to receive either nivolumab or docetaxel and the primary endpoint was overall survival. Although the results of the primary endpoint did not reveal any indication for a delayed treatment effect the secondary endpoint PFS shown in Figure 2 does (Wu et al., 2019).

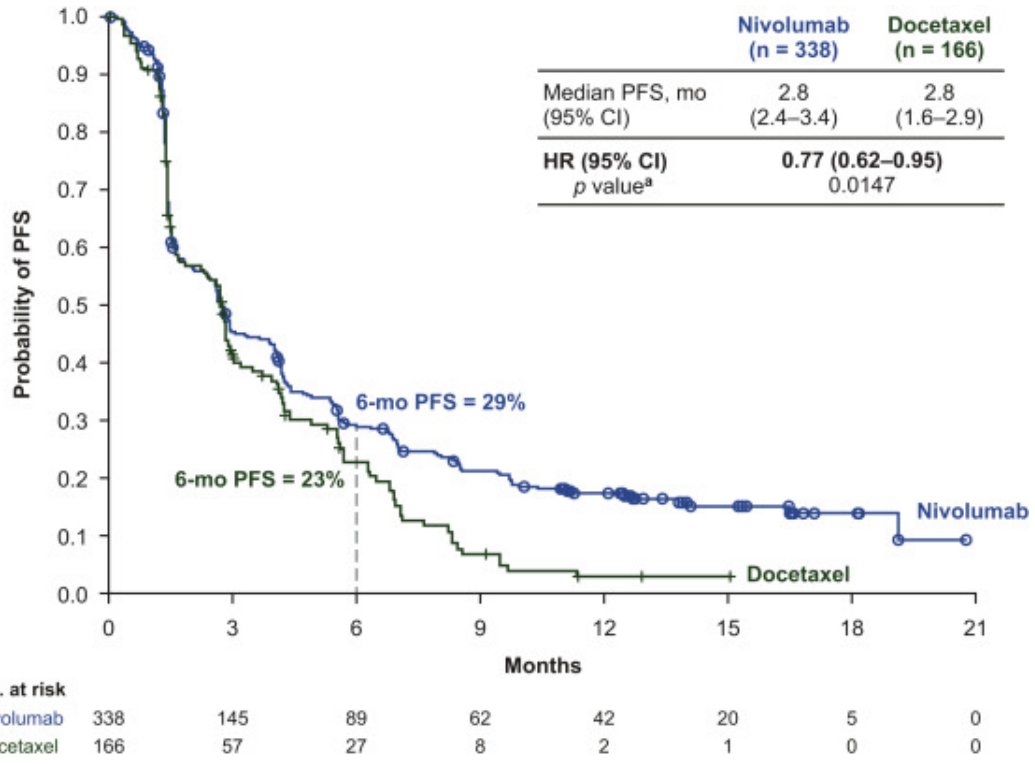

Figure 2: *Results of the CHECKMATE 078 trial for PFS (Wu et al., 2019). Figure used with permission of Elsevier.*

Based on this plot it seems as if the PH assumption is clearly violated for PFS where again a delay of approximately 3 months is revealed.

1.2 Previous work

In their groundbraking paper Peto and Peto (1972) investigated the close relationship between the logrank test and the PH assumption. They showed that a weighted logrank test is most powerful if the weights are chosen to be proportional to the logarithmic hazard ratio, implying that an unweighted logrank test achieves the highest power if the alternative under investigation has a constant hazard ratio over time, i.e. the proportional hazards assumption is fulfilled.

This work inspired many other authors to propose different weighting schemes adjusted to specific use cases of which the Fleming-Harrington class of weights is probably the best known (Fleming and Harrington, 1981). But not only weighted logrank tests but also other tests to deal with non-proportional hazards were developed based on a comparison of the Kaplan-Meier estimator, different regression approaches for time-to-event data and other tests.

Two articles (Ananthakrishnan et al., 2021; Klinglmüller et al., 2023) were published recently which give a good overview of the methods that are available at the moment. The first is a review by Ananthakrishnan et al. (2021) which is an excellent reference on suitable methods for designing and analyzing a trial when a deviation from the proportional hazards assumption is anticipated. However, although the nature of PH violations should be considered at the design stage of a trial, the authors point out that to their knowledge there is no clinical example in which this has been done yet. In a comprehensive simulation study by Klinglmüller et al. (2023) the performance of multiple methods under different general NPH scenarios such as a delayed onset of treatment effect, patient subgroups with heterogeneous treatment effects or heterogeneous patient frailty was investigated. They concluded that methods that are specifically tailored to specific deviations from the PH assumption can achieve higher power but lack robustness to different PH violations. The use of robust methods can therefore be adivsable if the focus is the testing of statistical hypothesis, but might fall short of an easily interpretable summary measure. However, in their simulation study the methods under consideration are restricted to the most important ones and the delayed onset scenarios are all based on constant hazards in the control arm.

1.3 Aim and structure of this thesis

The aim of this thesis is to investigate the performance of a variety of alternatives to the commonly used logrank test to compare a time-to-event endpoint between two groups. Motivated by the observed traits of the CHECKMATE trials the focus of this thesis lies solely on scenarios where the treatment effect is delayed. In contrast to the simulation study by Klinglmüller et al. (2023) the impact of the extent of delay will be investigated systematically. It is intended to give an overview of all methods that have been proposed in the literature and to elaborate on the families under which these methods can be subsumed so that connections between methods become more apparent.

In addition to that the power and type I error of these methods will be investigated in an extensive simulation study with particular attention to assessing how susceptible these methods are to violations of the proportional hazards assumption which is caused by the treatment effect being delayed. Furthermore, the impact of the form and extent of this delay is also under investigation. To ensure transparency a simulation plan was published in advance on zenodo.org (Behnisch, 2023). The overall motivation is to make different methods more accessible and to provide an informed basis to select an adequate method for the situation at hand.

This thesis is structured as follows. In Chapter 2, the approach for the systematic literature review is laid out and the basic concept of time-to-event data as well as the methods mentioned in this chapter and used within this thesis are provided in detail. The results are presented in Chapter 3. In Chapter 4, it is discussed how the results contribute to current research and limitations and directions for further research are outlined. In Chapter 5 and 6, a summary of this thesis is given, once in English and once in a translation to German. Reference to an online repository containing the used R program code is given in the Appendix.

Chapter 2

Methodology

This chapter first describes the literature search performed to assess the different methods that were already suggested to compare two-arm trials with time-to-event endpoints and to investigate which of these methods were already considered in simulation studies to assess their performance in a non-proportional hazards setting. Next, an introduction to the topic of time-to event endpoints will be given starting with general properties of time-to-event data and then different approaches to model such data will be presented. In the then following section, methods to analyze this kind of data are presented with a special focus on methods to handle non-proportional hazards that have been identified in the systematic literature search. Lastly, this chapter concludes with a description of the extensive simulation study to evaluate the performance of the different methods.

2.1 Literature search

As a first step a literature search had been performed to get an overview of all currently proposed methods for the analysis of time-to-event data in the presence of a delayed treatment effect or more generally of non-proportional hazard scenarios. In addition it was identified which methods have already been compared in simulation studies in this setting in order to see how an own extensive simulation study can contribute to the current research. In this section the process of identifying the relevant articles is described. The methods used in these articles will be explained later in Section 3.1. For the literature search performed on

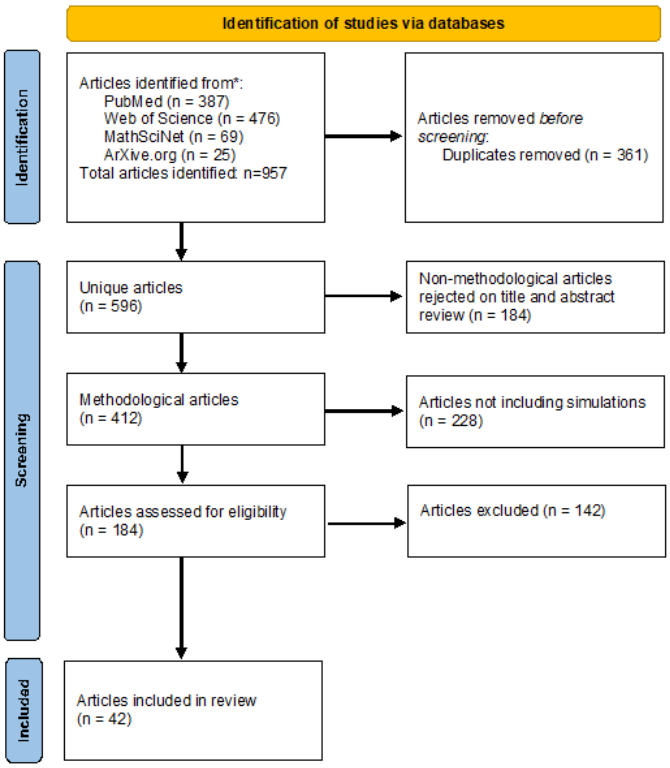

Figure 3: *Combined flowchart of the systematic literature search*

April 25th, 2022, 4 electronic databases were considered: Web of Science, PubMed, arXive and MathSciNet.

Since the delayed treatment effect is one of many examples for the violation of the proportional hazards assumption, the first step was to search for all publications dealing with the evaluation of non-proportional hazards (NPH). This was done because there are simulation studies in which methods for different NPH scenarios were compared without using the term "delayed treatment effect" in the abstract or title, but only representing a possible NPH scenario of the simulation. Thus, limiting the search to "delayed treatment effects" would fall short. In addition, a specific search for "delayed effects" and "immuno-oncology" or "immunotherapy" was performed to ensure that the search for "non-proportional hazards" did not miss any publications in this area. Figure 3 shows the combined results of both search strategies.

In total 596 unique articles were identified and further screened for eligibility. Abstract screening revealed that 184 of 596 ($\approx 31\%$) articles were clinical examples of non-proportional hazards or delayed treatment effects and did not include different statistical methods to

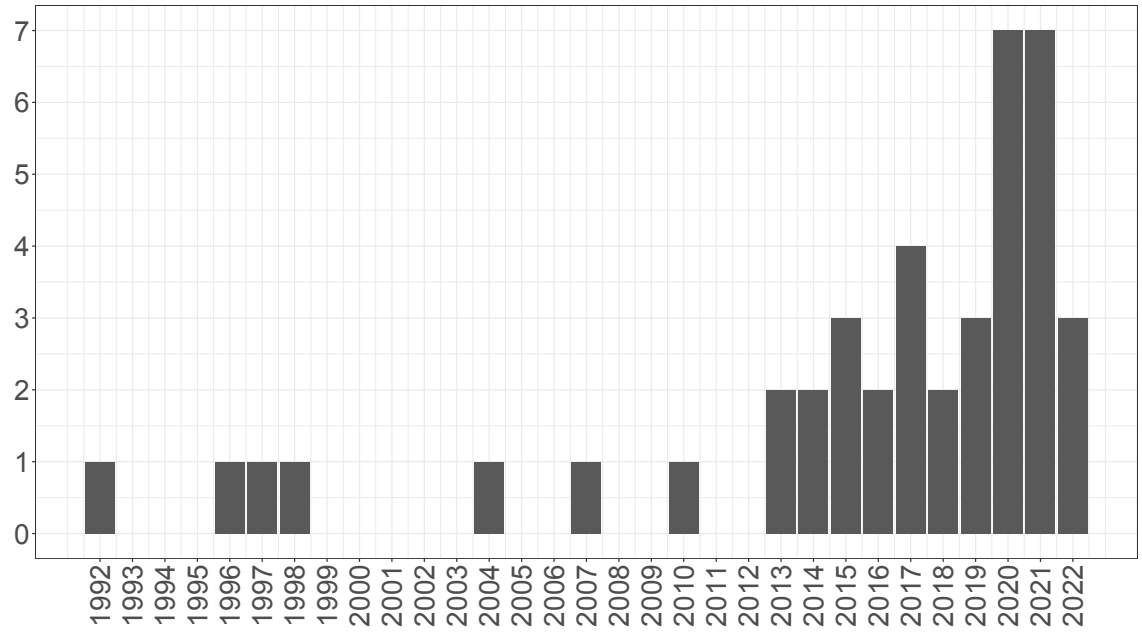

Figure 4: *Barplot of how many of the identified articles were published per year since 1992*

handle it. Of the remaining 412 articles, 228 ($\approx$ 55 %) did not compare the performance of the methods under consideration in simulations, resulting in a total of 184 articles assessed for eligibility. Reading the articles in detail revealed that 142 did not deal with hypothesis testing but with estimation approaches which resulted in an absolut number of 42 articles of which the methods used in the simulation studies were extracted and will be presented in the next chapter in Section 3.1. Figure 4 shows how many of the 42 articles were published in each year between 1992 and 2022. It can be seen that the need for methods to deal with non-proportional hazards has increased over the last 10 years which is possibly due to the advances in the field of immuno-oncology that were already described in the introduction of this thesis.

2.2 Time-to-event data

As the name suggests time-to-event data is defined as the time from an objectively defined starting point (e.g. time of randomization in RCTs, time of exposure in case-control studies, etc.) until an event of interest occurs. The best known types of time-to-event endpoints are survival endpoints, where the event of interest is the death of a patient (overall survival

(OS)) or a combined endpoint of death and progress of the underlying disease (progression-free survival (PFS)). These endpoints are commonly used in oncological studies as they provide meaningful evidence to measure the benefit of cancer therapy. The problem that arises with time-to-event data is that the event of interest cannot always be observed due to the limited observation period and hence incomplete data must be incorporated in the analyses since it carries information on how long a patient was event-free. This incomplete data is called a *censored observation*. To limit the amount of censoring the length of the observation period for each patient must be chosen with caution depending on the time necessary to recruit all patients, the rate the disease progresses and must be balanced with regulatory and economic considerations.

To formulate time-to-event data the following two latent/unobservable random variables need to be defined: Let T_{fail} denote the underlying time to the event of interest (often failure of a treatment) and T_{cens} the underlying censoring time. The following two assumptions on these two distributions are often imposed, which are crucial for analyzing time-to-event data and making valid inference:

1. **Independence:** T_{cens} is (stochastically) independent of T_{fail} conditional on possible explanatory variables.
2. **Uninformativeness:** T_{cens} contains no information on the distribution of T_{fail}

The observed data consists then of the observed time $T = \min(T_{fail}, T_{cens})$ together with the event indicator $D = \mathbb{1}(T = T_{fail})$. Moreover, let X denote the vector of covariates of each patient which in the setting of a two arm comparison is a one-dimensional vector containing only the treatment indicator, i.e. $X = 0$ indicates that the patient receives the control treatment and $X = 1$ indicates that the patient receives the experimental treatment. The n observations are then given by triples $(t_1, d_1, x_1), \ldots, (t_n, d_n, x_n)$.

In the following basic concepts of time-to-event data such as the survival and hazard function will be introduced. Furthermore, it will be illustrated how time-to-event data can be modelled by fully parametric distributions and in the framework of delayed treatment effects. The statistical methods to analyze time-to-event data with delayed treatment effects are then introduced and finally the structure of the simulation study is described in more detail.

2.2.1 Basic concepts

To describe the underlying failure time distribution of a time-to-event random variable T it is often convenient to use the survival function $S(t)$, which is complimentary to the cumulative distribution function $F(t)$, i.e. $S(t) = 1 - F(t)$. If T has a continuous distribution, which is often the case, the survival function $S(t)$ is continuous and differentiable. Another characteristic function that is used extensively in survival analysis is the hazard function $\lambda(t)$, which sometimes is also called (instantaneous) failure rate and is defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{P(T < t + dt | T \geqslant t)}{dt}.$$

It is easy to see that this is exactly the negative logarithmic derivative of the survival function,

$$\lambda(t) = -\frac{S'(t)}{S(t)} = -\frac{d \ln(S(t))}{dt},$$

and hence the survival function can be obtained from the hazard function via

$$S(t) = \exp\left(-\int_0^t \lambda(t')dt'\right).$$

The integral in the above formula is called the cumulative hazard function and is often denoted with $\Lambda(t)$. Note that although $S(t)$ is a monotonically decreasing function with $S(0) = 1$ and $S(\infty) = 0$ the hazard function $\lambda(t)$ can attain any positive value.

For the analysis of survival data one often assumes that the individual hazard functions $\lambda_i(t)$ are proportional to a baseline hazard function $\lambda_0(t)$ which is left unspecified. That is one assumes that there exists a constant $c_i > 0$ such that $\lambda_i(t) = c_i \lambda_0(t)$. A direct consequence of this assumption is that the hazard ratio is constant and for two individuals ($i \neq j$) the survival functions fulfill the following property:

$$S_j(t) = \exp\left(-c_j \Lambda_0(t)\right) = \exp\left(-\frac{c_j}{c_i} c_i \Lambda_0(t)\right) = \exp\left(-c_i \Lambda_0(t)\right)^{\frac{c_j}{c_i}} = S_i(t)^{\frac{c_j}{c_i}}.$$

Hence, for the proportional hazards assumption to hold, the survival functions can not cross nor be parallel (except for the case where they are identical). The appeal of this assumption,

that made it widely popular within the statistic community and the cornerstone of the famous Cox model, is that if it holds the hazard ratio is constant over time and hence it is sufficient to quantify the difference between the survival distributions. However, there are various reasons why this assumptions can be violated in a two arm trial such as: long-term survivors in the treatment arm, subgroups for which treatment is harmful or beneficial, treatment effect is diminishing over time or treatment effect is delayed.

The latter reason is the focus of this thesis and statistical methods to deal with it will be investigated more thoroughly.

2.2.2 Parametric distributions and their application

To model time-to-event data there exist a variety of parametric distributions on $[0, \infty)$ that can be used. These distributions and their properties shall be explained in more detail in the next sections.

(Piece-wise) Exponential distribution

Most frequently used is the exponential distribution ($\text{Exp}(\lambda)$) which is defined by the so called rate parameter $\lambda > 0$, which quantifies the rate at which the events of interest occur, and has the following survival function

$$S(t) = e^{-\lambda t}.$$

Intuitively, if λ is the rate at which events occur, the expected time between events is given by $\frac{1}{\lambda}$. The variance of this distribution is $\frac{1}{\lambda^2}$ so that the standard deviation equals the mean of this distribution. The median survival of this distribution, i.e. the time at which the survival probability is 50%, is then $\frac{\ln(2)}{\lambda}$. Figure 5 shows the survival and hazard function for three exemplary exponential distributions.

Based on the definition of the (cumulative) hazard one can easily see from the survival function that the hazard function is exactly the rate parameter λ and hence constant over time. This has the direct implication that for two different exponentially distributed survival times with rate parameter λ_1, λ_2 the hazards are always proportional and the proportionality factor is

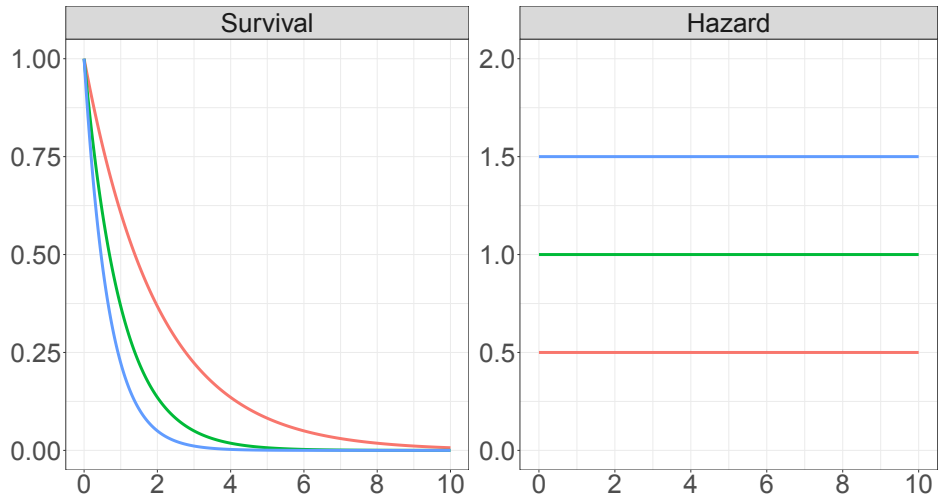

Figure 5: *Exemplary plot of the survival and hazard functions of different exponentially distributed survival times*

the ratio $\frac{\lambda_1}{\lambda_2}$. Hence the exponential distribution can only be used to simulate survival times that meet the proportional hazards assumption and the simulations could therefore not solely rely on this distribution except for the simulation of censoring times. To simulate failure times with non-proportional hazards piece-wise exponentially distributed failure times were used which are a simple extension of the exponential distribution and assume a piecewise constant hazard.

To define the piecewise exponential distribution let $0 = \tau_0 < \tau_1 < \ldots < \tau_J = \infty$ be a partitioning of $[0, \infty)$ into J intervals. Based on this partition the hazard is assumed constant within each interval, so that the hazard function is given by the following step function

$$\lambda(t) = \lambda_j, \text{for } t \text{ in } [\tau_{j-1}, \tau_j),$$

for given parameters $\lambda_1, \ldots, \lambda_J > 0$.

The cumulative hazard function is then by definition given as

$$\Lambda(t) = \int_0^t \lambda(s)ds = \sum_{j=1}^{J} \lambda_j(\min(t, t_j) - \min(t, t_{j-1}))$$

resulting in a survival function of $S(t) = \exp(-\Lambda(t))$.

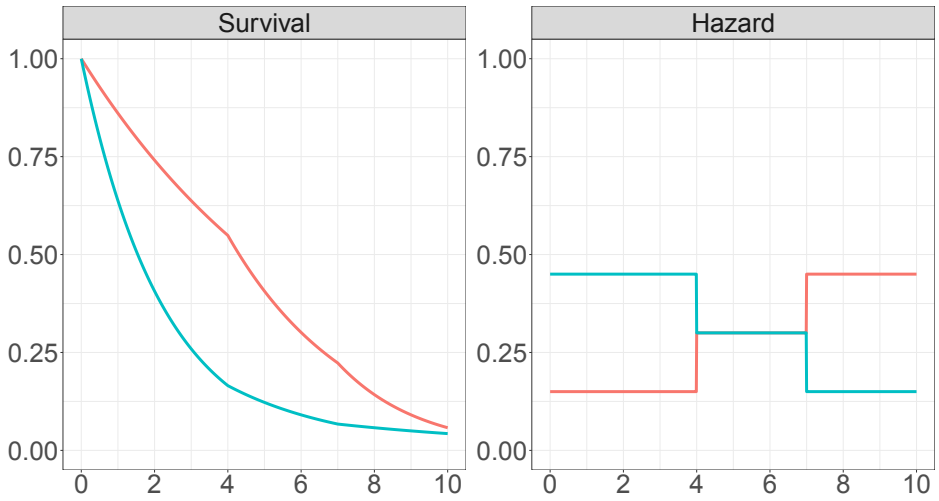

Figure 6: *Exemplary plot of survival and hazard functions of piecewise exponentially distributed survival times*

It is well known that every continuous and bounded function can be approximated by step functions so the piecewise exponential distribution can be used to approximate arbitrary hazard functions. Figure 6 shows the survival and hazard function for two exemplary piecewise exponential distributions.

Weibull distribution

In contrast to the exponential distribution the Weibull distribution, named after the swedish mathematician Waloddi Weibull and denoted with $\text{Wei}(\lambda, k)$, allows to model event times that do not necessarily have proportional hazards. The survival function of a Weibull distribution depends on two parameters, the scale parameter $\lambda > 0$ and the shape parameter $k > 0$, and is given by

$$S(t) = e^{-(\lambda t)^k}.$$

As one can see the choice of $k = 1$ reduces the Weibull distribution to an exponential distribution with the scale parameter λ as rate parameter. The cumulative hazard derived from this survival function is then $\Lambda(t) = (\lambda t)^k$ and hence the hazard function is

$$\lambda(t) = \frac{d}{dt}\Lambda(t) = \lambda k(\lambda t)^{k-1}.$$

This function is monotonically decreasing for $k < 1$, constant for $k = 1$ and monotonically increasing for $k > 1$. The ratio of the hazard of two different Weibull distributions with parameters (λ_1, k_1) and (λ_2, k_2) is given by

$$\frac{\lambda_1 k_1 (\lambda_1 t)^{k_1 - 1}}{\lambda_2 k_2 (\lambda_2 t)^{k_2 - 1}} = \frac{\lambda_1^{k_1} k_1}{\lambda_2^{k_2} k_2} t^{k_1 - k_2}$$

and hence it can be seen that the proportional hazards assumption is satisified if and only if these two distributions have the same shape parameter, i.e. $k_1 = k_2 = k$. In this case the hazard ratio is given by $\left(\frac{\lambda_1}{\lambda_2}\right)^k$.

Mean and variance of a Weibull distributed random variable T can not be expressed in a closed form but in terms of the Gamma function.

$$E[T] = \frac{1}{\lambda} \cdot \Gamma\left(1 + \frac{1}{k}\right),$$
$$\mathrm{Var}[T] = \frac{1}{\lambda^2}\left[\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right)\right]$$

The median survival can be expressed in terms of the shape and scale parameter as $\frac{\ln(2)^{1/k}}{\lambda}$. Survival and hazard functions for Weibull distributions with different shape parameters k can be seen in Figure 7.

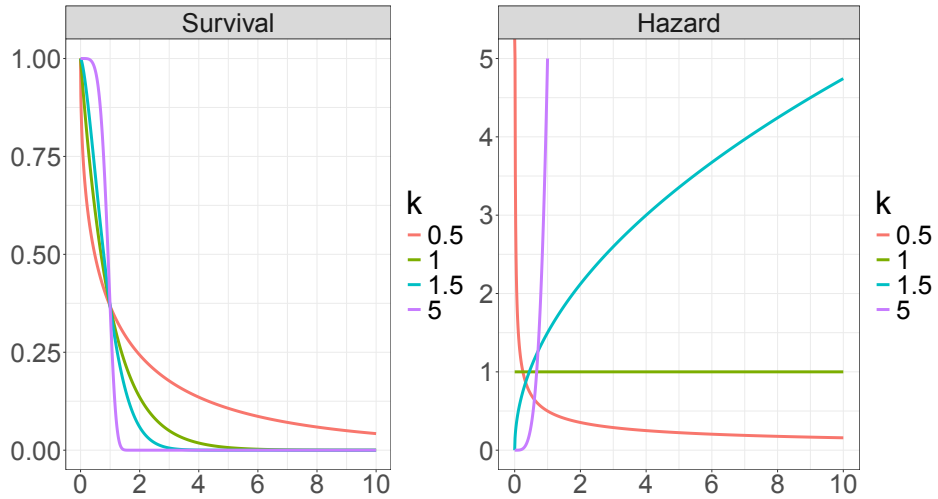

Figure 7: *Exemplary plot of survival and hazard functions of different Weibull distributed survival times*

The hazard and survival functions together with the mean and median of the three above mentioned distributions is summarized in Table 2.

Table 2: *Summary of the exponential, piecewise exponential and Weibull distribution*

	Exponential	piecewise Exponential	Weibull
Hazard $\lambda(t)$	λ	$\sum_{j=1}^{J} \lambda_j 1(\tau_{j-1} \leqslant t < \tau_j)$	$\lambda k(\lambda t)^{k-1}$
Survival $S(t)$	$e^{-\lambda t}$	$e^{-\sum_{j=1}^{J} \lambda_j [\min(t,\tau_j) - \min(t,\tau_{j-1})]}$	$e^{-(\lambda t)^k}$
Expected value	λ^{-1}		$\frac{1}{\lambda} \cdot \Gamma\left(1 + \frac{1}{k}\right)$
Median	$\frac{\ln(2)}{\lambda}$		$\frac{\ln(2)^{1/k}}{\lambda}$

2.2.3 Simulating delayed treatment effects data

As explained in the Introduction, one common feature of immuno-oncology trials is that the onset of the treatment effect is often delayed due to the indirect mechanism-of-action. For the analysis of time-to-event endpoints this poses a challenge since the often made assumption that the hazards are proportional is violated in this scenario. Two models that arise from this delayed onset of treatment effect and how these models can be used to simulate data for the extensive simulation study will be presented. Both models have in common that there is no treatment effect during the first phase of the study, i.e. hazard ratio equal to 1, which increases to the full treatment effect of a hazard ratio of $\theta < 1$ and stays constant thereafter.

Denote the hazard in the control arm with λ_C and the hazard in the experimental arm with λ_E. The simplest model to describe a delayed treatment effect is the **threshold lag model** (TLM) where one assumes that there is no effect up to a change point t^*, and then the full effect θ sets in, i.e. the hazard ratio is given as

$$\frac{\lambda_E(t)}{\lambda_C(t)} = \begin{cases} 1, & t \leqslant t^* \\ \theta, & t > t^* \end{cases}$$

with $\theta < 1$. This sudden onset of the full treatment effect is weakened in the **generalized linear lag model** (GLLM), where one assumes that after a first phase of no effect up to a lag time t_1^*, the effect linearly increases until it reaches the full effect θ at the delay time t_2^*.

This leads to the following hazard ratio function:

$$\begin{aligned} \frac{\lambda_E(t)}{\lambda_C(t)} &= \begin{cases} 1, & t \leqslant t_1^* \\ 1 - (1-\theta)\frac{t-t_1^*}{t_2^*-t_1^*}, & t_1^* < t \leqslant t_2^*. \\ \theta, & t > t_2^* \end{cases} \\ &= 1 - l(t) + \theta l(t), \end{aligned}$$

where $l(t) = \left(\frac{t-t_1^*}{t_2^*-t_1^*}\mathbb{1}(t_1^* \leqslant t < t_2^*) + \mathbb{1}(t \geqslant t_2^*)\right)$. As can be seen from this expression, the generalized linear lag model simplifies to the threshold lag model if the parameters t_1^* and t_2^*

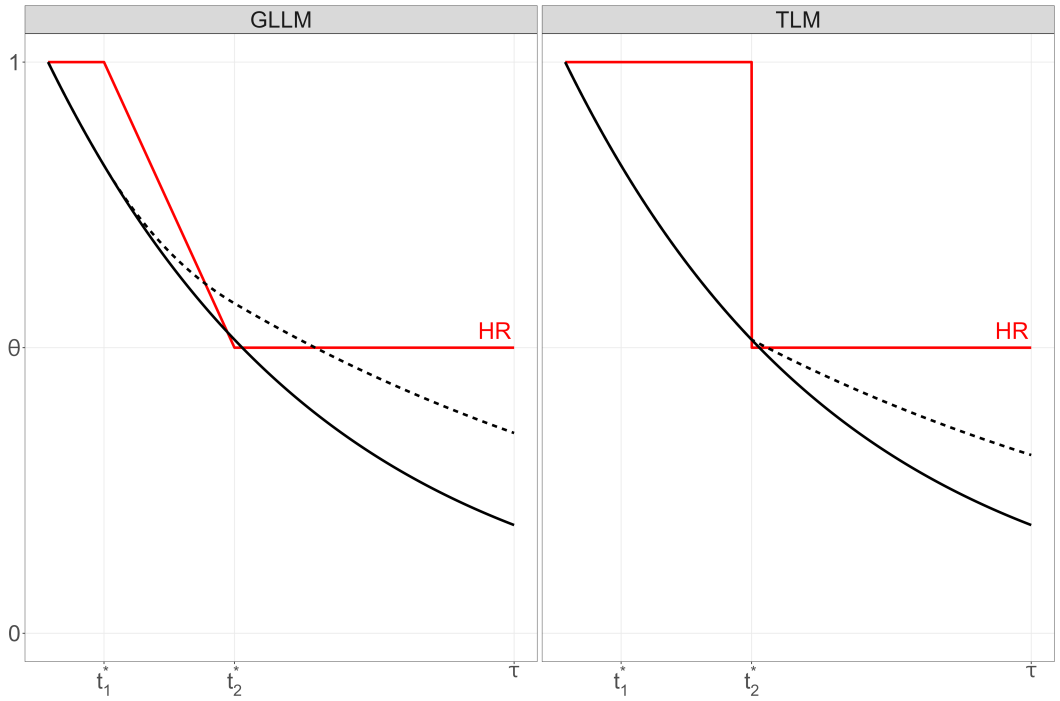

Figure 8: *Illustration of a generic GLLM and TLM scenario assuming equal full effect θ, study duration τ, lag t_2^* and changepoint t_1^*. The black lines are the survival in the control arm (solid line) and the experimental arm (dashed line) and the red line is the hazard ratio.*

coincide and to the proportional hazards model if both parameters are set to 0. The survival and hazard ratio functions resulting from these models are displayed in Figure 8 for a generic TLM and GLLM scenario.

To simulate data based on the generalized linear lag model the survival times in the treatment arm have to be generated based on the hazard that is derived from the control arm hazard by application of this model. For $t_2^* = 0$ the model reduces to a proportional hazards model $\lambda_E(t) = \theta \cdot \lambda_C(t)$. Assuming a general Weibull distribution $\text{Wei}(\lambda_C, k_C)$ for the survival times in the control arm and plugging in the hazard $\lambda_C(t) = \lambda_C^{k_C} k_C t^{k_C - 1}$ yields

$$\lambda_E(t) = \theta\lambda_C(t) = \theta\lambda_C^{k_C} k_C t^{k_C-1} = (\theta^{1/k_C}\lambda_C)^{k_C} k_C t^{k_C-1}$$

and hence the failure times in the experimental arm are $\text{Wei}(\theta^{1/k_C}\lambda_C, k_C)$ distributed. If the changepoint $t_2^* \neq 0$ then based on this model the hazard in the treatment arm can be

expressed as

$$\begin{aligned}
\lambda_E(t) &= [1 - l(t) + \theta l(t)]\lambda_C(t) \\
&= \left[1 - \left(\frac{t - t_1^*}{t_2^* - t_1^*} I(t_1^* \leqslant t < t_2^*) + I(t \geqslant t_2^*)\right) + \theta \left(\frac{t - t_1^*}{t_2^* - t_1^*} I(t_1^* \leqslant t < t_2^*) + I(t \geqslant t_2^*)\right)\right] \lambda_C(t) \\
&= \left[1 + (\theta - 1)\frac{t - t_1^*}{t_2^* - t_1^*} I(t_1^* \leqslant t < t_2^*) + (\theta - 1) I(t \geqslant t_2^*)\right] \lambda_C(t) \\
&= \left[\underbrace{\left(1 + (\theta - 1) I(t \geqslant t_2^*) - \frac{\theta - 1}{t_2^* - t_1^*} t_1^* I(t_1^* \leqslant t < t_2^*)\right)}_{=:A(t)} + \underbrace{\frac{\theta - 1}{t_2^* - t_1^*} I(t_1^* \leqslant t < t_2^*)\, t}_{=:B(t)}\right] \lambda_C(t) \\
&= (A(t) + B(t)t)\lambda_C(t),
\end{aligned}$$

where the newly defined time-dependent functions $A(t)$ and $B(t)$ can be rewritten as

$$A(t) = \begin{cases} 1 & , t < t_1^* \\ A := 1 - \frac{\theta - 1}{t_2^* - t_1^*} t_1^* & , t_1^* \leqslant t < t_2^* \\ \theta & , t \geqslant t_2^* \end{cases} \qquad B(t) = \begin{cases} B := \frac{\theta - 1}{t_2^* - t_1^*} & , t_1^* \leqslant t < t_2^*, \\ 0 & , \text{else} \end{cases}$$

where $A + Bt_1^* = 1$ and $A + Bt_2^* = \theta$ due to continuity of $\lambda_E(t)$. To simulate the survival times in the treatment arm, the inversion method (Kolonko, 2008, chapter 8, pg. 85-95) will be applied and to this end the inverse of the probability distribution function has to be derived. The probability distribution function is given as

$$F_E(t) = 1 - S_E(t) = 1 - \exp(-\Lambda_E(t))$$

and hence to find the inverse of $F_E(t)$ this equation needs to be solved for t. The above equation can be rearranged to

$$y = F_E(t) = 1 - \exp(-\Lambda_E(t)) \Leftrightarrow \exp(-\Lambda_E(t)) = 1 - y \Leftrightarrow \Lambda_E(t) = -\ln(1 - y)$$

so that the inverse of F_E evaluated at y is the same as the inverse of Λ_E evaluated at $\ln(1-y)$ and hence it is sufficient to find an inverse of the cumulative hazard function $\Lambda_E(t)$.

$$\begin{aligned}
\Lambda_E(t) &= \int_0^t \lambda_E(x)dx \\
&= \int_0^t (A(x) + B(x)x)\lambda_C(x)dx \\
&= \int_0^{\min(t,t_1^*)} \lambda_C(x)dx + \int_{\min(t,t_1^*)}^{\min(t,t_2^*)} (A + Bx)\lambda_C(x)dx + \int_{\min(t,t_2^*)}^{t} \theta\lambda_C(x)dx \\
&= \Lambda_C(\min(t, t_1^*)) + \theta(\Lambda_C(t) - \Lambda_C(\min(t, t_2^*))) + \int_{\min(t,t_1^*)}^{\min(t,t_2^*)} (A + Bx)\lambda_C(x)dx \\
&= \Lambda_C(\min(t, t_1^*)) + \theta(\Lambda_C(t) - \Lambda_C(\min(t, t_2^*))) + [(A + Bx)\Lambda_C(x)]_{\min(t,t_1^*)}^{\min(t,t_2^*)} - \int_{\min(t,t_1^*)}^{\min(t,t_2^*)} B\Lambda_C(x)dx \\
&= \Lambda_C(\min(t, t_1^*)) + \theta(H_C(t) - \Lambda_C(\min(t, t_2^*))) + (A + B\min(t, t_2^*))\Lambda_C(\min(t, t_2^*)) \\
&\quad - (A + B\min(t, t_1^*))\Lambda_C(\min(t, t_1^*)) - B\int_{\min(t,t_1^*)}^{\min(t,t_2^*)} \Lambda_C(x)dx
\end{aligned}$$

Assuming that the survival times in the control arm are $\text{Wei}(\lambda_C, k_C)$ distributed its cumulative hazard function is given by $\Lambda_C(t) = (\lambda_C \cdot t)^{k_C}$ and hence the last integral becomes

$$\begin{aligned}
\int_{\min(t,t_1^*)}^{\min(t,t_2^*)} \Lambda_C(x)dx = \int_{\min(t,t_1^*)}^{\min(t,t_2^*)} (\lambda_C x)^{k_C} dx &= \frac{\lambda_C^{k_C}}{k_C+1} x^{k_C+1}\Big|_{\min(t,t_1^*)}^{\min(t,t_2^*)} = \frac{1}{k_C+1} x\Lambda_C(x)\Big|_{\min(t,t_1^*)}^{\min(t,t_2^*)} \\
&= \frac{1}{k_C+1}\left(\min(t, t_2^*)\Lambda_C(\min(t, t_2^*)) - \min(t, t_1^*)\Lambda_C(\min(t, t_1^*))\right)
\end{aligned}$$

Putting this all together this expression can be reformulated piecewise as

$$\Lambda_E(t) = \begin{cases} \Lambda_C(t) & , t < t_1^* \\ (A + Bt)\Lambda_C(t) - \frac{B}{k_C+1}(t\Lambda_C(t) - t_1^*\Lambda_C(t_1^*)) & , t_1^* \leqslant t < t_2^*. \\ \theta\Lambda_C(t) - \frac{B}{k_C+1}(t_2^*\Lambda_C(t_2^*) - t_1^*\Lambda_C(t_1^*)) & , t \geqslant t_2^* \end{cases}$$

Combining all constant terms and all terms that depend on t the following equation needs to be solved:

$$z = \Lambda_E(t) = \begin{cases} \Lambda_C(t) & , t < t_1^* \\ \frac{B}{k_C+1} t_1^* \Lambda_C(t_1^*) + \left[(A + \frac{k_C}{k_C+1} Bt) \Lambda_C(t) \right] & , t_1^* \leqslant t < t_2^*. \\ -\frac{B}{k_C+1} (t_2^* \Lambda_C(t_2^*) - t_1^* \Lambda_C(t_1^*)) + \theta \Lambda_C(t) & , t \geqslant t_2^* \end{cases}$$

This equation can be solved analytically for $t < t_1^*$ and $t \geqslant t_2^*$ which yields:

$$t = \Lambda_E^{-1}(z) = \begin{cases} \frac{1}{\lambda_C} z^{1/k_C} & , z < \Lambda_T(t_1^*) \\ t \in [t_1^*, t_2^*) \colon (A + \frac{k_C}{k_C+1} Bt)(\lambda_C t)^{k_C} = z - \frac{B}{k_C+1} t_1^* (\lambda_C t_1^*)^{k_C} & , \Lambda_E(t_1^*) \leqslant z < \Lambda_E(t_2^*) \\ \frac{1}{\lambda_C} \left(\frac{z + \frac{B}{k_C+1} (t_2^* (\lambda_C t_2^*)^{k_C} - t_1^* (\lambda_C t_1^*)^{k_C})}{\theta} \right)^{1/k_C} & , z \geqslant \Lambda_E(t_2^*) \end{cases}$$

In the special case of a threshold lag model ($t_1^* = t_2^* = t^*$) this expressions simplify to

$$\Lambda_E(t) = \begin{cases} \Lambda_C(t) & , t < t^* \\ \theta \Lambda_C(t) + (1 - \theta) \Lambda_C(t^*) & , t \geqslant t^* \end{cases}$$

$$\Lambda_E^{-1}(z) = \begin{cases} \frac{1}{\lambda_C} z^{1/k_C} & , z < \Lambda_C(t^*) \\ \frac{1}{\lambda_C} \left(\frac{z - (1-\theta)(\lambda_C t^*)^{k_C}}{\theta} \right)^{1/k_C} & , z \geqslant \Lambda_C(t^*) \end{cases}$$

2.3 Analyzing time-to-event data

As explained in the Introduction of this thesis a problem that is often encountered in immuno-oncological studies with time-to-event endpoints is that the power of commonly used statistical methods for two arm comparisons can be heavily decreased due to the distinctive mechanism of action of these drugs. A particular mechanism, which was observed in many studies in this area and forms the basis of this thesis, is the delayed onset of treatment effect. In this section the methods that have been identified in the systematic literature search presented in the Section 2.1 shall be explained in more detail.

As for all analyses of time-to-event data the main challenge is to incorporate censored observations in the analysis since these observations carry at least in part information on the event of interest. To do so one has to acknowledge that time-to-event data is not rigid but changes over time and hence must be modelled as a stochastic process.

First, this stochastic process notation as well as basic concepts for analyzing time-to-event data will be presented and then the focus will be laid on inferential two arm comparisons as these shall be evaluated in the simulation study. It has to be noted that the different methods do not always target the same global null hypothesis $H_0^{\text{global}}: S_C(t) = S_E(t)$ of equal survival or equivalently equal hazard functions, but the null hypothesis related to each method is always a superset or implication of this global null.

2.3.1 Fundamental concepts and estimators

Stochastic process notation

Let $(t_1, d_1, x_1), \ldots, (t_n, d_n, x_n)$ denote the observations of n patients. The following processes describe how many patients (per arm) are at risk at time t.

$$\bar{Y}_E(t) = \sum_{i=1}^{n} \mathbb{1}(t_i \geqslant t, x_i = 1),$$
$$\bar{Y}_C(t) = \sum_{i=1}^{n} \mathbb{1}(t_i \geqslant t, x_i = 0),$$
$$\bar{Y}(t) = \bar{Y}_E(t) + \bar{Y}_C(t)$$

Furthermore, the processes

$$\bar{N}_E(t) = \sum_{i=1}^{n} \mathbb{1}(t_i \leqslant t, d_i = 1, x_i = 1),$$
$$\bar{N}_C(t) = \sum_{i=1}^{n} \mathbb{1}(t_i \leqslant t, d_i = 1, x_i = 0),$$
$$\bar{N}(t) = \bar{N}_E(t) + \bar{N}_C(t)$$

denote how many failures (per arm) occur up to and including time t. The number of failures in arm g *at* a given time t is denoted with $\Delta\bar{N}_g(t) = \bar{N}_g(t) - \bar{N}_g(t-)$. The total number in both arms will again be denoted without index. Theses two processes are right-continuous, non-decreasing step functions with $\bar{Y}(t)$ having jumps at $t_1, \ldots, t_n$ and $\bar{N}(t)$ having jumps at $\{t_i | d_i = 1, i = 1, \ldots, n\}$. As it is needed for future calculations the Lebesgue-Stieltjes integral is defined as follows:

Suppose $G(.)$ is a right-continuous, nondecreasing step function with jumps at $t_1, \ldots, t_n$. Then for any function $f(.)$ the Lebesgue-Stieltjes integral is defined as:

$$\int_a^b f(t)dG(t) := \sum_{a<t_i\leqslant b} f(t_i)(G(t_i) - G(t_i-)) = \sum_{a<t_i\leqslant b} f(t_i)\Delta G(t_i)$$

Kaplan-Meier estimator

The most common estimator for the survival function is the Kaplan-Meier estimator (Kaplan and Meier, 1958), which is obtained as the product limit of simple binomial proportions $1 - \frac{\Delta\bar{N}(t_i)}{\bar{Y}(t_i)}$ of patients alive just after t_{i-1} that survive beyond t_i and is defined as

$$\hat{S}_{KM}(t) = \prod_{t_i\leqslant t} \left(1 - \frac{\Delta\bar{N}(t_i)}{\bar{Y}(t_i)}\right).$$

For the standard error of this estimator Greenwood (1926) derived the following formula

$$\mathrm{Var}(\hat{S}_{KM}(t)) = \hat{S}_{KM}(t)^2 \sum_{t_i\leqslant t} \frac{\Delta\bar{N}(t_i)}{\bar{Y}(t_i)(\bar{Y}(t_i) - \Delta\bar{N}(t_i))}. \tag{2.1}$$

Deriving an estimate of the hazard function from the Kaplan-Meier estimator is difficult since the Kaplan-Meier estimator is a non-differentiable step function and hence some smoothing would be necessary. However, it is often satifactory to estimate the cumulative hazard function $\Lambda(t)$ which can be estimated by plugging the Kaplan-Meier estimator into the relation $\Lambda(t) = -\log(S(t))$.

This gives

$$\hat{\Lambda}_{KM}(t) = -\log(\hat{S}_{KM}(t)) = -\sum_{t_i \leqslant t} \log\left(1 - \frac{\Delta\bar{N}(t_i)}{\bar{Y}(t_i)}\right),$$

and the variance of this estimator can be obtained with the functional delta method from the Greenwood formula.

Nelson-Aalen estimator

Alternatively, Nelson (1969) and Aalen (1975) developed a more direct estimator of the cumulative hazard function and its variance as

$$\hat{\Lambda}_{NA}(t) = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)} = \sum_{t_i \leqslant t} \frac{\Delta\bar{N}(t_i)}{\bar{Y}(t_i)},$$
$$\mathrm{Var}(\hat{\Lambda}_{NA}(t)) = \sum_{t_i \leqslant t} \frac{\Delta\bar{N}(t_i)}{\bar{Y}^2(t_i)}.$$

In most settings the difference between these two estimators is small if the sample size is sufficiently large.

2.3.2 Basic inferential methods

The logrank test

The basic idea of the logrank test is similar to the χ^2-test, i.e. to compare the observed number of events with the expected number of events at each timepoint under the global null hypothesis H_0^{global}: $S_C(t) = S_E(t)$.

Let $t_{(1)} < \ldots < t_{(k)}$ denote the k distinct, ordered event times, allowing for tied events and hence $k \leqslant n$. At each failure time $t_{(i)}$ one can construct the following contingency table:

	Failure	No failure	Total
Control arm	$\Delta\bar{N}_C(t_{(i)})$	$\bar{Y}_C(t_{(i)}) - \Delta\bar{N}_C(t_{(i)})$	$\bar{Y}_C(t_{(i)})$
Experimental arm	$\Delta\bar{N}_E(t_{(i)})$	$\bar{Y}_E(t_{(i)}) - \Delta\bar{N}_E(t_{(i)})$	$\bar{Y}_E(t_{(i)})$
Total	$\Delta\bar{N}(t_{(i)})$	$\bar{Y}(t_{(i)}) - \Delta\bar{N}(t_{(i)})$	$\bar{Y}(t_{(i)})$

The number of failures in the control arm at time $t_{(i)}$ $\Delta\bar{N}_C(t_{(i)})$ given the total number of failures $\Delta\bar{N}(t_{(i)})$ then follows a hypergeometric distribution Hyp $\left(\bar{Y}(t_{(i)}), \bar{Y}_C(t_{(i)}), \Delta\bar{N}(t_{(i)})\right)$, with an expected number of failures of $\frac{\bar{Y}_C(t_{(i)})\Delta\bar{N}(t_{(i)})}{\bar{Y}(t_{(i)})}$ and a variance of

$$\Delta\bar{N}(t_{(i)})\frac{\bar{Y}_C(t_{(i)})}{\bar{Y}(t_{(i)})}\left(1-\frac{\bar{Y}_C(t_{(i)})}{\bar{Y}(t_{(i)})}\right)\frac{\bar{Y}(t_{(i)})-\Delta\bar{N}(t_{(i)})}{\bar{Y}(t_{(i)})-1},$$

where the last fraction is called finite population correction factor and cancels when there are no tied event times, i.e. no two patients have the event at the same time ($\Delta\bar{N}(t_{(i)}) = 1,$ for all $t_{(i)}$). Summing over all observed event times gives the well known logrank statistic:

$$\begin{aligned}
W &= \sum_{i=1}^{k}\left(\Delta\bar{N}_C(t_{(i)}) - \frac{\bar{Y}_C(t_{(i)})\Delta\bar{N}(t_{(i)})}{\bar{Y}(t_{(i)})}\right)\\
&= \sum_{i=1}^{k}\left(\Delta\bar{N}_C(t_{(i)}) - \frac{\bar{Y}_C(t_{(i)})\Delta\bar{N}_C(t_{(i)})}{\bar{Y}(t_{(i)})}\right) - \frac{\bar{Y}_C(t_{(i)})\Delta\bar{N}_E(t_{(i)})}{\bar{Y}(t_{(i)})}\\
&= \sum_{i=1}^{k}\frac{\bar{Y}_E(t_{(i)})\Delta\bar{N}_C(t_{(i)}) - \bar{Y}_C(t_{(i)})\Delta\bar{N}_E(t_{(i)})}{\bar{Y}(t_{(i)})}\\
&= \sum_{i=1}^{k}\frac{\bar{Y}_E(t_{(i)})\bar{Y}_C(t_{(i)})}{\bar{Y}(t_{(i)})}\left(\frac{\Delta\bar{N}_C(t_{(i)})}{\bar{Y}_C(t_{(i)})} - \frac{\Delta\bar{N}_E(t_{(i)})}{\bar{Y}_E(t_{(i)})}\right)\\
&= \int_0^\infty \frac{\bar{Y}_E(t)\bar{Y}_C(t)}{\bar{Y}(t)}\left(\frac{d\bar{N}_C(t)}{\bar{Y}_C(t)} - \frac{d\bar{N}_E(t)}{\bar{Y}_E(t)}\right). \qquad (2.2)
\end{aligned}$$

By the central limit theorem this statistic is asymptotically normally distributed under the global null hypothesis H_0^{global} with mean 0 and its variance can be estimated as

$$\hat{\sigma}^2 = \int_0^\infty \frac{\bar{Y}_E(t)\bar{Y}_C(t)}{\bar{Y}(t)}\left(1-\frac{\Delta\bar{N}(t)-1}{\bar{Y}(t)-1}\right)\left(\frac{d\bar{N}_C(t)}{\bar{Y}(t)} + \frac{d\bar{N}_E(t)}{\bar{Y}(t)}\right), \qquad (2.3)$$

and hence the standardised test statistic is $U = \frac{W}{\hat{\sigma}}$.

The Cox model

The Cox model belongs to the class of survival regression models that are based on modeling the hazard function

$$\lambda(t) = \lambda_0(t) \exp(\beta X),$$

making different assumptions on the baseline hazard function $\lambda_0(t)$.

The appeal of the Cox model is its flexibility which it gains from leaving the baseline hazard function completely unspecified. It was introduced by Cox (1972) who developed the notion of a partial likelihood which made estimation of the coefficients β possible. In his original paper Cox derived the partial likelihood using a conditioning argument which gets overly complicated in the presence of ties. In his discussion of Cox's paper Breslow suggested a profile likelihood approach to overcome this obstacle which will be presented here following the presentation by van Houwelingen and Stijnen in the Handbook of survival (van Houwelingen and Stijnen, 2020).

As censored observations do not contain information on when the event occurred but on how long the subject was event-free their contribution to the likelihood is the survival function S, whereas non-censored events contribute the full information which can be expressed as $f = \lambda \cdot S$. The overall likelihood is hence given as

$$L(\lambda_0, \beta) = \prod_{i=1}^{n} S(t_i) \cdot \lambda(t_i)^{d_i} = \prod_{i=1}^{n} \exp\left(-\Lambda_0(t_i) \exp(\beta x_i)\right) \cdot \left(\lambda_0(t_i) \exp(\beta x_i)\right)^{d_i}$$

and the overall log-likelihood as

$$l(\lambda_0, \beta) = \sum_{i=1}^{n} \left[-\Lambda_0(t_i) \exp(\beta x_i) + d_i(\ln(\lambda_0(t_i)) + \beta x_i)\right].$$

As time intervals in which no events occur do not provide any information it is plausible to concentrate all the risk in the event times and assume that the baseline hazard is discrete with non-zero mass at the event times. The cumulative hazard can then be expressed as the sum $\Lambda_0(t_i) = \sum_t Y_i(t) h_0(t)$, where $Y_i(t) = \mathbb{1}(t_i > t)$ is the individual at-risk indicator of

observation i. For the second summand instead of summing over all individual observations one can sum over time and weighting the log-baseline hazard at each time point by the number of events at the same timepoint, i.e. $\sum_i d_i \ln(\lambda_0(t_i)) = \sum_t \Delta\bar{N}(t) \ln(\lambda_0(t))$.

Plugging this into the overall log-likelihood gives

$$\begin{aligned} l(\lambda_0, \beta) &= \sum_{i=1}^{n} [-\Lambda_0(t_i) \exp(\beta x_i) + d_i(\ln(\lambda_0(t_i)) + \beta x_i)] \\ &= \sum_{i=1}^{n} -\Lambda_0(t_i) \exp(\beta x_i) + \sum_{i=1}^{n} d_i \ln(\lambda_0(t_i)) + \sum_{i=1}^{n} d_i \beta x_i \\ &= \sum_{i=1}^{n} - \sum_t Y_i(t) \lambda_0(t) \exp(\beta x_i) + \sum_t \Delta\bar{N}(t) \ln(\lambda_0(t)) + \sum_{i=1}^{n} \sum_t \Delta N_i(t) \beta x_i \\ &= \sum_t \left(-\lambda_0(t) \sum_{i=1}^{n} Y_i(t) \exp(\beta x_i) + \Delta\bar{N}(t) \ln(\lambda_0(t)) + \sum_{i=1}^{n} \Delta N_i(t) \beta x_i \right) \end{aligned}$$

Breslow (1974) derived an estimator for the baseline hazard which maximizes this expression for fixed value of β and is given by

$$\hat{\lambda}_0(t|\beta) = \frac{\Delta\bar{N}(t)}{\sum_{i=1}^{n} Y_i(t) \exp(\beta x_i)}.$$

The resulting profile log-likelihood for β is then

$$l(\hat{\lambda}_0(\beta), \beta) = \underbrace{\sum_{i=1}^{n} \int_0^\infty \ln \left(\frac{\exp(\beta x_i)}{\sum_{j=1}^{n} Y_j(t) \exp(\beta x_j)} \right) dN_i(t)}_{=\mathrm{pl}(\beta)} + \sum_t (-\Delta\bar{N}(t) + \ln(\Delta\bar{N}(t)))$$

with $\mathrm{pl}(\beta)$ being Cox's partial likelihood. Now to obtain an estimator for β this profile likelihood needs to be maximized. Note, that by definition of the integrating process $N_i(t)$ the integral simplifies to evaluating the integrand at t_i so that the derivative of the partial likelihood becomes:

$$\begin{aligned} \frac{\partial \,\mathrm{pl}(\beta)}{\partial \beta} &= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \ln \left(\frac{\exp(\beta x_i)}{\sum_{j=1}^{n} Y_j(t_i) \exp(\beta x_j)} \right) \\ &= \sum_{i=1}^{n} \left[x_i - \frac{\sum_{j=1}^{n} Y_j(t_i) x_j \exp(\beta x_j)}{\sum_{j=1}^{n} Y_j(t_i) \exp(\beta x_j)} \right] \end{aligned}$$

Setting this first partial derivative to zero gives the so called score equation. As for standard MLE it can be shown that these estimators are asymptotically normal and inference is usually done using the Wald test for β.

Another approach is to perform a score test, i.e. to use the test statistic

$$\frac{\partial \, \mathrm{pl}(0)}{\partial \beta} = \sum_{i=1}^{n} \left[x_i - \frac{\sum_{j=1}^{n} Y_j(t_i) x_j}{\sum_{j=1}^{n} Y_j(t_i)} \right]$$

which interestingly is exactly the test statistic of a logrank test, if the covariate x_i is simply a binary treatment indicator.

As can be seen from the model formulation the Cox model implicitly relies on the proportional hazards assumption and a variety of approaches - both graphical and testing - has been suggested to check this. As for linear regression models the graphical approaches rely on residual plots, which in case of time to event data were introduced by Schoenfeld (1982) and hence called the Schoenfeld residuals. The idea of these residuals is based on the partial (Cox, 1975) or marginal (Kalbfleisch and Prentice, 1973) likelihood argument used to make inference for the Cox model as explained above. The Schoenfeld residuals r_i are defined as the difference between the observed value of x_i and its conditional expectation given the patients at risk at this time and are given by the summands of the score function, i.e.

$$r_i(\beta) = \left[x_i - \frac{\sum_{j=1}^{n} Y_j(t_i) x_j \exp(\beta x_j)}{\sum_{j=1}^{n} Y_j(t_i) \exp(\beta x_j)} \right]$$

and can be estimated by substituting the MLE $\hat{\beta}$ for β. By construction of the Schoenfeld residuals $\hat{r}_i := r_i(\hat{\beta})$ their expectation is $\mathbb{E}[\hat{r}_i] = 0$ so to assess if the proportional hazards assumption holds the these residuals should be plotted against t_i and if it does should be centered about 0. Furthermore, Schoenfeld (1982) has shown that the variance covariance matrix of $\hat{r}_i$ and $\hat{r}_j$ is asymptotically $\delta_{ij}\hat{U}_i - \hat{U}_i\hat{U}^{-1}\hat{U}_j^T$ where $\hat{U}_i$ is the Jacobian matrix of $r_i(\beta)$ evaluated at $\hat{\beta}$, i.e.

$$\hat{U}_i = \frac{\partial r_i}{\partial \beta}\Big|_{\hat{\beta}} = \frac{\sum_{j=1}^{n} Y_j(t_i) x_j x_j^T \exp(\hat{\beta} x_j)}{\sum_{j=1}^{n} Y_j(t_i) \exp(\hat{\beta} x_j)} - \left[\frac{\sum_{j=1}^{n} Y_j(t_i) x_j \exp(\hat{\beta} x_j)}{\sum_{j=1}^{n} Y_j(t_i) \exp(\hat{\beta} x_j)} \right] \left[\frac{\sum_{j=1}^{n} Y_j(t_i) x_j \exp(\hat{\beta} x_j)}{\sum_{j=1}^{n} Y_j(t_i) \exp(\hat{\beta} x_j)} \right]^T$$

and $\hat{U} = \sum_{i=1}^{n} d_i \hat{U}_i$.

Based on the work of Schoenfeld, Grambsch and Therneau (1994) proposed a goodness-of-fit test to test the null hypothesis of proportional hazards versus the alternative of time-varying coefficients of the form

$$\beta(t) = \beta + \theta G(t),$$

where G is a $p \times p$ diagonal matrix with possibly different transformations of the time scale for each predictor as diagonal entries $G_{jj}(t) = g_j(t)$. Now to test the null hypothesis $H_0 : \theta = 0$ Grambsch and Therneau (1994) derived the following test statistic based on generalized least squares

$$T(G) = \left(\sum_i G(t_i)\hat{r}_i \right)^T D^{-1} \left(\sum_i G(t_i)\hat{r}_i \right),$$

where

$$D = \sum_i G(t_i)\hat{U}_i G(t_i)^T - \left(\sum_i G(t_i)\hat{U}_i \right) \hat{U}^{-1} \left(\sum_i G(t_i)\hat{U}_i \right)^T$$

and which under H_0 is asymptotically χ^2 distributed with p degrees of freedom. They note that this test statistic is the same as the score test statistic of $H_0 : (\beta, \theta) = (\hat{\beta}, 0)$. Commonly used transformations $g_j(t)$ of the survival times, that are also implemented in the *cox.zph* function in the *survival* package in R (Therneau, 2023), are the Kaplan-Meier transform, the rank transform and the identity transform. For the simulation study the Kaplan-Meier transform was used which is the default in R.

2.3.3 Advanced inferential methods

In this section more advanced inferential methods for two-arm comparisons are introduced and grouped into the five categories: weighted logrank tests (W1-W10), combinations of weighted logrank tests (C1-C12), Kaplan-Meier based tests (K1-K4), tests based on regression models (R1-R7) and other tests (O1-O7).

Each item in the list starts with the name of the test, followed by the abbreviation used in this thesis in parantheses, if this abbreviation differs from the methods name. The description of each test contains the test statistic together with its distribution under the null and, if applicable, the local null hypothesis as well as a measure for the treatment effect. For methods

that were not self implemented the R package that was used is referenced. R program code implementing all methods can be found in an online repository referenced in the Appendix.

Weighted logrank tests

Introducing a weight function $K(t)$ to the logrank statistic leads to a generalization which places different weights on each event time

$$W_K = \int_0^\infty K(t) \frac{\bar{Y}_E(t)\bar{Y}_C(t)}{\bar{Y}(t)} \left(\frac{d\bar{N}_C(t)}{\bar{Y}_C(t)} - \frac{d\bar{N}_E(t)}{\bar{Y}_E(t)} \right).$$

The variance of this test statistic under H_0^{global} can be estimated as

$$\hat{\sigma}_K^2 = \text{Var}[W_K] = \int_0^\infty K^2(t) \frac{\bar{Y}_E(t)\bar{Y}_C(t)}{\bar{Y}(t)} \left(1 - \frac{\Delta\bar{N}(t) - 1}{\bar{Y}(t) - 1} \right) \left(\frac{d\bar{N}_C(t)}{\bar{Y}(t)} + \frac{d\bar{N}_E(t)}{\bar{Y}(t)} \right)$$

leading to the standardised version of this test statistic $U_K = \frac{W_K}{\hat{\sigma}_K}$ which is asymptotically standard normal distributed under H_0. It can be easily seen that for $K \equiv 1$ this is exactly the standard logrank test, which will hence be denoted as U_1.

Although the introduction of weights makes these tests more sensitive to detect specific alternatives, the null hypothesis that is assessed with these tests is the global null of equal survival curves. For an estimator of treatment effects the most direct approach is to exploit the relationship between the Cox model and the logrank test as the score test based on the partial likelihood of this model. To do so Schemper (1992) describes how the weights of the weighted logrank test can be introduced into the estimation procedure of the Cox model which results in the average hazard ratio.

Alternatively, Lin and León (2017) proposed to derive an effect estimate from a Cox model with a time-varying covariate $X^*(t) = A(t)X$, where $A(t) = \frac{K(t)}{\max(K(t))}$ is called the effect adjustment factor. Based on this Cox model the exponentiated coefficient e^β represents the full effect and the hazard ratio for the comparison of two groups can be expressed as $\text{HR}(t) = (e^\beta)^{A(t)}$.

The following weighted logrank statistics were considered in this thesis:

(W1) **Fleming-Harrington (G(.,.)):**

The Fleming-Harrington family of weighted logrank tests is given by $G(\rho, \gamma) := U_K$ with weight $K(t) = \hat{S}(t-)^\rho \cdot (1 - \hat{S}(t-))^\gamma$ for $\rho, \gamma \geqslant 0$ where $\hat{S}(t-)$ is the Kaplan-Meier estimate in the pooled sample (Fleming and Harrington, 1981). Since the Kaplan-Meier estimator is a monotone decreasing function with a range in $[0, 1]$ a non-zero ρ places more weight on early timepoints whereas a non-zero γ places more weight on late timepoints and if both parameters are non-zero then middle timepoints are upweighted. The following weights were suggested in literature:

$$(\rho, \gamma) = (0, 1), (1, 0), (1, 1), (0, 0.5), (0, 2), (0.5, 0.5)$$

(W2) **Gray-Tsiatis test (G(-1,0)):**

Gray and Tsiatis (1989) proposed to use the inverse of the Kaplan- Meier estimator in the pooled sample as weight, i.e. $K(t) = \hat{S}(t-)^{-1}$, when studying diseases with a non-zero probability of being cured. This corresponds to the Fleming- Harrington test with weight $(-1, 0)$.

(W3) **Modestly weighted logrank test (MWLRT):**

In their paper Magirr and Burman (2019) pointed out that there is an interesting correspondence between weighted logrank tests and so called score tests. This correspondence was first shown by Letón and Zuluaga (2001). To construct a score test let again $t_{(1)} < \ldots < t_{(k)}$ denote the k distinct, ordered event times and let l_{gj} denote the number of censored observations in group g that fall in the interval $[t_{(j)}, t_{(j+1)})$ and $l_j = l_{0j} + l_{1j}$, i.e. based on the definition of the number at risk

$$l_j = \bar{Y}(t_{(j)}) - \bar{Y}(t_{(j+1)}) - \Delta\bar{N}(t_{(j)}).$$

The score test then gives scores c_j to each uncensored and C_j to each censored observation resulting in the following test statistic

$$S = \sum_{j=1}^{k} c_j \Delta\bar{N}_0(t_{(j)}) + \sum_{j=1}^{k} C_j l_j.$$

As shown by Letón and Zuluaga (2001) the score test and the weighted logrank test with weights $K(t_{(j)})$ are equivalent if the following equivalent conditions are met

$$C_j = -\sum_{i=1}^{j} K(t_{(i)}) \frac{\Delta \bar{N}(t_{(i)})}{\bar{Y}(t_{(i)})}, \quad K(t_{(j)}) = c_j - C_j \tag{2.4}$$

$$K(t_{(j+1)}) = (K(t_{(j)}) + c_{j+1} - c_j) \frac{\bar{Y}(t_{(j+1)})}{(\bar{Y}(t_{(j+1)}) - \Delta \bar{N}(t_{(j+1)}))}, \quad K(t_{(j)}) = c_j - C_j, \tag{2.5}$$

where the first condition is used to translate $K(t_{(j)}) \to C_j \to c_j$ and the last to translate $c_j \to K(t_{(j)}) \to C_j$. As can be seen from (2.4) the scores c_j and C_j associated with the standard logrank test are nonincreasing, which means that later events get lower, that is, "better" scores than early events.

In a more recent paper Magirr (2021) showed that the associated scores of the Fleming-Harrington test $G(0,1)$ are, contrary to the standard logrank test, increasing over time which is not acceptable.

To resolve this problematic behavior they proposed to start with a score statistic and keep the scores $c_j = 1$ fixed for events prior to a prespecified time t^*. After this time the weight is kept constant at $K(t_{(j^*)})$ where $t_{(j^*)} := \max\{t_{(j)} \mid t_{(j)} < t^*\}$. Using formula (2.4) this test can be formulated as a weighted logrank statistic and it downweights events before t^* and keeps the weight constant afterwards. This results in the weights K to be

$$K(t) = \frac{1}{\max\left(\hat{S}(t), \hat{S}(t^*)\right)},$$

where $\hat{S}$ is again the Kaplan-Meier estimate in the pooled sample. This method has been implemented in R within the *modestWLRT* package (Magirr, 2022).

(W4) **Threshold lag (Thres) and generalized linear lag (GenLin) test**:

Xu et al. (2017, 2018) investigated different modelling assumptions for the delay in cancer immunotherapies and derive optimal weights for these two models based on the result by Peto and Peto that optimal weights need to be proportional to the true logarithmic hazard ratio. The first model (Xu et al., 2017) assumes that the delay is

fixed at t_0 for all patients and the hazard ratio is of the form

$$\mathrm{HR}(t) = \frac{\lambda_E(t)}{\lambda_C(t)} = \begin{cases} 1, & t \leqslant t_0 \\ \theta, & t > t_0 \end{cases}$$

with the treatment effect $\theta < 1$. The optimal weights then result in the threshold lag test with $K(t) = \mathbb{1}(t \geqslant t_0)$ and the delay t_0 must be presepcified.

The second model (Xu et al., 2018) relaxes the assumption that the delay is fixed for all patients to a subject specific delay time t^*_{ind} which follows a distribution F_* on the interval $[T_1, T_2]$. Conditional on the observed subject-specific delay time the hazard ratio is the same as for the first model, i.e.

$$\mathrm{HR}(t \mid t^*_{\mathrm{ind}}) = \frac{\lambda_E(t \mid t^*_{\mathrm{ind}})}{\lambda_C(t \mid t^*_{\mathrm{ind}})} = \begin{cases} 1, & t \leqslant t^*_{\mathrm{ind}} \\ \theta, & t > t^*_{\mathrm{ind}}. \end{cases}$$

Integrating over the distribution of t^*_{ind} Xu et al. show that the marginal hazard ratio is given by

$$\mathrm{HR}(t) = \begin{cases} 1, & t \leqslant T_1 \\ \theta^{g(t)}, & T_1 < t < T_2 \\ \theta, & T_2 \leqslant t \end{cases}$$

where $g(t)$ is a monotone, increasing function that converges to the cumulative distribution function F_*. Hence, the asymptotically optimal weights are given by

$$K(t) = \begin{cases} 0, & t \leqslant T_1 \\ F_*(t), & T_1 < t < T_2 \\ 1, & T_2 \leqslant t \end{cases}$$

and in the special case of the delay time being uniformly distributed on $[T_1, T_2]$ the asymptotically optimal weight in this interval is given by $F_*(t) = \frac{t-T_1}{T_2-T_1}$. Based on this model the generalized linear lag test uses the weights $K(t) = \frac{t-T_1}{T_2-T_1} \cdot \mathbb{1}(T_1 < t < T_2) + \mathbb{1}(T_2 \leqslant t)$.

(W5) **Gehan-Breslow (GB)**:

The modified Wilcoxon test statistic was derived by Gehan (1965) and Breslow (1970) and places more weight on early timepoints by considering the number at risk in the pooled sample as weight function, i.e. $K(t) = \bar{Y}(t)$.

(W6) **Tarone-Ware (TW):**

The Tarone-Ware test is very similar to the Gehan-Breslow test but with $K(t) = \sqrt{\bar{Y}(t)}$ (Tarone and Ware, 1977).

(W7) **Peto-Peto (PP) and modified Peto-Peto (mPP):**

The weights of these tests are derived from the Kaplan-Meier product estimator and are chosen to be $K(t) = \prod_{i:\, t_i \leqslant t} \left(1 - \frac{\Delta N(t_i)}{\bar{Y}(t_i)+1}\right)$ or $K(t) = \prod_{i:\, t_i \leqslant t} \left(1 - \frac{\Delta N(t_i)\bar{Y}(t_i)}{(\bar{Y}(t_i)+1)^2}\right)$, respectively (Peto and Peto, 1972).

(W8) **Asymptotic logrank test (asymLR):**

Moreau et al. (1992) proposed the asymptotic logrank test with weights given by $K(t) = 1 + \log\left(-\log\left(\prod_{i:\, t_i \leqslant t} \frac{\bar{Y}(t_i)}{\bar{Y}(t_i)+\Delta N(t_i)}\right)\right)$

(W9) **Logit and modified Logit (mLogit):**

The logit weight function $K(t) = \text{Logit}_{a,\tau}(t) = \frac{\exp(a(t-\tau))}{(1+\exp(a(x-\tau)))}$ with parameters τ defined as the middle point of the prespecified transition period $[t_1, t_2]$ was introduced by Yu et al. (2021). The scaling parameter a is chosen such that the prespecified weight $w = 0.1$ at t_1 (and due to symmetry $1 - w = 0.9$ at t_2) is achieved. Based on this a modified logit weight version is defined as $K(t) = \frac{\text{Logit}_{a,\tau}(t) - \text{Logit}_{a,\tau}(t_1)}{\text{Logit}_{a,\tau}(t_2) - \text{Logit}_{a,\tau}(t_1)}$

(W10) **Maximin efficiency robust test (MERT):**

The MERT test tries to find the weighted logrank statistic which maximizes the minimum asymptotic relative efficiency over a class of lag functions (Ye and Yu, 2018). This class is denoted with $\mathcal{L}(\tilde{t}_1, \tilde{t}_2)$ and contains all lag functions that are 0 before $\tilde{t}_1$, monotone and nondecreasing in $[\tilde{t}_1, \tilde{t}_2)$ and 1 lateron. The asymptotic relative efficiency of W_K to the optimal test W_l is denoted as $\rho^2(W_K, W_l)$ and the weight of the MERT test as

$$W_{\tilde{t}_1,\tilde{t}_2} = \arg\max_K \min_{l \in \mathbf{L}(\tilde{t}_1,\tilde{t}_2)} \rho^2(W_K, W_l).$$

It was shown by Ye and Yu (2018) that this weight function is given as follows:

$$W_{\tilde{t}_1,\tilde{t}_2}(t) = \left(\frac{\Psi(\tau) - \Psi(t)}{\Psi(\tau) - \Psi(\tilde{t}_1)}\right)^{-1/2} \mathbb{1}(\tilde{t}_1 \leqslant t \leqslant \tilde{t}_2) + 2\left(\frac{\Psi(\tau) - \Psi(\tilde{t}_2)}{\Psi(\tau) - \Psi(\tilde{t}_1)}\right)^{-1/2} \mathbb{1}(t > \tilde{t}_2)$$

where τ is the study duration and $\Psi(t)$ is estimated by

$$\hat{\Psi}(t) = \frac{1}{n}\int_0^t \frac{\bar{Y}_C(u)\bar{Y}_E(u)}{\bar{Y}^2(u)} d\bar{N}(u).$$

Combinations of weighted logrank statistics

Considering a specific weight has the disadvantage that the test is tailored to detect a very specific alternative and has low power to detect a difference in situations where the distributions are different but follow another alternative. To overcome this drawback mutiple procedures have been suggested that combine different weighted logrank statistics. Most of these methods inccorporate statistics of the Fleming-Harrington family $G(\rho, \gamma)$ (W1) for different choices of ρ and γ.

To make inference in this situation one needs to know the multivariate distribution of the weighted logrank statistics under consideration.

By Theorem 7.5.1 in Fleming and Harrington (Fleming and Harrington, 1991, chapter 7, pg. 278-80) it can be shown that for different weight functions $K_1, \ldots, K_m$ the test statistics $(W_{K_1}, \ldots, W_{K_m})$ are m-variate normally distributed $\mathcal{N}(0, \Sigma_m = (\sigma_{lj})_{lj=1,\ldots,m})$ with estimated covariance between W_{K_l} and W_{K_j}

$$\hat{\sigma}_{lj} = \int_0^\infty K_l(t)K_j(t)\frac{\bar{Y}_E(t)\bar{Y}_C(t)}{\bar{Y}(t)}\left(1 - \frac{\Delta\bar{N}(t) - 1}{\bar{Y}(t) - 1}\right)\left(\frac{d\bar{N}_C(t)}{\bar{Y}(t)} + \frac{d\bar{N}_E(t)}{\bar{Y}(t)}\right)$$

under the null hypothesis of equality of the survival distributions H_0^{global}. Consequently, the standardised test statistics $(U_{K_1}, \ldots, U_{K_m})$ are also m-variate normally distributed $\mathcal{N}(0, P_m = (\rho_{lj})_{lj=1,\ldots,m})$ and the correlation can be estimated as $\hat{\rho}_{lj} = \frac{\hat{\sigma}_{lj}}{\hat{\sigma}_{ll}\hat{\sigma}_{jj}}$.

This multivariate normal distribution can be used to calculate the p-value or critical value of a test based on combinations of weighted logrank statistics.

In their article Roychoudhury et al. (2021) note that for the MaxCombo test the average hazard ratio can be calculated as an effect measure as explained for single weighted logrank tests. The weight used to calculate the AHR is then the weight corresponding the weighted

logrank statistic which maximized the MaxCombo test. This can be extended to every other combination test that uses the maximum of weighted logrank tests. However, the interpretation of this measure is difficult and not applicable to all combinations of weighted logrank tests so following the recommendation of Roychoudhury et al. (2021) one should use supportive measures such as the Kaplan-Meier plot and milestone survival rates.

(C1) **Adaptively weighted logrank test (YP):**

Yang and Prentice (2005) introduced the following semiparametric model for the hazards of two samples

$$\lambda_E(t) = \frac{\theta_1\theta_2}{\theta_1 + (\theta_2 - \theta_1)S_C(t)}\lambda_C(t) \quad (t < \tau_0),$$

where $\tau_0 = \sup\{t\colon S_C(t) > 0\}$ and $\theta_1, \theta_2 > 0$. Since the survival function has the property that $S_C(0) = 1$ and $S_C(\tau_0) = 0$ the parameter θ_1 and θ_2 can be interpreted as the short-term and long-term hazard ratios, respectively. To derive an estimator for this hazard ratio, the modelling equation will be reformulated in terms of the odds function of the control group defined as $R(t) = \frac{1-S_C(t)}{S_C(t)}$, whose derivative is given by

$$\frac{dR(t)}{dt} = \frac{-S_C'(t)S_C(t) - S_C'(t)(1 - S_C(t))}{S_C^2(t)} = \frac{\lambda_C(t)}{S_C(t)} = \lambda_C(t)(1 + R(t)).$$

Combining this the model translates to

$$\begin{aligned}
\lambda_E(t) &= \frac{\theta_1\theta_2}{\theta_1 + (\theta_2 - \theta_1)S_C(t)}\lambda_C(t) \\
&= \frac{\theta_1\theta_2}{\frac{\theta_1}{S_C(t)} + (\theta_2 - \theta_1)}\frac{\lambda_C(t)}{S_C(t)} \\
&= \frac{\theta_1\theta_2}{\theta_1(\frac{1}{S_C(t)} - 1) + \theta_2}R'(t) \\
&= \frac{\theta_1\theta_2}{\theta_1 R(t) + \theta_2}R'(t) \\
&= \left(\frac{R(t)}{\theta_2} + \frac{1}{\theta_1}\right)^{-1}R'(t) \\
&= \frac{1}{(\exp(-\beta_2)R(t) + \exp(-\beta_1))}R'(t),
\end{aligned}$$

where $\beta_j = \log\theta_j$ for $j = 1, 2$.

Based on this Yang and Prentice derive pseudo maximum likelihood estimators for $\beta = (\beta_1, \beta_2)$ by constructing an estimator $\hat{R}(t;\beta)$ for fixed β. Now if $L(\beta, R)$ denotes the likelihood function based on this model, the pseudo maximum likelihood estimator for β is defined as $\hat{\beta} = \arg\max_\beta L(\beta, \hat{R}(t;\beta))$ or equivalently as the zero of the score

$$Q(\beta) = \frac{\partial \log L(\beta, R)}{\partial \beta}\Big|_{R(t)=\hat{R}(t;\beta)}.$$

This approach is very similar to the approach of the Cox model where first the unknown baseline hazard in the likelihood function is replaced by the Breslow estimator for fixed coefficients β and then maximized in β. The difference is, however, that in the approach for the Cox model the Breslow estimator maximizes the likelihood for fixed β, which leads to a profile likelihood, and the estimator for the odds function R in the Yang and Prentice model has no such optimality property. Nevertheless, Yang and Prentice showed that under certain regularity conditions, the pseudo maximum likelihood estimator $\hat{\beta}$ is consistent for β and that first the score $\sqrt{n}Q(\beta)$ converges in distribution to a zero mean bivariate normal distribution with covariance matrix $V(\beta)$ and second $\sqrt{n}(\hat{\beta} - \beta)$ converges in distribution to a bivariate normal distribution with zero mean and covariance matrix $U(\beta)$.

Based on this model Yang and Prentice (2010) suggested to use an adaptively weighted logrank test, with the estimated hazard ratio as weight, i.e. the weights are

$$\Phi_2(t) = \frac{1 + \hat{R}(t;\hat{\beta})}{\exp(-\hat{\beta}_1) + \exp(-\hat{\beta}_2)\hat{R}(t;\hat{\beta})} \text{ and}$$
$$\Phi_1(t) = \tfrac{1}{\Phi_2(t)}.$$

The test then uses a combination of the weighted logrank statistics U_{Φ_1} and U_{Φ_2}.

Since under non-proportional hazards either one of these weights can be expected to be more sensitive to departures from the null hypothesis, they proposed the following combination of these statistics $\max(|U_{\Phi_1}|, |U_{\Phi_2}|)$. Estimation procedure of this model has been implemented in the *YPmodel* package in R (Yang and Prentice, 2011) and was used to construct the test statistics.

(C2) **Maximum combination test (MaxCombo):**

The Max-Combo test is defined as $Z_{\max} = \max(|G(0,0)|, |G(0,1)|, |G(1,0)|, |G(1,1)|)$.

Based on the observed realization $Z_{\max} = z$ of this test statistic the p-value is given by

$$\begin{aligned} p &= \mathbb{P}_{H_0}[\max(|G(0,0)|, |G(0,1)|, |G(1,0)|, |G(1,1)|) \geqslant z] \\ &= 1 - \mathbb{P}_{H_0}[\max(|G(0,0)|, |G(0,1)|, |G(1,0)|, |G(1,1)|) \leqslant z] \\ &= 1 - \mathbb{P}_{H_0}[|G(0,0)| \leqslant z, |G(0,1)| \leqslant z, |G(1,0)| \leqslant z, |G(1,1)| \leqslant z] \\ &= \int_{-z}^{z}\int_{-z}^{z}\int_{-z}^{z}\int_{-z}^{z} \phi_{(0,P_4)}(x_1, x_2, x_3, x_4)dx_1dx_2dx_3dx_4, \end{aligned}$$

where $\phi_{(\mu,\Sigma)}$ denotes the density of a multivariate normal distribution with mean vector μ and covariance matrix Σ.

(C3) **Modified MaxCombo test (mMaxCombo):**

Cheng and He (2021) proposed a modified version of the MaxCombo test to improve power in detecting crossing hazards by replacing $G(1,1)$ by a weighted logrank test with weight

$$K_\theta(t) = \mathbb{1}(\hat{S}(t-) \leqslant \theta)\frac{\hat{S}(t-) - \theta}{\theta} + \mathbb{1}(\hat{S}(t-) > \theta)\frac{\hat{S}(t-) - \theta}{1-\theta}.$$

Here the parameter $\theta \in (0,1)$ represents the survival rate at the point where the crossing occurs. If no previous data is available one should chose the uninformative parameter $\theta = 0.5$, reducing to the weight $K_{0.5}(t) = 2\hat{S}(t-) - 1$.

(C4) **Zm3 test:**

The Zm3 test is defined as $Z_{m_3} = \max(|G(0,0)|, |G(0,1)|, |G(1,0)|)$ (Karrison, 2016). Similar to the Max-Combo test the p-value is given by

$$p = \int_{-z}^{z}\int_{-z}^{z}\int_{-z}^{z} \phi_{(0,P_3)}(x_1, x_2, x_3)dx_1dx_2dx_3.$$

(C5) **Modified Zm3 test (mZm3):**

Royston and Parmar (2020) proposed a modified version of Z_{m_3}, which they call "modified versatile weighted logrank test" replacing $G(1,0)$ by a weighted logrank test with weight $K(t) = \max\left(0.001, \frac{\hat{S}(t-) - \min(\hat{S}(t-))}{1 - \min(\hat{S}(t-))}\right)$. The reason to do so is that if an early effect with a low event rate is present the weights of $G(1,0)$ are too close to 1 and hence $G(1,0)$ closely resembles the standard logrank test which diminishes the gain in power.

(C6) **Versatile tests by Lee (Lee1, Lee2, Lee3):**

Lee (2007) proposed the following versatile combinations of the Fleming-Harrington test statistics

$$\frac{|G(0,1)+G(1,0)|}{2},$$
$$\frac{|G(0,1)|+|G(1,0)|}{2}, \text{ and}$$
$$\max(|G(0,1)|,|G(1,0)|),$$

which are abbreviated with Lee1, Lee2 and Lee3, respectively.

The critical values or equivalently p-values can be derived from the multivariate distribution as follows.

- Let z be the observed realization of the Lee1 test statistic. With ρ being the correlation between $G(0,1)$ and $G(1,0)$ it then holds asymptotically that $\binom{G(0,1)}{G(1,0)} \sim \mathcal{N}\left(\binom{0}{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. Hence by the transformation rule (asymptotically)

$$G(0,1)+G(1,0) = (1,1)\cdot\begin{pmatrix} G(0,1) \\ G(1,0) \end{pmatrix} \sim \mathcal{N}\left(0, 2(1+\rho)\right).$$

The p-value is then given by

$$\begin{aligned} p &= \mathbb{P}_{H_0}\left[\frac{|G(0,1)+G(1,0)|}{2} \geqslant z\right] \\ &= 1-\mathbb{P}_{H_0}[|G(0,1)+G(1,0)| \leqslant 2z] \\ &= 1-\mathbb{P}_{H_0}[-2z \leqslant G(0,1)+G(1,0) \leqslant 2z] \\ &= 1-\Phi\left(\frac{2z}{\sqrt{2(1+\rho)}}\right)+\Phi\left(\frac{-2z}{\sqrt{2(1+\rho)}}\right) \\ &= 2\Phi\left(\frac{-2z}{\sqrt{2(1+\rho)}}\right) \end{aligned}$$

- Let z be the observed realization of the Lee2 test statistic and let ρ again denote the correlation coefficient between $G(0,1)$ and $G(1,0)$. The conditional distribution

in this case is then given by

$$G(0,1) \mid G(1,0) \sim \mathcal{N}(\rho G(1,0), 1-\rho^2)$$

Applying the law of total probability the p-value can then be calculated as

$$\begin{aligned}
p &= \mathbb{P}\left[\frac{|G(0,1)| + |G(1,0)|}{2} \geqslant z\right] \\
&= 1 - \mathbb{P}[|G(0,1)| + |G(1,0)| \leqslant 2z] \\
&= \int_{-2z}^{2z} \mathbb{P}[|G(0,1)| + |G(1,0)| \leqslant 2z \mid G(1,0) = w] \cdot \varphi(w) dw \\
&= \int_{-2z}^{2z} \mathbb{P}[-2z + |w| \leqslant |G(0,1)| \leqslant 2z - |w| \mid G(1,0) = w] \cdot \varphi(w) dw \\
&= \int_{-2z}^{2z} \left[\Phi\left(\frac{2z - |w| - \rho w}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{-2z + |w| - \rho w}{\sqrt{1-\rho^2}}\right)\right] \cdot \varphi(w) dw
\end{aligned}$$

- Similar to the Max-Combo test the p-value for an observed realization z of the Lee3 test is given by

$$p = \int_{-z}^{z} \int_{-z}^{z} \phi_{(0,P_2)}(x_1, x_2) dx_1 dx_2.$$

(C7) **Modified Lee tests (mLee2, mLee3):**

Yang and Zhao (2007) considered the sum $|G(0,0)| + |G(1,0)|$ as a modification of the Lee2 test. Callegaro and Spiessens (2017) proposed to use the standard logrank test $G(0,0)$ instead of the $G(1,0)$ statistic in the versatile test Lee3.

(C8) **Projection test (ProjTest):**

Additionally to the modified Max-Combo test Cheng and He (2021) proposed a projection-type test based on the weighted logrank statistics $(G(0,0), G(1,0), U_{K_{0.5}})$ which are trivariate normally distributed and let P denote their correlation matrix. The idea of projection-type tests was introduced by Brendel et al. (2014) and the test can be constructed as

$$(G(0,0), G(1,0), W_{K_{0.5}}) \cdot \hat{P}^{-1} \cdot (G(0,0), G(1,0), W_{K_{0.5}})^t$$

which is $\chi^2_{\text{rank}(\hat{P}^{-1})}$ distributed and $\hat{P}^{-1}$ denotes the Moore-Penrose inverse of $\hat{P}$.

(C9) **Modified Score test (mScore):**

Similarly, one can construct a projection test based on the bivariate normal distribution of the logrank test and the weighted logrank test with weight $K(t) = \log(1+\hat{\Lambda}_{NA}(t-))$. This approach was proposed by Bagdonavicius et al. (2004).

(C10) **V0 test:**

As explained above, Ye and Yu (2018) have derived the MERT test over the class $\mathcal{L}(\tilde{t}_1, \tilde{t}_2)$ of lag functions. In a previous work Zucker and Lakatos (1990) derived the maximin efficiency robust statistic V^* over a different family $\mathcal{L}(t_2^*)$ of all functions which are monotone and nondecreasing on $[0, t_2^*]$ and equal to one afterwards. They had shown that this statistic can be expressed as the limit as $k \to \infty$ of the linear combination statistic:

$$V_k = U_1 + U_{w_{t_2^*}} + (1 - \rho^{1/2^k}) \sum_{j=1}^{2^k-1} U_{w_{t_{kj}}},$$

where $\rho^2 = \rho^2(U_1, U_{w_{t_2^*}}) = \frac{\Psi(\tau)-\Psi(t_2^*)}{\Psi(\tau)}$, $t_{kj} = \Psi^{-1}(\Psi(\tau)(1-\rho^{j/2^{k-1}}))$ and $U_{w_{\tilde{t}}}$ denotes the test statistic of a weighted logrank test with threshold lag $w_{\tilde{t}}(t) = \mathbb{1}(t > \tilde{t})$. This representation inspired Ding and Wu (2020) to consider the sum of logrank and threshold lag test as an approximated version of the MERT test called V0 test.

(C11) **Partially grouped logrank test (ParGroup):**

The partially grouped logrank test proposed by Sposto et al. is a hybrid between the fixed-time analysis and the usual logrank statistics (Sposto et al., 1997). Given a predetermined grouping time t_c and the numbers N_E, N_C initially at risk one defines an analogous to the number of events at t_c as:

$$\Delta\tilde{N}_E(t_c) = N_E \cdot (1 - \hat{S}_E(t_c))$$
$$\Delta\tilde{N}_C(t_c) = N_C \cdot (1 - \hat{S}_C(t_c))$$

As for the logrank statistic one can determine that the distribution of $\Delta\tilde{N}_C(t_c)$ given $\Delta\tilde{N}(t_c)$ is approximately normal. For times after t_c the logrank statistic is employed, i.e. a weighted logrank test with threshold lag $K_{t_c}(t) = \mathbb{1}(t \geqslant t_c)$. The total test

statistic is then given as

$$\text{ParGroup} = \frac{\left[\left(\Delta\tilde{N}_C(t_c) - E\left(\Delta\tilde{N}_C(t_c) \mid \Delta\tilde{N}(t_c)\right)\right) + W_{K_{t_c}}\right]^2}{\text{Var}\left(\Delta\tilde{N}_C(t_c) \mid \Delta\tilde{N}(t_c)\right) + \text{Var}\left(W_{K_{t_c}}\right)}$$

(C12) **Kolmogorov-Smirnov tests (KS LR, KS FH, KS GB, KS Cheng):**
Instead of taking the integral over the whole time period, one could also restrict the integration interval to $[0, t]$, which makes W_K a time-dependent test statistic $W_K(t)$. The idea of the so called Kolmogorov-Smirnov or Renyi type test statistics is to take the supremum over all these test statistics $\sup_t W_K(t)/\hat{\sigma}_K$, which can be shown to be asymptotically distributed as the supremum of a Brownian motion $\sup_{0\leqslant t\leqslant 1} B(t)$ (Fleming et al., 1987). Inference can then be made using this asymptotic distribution and the following result given by Fleming et al. (1987). Given a Brownian motion B then

$$P\left[\sup_{0\leqslant t\leqslant 1} |B(t)| > x\right] = 1 - \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\exp\left(-\frac{\pi^2(2k+1)^2}{8x^2}\right)$$

Alternatively, one can choose a permutation approach. Cheng and He (2021) considered the following weighted logrank tests: standard logrank, Gehan-Breslow (GB), $G^{1,0}$ and the modified weighted logrank test $K_{0.5}(t)$ introduced by Cheng for the modified MaxCombo test.

Kaplan-Meier based tests

These tests are all based on the Kaplan-Meier estimator of the survival function

$$\hat{S}(t) = \prod_{s \leqslant t} \left(1 - \frac{\Delta \bar{N}(s)}{\bar{Y}(s)}\right),$$

which was already introduced at the beginning of Section 2.2.1. It has to be noted that this estimator drops to 0 only if at the last event time all patients in the risk set have an event. Otherwise, the Kaplan-Meier estimator is only defined up to this point and extensions beyond this point are invalid. The variance of the Kaplan-Meier estimator is given by Greenwood's formula (2.1). Based on this estimator and its variance one can construct the following tests:

(K1) **Milestone survival test (Mile, MileCLL):**

Milestone survival tests assess the null hypothesis H_0^{Mile}: $S_C(t_0) = S_E(t_0)$ of equal survival rates at a fixed timepoint t_0 by using Kaplan-Meier estimates: $\hat{S}_C(t_0) - \hat{S}_E(t_0)$. The test statistic is then given by

$$\frac{(\hat{S}_C(t_0) - \hat{S}_E(t_0))^2}{\widehat{\text{Var}}[\hat{S}_C(t_0)] + \widehat{\text{Var}}[\hat{S}_E(t_0)]}$$

which under the null has an asymptotic χ^2 distribution with one degree of freedom. Klein et al. (2007) have been shown that applying a transformation to the Kaplan-Meier estimates can give a better asymptotic and hence the identity (Mile) and cloglog (MileCLL) transformation were considered. The natural effect measure is then the difference of the survival rates.

(K2) **Weighted Kaplan-Meier test (WKM):**

The weighted Kaplan-Meier test compares the integrated weighted difference between the Kaplan-Meier estimates, i.e. for a given estimator $\hat{w}$ of a deterministic weight function $w(t)$ and a truncation time τ the test statistic is given as

$$\text{WKM}(\tau) = \sqrt{\frac{n_0 n_1}{n}} \int_0^\tau \hat{w}(t) \left(\hat{S}_E(t) - \hat{S}_C(t)\right) dt.$$

Pepe and Fleming (1989) showed that under certain constraints on the weight function and its estimator, that will ensure stability of the WKM statistic, it is asymptotically

normally distributed under the global null hypothesis H_0^{global}: $S_C(t) = S_E(t) = S(t)$, i.e. $\text{WKM}(\tau) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\tau))$, where

$$\sigma^2(\tau) = -\int_0^\tau \frac{\left(\int_t^\tau w(u)S(u)du\right)^2}{S^2(t)} \frac{n_C S_{\text{cens},E}(t) + n_E S_{\text{cens},C}(t)}{nS_{\text{cens},C}(t)S_{\text{cens},E}(t)} dS(t)$$

The weight function considered by Pepe and Fleming (1989) is based on the Kaplan-Meier estimator for the censoring times $\hat{S}_{\text{cens}}$ in both groups

$$w(t) = \frac{n\hat{S}_{\text{cens},C}\hat{S}_{\text{cens},E}}{n_C\hat{S}_{\text{cens},E} + n_E\hat{S}_{\text{cens},C}}.$$

Within this thesis the implemented *wkm.Stat* function in the *nphsim* package in R (Wang et al., 2017) was used.

(K3) **Restricted mean survival time tests (RMST, MaxRMST):**

In the special case of $w \equiv 1$ the weighted Kaplan-Meier statistic reduces to the difference of the estimates of the so called restricted mean survival time $\text{RMST}(\tau)$.

The RMST can be interpreted as the mean of the survival time $\min(T, \tau)$ and hence be expressed as $\text{RMST}(\tau) = E[\min(T, \tau)] = \int_0^\tau S(t)dt$ and naturally estimated by plugging in the Kaplan-Meier estimator. As outlined by Huang and Kuan (2018) the estimator $\widehat{\text{RMST}}(\tau) = \int_0^\tau \hat{S}(t)dt$ is approximately normal and making use of the Greenwood formula its variance can be estimated by

$$\text{Var}[\widehat{\text{RMST}}(\tau)] = \sum_{t_i \leqslant \tau} \left[\int_{t_i}^\tau \hat{S}(t)dt\right]^2 \frac{\Delta\bar{N}(t_i)}{\bar{Y}(t_i)(\bar{Y}(t_i) - \Delta\bar{N}(t_i))}.$$

Taking all this together the null hypothesis of equal restricted mean survival times H_0^{RMST}: $\text{RMST}_C(\tau) = \text{RMST}_E(\tau)$ can be tested with the test statistic

$$\frac{\int_0^\tau \left(\hat{S}_E(t) - \hat{S}_C(t)\right) dt}{\text{Var}[\widehat{\text{RMST}}_C(\tau)] + \text{Var}[\widehat{\text{RMST}}_E(\tau)]},$$

which is asymptotically standard normally distributed under the null hypothesis. The appeal of the RMST approach, which made it popular in the statistics community lately, is that the difference in RMST is an easily interpretable and natural effect measure.

For the simulation study the *rmst2* function of the *survRM2* package (Uno et al., 2022) has been adapted.

Furthermore, Royston and Parmar (2016) suggested to use the maximum RMST difference over all possible cutoff values τ and make inference through a permutation test approach. To obtain the maximum RMST difference they found that a grid search within the interval of the 30th percentile of the event times and the largest (uncensored) event time gives reasonable results and recommend a number of 10 points for this grid.

(K4) **Area between curves test (ABC):**

The area between curves test statistic is very similar to that of the RMST and defined by

$$T_n = \sqrt{n} \int_0^\tau \left| \hat{S}_E(t) - \hat{S}_C(t) \right| dt,$$

where τ denotes the largest censoring time. Lin and Xu (2010) postulated asymptotic normality of the standardized statistic $(T_n - \hat{E}(T_n))/\sqrt{\widehat{\mathrm{Var}}(T_n)}$ under the global null hypothesis of equal survival curves H_0^{global}. Liu et al. (2020) used a resampling approach to approximate the residuals $\hat{S}_g(t) - S_g(t)$ in each group ($g = E, C$) and with that estimated the test statistic T_n, which - under the null hypothesis - can be written as

$$T_n = \sqrt{n} \int_0^\tau \left| \left(\hat{S}_E(t) - S_E(t) \right) - \left(\hat{S}_C(t) - S_C(t) \right) \right| dt.$$

For the simulation study the *LinStatABC* function of the *RBT4TCSC* package available on github (`https://github.com/LTTGH/RBT4TCSC.git`) has been adapted.

Tests based on regression models

Many of the following models are based on modeling the hazard function as

$$\lambda(t) = \lambda_0(t) \exp(\beta X).$$

As elaborated in Section 2.3.2 the Cox model leaves the baseline hazard function $\lambda_0(t)$ completely unspecified and the coefficients β are estimated via the partial likelihood approach. There are various approaches to extend or generalize the standard Cox model by specifying a baseline hazard function, introducing a function to weigh the observations or allowing the covariates or coefficients to vary with time. In addition, other alternatives are considered that model either the cumulative hazard function or the failure time directly.

(R1) **Average hazard ratio test (AHR):**

The AHR is obtained by introducing a weight function $w(t)$ to weigh the contributions to the log partial likelihood and derive the weighted maximum likelihood estimates as solutions to the weighted score equations

$$\sum_{i=1}^{n} w(t_i) \left[x_i - \frac{\sum_{j=1}^{n} \mathbb{1}(t_j > t_i) x_j \exp(\beta x_j)}{\sum_{j=1}^{n} \mathbb{1}(t_j > t_i) \exp(\beta x_j)} \right] = 0.$$

Schemper et al. (2009) proposed to use $w(t) = \frac{\hat{S}(t)}{\hat{S}_{cens}(t)}$, where the inverse probability of censoring weighting is used to compensate the attenuation in observed events due to earlier censoring. The MLE $\hat{\beta}$ is then an estimator of the average hazard ratio (AHR) and inference of the null hypothesis $H_0 \colon \beta = 0$ which is equivalent to the global null hypothesis H_0^{global} can be made by Wald-type tests for which robust estimators for the variance of $\hat{\beta}$ are available. This approach is implemented in the *coxphw* function in the *coxphw* package in R (Dunkler et al., 2018).

(R2) **Landmark test:**

The landmark model considers only events after a prespecified assumed landmark time t_{Landmark} by resetting the origin to this time, i.e. only observations with $t_i > t_{\text{Landmark}}$ are considered with observed time $\tilde{t}_i = t_i - t_{\text{Landmark}}$. Based on this transformed data the usual inference is made with the Wald test of the Cox model targeting the local null

hypothesis that the survival distributions are equal beyond the Landmark timepoint H_0^{Landmark}: $S_C(t) = S_E(t)$ for all $t > t_{\text{Landmark}}$.

(R3) **Time-dependent treatment effect test (CoxTD):**

If it can not be assumed that the effect of X is constant over time, one can employ a time-varying effect $\beta(t)$. by adding an interaction between X and $\log(t+1)$ which yields the model

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 X + \beta_2 \log(t+1) X).$$

This then becomes a model with a time-dependent coefficient $\beta(t) = \beta_1 + \beta_2 \log(t+1)$ which can be estimated using the partial likelihood approach. The resulting estimator $\hat{\beta}(t) = \hat{\beta}_1 + \hat{\beta}_2 \log(t+1)$ represents the time-dependent logarithmic hazard ratio. In this time-dependent Cox model inference can be made by using the likelihood ratio test to test the null hypothesis H_0: $(\beta_1, \beta_2) = 0$ which is equivalent to the global null hypothesis H_0^{global}. This can be done in R by using the time-transform functionality of the *coxph* function in the *survival* package (Therneau, 2023).

(R4) **Piecewise exponential (lag) model test (PWExp, PWExpLag):**

For the piecewise exponential models one assumes that given a partition of the time interval $0 = t_0 < t_1 < t_2 = \infty$, the baseline hazard is constant within the interval $[t_{j-1}, t_j)$ $(j = 1, 2)$. The resulting model is then given by

$$\lambda(t) = \left(\sum_{j=1}^{2} \lambda_j \mathbb{1}(t_{j-1} \leqslant t < t_j) \right) \exp(\beta X)$$

which is fully parametric. An estimate of the treatment effect $\hat{\beta}$ can hence be obtained by usual maximum likelihood methods and inference can be made with Wald-type tests to assess the global null hypothesis H_0^{global}. Additionally, a lagged piecewise exponential model, which only considers events after a prespecified assumed lag time t_{lag} by resetting the origin to this time, is considered, assessing the same null hypothesis as the Landmark model, i.e. H_0^{Landmark}: $S_C(t) = S_E(t)$ for all $t > t_{\text{lag}}$. It is implemented in the *pchreg* function in the *eha* package in R (Broström, 2020).

(R5) **Additive hazard model test(Aalen):**

The Aalen additive model assumes that the hazard function $\lambda(t)$ can be expressed as

$$\lambda(t) = \beta(t)X.$$

Estimators for the cumulative coefficient $B(t) = \int_0^t \beta(s)ds$ can be obtained by a least squares approach making use of the counting process formulation as shown by Scheike and Martinussen (2006, Chapter 5). It has been shown that these estimators are asymptotically normal and hence inference on the global null hypothesis H_0^{global} can be made using the Wald test. This approach is implemented in R in the *aalen* function in the *timereg* package (Scheike and Martinussen, 2006).

(R6) **Royston/Parmar spline model test (RP.PH/RP.TD):**

Royston and Parmar (2002) proposed to model the logarithmic cumulative hazard function using restricted cubic splines

$$s(\log(t)|\gamma, k) = \gamma_0 + \gamma_1 \log(t) + \gamma_2 v_1(\log(t)) + \ldots + \gamma_k v_{k-1}(\log(t)),$$

with a prespecified number of knots k and where $\gamma_0, \ldots, \gamma_k$ denote the spline coefficients and $v_j(x)$ the j-th basis function. They found that $k_0 = 4$ knots are sufficiently flexible which results in the following PH model

$$\log(\Lambda(t)) = \gamma_0 + \gamma_1 \log(t) + \gamma_2 v_1(\log(t)) + \gamma_3 v_2(\log(t)) + \gamma_4 v_3(\log(t)) + \beta X.$$

If again the treatment effect can not be assumed to be independent of time, the spline coefficient γ_1 can be modelled to depend on treatment X, i.e. $\gamma_1 = \gamma_1(X) = \gamma_{10} + \gamma_{11}X$. To assess the global null hypothesis H_0^{global} which is equivalent to $H_0 \colon \beta = 0$ in the PH model and to $H_0 \colon (\beta, \gamma_{11}) = 0$ in the time-dependent model a likelihood ratio test is used. This functionality is implemented in the *flexsurvspline* function in the *flexsurv* package (Jackson, 2016).

(R7) **Accelerated failure time model test (AFT):**

The (semi-)parametric accelerated failure time models pose a well-known alternative to the Cox model and the effect of the covariates is assumed to accelerate or decelerate

the time to event relative to a baseline distribution S_0, which yields the model

$$S(t|X) = S_0(te^{\beta X}).$$

In case of the semiparametric AFT model the distribution of S_0 is left unspecified and two different approaches have been developed to tackle this problem. The first approach is based on another commonly used reformulation of the model in terms of the logarithmic failure time

$$\log(T) = \beta X + \varepsilon.$$

Based on this formulation the least squares approach of classical linear regression was extended to time-to-event analysis by incorporating censoring. This was first done by Miller (1976) and later by Buckley and James (1979) and many other authors.
The second approach is based on the hazard which can be expressed as

$$\lambda(t|X) = \lambda_0(te^{\beta X})e^{\beta X}.$$

Inference on β is then based on a partial likelihood similar to the Cox model, but can also be motivated by classical testing theory (Tsiatis, 1990).
For the fully parametric approach one assumes different distributions for S_0 (e.g. Weibull, exponential) resulting in AFT Wei, AFT exp, etc. These parametric models are all considered in the simulations and for inference the Wald test is used to assess the global null hypothesis H_0^{global}.
These methods are implemented in the *survreg* function in the *survival* package (Therneau, 2023).

Other tests

This section comprises all methods that cannot be subsumed under the previous sections.

(O1) **Milestone test based on the Nelson-Aalen estimator (MileNA):**
This Milestone survival test also assesses the null hypothesis H_0^{Mile} but based on Nelson-Aalen estimates of the cumulative hazard rates at a fixed timepoint t_0. Together with

the variance of the Nelson-Aalen estimator (2.3.1) the test statistic is then

$$Z_{NA}(t_0) = \frac{\hat{\Lambda}_{\mathrm{NA},C}(t_0) - \hat{\Lambda}_{\mathrm{NA},E}(t_0)}{\sqrt{\widehat{\mathrm{Var}}[\hat{\Lambda}_{\mathrm{NA},C}(t_0)] + \widehat{\mathrm{Var}}[\hat{\Lambda}_{\mathrm{NA},E}(t_0)]}},$$

which is asymptotically standard normal distributed under the null hypothesis. The estimated difference of cumulative hazard rates at t_0 can be translated into a ratio of survival proportions by exploiting the relationship $S(t) = \exp(-\Lambda(t))$ as

$$\exp\left(\hat{\Lambda}_{\mathrm{NA},C}(t_0) - \hat{\Lambda}_{\mathrm{NA},E}(t_0)\right) = \frac{\hat{S}_{\mathrm{NA},E}(t_0)}{\hat{S}_{\mathrm{NA},C}(t_0)}.$$

(O2) **Linear and Quadratic combination test (LLRNA, QLRNA):**
Logan et al. (2015) proposed a combination test by decomposing the global null hypothesis into two subhypotheses and applying a test to each test statistic. Their decomposition of the global null hypothesis is

$$\{H_{01}\colon S_E(t_0) = S_C(t_0)\} \cap \{H_{02}\colon \lambda_E(t) = \lambda_C(t), t > t_0\}.$$

Now the first subhypothesis is tested using the Milestone survival test based on Nelson-Aalen estimates and for the second subhypothesis the logrank test left truncated at t_0, i.e. the integration in (2.2) starting at t_0, is used. Denoting the standardized test statistic of the latter with $Z_{LR}(t_0)$ both a linear combination $\frac{Z_{NA}(t_0)+Z_{LR}(t_0)}{\sqrt{2}}$ and a quadratic combination $Z^2_{NA}(t_0) + Z^2_{LR}(t_0)$ were proposed. The former is asymptotically standard normally distributed under H_{01} and the latter is χ^2_2 distributed under H_{02}. To quantify the survival differences the effects for both subhypotheses can be used.

(O3) **Checking PH approach test (CheckPH):**
Campbell and Dean (2014) proposed a two-stage approach to assess the global null hypothesis H_0^{global}:

- Test the PH assumption using Grambsch-Therneau (GT) test, which was introduced in section 2.3.2, at level α_{GT}
- If GT is not significant use Cox, otherwise CoxTD, to test for a treatment effect at level $\alpha = 0.05$.

The performance of this approach depends on the choice of α_{GT}, which for this simulation was taken to be 0.05. This is expected to lead to an inflated type I error due to multiple testing, and Campbell and Dean (2014) suggest that this can be remedied by a permutation test approach. However, this was not implemented in the simulation study as permutational approaches lead to a substantial increase of run-time which was not feasible.

(O4) **Joint test (jointTest):**

Royston and Parmar (2014) proposed a joint test to assess the global null hypothesis H_0^{global} as the sum of the Grambsch-Therneau test statistic and the Cox test statistic. They investigated that these two test statistics are independent and hence the joint test is χ^2 distributed with 2 degrees of freedom.

(O5) **Combined test (combTest):**

In a later paper Royston and Parmar proposed another test to assess the global null hypothesis H_0^{global} called the combined test (Royston and Parmar, 2016). This test is based on the maximum RMST which was described in the Kaplan-Meier based methods section. However, instead of calculating the p-value (p_{perm}) by a permutation approach an approximation based on the naive χ^2 p-value ($p_{\max}$) of the maximum RMST is used. Fitting a Box-Tidwell model they derived the formula

$$p_{\text{perm}} = 1.762 p_{\max}^{0.885} - 0.802 p_{\max}^{2.547}.$$

In a next step this p-value is combined with the p-value from a Cox model (p_{Cox}) by taking the minimum $p_{\min} = \min(p_{\text{perm}}, p_{\text{Cox}})$. Finally, the null distribution of $p_{\min}$ is approximated by a Beta distribution yielding the final p-value $p_{\text{comb}} = F_{\text{Beta}(1,1.5)}(p_{\min})$

(O6) **Cauchy changepoint test (CauchyCP):**

Zhang et al. (2021) proposed the CauchyCP method. This method is based on the single changepoint Cox model, which assumes that the effect of treatment is time dependent and changes at a pre-specified changepoint t^*, i.e. $\beta(t) = \beta_1^* 1_{[0,t^*)}(t) + \beta_2^* 1_{[t^*,\infty)}(t)$. For the CauchyCP test one combines multiple such single changepoint Cox models as follows:

(a) For prespecified timepoints $t_1^*, \ldots, t_m^*$, fit a single changepoint model with changepoint t_i^*

(b) To test the null hypothesis that $\beta^*_{i1} = \beta^*_{i2} = 0$ which is equivalent to the global null hypothesis H_0^{global} perform a likelihood ratio test. The p-value of the test corresponding to changepoint t_i^* is denoted with p_i.

(c) An omnibus test statistic is constructed using the Cauchy combination method as $CCP = \frac{1}{m}\sum_{i=1}^{m} \tan(\pi(0.5 - p_i))$ and the final p-value is then calculated as $p_{CCP} = 0.5 - \arctan(CCP)/\pi$.

Zhang et al. note that specifying too many candidate changepoints without prior knowledge might lower statistical power and suggest to use $t_1^* = 0$ and the 25th, 50th and 75th percentiles of the event times. This method is implemented in the *CauchyCP* package in R (Zhang, 2022).

Table 3 summarizes the methods that were introduced above.

Table 3: *Structured overview of all methods used in this thesis and their abbreviations*

Class	No.	Abbreviation	Name
Basic	B1	Logrank	logrank test
	B2	Cox	Cox model test
Weighted logrank tests	W1	$G(\rho, \gamma)$	Fleming-Harrington test
	W2	$G(-1, 0)$	Gray-Tsiatis test
	W3	MWLRT	modestly weighted logrank test
	W4	Thres/GenLin	threshold test or generalized linear lag test
	W5	GB	Gehan-Breslow test
	W6	TW	Tarone-Ware test
	W7	PP/mPP	(modified) Peto-Peto test
	W8	asymLR	asymptotic logrank test
	W9	Logit/mLogit	(modified) logit test
	W10	MERT	maximin efficiency robust test
Combinations of weighted logrank statistics	C1	YP	adaptively weighted logrank test
	C2	MaxCombo	maximum combination of weighted logrank tests
	C3	mMaxCombo	modified MaxCombo test
	C4	Zm3	Zm3 test
	C5	mZm3	modified Zm3 test
	C6	Lee1/Lee2/Lee3	versatile tests by Lee
	C7	mLee2/mLee3	modified versatile tests by Lee
	C8	ProjTest	projection test
	C9	mScore	modified score test
	C10	V0	approximation of MERT
	C11	ParGroup	partially grouped logrank test
	C12	KS LR/FH/GB/Cheng	Kolmogorov-Smirnov tests based on the logrank / Fleming-Harrington / Gehan-Breslow / Cheng test statistic

Class	No.	Abbreviation	Name
Kaplan-Meier based tests	K1	Mile/MileCLL	milestone survival test
	K2	WKM	weighted Kaplan-Meier test
	K3	RMST	restricted mean survival time test
	K4	ABC	area between curves test
Tests based on regression models	R1	AHR	average hazard ratio test
	R2	Landmark	landmark test
	R3	CoxTD	time-dependent treatment effect test
	R4	PWExp/PWExpLag	piecewise exponential (lag) model test
	R5	Aalen	additive hazard model test
	R6	RP.PH/RP.TD	Royston-Parmar spline model test(PH or time-dependent)
	R7	AFT	accelerated failure time model test
Other tests	O1	MileNA	milestone test based on Nelson-Aalen estimator
	O2	LLRNA/QLRNA	linear and quadratic combination test
	O3	CheckPH	checking PH approach
	O4	jointTest	joint test by Royston and Parmar
	O5	combTest	combined test by Royston and Parmar
	O6	CauchyCP	Cauchy changepoint test

2.4 Design of the simulation study

As outlined in the previous section, a plethora of methods exists to analyze time-to-event data some of which are specifically tailored to handle non-proportional hazards and some are not. Although not all methods are tailored to target the global null hypothesis

$$H_0^{\text{global}}\colon S_C(t) = S_E(t) \text{ for all } t,$$

their individual null hypothesis is a superset or implication of it, e.g. the null hypothesis $H_0^{\text{Mile}}\colon S_C(t_0) = S_E(t_0)$ for $t_0 > 0$ of the Milestone methods. Hence, the rejection of these individual null hypotheses can be seen as a surrogate for the rejection of H_0^{global}, so that in the simulation study the control of type I error and the influence of the delayed onset of the treatment effect on the power of testing this global null hypothesis can be investigated. The simulation study is planned in accordance with the ADEMP structure as introduced by Morris et al. (2019) and to ensure transparency a simulation plan was published in advance on zenodo.org (Behnisch, 2023). The acronym ADEMP stands for the five aspects that should be covered in the description of a simulation study: Aims, Data-generating mechanisms, Methods, Estimands and Performance measures. The next sections describe the ADEMP structure for the examination of power and type I error of the methods. Here, the structure of the simulation study to assess the power is described first since the scenarios of the data-generating mechanisms for the type 1 error are deduced from the scenarios considered for the assessment of the power.

2.4.1 Assessment of Power

In the following the ADEMP structure of the power assessment is described in detail.

Aim

To systematically assess the power of the different methods in settings with a delayed treatment effect and investigate how the duration of the delay impacts the power of the method.

Data-generating mechanism

For the data-generating mechanism the following parameters were chosen:

- As can be seen from results of clinical trials in the field of immuno-oncology the study duration and accrual is quite heterogeneous. For example in the CHECKMATE 078 trial patients were enrolled for 12 months (December 2015 to November 2016) and the database was locked on October 27, 2017, so that the total duration of the study was 23 months and hence the proportion of accrual on the study duration was approx. 50%. In contrast the CHECKMATE 067 trial enrolled patients for 9 months (July 3, 2013 to March 31, 2014) but had a total study duration of 58 months (database lock on May 10, 2018) giving proportion of approx. 16% if the accrual on the total study duration. To take this into account the study duration was taken to be

$$\tau = 12, 24, 48, 60 \text{ months}$$

- Uniform accrual over the interval $[0, a]$, where a is expressed as fraction of the total study duration τ, i.e. $a = \text{acc} \cdot \tau$ where $\text{acc} \in \{0.2, 0.4\}$.
- The delayed treatment effect occurs with a delay of $t_2^* = \text{lag}_2 \cdot \tau$, where $\text{lag}_2 \in \{0, 0.1, 0.2, 0.3, 0.4\}$.
- The hazard ratio θ of the full treatment effect varies $\theta = 0.5, 0.6, 0.7, 0.8$
- The hazards $\lambda_C(t)$ and $\lambda_E(t)$ of the control and experimental arm, respectively, can be expressed in terms of a general lag model, i.e.

$$\lambda_E(t) = [1 - l(t) + \theta l(t)]\lambda_C(t),$$

  with a monotone non-decreasing lag function $0 \leqslant l(t) \leqslant 1$. The lag function is taken to represent the following three delay patterns:

  - Threshold lag with a change point t^*: $l(t) = I(t \geqslant t^*)$,
  - Linear lag with change point t^*: $l(t) = \frac{t}{t^*} I(t < t^*) + I(t \geqslant t^*)$ or
  - Generalized linear lag with change points t_1^*, t_2^*: $l(t) = \frac{t - t_1^*}{t_2^* - t_1^*} I(t_1^* \leqslant t < t_2^*) + I(t \geqslant t_2^*)$.

  It can easily be seen that the linear lag and the threshold lag are contained in the generalized linear lag pattern by choosing $t_1^* = 0$ or $t_1^* = t_2^*$, respectively. The change point t_1^* will be chosen as proportion of the lag t_2^*, meaning

$$t_1^* = \text{lag}_1 \cdot t_2^*, \text{ with } \text{lag}_1 \in \{0.3, 0.7, 1\}.$$

- The distribution of survival times in the control group is chosen to be Weibull distributed $\text{Wei}(\lambda_C, k_C)$ with shape parameter k_C reflecting the following types of hazards:
  - decreasing hazard: $k_C = 0.5$
  - constant hazard (exponential): $k_C = 1$
  - (linearly) increasing hazard: $k_C = 2$

  and scale parameter λ_C chosen in dependence on the aspired median survival times. Since the observed median survival times in the control arms of the above mentioned CHECKMATE trials ranges from 2.3 to 4.2 months in PFS and from 5.1 to 19.9 months in OS, we considered median survival times of

$$\text{med}_C = 5, 15, 20$$

  and calculated the scale parameter as $\lambda_C = \frac{(\ln(2))^{1/k_C}}{\text{med}_C}$.

To calculate the total number of scenarios one has to keep in mind that if $t_2^* = 0$, i.e. no delay is assumed, there will also be no changepoint present and hence $t_1^* = 0$. Thus the total number of simulated scenarios described above will be $4 \cdot 2 \cdot 4 \cdot 4 \cdot 3 \cdot 3 \cdot 3 + 4 \cdot 2 \cdot 4 \cdot 3 \cdot 3 = 3744$. For this thesis it is assumed that only administrative censoring is present in the data.

Sample size calculation

To set a benchmark for the analysis sample size calculation was performed using the rpact package based on the sample size calculation method by Schoenfeld (Wassmer and Pahlke, 2023). In case of PH scenarios this yields the necessary sample size to achieve a power of 80% based on the full effect θ.

In case of NPH scenarios two different strategies for sample size calculation were investigated based on different calculations of the following overall treatment effect:

- the naive average effect $\frac{1}{\tau}\int_0^\tau \frac{\lambda_E(t)}{\lambda_C(t)}dt$
- the average effect $\frac{1}{\tau}\int_0^\tau \frac{\lambda_E(t)}{\lambda_C(t)}S_p(t)dt$ based on the pooled survival function in both arms, i.e. $S_p(t) = \frac{1}{2}(S_E(t) + S_C(t))$

Examples of the resulting effect sizes are displayed in Figures 9, 10 and 11, where the first plot illustrates the behavior in case of decreasing hazards ($k_C = 0.5$), the second plot in case of constant hazards ($k_C = 1$) and the last plot in case of increasing hazards ($k_C = 2$). For all three plots a study duration of $\tau = 24$ months, a maximum effect of $\theta = 0.5$ and a median survival of $\text{med}_C = 15$ months in the control arm was assumed.

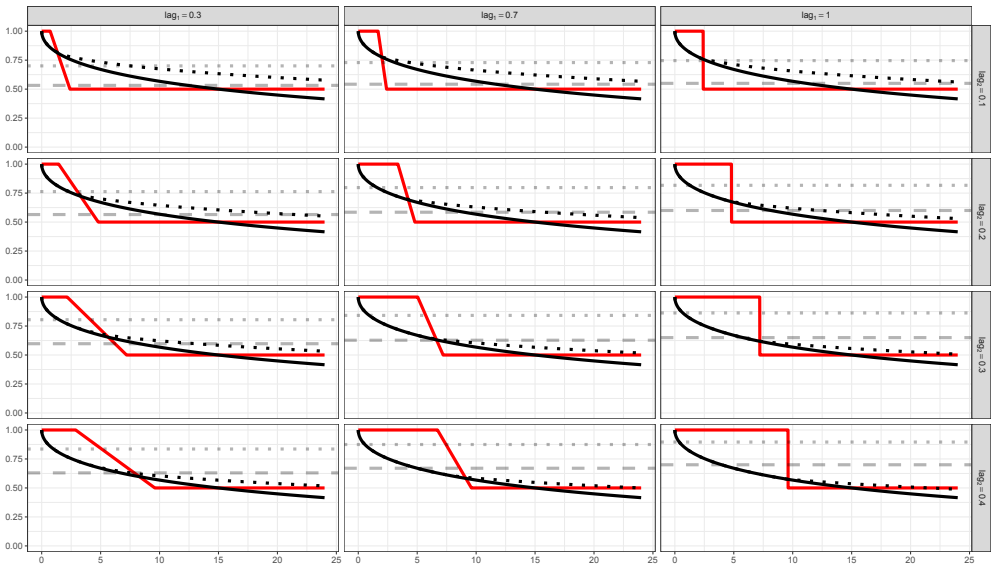

Figure 9: *Averaged treatment effects in gray for a scenario with decreasing hazard, a study duration of 24 months and a median survival of 15 months in the control arm (naive average HR = dashed, average HR = dotted). Black lines represent the survival curves in the experimental (dashed) and control arm (solid) and the red line represents the hazard ratio.*

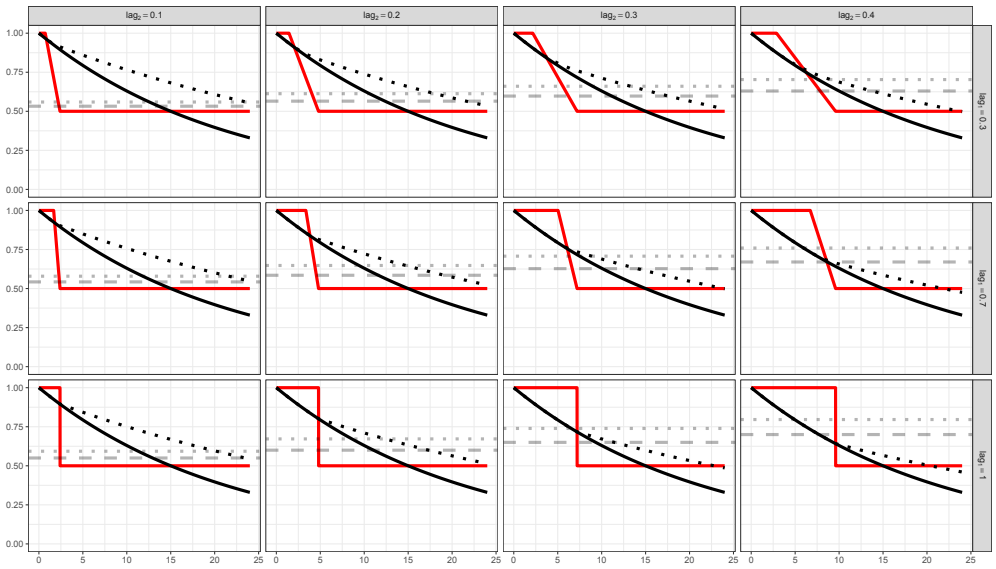

Figure 10: *Averaged treatment effects in gray for a scenario with constant hazard, a study duration of 24 months and a median survival of 15 months in the control arm (naive average HR = dashed, average HR = dotted). Black lines represent the survival curves in the experimental (dashed) and control arm (solid) and the red line represents the hazard ratio.*

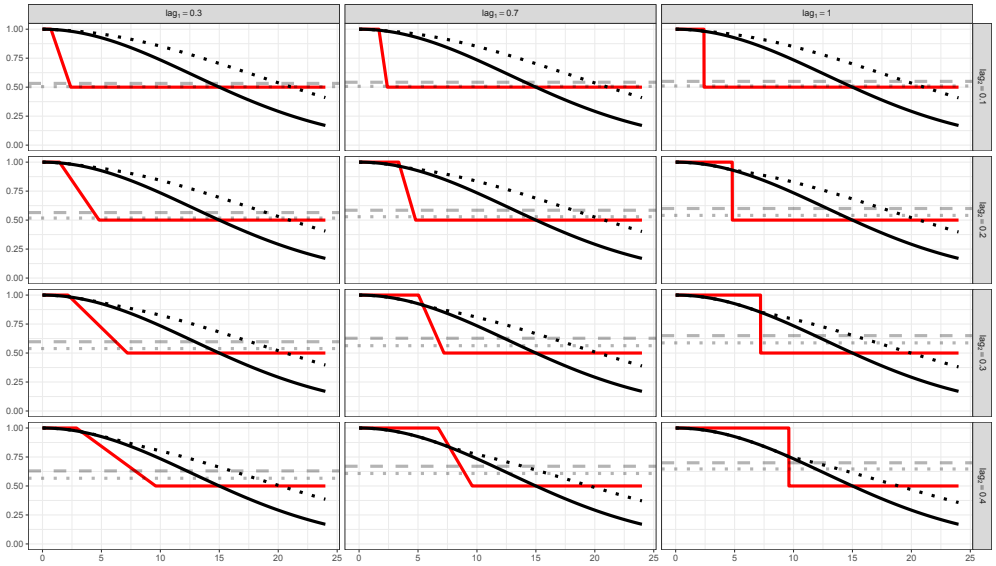

Figure 11: *Averaged treatment effects in gray for a scenario with increasing hazard, a study duration of 24 months and a median survival of 15 months in the control arm (naive average HR = dashed, average HR = dotted). Black lines represent the survival curves in the experimental (dashed) and control arm (solid) and the red line represents the hazard ratio.*

Based on this average hazard ratio the necessary number of observations n_{obs} needed to achieve a power of 80% was then again calculated using the Schoenfeld formula and assuming

proportional hazards. As these exemplary plots suggest, the calculation based on the average effect led to higher sample sizes than for the naive average approach in a majority of scenarios ($2713/3744 \approx 72\%$) and only in cases with increasing hazards ($k_C = 2$) the sample size was smaller ($734/3744 \approx 20\%$). The calculated sample size based on the naive average hazard ratio approach ranged from 66 to 13604 whereas the sample size for the average hazard ratio approach reached up to $6.9 \cdot 10^{12}$.

As the power in time-to-event analyses is mainly driven by the number of events it had to be expected that the sample size is high in scenarios with a combination of a short study duration and a big median survival in the control group as only a little proportion of subjects will have an event during the short observation period. In case of the average hazard ratio the converse was also unfavorable for the sample size calculation. Here, a long study duration combined with a short median survival in the control group, which results in a fast declining pooled survival curve, leads to an downweighting of later timepoints and hence the average hazard ratio is shifted more towards the early effect which for the delayed treatment effect scenarios is the null effect.

As the sample sizes based on the average hazard ratio approach tend to be much bigger or even computationally not feasible in some scenarios and the focus of this simulation is not the compensation of power loss but the comparison of the power of different methods, the simulation study is based on the naive average effect.

Sample size calculation in PH scenarios

Based on the structure of the data-generating process the PH scenarios correspond to the scenarios where neither a delay nor a lag is present, i.e. $\text{lag}_2 = \text{lag}_1 = 0$. This results in a total of 288 scenarios. First the Schoenfeld formula calculates the number of events (n_{events}) necessary to detect the assumed treatment effect. In the next step the distributional assumption is exploited to calculate the number of observations (n_{obs}) needed to observe the necessary number of events. Hence the results of the sample size calculation are split into four tables, one for each assumed treatment effect θ and shown in tables 4, 5, 6 and 7.

Table 4: *Sample size for PH scenarios with a hazard ratio of 0.5 to achieve a power of 80%.* med_C *= median survival in control arm,* acc *= accrual proportion,* k_C *= Weibull shape parameter in control arm,* n_{events} *= number of events,* n_{obs} *= number of observations,* τ *= overall study duration*

med_C	acc	k_C	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
5	0.2	0.5	66	126	104	88	84
5	0.4	0.5	66	132	108	90	86
15	0.2	0.5	66	188	146	116	108
15	0.4	0.5	66	198	152	120	112
20	0.2	0.5	66	210	162	126	118
20	0.4	0.5	66	222	168	132	124
5	0.2	1.0	66	102	76	68	68
5	0.4	1.0	66	108	80	70	68
15	0.2	1.0	66	214	128	88	82
15	0.4	1.0	66	236	140	94	86
20	0.2	1.0	66	272	156	102	92
20	0.4	1.0	66	302	172	108	96
5	0.2	2.0	66	76	66	66	66
5	0.4	2.0	66	82	66	66	66
15	0.2	2.0	66	280	104	68	66
15	0.4	2.0	66	342	120	72	68
20	0.2	2.0	66	468	148	76	70
20	0.4	2.0	66	576	176	82	72

Table 5: *Sample size for PH scenarios with a hazard ratio of 0.6 to achieve a power of 80%.* med_C *= median survival in control arm,* acc *= accrual proportion,* k_C *= Weibull shape parameter in control arm,* n_{events} *= number of events,* n_{obs} *= number of observations,* τ *= overall study duration*

med_C	acc	k_C	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
5	0.2	0.5	121	220	180	154	148
5	0.4	0.5	121	230	186	158	152
15	0.2	0.5	121	326	252	202	190
15	0.4	0.5	121	342	264	210	196
20	0.2	0.5	121	364	280	220	206
20	0.4	0.5	121	384	292	230	214
5	0.2	1.0	121	176	136	124	122
5	0.4	1.0	121	190	142	124	122
15	0.2	1.0	121	370	224	156	144
15	0.4	1.0	121	408	242	164	150
20	0.2	1.0	121	470	272	176	160
20	0.4	1.0	121	520	298	190	168
5	0.2	2.0	121	134	122	122	122
5	0.4	2.0	121	144	122	122	122
15	0.2	2.0	121	486	182	124	122
15	0.4	2.0	121	592	208	128	122
20	0.2	2.0	121	808	258	134	124
20	0.4	2.0	121	994	306	144	130

Table 6: *Sample size for PH scenarios with a hazard ratio of 0.7 to achieve a power of 80%.* med_C *= median survival in control arm,* acc *= accrual proportion,* k_C *= Weibull shape parameter in control arm,* n_{events} *= number of events,* n_{obs} *= number of observations,* τ *= overall study duration*

med_C	acc	k_C	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
5	0.2	0.5	247	430	354	304	292
5	0.4	0.5	247	448	366	312	298
15	0.2	0.5	247	632	492	396	372
15	0.4	0.5	247	664	514	410	384
20	0.2	0.5	247	708	544	430	402
20	0.4	0.5	247	744	570	448	418
5	0.2	1.0	247	348	272	250	248
5	0.4	1.0	247	372	280	252	250
15	0.2	1.0	247	720	436	306	284
15	0.4	1.0	247	794	474	324	296
20	0.2	1.0	247	910	530	348	314
20	0.4	1.0	247	1010	578	372	332
5	0.2	2.0	247	268	248	248	248
5	0.4	2.0	247	288	248	248	248
15	0.2	2.0	247	942	356	250	248
15	0.4	2.0	247	1146	408	258	250
20	0.2	2.0	247	1564	502	268	252
20	0.4	2.0	247	1922	594	288	262

Table 7: *Sample size for PH scenarios with a hazard ratio of 0.8 to achieve a power of 80%.* med_C *= median survival in control arm,* acc *= accrual proportion,* k_C *= Weibull shape parameter in control arm,* n_{events} *= number of events,* n_{obs} *= number of observations,* τ *= overall study duration*

med_C	acc	k_C	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
5	0.2	0.5	631	1056	872	754	728
5	0.4	0.5	631	1098	900	772	744
15	0.2	0.5	631	1540	1202	970	914
15	0.4	0.5	631	1614	1254	1006	946
20	0.2	0.5	631	1720	1328	1056	988
20	0.4	0.5	631	1808	1388	1098	1026
5	0.2	1.0	631	858	680	636	632
5	0.4	1.0	631	914	702	638	634
15	0.2	1.0	631	1748	1070	760	710
15	0.4	1.0	631	1926	1158	800	738
20	0.2	1.0	631	2210	1292	858	780
20	0.4	1.0	631	2448	1410	914	822
5	0.2	2.0	631	672	632	632	632
5	0.4	2.0	631	718	632	632	632
15	0.2	2.0	631	2284	876	636	632
15	0.4	2.0	631	2776	1002	650	636
20	0.2	2.0	631	3786	1226	672	640
20	0.4	2.0	631	4650	1450	718	660

Sample size calculation in NPH scenarios

The naive average hazard ratio $\bar{\theta}_{\text{naive}}$ is obtained by integrating the assumed general linear lag model function, i.e.

$$\begin{aligned}\bar{\theta}_{\text{naive}} = \frac{1}{\tau}\int_0^\tau \frac{\lambda_E(t)}{\lambda_C(t)}dt &= \frac{1}{\tau}\int_0^\tau [1 - l(t) + \theta l(t)]dt \\ &= \frac{1}{\tau}\int_0^\tau \left[1 + (\theta - 1)\left(\frac{t - t_1^*}{t_2^* - t_1^*}I(t_1^* \leqslant t < t_2^*) + I(t \geqslant t_2^*)\right)\right]dt \\ &= 1 + \frac{\theta - 1}{\tau}\int_{t_1^*}^{t_2^*}\frac{t - t_1^*}{t_2^* - t_1^*}dt + \frac{\theta - 1}{\tau}\int_{t_2^*}^{\tau} dt \\ &= 1 + \frac{(\theta - 1)(t_2^* - t_1^*)}{2\tau} + \frac{(\theta - 1)(\tau - t_2^*)}{\tau}.\end{aligned}$$

Hence the calculation of the naive average hazard ratio depends on the maximum effect θ, the delay t_2^* and the changepoint t_1^*. As the delay and changepoint are expressed as proportions of the overall study duration τ the naive average hazard ratio is independent of it. The sample size was then calculated with the Schoenfeld formula based on this naive average hazard ratio for all 3456 NPH scenarios. The results are summarized in the following Tables 8, 9 and 10 for decreasing, constant and increasing hazards, respectively. In these tables for each combination of θ, t_2^* and t_1^* the naive average hazard ratio and the number of events are presented. The resulting sample sizes are then summarized per study duration and only the range of sample sizes over median survival med_C and accrual proportion acc is given.

Table 8: *Results of the sample size calculation based on the naive average hazard ratio* $\bar{\theta}_{\text{naive}}$ *in NPH scenarios with decreasing hazard.* θ *= maximum treatment effect,* lag_2 *= delay proportion,* lag_1 *= changepoint proportion,* n_{events} *= number of events,* n_{obs} *= number of observations,* τ *= overall study duration*

θ	lag_2	lag_1	$\bar{\theta}_{\text{naive}}$	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
0.5	0.1	0.3	0.5325	80	150 - 262	122 - 200	104 - 156	100 - 146
0.5	0.1	0.7	0.5425	84	158 - 276	130 - 212	110 - 166	106 - 154
0.5	0.1	1.0	0.5500	88	166 - 288	136 - 220	114 - 172	110 - 160

θ	lag_2	lag_1	$\bar{\theta}_{\text{naive}}$	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
0.5	0.2	0.3	0.5650	97	180 - 314	148 - 240	126 - 188	120 - 174
0.5	0.2	0.7	0.5850	110	202 - 350	166 - 268	140 - 210	134 - 196
0.5	0.2	1.0	0.6000	121	220 - 384	180 - 292	154 - 230	148 - 214
0.5	0.3	0.3	0.5975	119	218 - 378	178 - 288	152 - 226	146 - 210
0.5	0.3	0.7	0.6275	145	262 - 454	214 - 346	182 - 272	176 - 254
0.5	0.3	1.0	0.6500	170	302 - 524	248 - 400	212 - 314	204 - 294
0.5	0.4	0.3	0.6300	148	266 - 460	218 - 352	186 - 276	178 - 258
0.5	0.4	0.7	0.6700	196	346 - 600	284 - 458	244 - 360	234 - 336
0.5	0.4	1.0	0.7000	247	430 - 744	354 - 570	304 - 448	292 - 418
0.6	0.1	0.3	0.6260	144	258 - 448	212 - 344	182 - 270	174 - 252
0.6	0.1	0.7	0.6340	152	272 - 472	224 - 360	190 - 284	184 - 264
0.6	0.1	1.0	0.6400	158	284 - 490	232 - 376	198 - 294	190 - 274
0.6	0.2	0.3	0.6520	172	306 - 530	252 - 406	216 - 320	206 - 298
0.6	0.2	0.7	0.6680	193	342 - 590	280 - 452	240 - 356	230 - 332
0.6	0.2	1.0	0.6800	212	372 - 642	306 - 492	262 - 388	252 - 360
0.6	0.3	0.3	0.6780	208	366 - 634	302 - 484	258 - 382	248 - 356
0.6	0.3	0.7	0.7020	251	438 - 754	360 - 578	308 - 456	296 - 424
0.6	0.3	1.0	0.7200	291	504 - 868	414 - 664	356 - 524	342 - 488
0.6	0.4	0.3	0.7040	255	444 - 766	366 - 586	314 - 462	302 - 432
0.6	0.4	0.7	0.7360	335	574 - 988	472 - 758	408 - 598	392 - 558
0.6	0.4	1.0	0.7600	417	710 - 1218	584 - 934	504 - 738	486 - 688
0.7	0.1	0.3	0.7195	290	502 - 864	412 - 662	354 - 522	342 - 486
0.7	0.1	0.7	0.7255	305	526 - 906	434 - 694	372 - 548	358 - 510
0.7	0.1	1.0	0.7300	317	546 - 940	450 - 720	386 - 568	372 - 530
0.7	0.2	0.3	0.7390	344	590 - 1014	486 - 776	418 - 612	402 - 572
0.7	0.2	0.7	0.7510	383	654 - 1124	538 - 862	464 - 680	448 - 634
0.7	0.2	1.0	0.7600	417	710 - 1218	584 - 934	504 - 738	486 - 688

θ	lag_2	lag_1	$\bar{\theta}_{\text{naive}}$	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
0.7	0.3	0.3	0.7585	411	700 - 1202	576 - 922	498 - 728	480 - 680
0.7	0.3	0.7	0.7765	491	830 - 1422	684 - 1092	590 - 862	570 - 806
0.7	0.3	1.0	0.7900	566	950 - 1628	784 - 1250	678 - 988	654 - 922
0.7	0.4	0.3	0.7780	499	842 - 1442	694 - 1108	600 - 876	578 - 818
0.7	0.4	0.7	0.8020	645	1078 - 1846	892 - 1418	772 - 1122	744 - 1048
0.7	0.4	1.0	0.8200	798	1324 - 2264	1096 - 1740	950 - 1376	916 - 1286
0.8	0.1	0.3	0.8130	733	1220 - 2086	1010 - 1604	874 - 1268	844 - 1186
0.8	0.1	0.7	0.8170	769	1278 - 2186	1058 - 1680	916 - 1330	884 - 1242
0.8	0.1	1.0	0.8200	798	1324 - 2264	1096 - 1740	950 - 1376	916 - 1286
0.8	0.2	0.3	0.8260	860	1424 - 2432	1180 - 1870	1022 - 1480	986 - 1384
0.8	0.2	0.7	0.8340	953	1576 - 2688	1304 - 2066	1132 - 1638	1092 - 1530
0.8	0.2	1.0	0.8400	1033	1704 - 2906	1412 - 2234	1224 - 1770	1182 - 1656
0.8	0.3	0.3	0.8390	1019	1682 - 2868	1392 - 2206	1208 - 1748	1166 - 1634
0.8	0.3	0.7	0.8510	1207	1982 - 3376	1644 - 2598	1426 - 2060	1378 - 1926
0.8	0.3	1.0	0.8600	1381	2260 - 3848	1876 - 2962	1630 - 2348	1574 - 2196
0.8	0.4	0.3	0.8520	1224	2010 - 3424	1666 - 2634	1448 - 2088	1398 - 1954
0.8	0.4	0.7	0.8680	1567	2558 - 4352	2124 - 3350	1846 - 2658	1784 - 2486
0.8	0.4	1.0	0.8800	1922	3124 - 5310	2596 - 4090	2258 - 3246	2182 - 3036

Table 9: *Results of the sample size calculation based on the naive average hazard ratio* $\bar{\theta}_{\text{naive}}$ *in NPH scenarios with constant hazard.* θ *= maximum treatment effect,* lag_2 *= delay proportion,* lag_1 *= changepoint proportion,* n_{events} *= number of events,* n_{obs} *= number of observations,* τ *= overall study duration*

θ	lag_2	lag_1	$\bar{\theta}_{\text{naive}}$	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
0.5	0.1	0.3	0.5325	80	120 - 356	92 - 204	82 - 128	80 - 114
0.5	0.1	0.7	0.5425	84	128 - 376	96 - 214	86 - 136	86 - 122
0.5	0.1	1.0	0.5500	88	132 - 392	102 - 224	90 - 142	90 - 126
0.5	0.2	0.3	0.5650	97	144 - 426	110 - 244	100 - 154	98 - 138
0.5	0.2	0.7	0.5850	110	162 - 478	124 - 272	112 - 174	110 - 154
0.5	0.2	1.0	0.6000	121	176 - 520	136 - 298	124 - 190	122 - 168
0.5	0.3	0.3	0.5975	119	174 - 514	134 - 294	122 - 186	120 - 166
0.5	0.3	0.7	0.6275	145	210 - 616	162 - 352	148 - 224	146 - 200
0.5	0.3	1.0	0.6500	170	244 - 712	188 - 406	172 - 260	170 - 232
0.5	0.4	0.3	0.6300	148	214 - 626	164 - 358	150 - 228	148 - 204
0.5	0.4	0.7	0.6700	196	280 - 814	216 - 466	198 - 298	198 - 266
0.5	0.4	1.0	0.7000	247	348 - 1010	272 - 578	250 - 372	248 - 332
0.6	0.1	0.3	0.6260	144	208 - 610	160 - 348	146 - 222	144 - 198
0.6	0.1	0.7	0.6340	152	218 - 642	170 - 366	154 - 234	152 - 210
0.6	0.1	1.0	0.6400	158	228 - 666	176 - 382	160 - 244	160 - 218
0.6	0.2	0.3	0.6520	172	246 - 720	192 - 412	174 - 264	174 - 236
0.6	0.2	0.7	0.6680	193	276 - 802	214 - 460	196 - 294	194 - 264
0.6	0.2	1.0	0.6800	212	300 - 872	234 - 500	214 - 320	212 - 286
0.6	0.3	0.3	0.6780	208	296 - 860	230 - 492	212 - 316	210 - 282
0.6	0.3	0.7	0.7020	251	354 - 1024	276 - 588	254 - 376	252 - 338
0.6	0.3	1.0	0.7200	291	406 - 1178	318 - 676	294 - 434	292 - 390
0.6	0.4	0.3	0.7040	255	358 - 1040	280 - 596	258 - 382	256 - 342
0.6	0.4	0.7	0.7360	335	464 - 1340	364 - 770	338 - 496	336 - 444
0.6	0.4	1.0	0.7600	417	574 - 1652	452 - 950	420 - 612	418 - 550

θ	lag_2	lag_1	$\bar{\theta}_{\text{naive}}$	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
0.7	0.1	0.3	0.7195	290	406 - 1172	318 - 672	294 - 432	292 - 388
0.7	0.1	0.7	0.7255	305	426 - 1230	334 - 706	308 - 454	306 - 408
0.7	0.1	1.0	0.7300	317	442 - 1276	346 - 732	320 - 472	318 - 422
0.7	0.2	0.3	0.7390	344	476 - 1374	374 - 790	346 - 508	346 - 456
0.7	0.2	0.7	0.7510	383	530 - 1524	416 - 876	386 - 564	384 - 506
0.7	0.2	1.0	0.7600	417	574 - 1652	452 - 950	420 - 612	418 - 550
0.7	0.3	0.3	0.7585	411	566 - 1630	446 - 936	414 - 604	412 - 542
0.7	0.3	0.7	0.7765	491	672 - 1928	532 - 1110	494 - 716	492 - 644
0.7	0.3	1.0	0.7900	566	770 - 2204	610 - 1270	570 - 822	568 - 738
0.7	0.4	0.3	0.7780	499	682 - 1956	540 - 1126	502 - 728	500 - 654
0.7	0.4	0.7	0.8020	645	876 - 2502	696 - 1442	650 - 934	646 - 840
0.7	0.4	1.0	0.8200	798	1078 - 3064	858 - 1768	802 - 1148	800 - 1034
0.8	0.1	0.3	0.8130	733	992 - 2826	788 - 1630	738 - 1056	734 - 952
0.8	0.1	0.7	0.8170	769	1040 - 2960	826 - 1706	774 - 1108	770 - 998
0.8	0.1	1.0	0.8200	798	1078 - 3064	858 - 1768	802 - 1148	800 - 1034
0.8	0.2	0.3	0.8260	860	1158 - 3294	924 - 1900	864 - 1234	862 - 1112
0.8	0.2	0.7	0.8340	953	1282 - 3638	1022 - 2100	958 - 1366	956 - 1230
0.8	0.2	1.0	0.8400	1033	1388 - 3932	1108 - 2272	1038 - 1476	1036 - 1332
0.8	0.3	0.3	0.8390	1019	1368 - 3882	1092 - 2242	1024 - 1458	1022 - 1314
0.8	0.3	0.7	0.8510	1207	1614 - 4568	1292 - 2640	1212 - 1718	1208 - 1550
0.8	0.3	1.0	0.8600	1381	1844 - 5206	1476 - 3010	1388 - 1962	1382 - 1770
0.8	0.4	0.3	0.8520	1224	1638 - 4634	1310 - 2678	1230 - 1744	1226 - 1572
0.8	0.4	0.7	0.8680	1567	2088 - 5886	1674 - 3406	1574 - 2220	1570 - 2004
0.8	0.4	1.0	0.8800	1922	2552 - 7178	2048 - 4156	1930 - 2714	1924 - 2450

Table 10: *Results of the sample size calculation based on the naive average hazard ratio* $\bar{\theta}_{\text{naive}}$ *in NPH scenarios with increasing hazard.* θ *= maximum treatment effect,* lag_2 *= delay proportion,* lag_1 *= changepoint proportion,* n_{events} *= number of events,* n_{obs} *= number of observations,* τ *= overall study duration*

θ	lag_2	lag_1	$\bar{\theta}_{\text{naive}}$	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
0.5	0.1	0.3	0.5325	80	90 - 682	80 - 210	80 - 98	80 - 88
0.5	0.1	0.7	0.5425	84	96 - 718	84 - 220	84 - 104	84 - 92
0.5	0.1	1.0	0.5500	88	100 - 748	88 - 230	88 - 108	88 - 96
0.5	0.2	0.3	0.5650	97	108 - 814	98 - 250	98 - 118	98 - 106
0.5	0.2	0.7	0.5850	110	122 - 910	110 - 280	110 - 132	110 - 118
0.5	0.2	1.0	0.6000	121	134 - 994	122 - 306	122 - 144	122 - 130
0.5	0.3	0.3	0.5975	119	132 - 980	120 - 302	120 - 142	120 - 128
0.5	0.3	0.7	0.6275	145	160 - 1174	146 - 362	146 - 172	146 - 156
0.5	0.3	1.0	0.6500	170	186 - 1356	170 - 418	170 - 200	170 - 182
0.5	0.4	0.3	0.6300	148	162 - 1194	148 - 368	148 - 174	148 - 158
0.5	0.4	0.7	0.6700	196	214 - 1552	196 - 478	196 - 230	196 - 208
0.5	0.4	1.0	0.7000	247	268 - 1922	248 - 594	248 - 288	248 - 262
0.6	0.1	0.3	0.6260	144	158 - 1164	144 - 358	144 - 170	144 - 154
0.6	0.1	0.7	0.6340	152	166 - 1224	152 - 376	152 - 180	152 - 162
0.6	0.1	1.0	0.6400	158	174 - 1272	158 - 392	158 - 186	158 - 170
0.6	0.2	0.3	0.6520	172	188 - 1374	172 - 424	172 - 202	172 - 184
0.6	0.2	0.7	0.6680	193	210 - 1530	194 - 472	194 - 226	194 - 206
0.6	0.2	1.0	0.6800	212	230 - 1662	212 - 514	212 - 248	212 - 224
0.6	0.3	0.3	0.6780	208	226 - 1640	208 - 506	208 - 244	208 - 222
0.6	0.3	0.7	0.7020	251	272 - 1952	252 - 604	252 - 292	252 - 266
0.6	0.3	1.0	0.7200	291	314 - 2240	292 - 694	292 - 338	292 - 308
0.6	0.4	0.3	0.7040	255	276 - 1980	256 - 612	256 - 296	256 - 270
0.6	0.4	0.7	0.7360	335	360 - 2550	336 - 792	336 - 386	336 - 352
0.6	0.4	1.0	0.7600	417	448 - 3140	418 - 976	418 - 478	418 - 438

θ	lag_2	lag_1	$\bar{\theta}_{\text{naive}}$	n_{events}	n_{obs} for $\tau = 12$	n_{obs} for $\tau = 24$	n_{obs} for $\tau = 48$	n_{obs} for $\tau = 60$
0.7	0.1	0.3	0.7195	290	314 - 2232	290 - 692	290 - 336	290 - 306
0.7	0.1	0.7	0.7255	305	330 - 2342	306 - 726	306 - 352	306 - 322
0.7	0.1	1.0	0.7300	317	342 - 2428	318 - 752	318 - 366	318 - 334
0.7	0.2	0.3	0.7390	344	370 - 2616	344 - 812	344 - 396	344 - 362
0.7	0.2	0.7	0.7510	383	412 - 2900	384 - 900	384 - 440	384 - 404
0.7	0.2	1.0	0.7600	417	448 - 3140	418 - 976	418 - 478	418 - 438
0.7	0.3	0.3	0.7585	411	440 - 3098	412 - 962	412 - 472	412 - 432
0.7	0.3	0.7	0.7765	491	524 - 3664	492 - 1140	492 - 562	492 - 514
0.7	0.3	1.0	0.7900	566	604 - 4190	566 - 1306	566 - 644	566 - 592
0.7	0.4	0.3	0.7780	499	532 - 3718	500 - 1158	500 - 570	500 - 522
0.7	0.4	0.7	0.8020	645	688 - 4752	646 - 1482	646 - 734	646 - 674
0.7	0.4	1.0	0.8200	798	848 - 5820	798 - 1818	798 - 904	798 - 832
0.8	0.1	0.3	0.8130	733	780 - 5366	734 - 1676	734 - 832	734 - 766
0.8	0.1	0.7	0.8170	769	818 - 5618	770 - 1754	770 - 872	770 - 802
0.8	0.1	1.0	0.8200	798	848 - 5820	798 - 1818	798 - 904	798 - 832
0.8	0.2	0.3	0.8260	860	912 - 6252	860 - 1954	860 - 974	860 - 896
0.8	0.2	0.7	0.8340	953	1010 - 6906	954 - 2160	954 - 1078	954 - 994
0.8	0.2	1.0	0.8400	1033	1094 - 7462	1034 - 2334	1034 - 1166	1034 - 1076
0.8	0.3	0.3	0.8390	1019	1080 - 7364	1020 - 2304	1020 - 1152	1020 - 1062
0.8	0.3	0.7	0.8510	1207	1276 - 8666	1208 - 2714	1208 - 1360	1208 - 1256
0.8	0.3	1.0	0.8600	1381	1460 - 9872	1382 - 3094	1382 - 1554	1382 - 1436
0.8	0.4	0.3	0.8520	1224	1296 - 8788	1224 - 2752	1224 - 1380	1224 - 1274
0.8	0.4	0.7	0.8680	1567	1656 - 11160	1568 - 3502	1568 - 1762	1568 - 1628
0.8	0.4	1.0	0.8800	1922	2028 - 13604	1922 - 4272	1922 - 2156	1922 - 1996

Estimand and Methods

This simulation study is designed to assess the power of the different methods to compare two groups that were introduced in Section 2.3. Some of these methods require that parameters for analysis are prespecified and hence different restrictions on the choice of these parameters must be met:

For the RMST and milestone methods, as mentioned above, the null hypothesis is only a superset of the global null hypothesis H_0^{global} so that for the scenarios under consideration in this simulation study, the parameters of these methods must be chosen carefully ensuring that the scenario is not part of the local null hypothesis. Take for example the threshold lag scenario with a study duration of 60 months, where the survival functions are equal during the first $0.4 \cdot 60 = 24$ months and then start to diverge. This is a scenario of the global alternative hypothesis $H_1^{\text{global}}\colon S_C(t) \neq S_E(t)$ for some t, but when considering the 1-year or 2-year milestone survival the survival rates are identical and hence it is a scenario under the local null hypothesis. Hence, for the scenario under consideration to fall under the local alternative the truncation time τ and the time point t_0 must be chosen to be bigger than the changepoint t_1^*. A further restriction is that these methods are based on the Kaplan-Meier estimator, which poses a practical difficulty since the Kaplan-Meier estimate is only defined up to the longest observed time if this time is censored. Hence both the truncation time τ for the RMST as well as the timepoint t_0 used for rate comparisons must be chosen to be before the minimum of the last observed times in each treatment group of each simulated dataset.

The Landmark and GenLin methods downweight early events to improve the power to detect specific alternatives and hence target null hypotheses of the form $H_0\colon S_C(t) = S_E(t)$ for all $t > t_0$ based on a prespecified timepoint t_0. As all scenarios in this simulation study are scenarios with delayed treatment effect this imposes no restriction on the choice of the timepoint t_0. However, calculation of these methods can be problematic when no or only few events occur after the timepoint t_0 so it should be chosen reasonably small. Based on the choice of t_0 these methods can appear to be superior to all other methods if the weights chosen correspond to the true alternative under consideration and it should hence be investigated whether they are susceptible to parameter misspecification. Now, to take all of this into account, the following choices have been implemented and are summarized in Table 11.

- **RMST and milestone:** As mentioned before for each simulated dataset $j = 1, \ldots, n_{\text{sim}}$ the truncation time for RMST and the time point for rate comparisons should lie within the interval $(t_1^*, \min(t^j_{\max,C}, t^j_{\max,T})]$, where $t^j_{\max,C}, t^j_{\max,T}$ denote the maximum observed times in the respective group of dataset j. As the power of the test increases the more information after the assumed changepoint is taken into account a value was chosen that was close to the upper limit of all of these n_{sim} intervals. To do so, for each simulated dataset this upper limit was calculated and the minimum over all datasets of the same scenario was taken, i.e., the truncation time and milestone time was chosen to be:

$$\min_{j=1,\ldots,n_{\text{sim}}} \left(\min \left(t^j_{\max,C}, t^j_{\max,T} \right) \right)$$

- **Landmark analysis:** Landmark analysis considers only those individuals who have survived until a pre-specified landmark time and ignores the data before that time. Since the survival curves are identical until the change point t_1^*, the data before this time should not contribute to an estimated difference between treatments. Hence, for the choice of the landmark times in the main analysis the true change point t_1^* is used in scenarios in which a delay exists and 0.2τ in scenarios with proportional hazards.
  As a sensitivity analysis misspecification of this parameter was investigated by choosing the landmark time too early or too late in scenarios in which a delay exists or by ignoring more of the early events in proportional hazards scenarios without delay. This yields the following parameters for sensitivity analyses:
  - PH scenarios: $0.1\tau, 0.3\tau$
  - Non-PH scenarios: $0.9t_1^*$ (too early) and $1.1t_1^*$ (too late)

- **Generalized linear lag model:** The aim of the models in this class is to mimick the course of the hazard ratio by downweighting early events, increasing the weights during a transition period and giving full weight for all events thereafter. To do so the assumed start and end of the transition period must be prespecified and for the main analysis was taken to be close to the true parameter, i.e.
  - PH scenarios ($t_2^* = 0$): $[0, 0.2\tau]$
  - Threshold lag scenarios ($t_2^* \neq 0, t_1^* = t_2^*$): $[0.9t_2^*, 1.1t_2^*]$
  - Generalized linear lag scenarios ($t_2^* \neq 0, t_1^* \neq t_2^*$): $[t_1^*, t_2^*]$.

To assess the robustness of the analysis against misspecification the following intervals are specified as sensitivity analyses:

- PH scenarios ($t_2^* = 0$): $[0, 0.3\tau], [0.1, 0.3\tau]$
- Threshold lag scenarios ($t_2^* \neq 0, t_1^* = t_2^*$): $[0.8t_2^*, 1.2t_2^*], [0.7t_2^*, 1.3t_2^*]$
- Generalized linear lag scenarios ($t_2^* \neq 0, t_1^* \neq t_2^*$):
  1. Too big: Take the midpoint of the interval $\frac{t_1^*+t_2^*}{2}$ and increase the width by 10%. This gives the interval $\left[\frac{t_1^*+t_2^*}{2} \pm 1.1(t_2^* - t_1^*)\right]$.
  2. Too small: Take the midpoint of the interval $\frac{t_1^*+t_2^*}{2}$ and decrease the width by 10%. This gives the interval $\left[\frac{t_1^*+t_2^*}{2} \pm 0.9(t_2^* - t_1^*)\right]$.

Table 11: *Parameters for analysis in each power scenario.* $t^j_{\max,g}$ *= maximum observed time in simulated dataset* j *and treatment arm* g*,* τ *= overall study duration,* t_1^* *= changepoint in GLLM,* t_2^* *= delay in GLLM*

Method	Also applied to	Parameter	
RMST	Mile, ABC	$\lfloor \min_{j=1,\ldots,n_{\text{sim}}} \left(\min\left(t^j_{\max,C}, t^j_{\max,T}\right)\right) \rfloor$	
Landmark	MWLRT, Thres, V0, PWExp, ParGroup, LLRNA, QLRNA	PH scenarios:	$\begin{cases} 0.1\tau, \\ 0.2\tau, \\ 0.3\tau \end{cases}$
		NPH scenarios:	$\begin{cases} 0.9 \cdot t_1^*, \text{ too early} \\ t_1^*, \text{ correct} \\ 1.1 \cdot t_1^*, \text{ too late} \end{cases}$
GenLin	PWExpLag, MERT, (m)Logit	PH scenarios:	$\begin{cases} [0, 0.3 \cdot \tau], \\ [0, 0.2 \cdot \tau], \\ [0.1 \cdot \tau, 0.3 \cdot \tau] \end{cases}$
		Threshold lag scenarios:	$\begin{cases} [0.9 \cdot t_2^*, 1.1 \cdot t_2^*] \\ [0.8 \cdot t_2^*, 1.2 \cdot t_2^*] \\ [0.7 \cdot t_2^*, 1.3 \cdot t_2^*] \end{cases}$
		Linear lag scenarios:	$\begin{cases} \left[\frac{t_1^*+t_2^*}{2} \pm 1.1 \cdot (t_2^* - t_1^*)\right] \\ [t_1^*, t_2^*] \\ \left[\frac{t_1^*+t_2^*}{2} \pm 0.9 \cdot (t_2^* - t_1^*)\right] \end{cases}$

Performance measure

The power was chosen as performance measure to evaluate the different methods and calculated as the proportion of rightly rejected null hypotheses within each simulated scenario. To

control the Monte-Carlo standard error the number of simulated datasets n_{sim} was calculated as described by Morris et al. (2019).

The Monte-Carlo standard error for the power is given by

$$\sqrt{\frac{\hat{\beta}(1-\hat{\beta})}{n_{sim}}}.$$

As it is the goal of this thesis to assess the power of the different methods in the different scenarios, the power is unknown a priori and hence the number of simulation runs is determined for the worst case of 50% power. To keep the Monte-Carlo standard error below 1% a total of $n_{sim} = 2500$ simulated datasets are needed in this case.

2.4.2 Assessment of Type I error

In the following the ADEMP structure of the power assessment is described in detail.

Aim

To assess the type I error of the different methods in scenarios comparable to the scenarios considered in the power simulation.

Data-generating mechanism

The null scenarios, i.e. scenarios with equally distributed survival times in both arms, were chosen to be comparable to the power scenarios of the previous section. This reduces the parameters to

- the study duration $\tau = 12, 24, 48, 60$ months
- uniform accrual over the interval $[0, a]$, where $a = 0.2\tau$, 0.4τ.
- Weibull distributed $\text{Wei}(\lambda_C, k_C)$ survival times with shape parameter $k_C = 0.5, 1, 2$ and scale parameter λ_C chosen in dependence on the median survival times of $\text{med}_C = 5, 15, 20$.

This results in total number of $4 \cdot 2 \cdot 3 \cdot 3 = 72$ scenarios. n_{obs} was chosen to be the same as the sample size calculated for the PH scenarios of the maximal treatment effect $\theta = 0.5$.

Estimand and Methods

The choice of the parameters for analysis is somewhat arbitrary since there is no meaningful choice for null scenarios where survival times are equally distributed in both arms. However, as the choice of the parameters for analysis should not influence type I error they were set based on the choice of the parameters in the power scenario and are summarized in Table 12.

Performance measure

The type I error rate α was chosen as performance measure to evaluate the different methods and as for the power the number of simulated datasets was chosen to be $n_{sim} = 2500$.

Table 12: *Parameters for analysis in each null scenario.* $t^j_{\text{max},g}$ *= maximum observed time in simulated dataset* j *and treatment arm* g*,* τ *= overall study duration*

Method	Also applicable to	Parameter
RMST	Mile, ABC	$\lfloor \min_{j=1,\ldots,n_{\text{sim}}} \left(\min \left(t^j_{\text{max},C}, t^j_{\text{max},T} \right) \right) \rfloor$
Landmark	MWLRT, Thres, PWExp, V0, ParGroup, LLRNA, QLRNA	0.2τ
GenLin Model	PWExpLag, MERT, (m)Logit	$[0, 0.2 \cdot \tau]$

The Monte-Carlo standard error is then sufficiently small and given as

$$\sqrt{\frac{\widehat{\alpha}(1-\widehat{\alpha})}{n_{sim}}} \approx 0.0044,$$

in case all methods control the assumed two-sided level of $\alpha = 0.05$. If methods do not control type 1 error the Monte-Carlo standard error is bounded by 1% as outlined for the power above.

Chapter 3

Results

In this chapter the results of the objectives under investigation for this thesis are presented. In the first section the results of the systematic literature search are shown. The following sections present the results of the simulation study elaborating on the impact of a potentially delayed treatment effect on the rejection rate of the different methods identified in the systematic literature search, which were explained in detail in the previous chapter. At first the results for the type I error are presented and discussed in Section 3.2. Then the results for the power of the different methods are shown and examined under various aspects in Section 3.3. Lastly, a ranking system summarizes the performance of the methods under different assumptions to facilitate the choice of an appropriate method at the planning stage in Section 3.4. For ease of comprehension, methods are always given with their abbreviation, that is used in the plots and was summarized in Table 3, in italic throughout this chapter.

3.1 Literature search

The aim of the literature search was to identify methods that have already been proposed to analyze time-to-event data in non-proportional hazard scenarios. To see how the present extensive simulation study can contribute to the current research only articles have been included in which the methods have already been compared in simulation studies in NPH settings. In Section 2.1 the general approach of the literature search and the number of identified and further screened articles have already been described. In this section the focus

lies on the presentation of the methods that have been examined in the simulation study of the final 42 articles and to describe which of these methods were further considered in the simulation study of this thesis and why. A description of the used methods was given in the previous chapter in Section 2.3.3. The following Figure 12 shows the methods used in each article and if a delayed treatment effect scenario was considered in the simulation studies of the article or not.

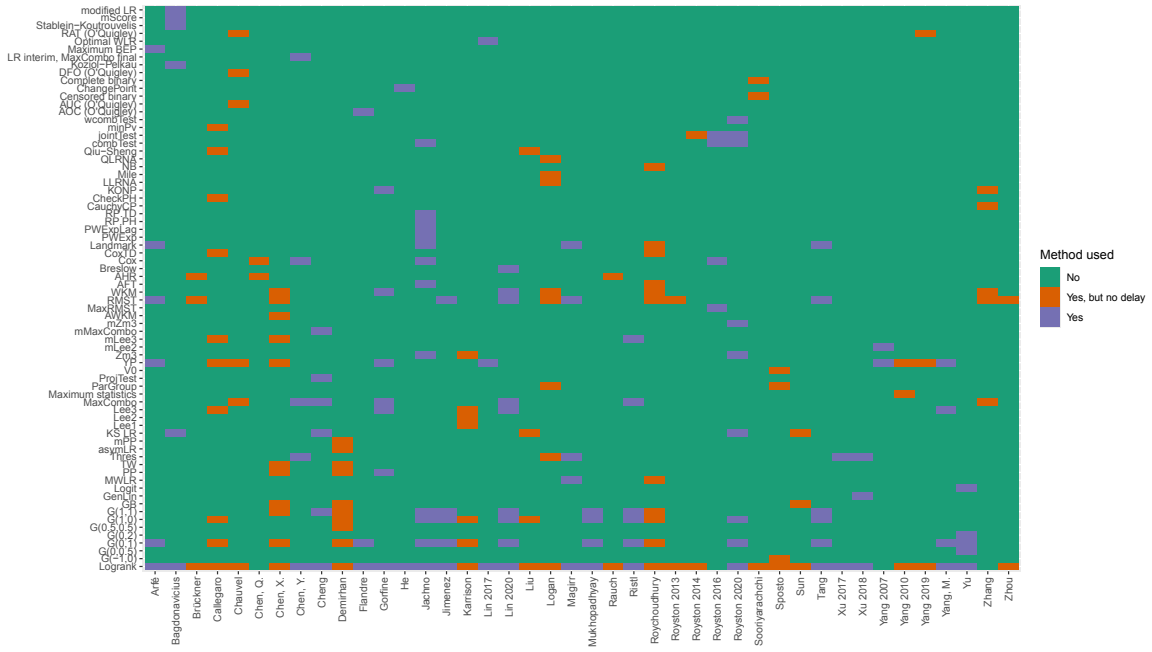

Figure 12: *Results of the systematic literature search showing which method has been used in which publication and if delayed treatment effect scenarios have been considered*

As can be seen the standard *logrank* test was considered in almost all (38/42) of the simulation studies. With 50% the second most used class of methods were combination tests, i.e. tests based on the combination of different weighted logrank statistics (*KS LR, Lee1, Lee2, Lee3, mLee2, mLee3, MaxCombo, mMaxCombo, mZm3, ParGroup, ProjTest, V0, YP, Zm3*). A third of the publications used Kaplan-Meier based tests (*AWKM, MaxRMST, RMST, WKM*) and in 26% of the publications tests based on regression models (*AFT, AHR, Breslow, Cox, CoxTD, Landmark, PWExp, PWExpLag, RP.PH, RP.TD*) were considered.

Implications for simulation study: Although it was endeavored to include all methods in this simulation study, the following three were not applicable in this simulation setting.

Firstly, the *Stablein-Koutrouvelis* method by Stablein and Koutrouvelis (1985) is only applicable for singly censored data, which is data where only the first $r < n$ of the ordered survival times are observed. This can, for example, occur if all subjects are recruited at the same time and observed for a fixed study duration without any drop-outs. In this simulation it was assumed that subjects are recruited linearly over a given accrual period and are then censored at the end of the study given by the time after start of accrual at which the study is terminated. Hence there are indeed no censorings of early survival times, but it can and will happen that the censorings occur although the highest survival time itself is uncensored. Secondly, the *Maximum BEP* method by Arfè et al. (2021) was developed as a test that maximizes the Bayesian expected power (BEP) given early-stage data, which was not available in this simulation study. Lastly, the *LR interim, MaxCombo final* method by Chen et al. (2022) was not applicable as it is designed for group-sequential trials where at interim a standard logrank test and for the final analysis the MaxCombo test is used.

Furthermore, some methods were excluded as they were too complex and not implemented in any statistical software. These methods include all the methods based on the approach by John O'Quigley (*AOC (O'Quigley), AUC (O'Quigley), DFO (O'Quigley), RAT (O'Quigley)*) (Chauvel and O'Quigley, 2014; Flandre and O'Quigley, 2019), the *Koziol-Petkau* method by Koziol (1978), the *complete binary* and *censored binary* method by Sooriyarachchi and Whitehead (1998), the optimally weighted logrank test (*Optimal WLR*) by Lin and León (2017), the two-stage change point approach (*ChangePoint*) by He and Su (2015) and modified logrank type test (*modified LR*) by Bagdonavicius et al. (2004).

3.2 Type I error

As explained in Section 2.4.2 the assessment of type 1 error rate was done for 72 scenarios comparable to the power scenarios. For each of these scenarios $n_{\text{sim}} = 2500$ datasets have been simulated to keep the Monte-Carlo standard error below 0.44% and the methods have been applied to each dataset. However, some of the methods were not evaluable in some of the simulated datasets. This happened primarily for methods for which parameters for analysis had to be chosen if the data was incompatible with the chosen parameter. For example, in case of the methods of the Landmark type this happened if no event occurred after or in case of the Milestone type methods if the Kaplan-Meier estimator was not evaluable at the specified timepoint. If at least 25% of the simulated datasets were evaluable this was considered sufficient as the Monte-Carlo standard error then doubles and is still below 1%.

The following Table 13 shows for each shape of the hazard function in how many of the 24 scenarios the method was non-evaluable at least once. In parentheses the range of the number of non-evaluable datasets is given.

For the 24 decreasing hazard scenarios all methods could be evaluated sufficiently often, which is defined as being evaluable in more than 1875, i.e. 75% of the 2500 simulated datasets. The *AHR, RP.PH* and *RP.TD* method produced missings in all 24 scenarios ranging from 94 (3.76%) to 210 (8.4%) missings for the AHR method and from 14 (0.56%) to 29 (1.16%) missings for the Royston-Parmar models. Milestone survival (*Mile, MileCLL*) was not evaluable for 2 (0.08%) datasets in the scenario with a long study duration of $\tau = 60$ months, a median survival of $\text{med}_C = 5$ months in the control arm and an accrual period of $0.4 \cdot \tau = 24$ months.

As for decreasing hazards all methods could be evaluated sufficiently often for constant hazard scenarios. Missings occurred less often than for decreasing hazards ranging from 1 (0.04%) to 8 (0.32%) datasets where missing results were produced. In contrast to decreasing hazards this only happened in at most 10 of the 24 scenarios and not in all.

With increasing hazards in the control group methods produced missing results more often than in the decreasing and constant hazards scenarios. This is due to the fact that the

Table 13: *Number of scenarios in which each method was non-evaluable at least once for the assessment of type 1 error. Additionally, the number of scenarios in which more than 75% of datasets were not evaluable is given, indicated by an asterisk. Given in parentheses is the range for the number of non-evaluable datasets.*

Method	Decreasing hazard (n=24)	Constant hazard (n=24)	Increasing hazard (n=24)
AHR	24 (94 - 210)	1 (1)	-
LLRNA	-	2 (5 - 8)	4/2* (334 - 1993)
Landmark	-	2 (5 - 8)	4/2* (334 - 1993)
MERT	-	2 (5 - 8)	11/2* (1 - 1993)
MWLRT	-	-	4 (7 - 721)
Mile	1 (2)	4 (1 - 1)	5 (1 - 2)
MileCLL	1 (2)	4 (1 - 1)	5 (1 - 2)
PWExp	-	-	7 (1 - 25)
ParGroup	-	2 (5 - 8)	4/2* (334 - 1993)
QLRNA	-	2 (5 - 8)	4/2* (334 - 1993)
RMST	-	1 (1)	-
RP.PH	24 (14 - 29)	10 (1 - 2)	11 (1 - 4)
RP.TD	24 (14 - 29)	9 (1 - 2)	13 (1 - 4)
Thres	-	2 (5 - 8)	4/2* (334 - 1993)
V0	-	2 (5 - 8)	4/2* (334 - 1993)

parameters for analysis are chosen based on the overall study duration τ and the delay and changepoint parameters t_2^*, t_1^* and not taking the failure time distribution into account. As the failure rate accelerates over time events occur at early timepoints and therefore in many cases before the parameter of analysis which causes the methods to be incalculable.

In total missings occurred in at most 13 of the 24 scenarios and in case of low median survival of 5 months in the control group and a long overall study duration of $\tau = 60$ months 8 methods produced more than 25% missings ranging from approx. 700 not evaluable datasets for the *MWLRT* method up to approx. 2000 non-evaluable datasets for the *Landmark, LLRNA, QLRNA, MERT, ParGroup, Thres* and *V0* method. All these methods have in common that the chosen parameter for analysis was 12 months, which corresponds to 20% of the overall study duration, but is also the 98% quantile of the survival time distribution. Furthermore, all of these methods produced missings in 2 other scenarios ($\tau = 48$ and $\text{med}_C = 5$) with the exception of the *MERT* method which produced missings in 9 other scenarios. The number of missings produced ranged from 1 to 334 for the *MERT* method, the *MWLRT*

produced 7 and 18 missings and all other methods 334 in both additional scenarios. The remaining methods produced almost no missings: 5 scenarios with up to 2 (0.08%) missings for the Milestone survival rate (*Mile, MileCLL*), 7 scenarios with up to 25 (1%) missings for the piecewise exponential model (*PWExp*) and 11 and 13 scenarios with up to 4 (0.16%) missings for the Royston-Parmar model with constant and time-dependent treatment effect (*RP.PH, RP.TD*), respectively.

The following plots show the type I error for the null scenarios considered separately for decreasing (Figure 13), constant (Figure 14) and increasing (Figure 15) hazards. In each plot the calculated type I error values are displayed by a boxplot together with the 24 points for each scenario. Within each plot the logrank test is highlighted by a vertical dashed reference line and the nominal α level of 5% by a horizontal solid reference line. Type I error values for methods that were not evaluable in more than 25% of the simulated datasets are indicated by a square instead of a dot as this corresponds to a twice as high Monte Carlo standard error. The colors show if the method was conservative (yellow), i.e. stayed below the nominal level by more than the Monte-Carlo standard error, had inflated type I error (red), i.e. the estimated type 1 error exceeds the nominal level by more than the Monte-Carlo standard error, or controlled type I error (green).

3.2.1 Decreasing hazards

As can be seen in Figure 13 the *Aalen* additive model and the Milestone rate comparison based on the Nelson-Aalen estimator (*MileNA*) tend to be more conservative than the other methods. This is especially the case if the sample size is low. In contrast, the AFT model based on the exponential and logarithmic error distribution (*AFT (exp), AFT (log)*), the naive PH check procedure (*CheckPH*), the Yang and Prentice model (*YP*), the Royston-Parmar model with time-dependent treatment effect (*RP.TD*), the *ABC* method as well as the piecewise exponential method (*PWExp*) tend to inflate type I error.

In case of the *AFT (exp)* model this is the most severe with a type 1 error inflation in all 24 scenarios ranging from 7.2% up to 20.5% followed by the piecewise exponential model (*PWExp*) exceeding the nominal level in 23 scenarios ranging from 5.2% to 11.9% and the

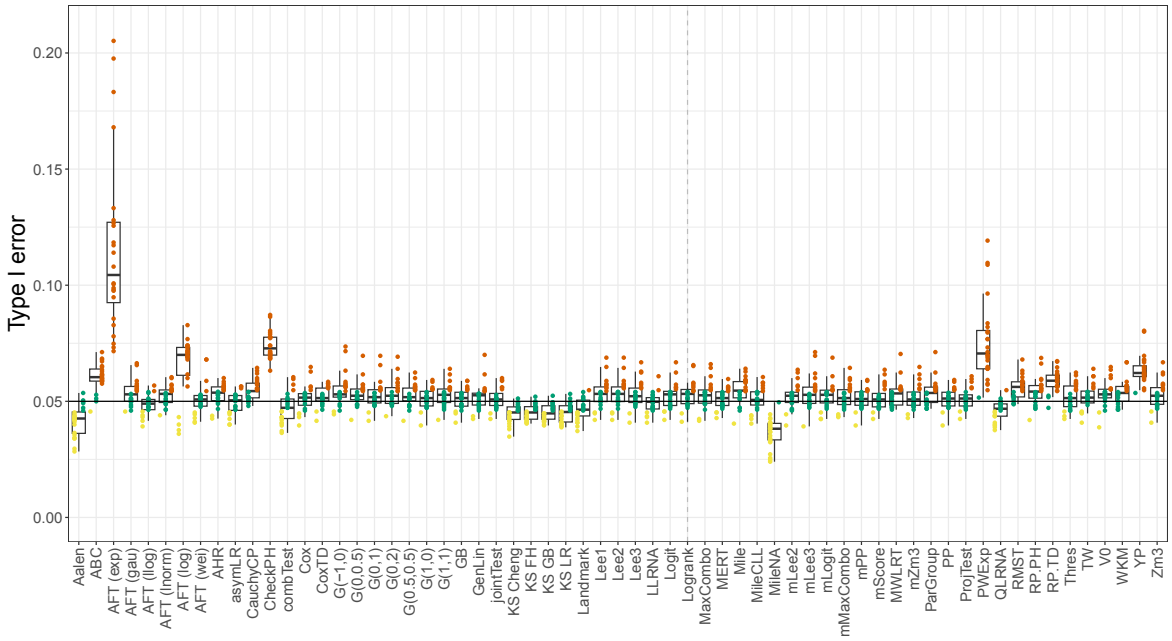

Figure 13: *Type 1 error rates in scenarios with decreasing hazard. The logrank test is highlighted by a vertical dashed reference line and the nominal α level of 5% by a horizontal solid reference line. The colors indicate if type I error is controlled taking the Monte-Carlo standard error into account (yellow=conservative, red=inflated, green=controlled)*

naive PH check procedure (*CheckPH*) ranging from 6.3% to 8.7% in all scenarios. The other three methods showed a type 1 error inflation of at most 8.3%.

3.2.2 Constant hazards

As shown in Figure 14 the *ABC* method and the Milestone rate comparison based on the Nelson-Aalen estimator (*MileNA*) tend to be more conservative than the other methods. In contrast to decreasing hazard scenarios, the AFT model based on the exponential error distribution (*AFT (exp)*) performs well as the distributional assumption of this model is satisfied. Again the naive PH check procedure (*CheckPH*), the Yang and Prentice model (*YP*), the Royston-Parmar model with time-dependent treatment effect (*RP.TD*) and the AFT model with logarithmic error distributions (*AFT (log)*) had a slightly increased type 1 error. This occurred in 24, 22, 19 and 17 scenarios with a maximal type I error of 8.7%, 7.9%, 6.8% and 7.3%, respectively. Additionally, the Gray-Tsiatis test (*G(-1,0)*) revealed a type 1 error inflation in 12 scenarios with the most severe type 1 error increase of up to 11.4%.

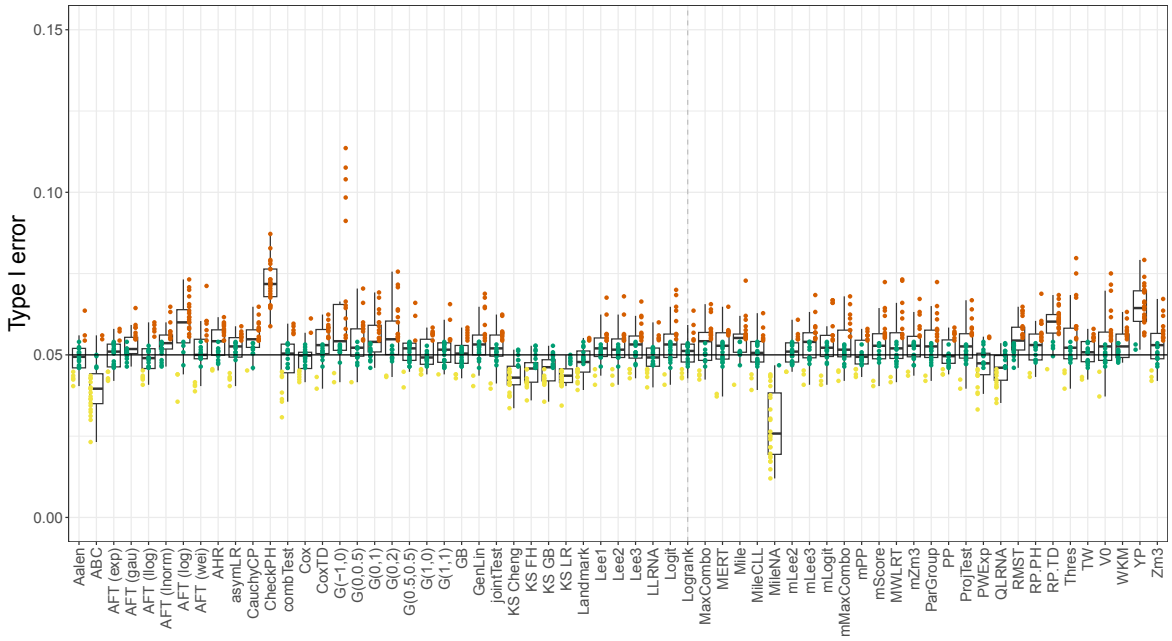

Figure 14: *Type 1 error rates in scenarios with constant hazard. The logrank test is highlighted by a vertical dashed reference line and the nominal α level of 5% by a horizontal solid reference line. The colors indicate if type I error is controlled taking the Monte-Carlo standard error into account (yellow=conservative, red=inflated, green=controlled)*

3.2.3 Increasing hazards

Concerning type 1 error the AFT model with exponentially distributed error terms (*AFT (exp)*), the *QLRNA* test and the piecewise exponential model (*PWExp*) tend to be more conservative than the other methods as shown in Figure 15. Additionaly this is also the case for the *ABC* method and the Milestone rate comparison based on the Nelson-Aalen estimator (*MileNA*), which has already been observed in the decreasing and constant hazards scenarios.

The naive PH check procedure (*CheckPH*), the Royston-Parmar model with time-dependent treatment effect (*RP.TD*), the Yang and Prentice model (*YP*) and the AFT model with logarithmic distributed error terms (*AFT (log)*) had inflated type 1 error in at least 20 of the 24 scenarios with a type 1 error of up to 8.4%. Two methods occassionally had type 1 error over 10%, which are the modestly weighted logrank test (*MWLRT*) with type 1 error of 9.4% to 11.7% in 4 scenarios and the Gray-Tsiatis test (*G(-1,0)*) with type 1 error of 8.8% to 11.4% in 13 scenarios.

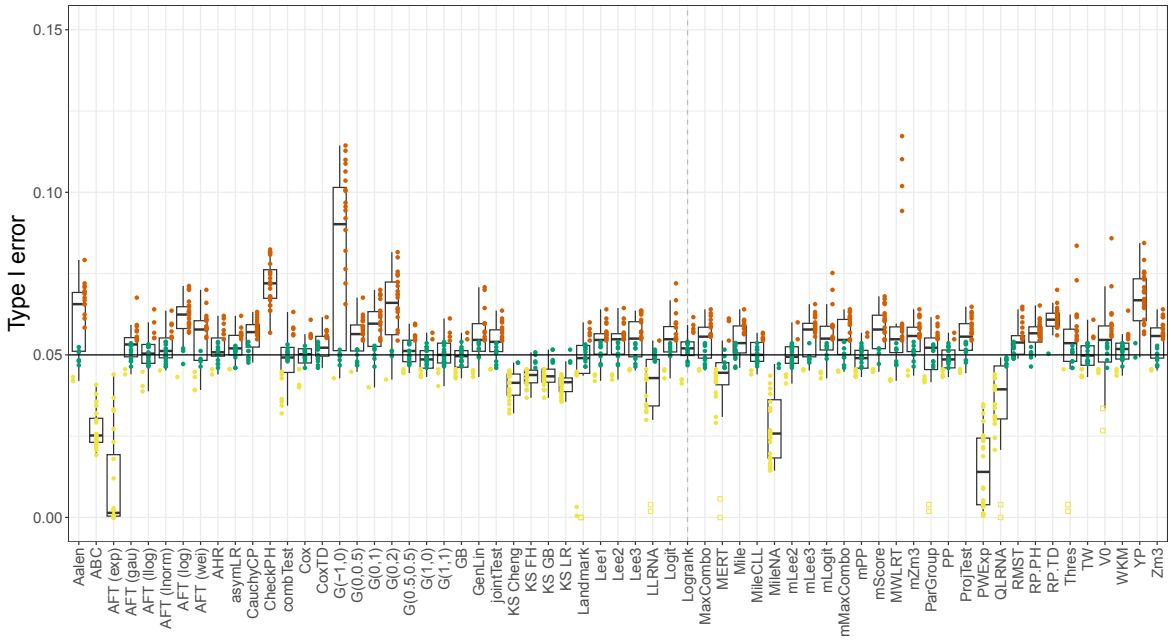

Figure 15: *Type 1 error rates in scenarios with constant hazard. The logrank test is highlighted by a vertical dashed reference line and the nominal α level of 5% by a horizontal solid reference line. Square shape indicates that the method was not evaluable in more than 75% of scenarios. The colors indicate if type I error is controlled taking the Monte-Carlo standard error into account (yellow=conservative, red=inflated, green=controlled)*

3.3 Power

In this section the results of the assessment of power of the methods will be presented. This was done for the 3744 scenarios outlined in Section 2.4.1 of the previous chapter. These 3744 scenarios can be further divided into 1248 scenarios for decreasing, constant and increasing hazard respectively and further into 96 PH scenarios and 1152 NPH scenarios. As for the assessment of type 1 error some of the methods were not evaluable in some of the simulated datasets. Again methods were considered sufficiently often evaluable if the number of missings did not exceed 625, i.e. 25% of the simulated datasets. The following tables present for each shape of the hazard function in how many of the scenarios the method was non-evaluable at least once for PH scenarios (Table 14) and NPH scenarios (Table 15). In parentheses the range of the number of non-evaluable datasets is given. In contrast to the assessment of type 1 error, Milestone comparison of survival rates (*Mile, MileCLL*) and *RMST* were always evaluable and did not produce any missings in any of the scenarios.

Table 14: *Number of PH scenarios in which each method was non-evaluable at least once for the assessment of power. Given in parentheses is the range of the number of non-evaluable datasets.*

Method	Decreasing hazard (n=96)	Constant hazard (n=96)	Increasing hazard (n=96)
AHR	96 (67 - 1264)	15 (1 - 3)	-
LLRNA	-	2 (2 - 3)	12 (6 - 1393)
Landmark	-	2 (2 - 3)	12 (6 - 1393)
MERT	-	2 (2 - 3)	25 (1 - 1393)
MWLRT	-	-	4 (3 - 11)
PWExp	-	-	13 (1 - 68)
PWExpLag	-	-	13 (1 - 68)
ParGroup	-	2 (2 - 3)	12 (6 - 1393)
QLRNA	-	2 (2 - 3)	12 (6 - 1393)
RP.PH	96 (9 - 185)	21 (1 - 3)	15 (1 - 4)
RP.TD	96 (9 - 185)	21 (1 - 3)	15 (1 - 5)
Thres	-	2 (2 - 3)	12 (6 - 1393)
V0	-	2 (2 - 3)	12 (6 - 1393)

In case of constant hazard in the control arm all methods were almost always evaluable and missings occurred in not more than 3 (0.12%) datasets of any scenario. The Royston-Parmar

models (*RP.PH, RP.TD*) were non-evaluable in 21 scenarios and the *AHR* was non-evaluable in 15 scenarios. All other methods were non-evaluable only in the 2 most extreme scenarios with overall study duration of $\tau = 60$ months, a median survival of $\text{med}_C = 5$ months in the control arm and an treatment effect of $\theta = 0.5$.

For decreasing hazard only 3 methods (*AHR, RP.PH, RP.TD*) produced missings which happened in all 96 scenarios. For the Royston-Parmar models the number of non-evaluable datasets was at most 185 (7.4%) and therefore sufficiently small. The *AHR* method on the other hand was more difficult to evaluate and produced missings in up to 1264 (50.56%) datasets. This was negatively correlated with the treatment effect, i.e. the higher the hazard ratio θ , that is the lower the treatment effect, the more datasets were non-evaluable. For the lowest treatment effect of $\theta = 0.8$ the number of non-evaluable datasets ranged from 611 (24.44%) to the maximum of 1264 (50.56%).

With increasing hazard there were seven methods with a very high number of non-evaluable datasets of up to 1393 (55.72%): *LLRNA, Landmark, MERT, ParGroup, QLRNA, Thres, V0*. Such high numbers of missings above 25% of all simulated datasets occurred only in four scenarios, which were the most extreme scenarios with overall study duration of $\tau = 60$ months, a median survival of $\text{med}_C = 5$ months in the control arm and a treatment effect of $\theta = 0.5$ or $\theta = 0.6$.

Now for NPH scenarios the pattern is very similar. Again for constant hazard scenarios the number of non-evaluable datasets was sufficiently low for all methods with at most 332 (13.28%) for the *MERT* method. For decreasing hazard *AHR, RP.PH* and *RP.TD* were non-evaluable in at least one dataset in all 1152 NPH scenarios. The number of non-evaluable datasets was at most 553 (22.12%) for the Royston-Parmar models whereas the *AHR* method exceeded the 25% threshold in 521 of the 1152 NPH scenarios. The number of non-evaluable datasets increased the higher the hazard ratio starting with 114 (4.56%) to 673 (26.92%) non-evaluable datasets for $\theta = 0.5$ up to 779 (31.16%) to 2257 (90.28%) non-evaluable datasets for $\theta = 0.8$. As before increasing hazard scenarios revealed to be more difficult to handle for many methods. For the Royston-Parmar models (*RP.PH, RP.TD*) the number of non-evaluable datasets was very small and at most 3 (0.12 %). The most troublesome method was the *MERT* method which was not evaluable in 303 scenarios at least once and completely

Table 15: *Number of NPH scenarios in which each method was non-evaluable at least once for the assessment of power. Additionally, the number of scenarios in which more than 75% of datasets were not evaluable is given, indicated by an asterisk. Given in parentheses is the range of the number of non-evaluable datasets.*

Method	Decreasing hazard (n=1152)	Constant hazard (n=1152)	Increasing hazard (n=1152)
AHR	1152/30* (114 - 2257)	428 (1 - 9)	-
GenLin	-	3 (1 - 1)	60/28* (1 - 2500)
LLRNA	-	8 (1 - 52)	69/38* (1 - 2500)
Landmark	-	8 (1 - 52)	69/38* (1 - 2500)
MERT	60 (1 - 13)	79 (1 - 332)	303/92* (1 - 2500)
MWLRT	-	1 (1)	58/30* (1 - 2500)
PWExp	-	24 (1 - 53)	293/13* (1 - 2210)
PWExpLag	60 (1 - 13)	50 (1 - 74)	226/28* (1 - 2500)
ParGroup	-	8 (1 - 52)	69/38* (1 - 2500)
QLRNA	-	8 (1 - 52)	69/38* (1 - 2500)
RP.PH	1152 (14 - 553)	56 (1 - 2)	55 (1 - 3)
RP.TD	1152 (14 - 553)	63 (1 - 2)	59 (1 - 3)
Thres	-	8 (1 - 52)	69/38* (1 - 2500)
V0	-	8 (1 - 52)	69/38* (1 - 2500)
mLogit	-	3 (1 - 1)	60/28* (1 - 2500)

non-evaluable for all datasets in 51 scenarios. For the piecewise exponential models (*PWExp, PWExpLag*) the number of scenarios in which at least one dataset was non-evaluable was also high, but it only happened in six scenarios that all datasets were not evaluable and this were the most extreme threshold lag scenarios with overall study duration of $\tau = 60$ months and median survival of $\text{med}_C = 5$ months in the control arm. All other methods in Table 15 except *AHR* produced at least one missing in less than 70 scenarios of which at most in 15 scenarios no dataset was evaluable.

Before the results of the power evaluation are presented the impact of parameter misspecification will be provided (Section 3.3.1). The choice of the parameters for analysis were explained in the paragraph on "Estimand and Methods" within Section 2.4.1 and summarized in Table 11.

3.3.1 Parameter misspecification

As outlined in Section 2.4.1 it is distinguished between methods of the *Landmark* type for which only one parameter needs to be prespecified for the analysis and methods of the *GenLin* type for which two parameters need to be prespecified. For the former methods a timepoint t_{Landmark} is specified and all events prior to this timepoint are ignored whereas all events after that timepoint contribute fully to the test statistic. This is reasonable when a delayed treatment effect is anticipated, but problematic when none is present. Hence for PH scenarios there is no right choice of the Landmark parameter and therefore the timepoints were chosen arbitrarily as 10%, 20% and 30% of the overall study duration τ. For these scenarios the impact of the different parameter choices is shown in the following boxplot (Figure 16).

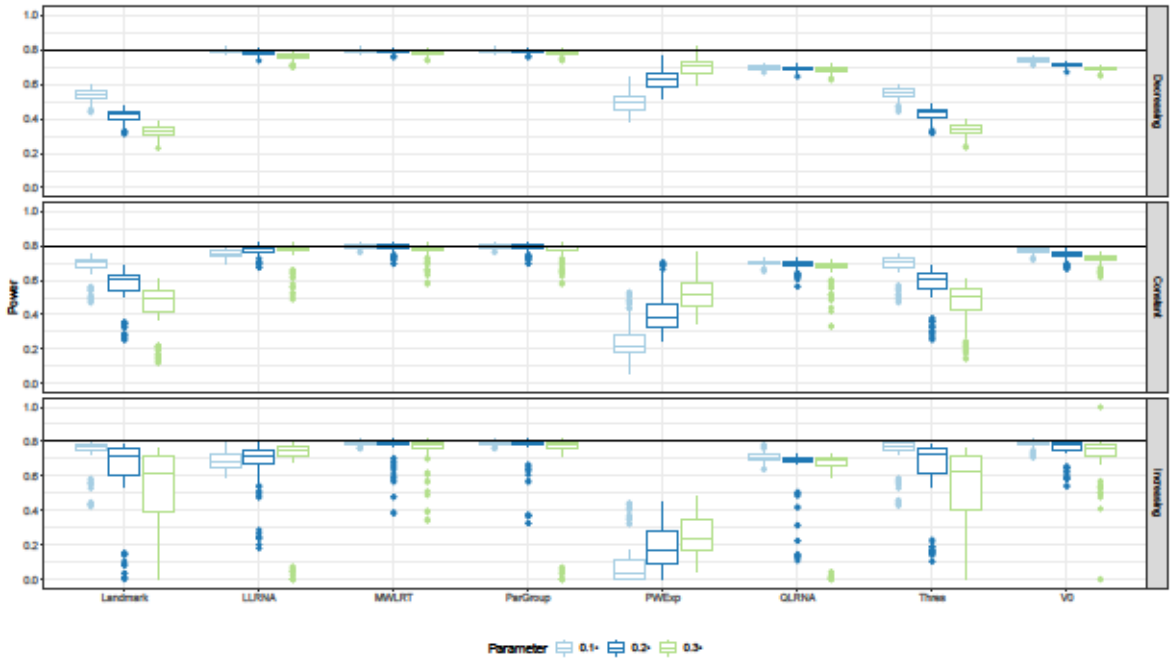

Figure 16: *Boxplot over all scenarios of the power of the Landmark type methods based on the different choices of the parameter* t_{Landmark} *in PH scenarios arranged by the shape of the hazard.*

As can be seen in this plot the impact of the different choices of the parameter t_{Landmark} on the modestly weighted logrank test (*MWLRT*) and the partially grouped logrank test (*ParGroup*) is negligible and the impact on the *V0* test and the linear or quadratic combination of logrank and Nelson-Aalen estimator (*LLRNA*, *QLRNA*) is also minor and only occurs in scenarios with a very low median survival in the control arm of $\text{med}_C = 5$ months in combination

with a very high overall study duration of $\tau = 48$ or $\tau = 60$ months. The impact on the *Landmark* and *Thres* method is as expected where the power of the single methods is reduced the more early events are ignored. The piecewise exponential model shows a somewhat contradictory behavior of increasing power the later the specified Landmark parameter. This can be explained by the model structure, which does not simply ignore early events but allows different constant hazards before and after the prespecified parameter. To estimate these parameters, events before the Landmark timepoint must be observed which is improved by choosing later timepoints for the analysis. Next for the NPH scenarios it makes more sense to assess parameter misspecification as the true parameter is the changepoint t_1^* up to which survival in both arms is equal and then starts to diverge. To assess the impact of misspecification the Landmark parameter was chosen to under- and overestimate the true changepoint by 10%.

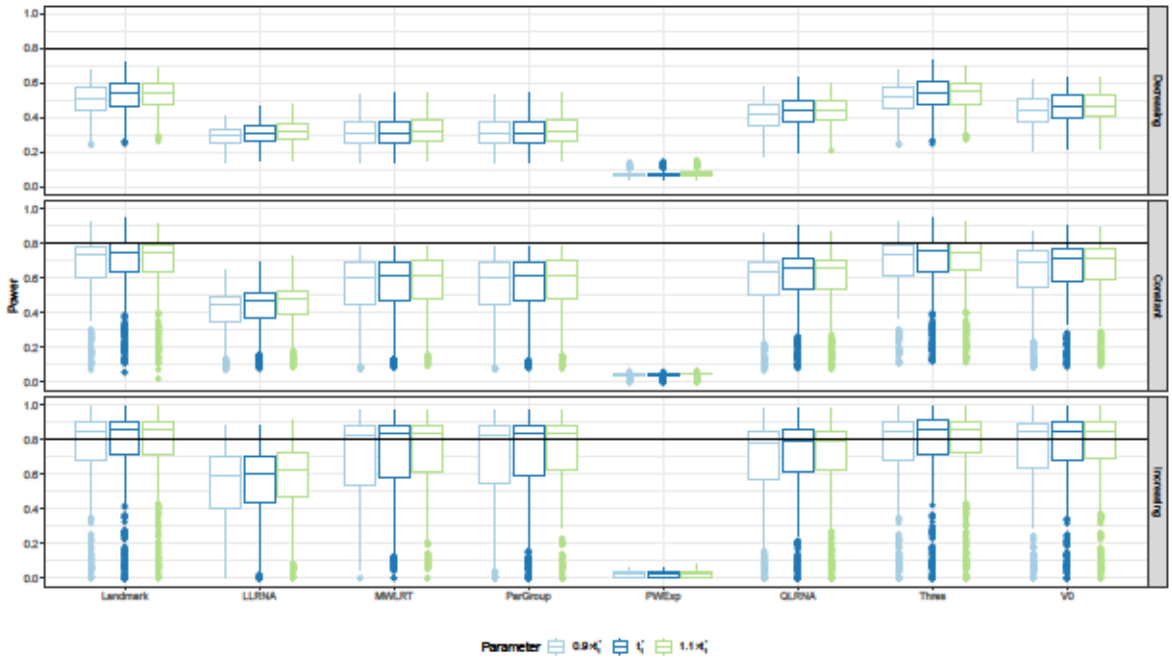

Figure 17: *Boxplot over all scenarios of the power of the Landmark type methods based on different choices of the parameter* t_{Landmark} *in NPH scenarios arranged by the shape of the hazard.*

In Figure 17 it can be seen that the impact of misspecifying the parameter is negligible for all methods. Detailed investigation of the methods showed that differences become more apparent when the delay and changepoint proportion are bigger as this leads to an greater changepoint t_1^* and with that the misspecification becomes more severe. In these cases un-

derestimation of the true changepoint t_1^* led to an decrease in power so that it is beneficial to ignore slightly more events than necessary.

For the GenLin type methods an interval $[t_{\text{low}}, t_{\text{up}}]$ is specified and all events prior to timepoint t_{low} are ignored while the weights for events within the interval are increasing and events after t_{up} get full weight. The correct choice of these parameters is $[t_1^*, t_2^*]$ in linear lag scenarios but somewhat arbitrary in PH and Threshold lag scenarios. For PH scenarios the impact of the different parameter choices is shown in the following boxplot (Figure 18).

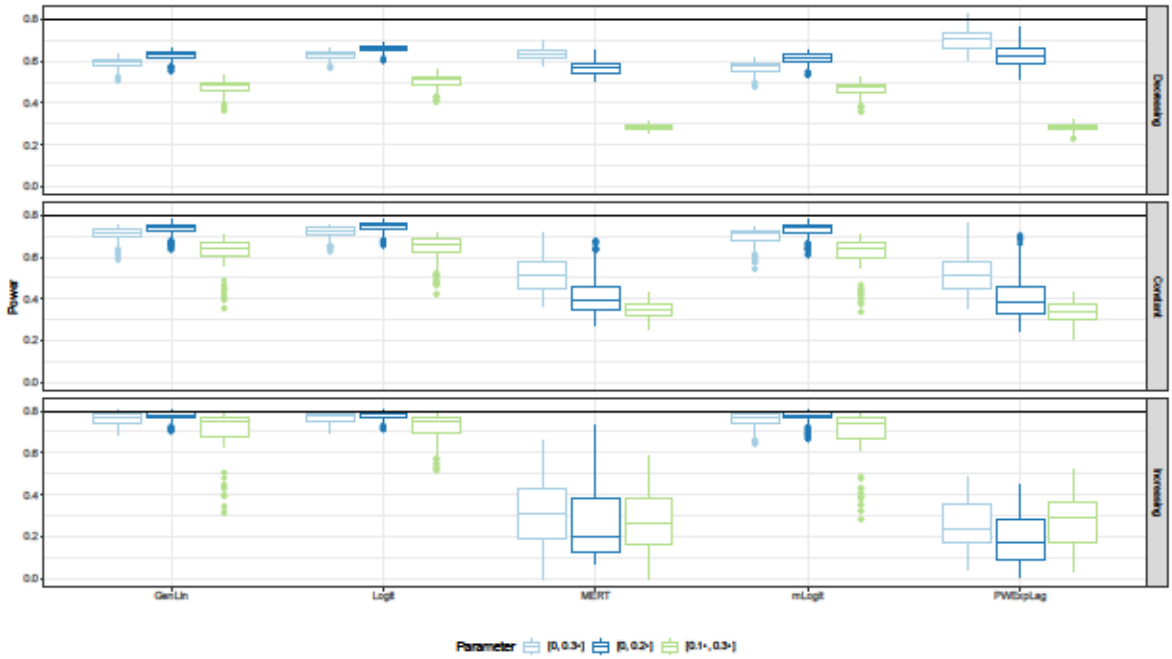

Figure 18: *Boxplot over all scenarios of the power of the GenLin type methods based on the different choices of the interval* $[t_{\text{low}}, t_{\text{up}}]$ *in PH scenarios arranged by the shape of the hazard.*

For the *GenLin*, *Logit* and modified logit method (*mLogit*) the worst choice of parameters is the one where events within the first 10% of the overall study duration are ignored ($[0.1\tau, 0.3\tau]$). Of the other two choices where no events are ignored the one performs better where the phase of increasing weights is smaller, i.e. the choice $[0, 0.2\tau]$. The remaining two methods (*MERT, PWExpLag*) showed a different effect. For decreasing and constant hazards $[0.1\tau, 0.3\tau]$ also was the worst choice of parameters followed by $[0, 0.2\tau]$ and then $[0, 0.3\tau]$, which is in line with the observation for the piecewise exponential model previously. For increasing hazards the two parameter choices with the later upper GenLin parameter of $t_{\text{up}} = 0.3\tau$ were very similar and better than the choice $t_{\text{up}} = 0.2\tau$ except for scenarios

with very low median survival in the control arm of $\text{med}_C = 5$ months and late overall study duration of $\tau = 48, 60$ months. For threshold lag scenarios intervals of increasing width were chosen centered around the true delay t_2^*. The impact of these choices on the power of the GenLin type methods is displayed in Figure 19.

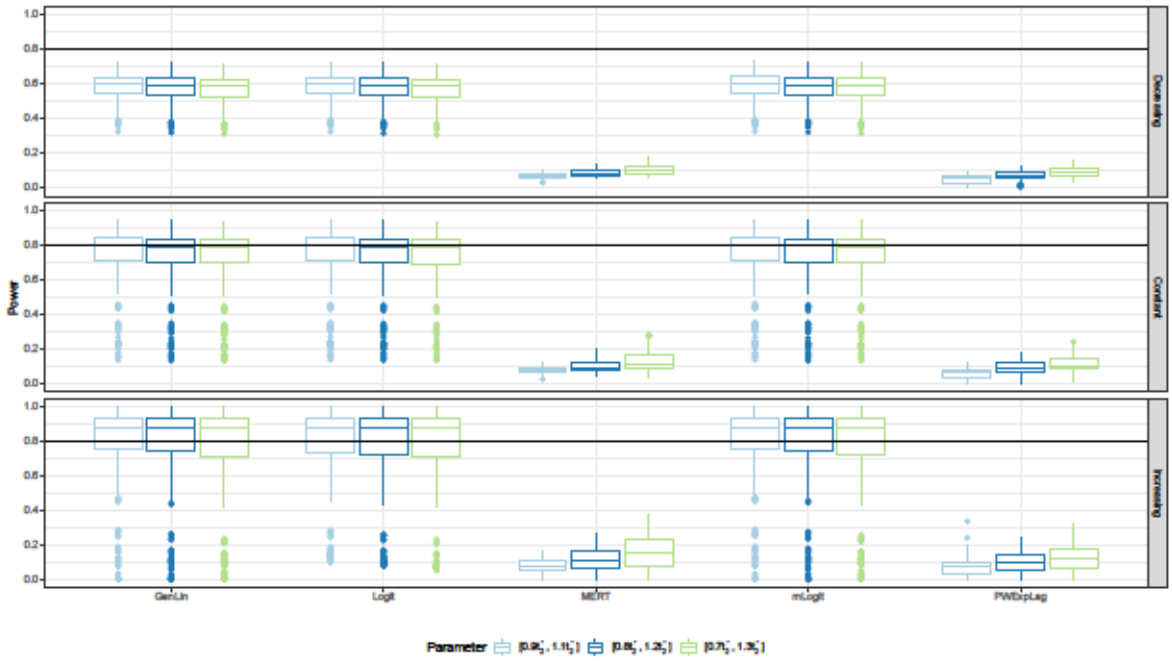

Figure 19: *Boxplotover all scenarios of the power of the GenLin type methods based on the different choices of the interval* $[t_{\text{low}}, t_{\text{up}}]$ *in threshold lag scenarios arranged by the shape of the hazard.*

As for PH scenarios the *GenLin*, *Logit* and modified logit method (*mLogit*) performed very similar but were not affected by the different choices of parameters. The *MERT* and*PWExpLag* method showed again to benefit from a wider interval, i.e. a wider timeinterval to estimate the initial exponential parameter. Lastly, the power of the methods in NPH scenarios was not impacted by the different choices of analysis parameters as shown in Figure 20.

The results for the assessment of the power of the methods are then structured by the shape of the underlying hypothetical distribution of the survival times in the control arm distinguishing between decreasing (Section 3.3.2), constant (Section 3.3.3) and increasing hazards (Section 3.3.4). Within each part the PH scenarios are reported first as they can be understood as benchmark for all other scenarios where the proportional hazards assumption is violated.

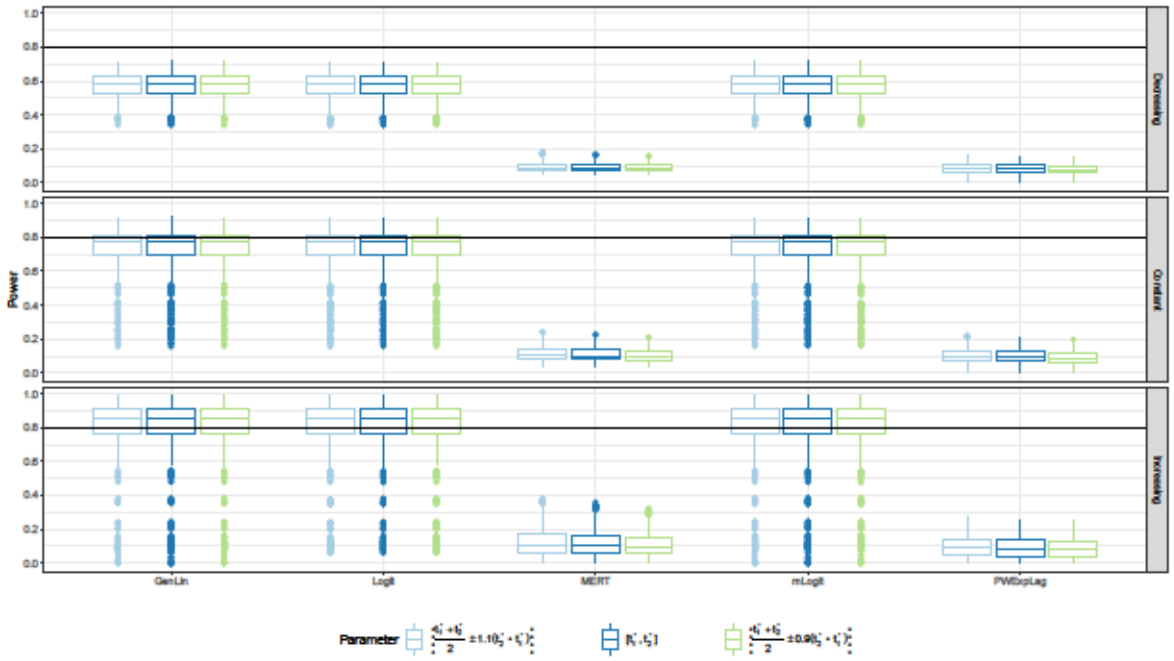

Figure 20: *Boxplot over all scenarios of the power of the GenLin type methods based on the different choices of the interval* $[t_{\text{low}}, t_{\text{up}}]$ *in linear lag scenarios arranged by the shape of the hazard.*

3.3.2 Decreasing hazards

In this section the power of the methods is compared if the hazard in the control arm is decreasing. The section commences with a direct comparison of the methods to the logrank test by calculating in how many of the 1248 simulated scenarios each method exceeds the power of the logrank test by more than the Monte-Carlo standard error given the number of evaluable datasets. This is displayed in a bar chart in Figure 21. A green bar indicates that the method exceeds the logrank test in all NPH scenarios and a proportion of zero indicates that the method has lower power than the logrank test in all NPH scenarios.

The AFT model with exponentially distributed error terms (*AFT (exp)*) outperforms the logrank test in all scenarios - even in the 96 PH scenarios - which can be explained by the inflated type 1 error of this method. As has been seen in the previous section the Yang and Prentice model (*YP*) and the naive PH check procedure (*CheckPH*) have also inflated type 1 error. Expectedly, they outperform the logrank test in 22 and 16 PH scenarios, respectively.

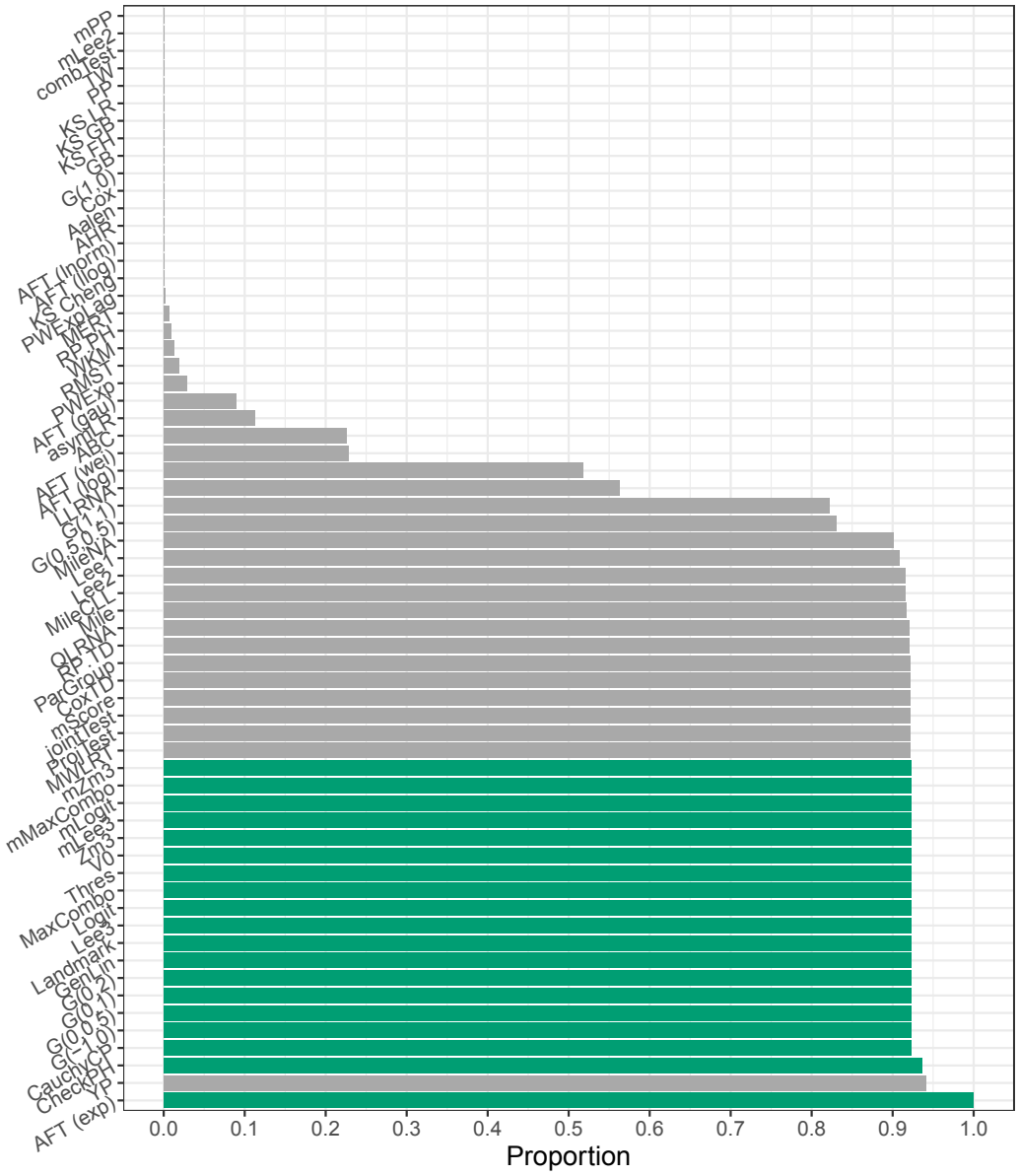

Figure 21: *Proportion of 1248 scenarios in which each method exceeds the power of the logrank test in decreasing hazard scenarios. Green bars indicate that method outperforms the logrank test in all NPH scenarios.*

Of the remaining methods which have been seen to control type 1 error, none had higher power than the logrank test in the PH scenarios, which is in line with the result by Peto and Peto (1972). The following methods, however, outperform the logrank test in all NPH scenarios: *CauchyCP*, Gray-Tsiatis test (*G(-1,0)*), Fleming-Harrington test for late differences (*G(0,0.5)*, *G(0,1)*, *G(0,2)*), Generalized linear model (*GenLin*), *Landmark* analysis,

(modified) *Lee3*, (modified) *Logit*, (modified) *MaxCombo* test, Threshold test (*Thres*), *V0* and (modified) *Zm3*.

These methods are followed closely by the modestly weighted logrank test (*MWLRT*), projection test (*ProjTest*), joint test (*jointTest*), modified Score test (*mScore*), time-dependent Cox model (*CoxTD*), partially grouped test (*ParGroup*), time-dependent Royston-Parmar model (*RP.TD*), quadratic combination of logrank and Nelson-Aalen (*QLRNA*), Milestone survival (*Mile, MileCLL, MileNA*) and the other combinations by Lee (*Lee1, Lee2*) all exceeding the power of the logrank test in more than 90% of scenarios.

3.3.2.1 PH scenarios

Based on the data-generating process the PH scenarios are those scenarios where no delay and hence no changepoint is present ($t_2^* = t_1^* = 0$). Considering only the scenarios with decreasing hazard results in 96 scenarios defined by the combination of median survival in the control arm med_C, the accrual proportion acc, the overall study duration τ and the maximum treatment effect θ. For the methods for which parameters had to be chosen the results displayed here are based on the moderate parameter value, i.e. 20% of the overall study duration for the methods using the Landmark cutoff and an interval of $[0, 0.2\tau]$ for the methods using the GenLin parameter.

As a sample size calculation tailored to this test has been performed, the logrank test should achieve approximately 80% power in each scenario. Figure 22 shows all methods on the x-axis and the power of the method in all 96 scenarios summarized in a boxplot with the Logrank test highlighted in red, which - as expected - achieves the targeted power. It can be seen that for most of the methods the power is similar throughout the different scenarios, but the asymptotic logrank test (*asymLR*), the Kolmogorov-Smirnov type test based on Cheng (*KS Cheng*), the Milestone survival (*Mile, MileCLL, MileNA*) and the piecewise exponential model (*PWExp*) have more spread out boxplots with an interquartile range of more than 5 percentage points. The former two having really low power and even less than 50% power in some of the 96 scenarios. The Landmark and Threshold test stay below 50% power throughout all scenarios, which shows that for proportional hazards the loss in power can be severe when early events are ignored.

The best of the Kolmogorov-Smirnov type tests is based on the logrank statistic (*KS LR*) which has a stable power of approximately 74% to 79%.

Royston-Parmar PH model (*RP.PH*) performs really good with a power of approx. 78% to 82% which is always higher than the Royston-Parmar TD model (*RP.TD*, 69% to 73%). Power of rate comparisons (*Mile, MileCLL, MileNA*) ranges from approx. 55% to 78%. *RMST* obtained a power of approx. 80% throughout making it competitive with the logrank test.

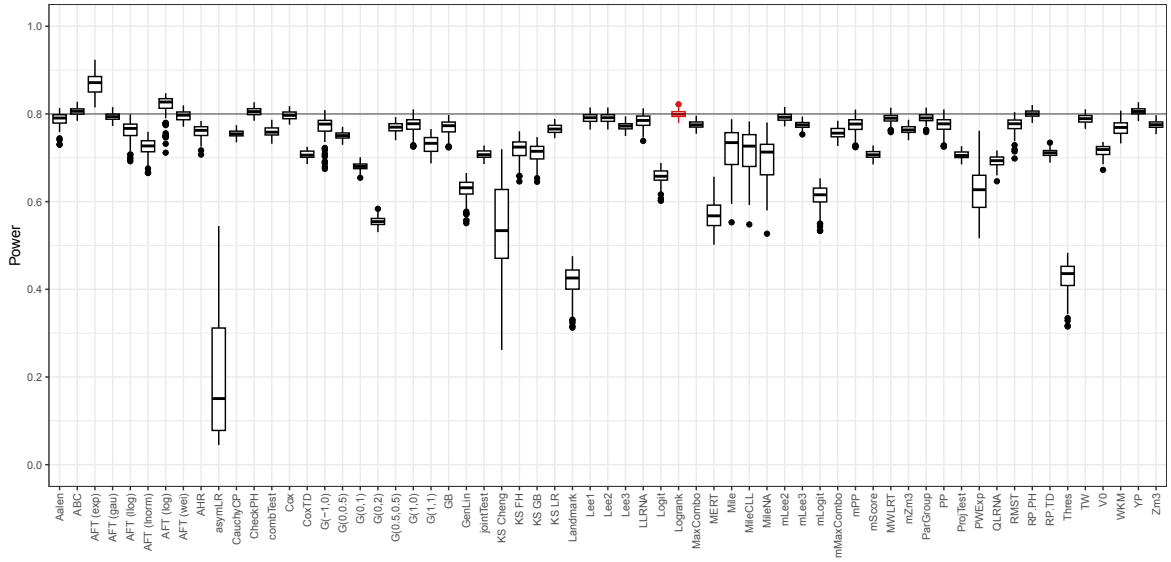

Figure 22: *Boxplot of power in PH scenarios with decreasing hazard for all 96 scenarios. Logrank is highlighted in red.*

For the aforementioned methods with very variable power, i.e. an interquartile range of more than 5 percentage points, a nested loop plot was created to identify the cause of this behavior (Figure 23). A nested loop plot is a tool to visualize the performance measure of interest for all combinations of parameters of the data-generating process. In this plot the power of the method is displayed on the y-axis with a reference line for the target power of 80% and the x-axis is defined by the overall study duration clustered by the remaining simulation parameters med_C and acc, whose values are displayed in the bottom part of the plot as steps. For each treatment effect θ a plot was created and these four plots are combined in a grid. Due to the performed sample size calculation n_{obs} increases the larger the median survival in

the control group and for fixed median survival it gets smaller the longer the study duration and the smaller the accrual period.

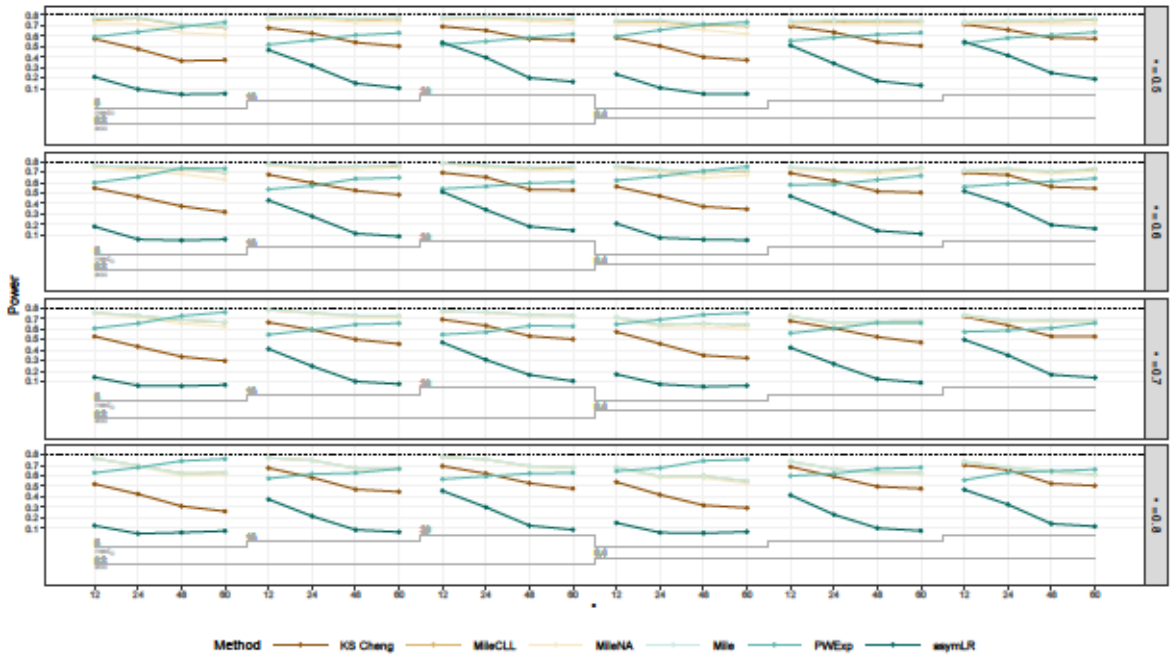

Figure 23: *Nested loop plot of power in PH scenarios with decreasing hazard for methods with variable power with overall study duration τ on the x-axis and clustered by median survival in the control arm* med_C *and duration of accrual* acc. *The panels are arranged by the treatment effect θ.*

The worst performance revealed the asymptotic logrank test (*asymLR*) of which the power increased with increasing sample size, but stayed below the power of all other tests throughout. Its variability is affected by all simulation parameters except the accrual proportion acc. As can be seen the range of the power values increases with lower study duration τ, greater treatment effect or equivalently lower θ and greater med_C. The Kolmogorov-Smirnov type test based on the weighted logrank test by Cheng (*KS Cheng*) has been seen to perform worse than the other Kolmogorov-Smirnov type tests, but gains power with increasing sample size up to a power of approximately 72%. The piecewise exponential model (*PWExp*) revealed a contradictory behavior in that the power increases the longer the study duration and the shorter the median survival in the control group, which corresponds to a decrease in sample size. This might be due to the fact that the piecewise exponential model assumes that the hazard is piecewise constant and changes at the prespecified parameter, which in case of PH

scenarios was taken to be 20% of the overall study duration. For the estimation of the hazard before this time enough events must have occurred. This number of events increases the longer the study duration and with that the later this timepoint or the shorter the median survival and with that the more likely the occurence of early events. These patterns are the same within each row and hence the effect of θ is negligible. The variability of *KS Cheng* and *PWExp* is mainly driven by the study duration and increases with bigger τ. Milestone rate comparisons (*Mile, MileCLL, MileNA*) perform better than the other methods except for the piecewise exponential model. The variability of these methods is mainly caused by the treatment effect θ, which can be explained by the fact that these methods compare the estimated survival rates at the latest possible timepoint which is biggest if the treatment effect is big.

3.3.2.2 NPH scenarios

In total there are 1152 NPH scenarios with decreasing hazard uniquely defined by the combination of median survival in the control arm med_C, the accrual proportion acc, the overall study duration τ, the maximum treatment effect θ and the delay t_2^* and changepoint t_1^*. For the methods for which parameters had to be chosen the results displayed here are based on the correct specified parameter, i.e. t_1^* for the methods using the Landmark cutoff and an interval of $[t_1^*, t_2^*]$ in case of linear lag scenarios for the methods using the GenLin parameter. In case of threshold lag scenarios ($t_1^* = t_2^*$) there is no correct interval for the latter methods and the interval $[0.8 \cdot t_2^*, 1.2 \cdot t_2^*]$ was chosen.

Overall power of all methods: Figure 24 shows all methods on the x-axis and the power of the method in all NPH scenarios summarized in a boxplot with the logrank test highlighted in red. For comparison grey boxplots are added that represent the power in PH scenarios as shown in the preceding section. It can be seen that the boxplots are much wider than in the PH scenarios since the accumulated scenarios are much more heterogeneous and differ in the extent they deviate from the PH assumption. Interestingly, this is not the case for the asymptotic logrank test (*asymLR*) the Kolmogorov-Smirnov type test based on the weighted logrank statistic by Cheng (*KS Cheng*) and the piecewise exponential models (*PWExp, PWExpLag*). These four methods were already seen in the previous section to have very variable power values which is now reduced in case of NPH. As expected *Landmark* and

Threshold test (*Thres*) achieve higher power than in the PH scenarios as ignoring early events in the phase where the groups are equal enhances power. For all other methods the violation of the PH assumption results in a sometimes drastic loss in power, with the power in the NPH scenarios being always below the power in the PH scenarios except for the Fleming-Harrington test for late difference (*G(0,2)*), the generalized linear model (*GenLin*), and the Logit tests (*Logit, mLogit*) for which both boxplots overlap.

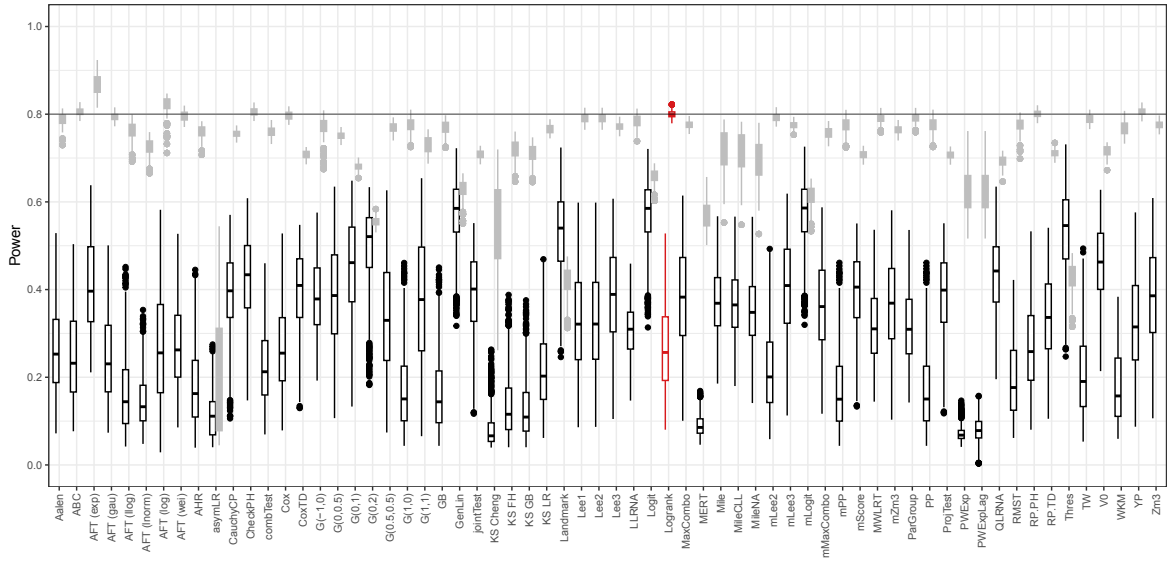

Figure 24: *Boxplot of power in NPH scenarios with decreasing hazard in black and for comparison the power in PH scenarios in grey. Highlighted in red is the logrank test.*

To disentangle the effect of the different delay and changepoint combinations the following Figure 25 is similar to the previous one but with the NPH scenarios arranged by delay and changepoint proportion. The grey boxplots of PH scenarios are the same in each panel. Similar to Figure 24 the *Landmark* and *Thres* method perform better than in PH scenarios and always outperform the logrank test. The power of each method decreases with increasing delay t_2^* and changepoint t_1^* but as before the power of the *GenLin* and (modified) *Logit* stays relatively stable.

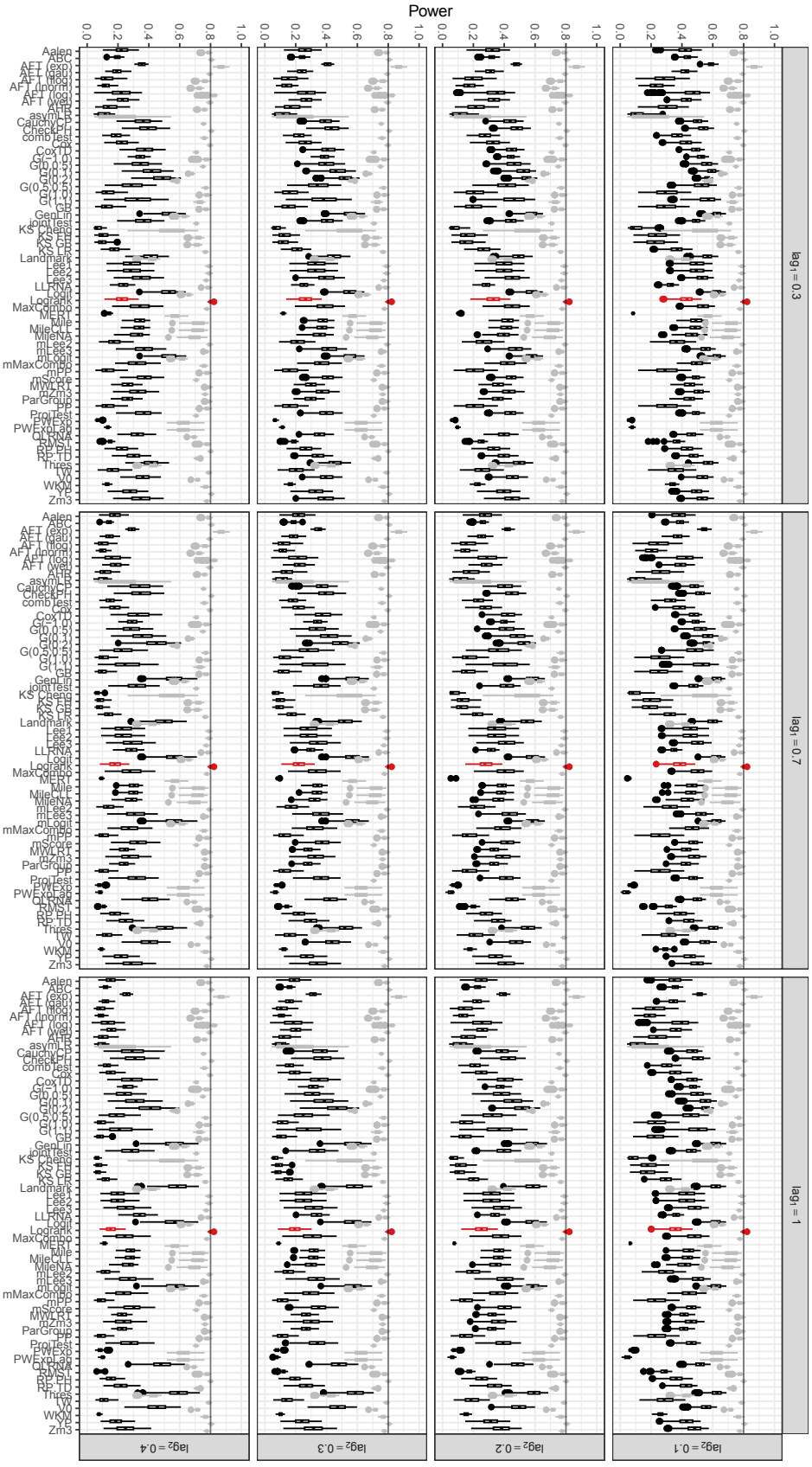

Figure 25: *Boxplot of power in NPH scenarios with decreasing hazard arranged by delay proportion* lag_2 *and changepoint proportion* lag_1*. For comparison the power in PH scenarios is given in grey and the same in each panel and the logrank test is highlighted in red.*

Power of single methods: The impact of the different simulation parameter on the power of the single methods is displayed in a nested loop plot. In this plot the power of the method is displayed on the y-axis and the x-axis is defined by the overall study duration τ clustered by the remaining simulation parameters med_C, acc, lag_2 and lag_1, whose values are displayed in the bottom part of the plot as steps. A reference line is given at the target power of 80%. For each maximum treatment effect θ a plot was created and these four plots are combined in a grid.

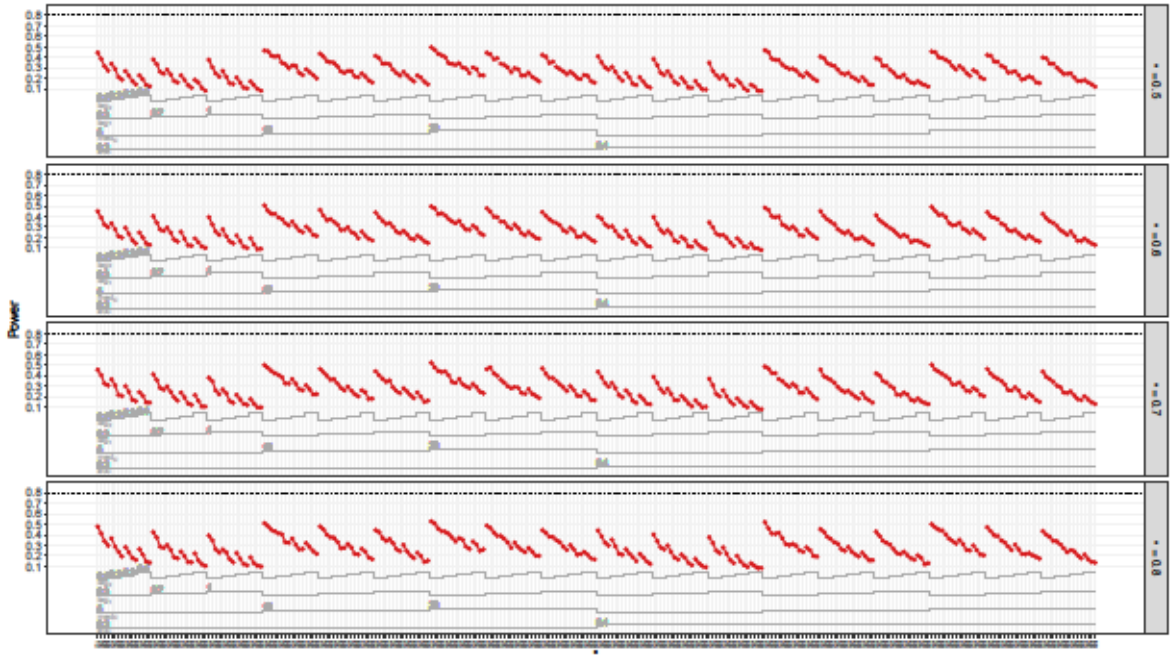

Figure 26: *Power of the logrank test in NPH scenarios for decreasing hazards with the overall study duration τ on the x-axis and further clustered by median survival time in the control arm* med_C, *the accrual proportion* acc, *the delay proportion* lag_2 *and changepoint proportion* lag_1. *The panels are arranged by the maximum treatment effect θ.*

Figure 26 displays the behavior of the well established logrank test. Although the power of the logrank test has been seen to be relatively stable at 80% in PH scenarios this drops to approx. 53% to 8% in NPH scenarios. With increasing study duration and therefore decreasing sample size the power is reduced drastically. This reduction is further driven by the delay proportion lag_2, whereas the changepoint proportion lag_1, the median survival in the control group med_C and the accrual proportion acc play a minor role.

As has been shown at the beginning of section 3.3.2 there are 17 methods which control type 1 error and outperform the logrank test in all NPH scenarios. Of these, four methods dominate the other methods in power, i.e. achieve the highest power among these methods. These methods are all parametric methods: the generalized linear model (*GenLin*), the logit and modified logit model (*Logit, mLogit*) and the threshold model (*Thres*). As expected the latter has the highest power in threshold lag scenarios ($\text{lag}_1 = 1$) and falls short against the other methods in other scenarios which all perform very similar. The difference between the power of the logrank test and these methods grows bigger the greater the changepoint t_1^*. The four methods are displayed against the logrank test in the following nested loop plot for NPH scenarios (Figure 27).

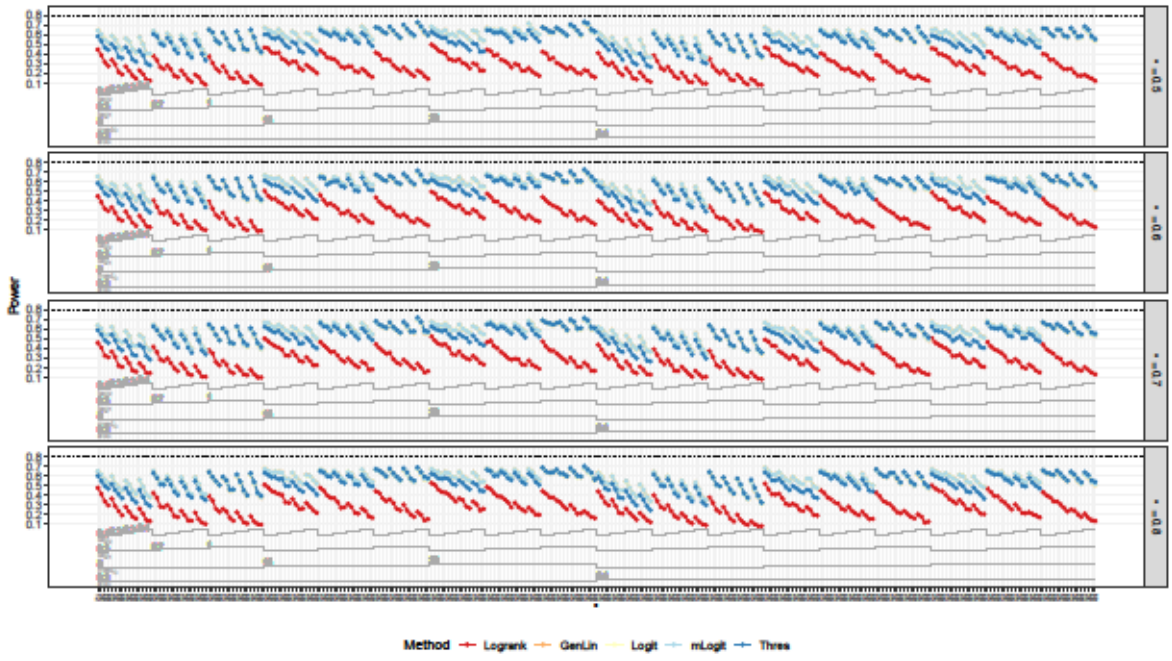

Figure 27: *Power in NPH scenarios of the parametric methods that outperform the logrank test in all NPH scenarios for decreasing hazards together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

Although these methods perform very well in NPH scenarios with a power between 24.7% and 73.1%, their performance in PH scenarios is poor since all of these methods ignore or downweight data based on the prespecified parameters of analysis. To see how much the

power is reduced if one uses these methods although the PH assumption is not violated, their performance in PH scenarios is shown in Figure 28.

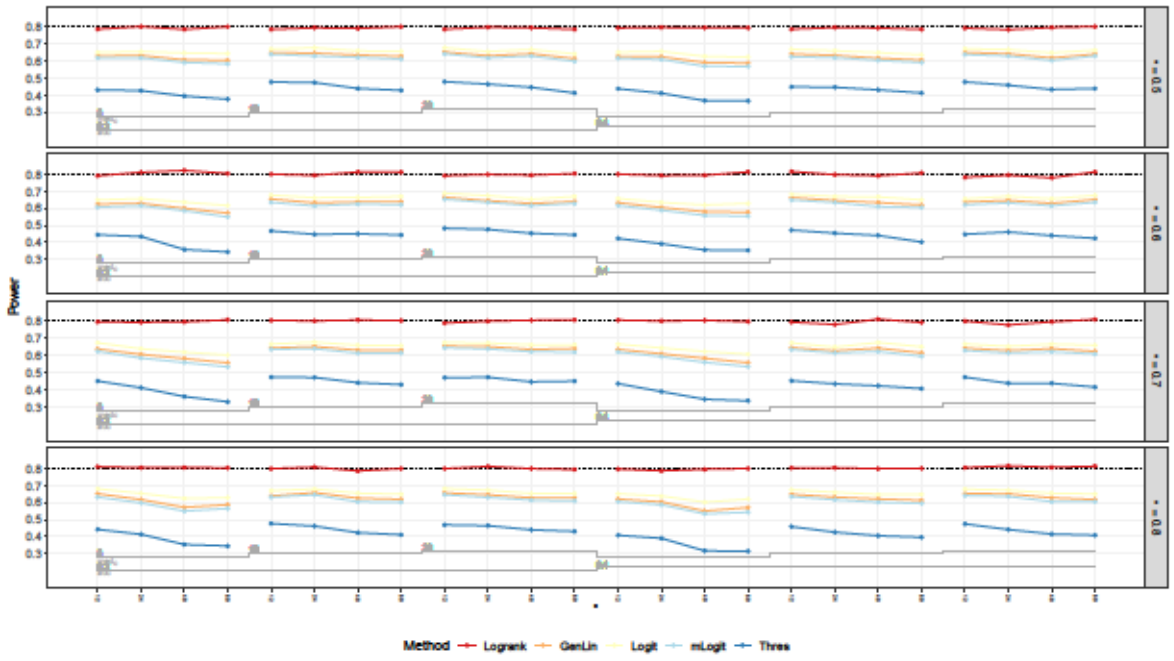

Figure 28: *Power in PH scenarios of the parametric methods that outperform the logrank test in all NPH scenarios for decreasing hazards together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

The power of the Threshold test (*Thres*) is the lowest throughout all PH scenarios and ranges from 31.5% to 48.3%. *GenLin* and both *Logit* models have very similar power between 53.3% and 68.8% but with the power of the *Logit* model being higher than the power of the *GenLin* model which in turn is higher than the power of the *mLogit* model.

Another drawback of these methods is that their performance depends on the parameters chosen for the analysis and they can be susceptible to misspecification as shown in Section 3.3.1. The following plot (Figure 29) shows nonparametric alternatives which also outperformed the logrank test in all NPH scenarios but do not achieve such a high power than the correctly specified parametric methods.

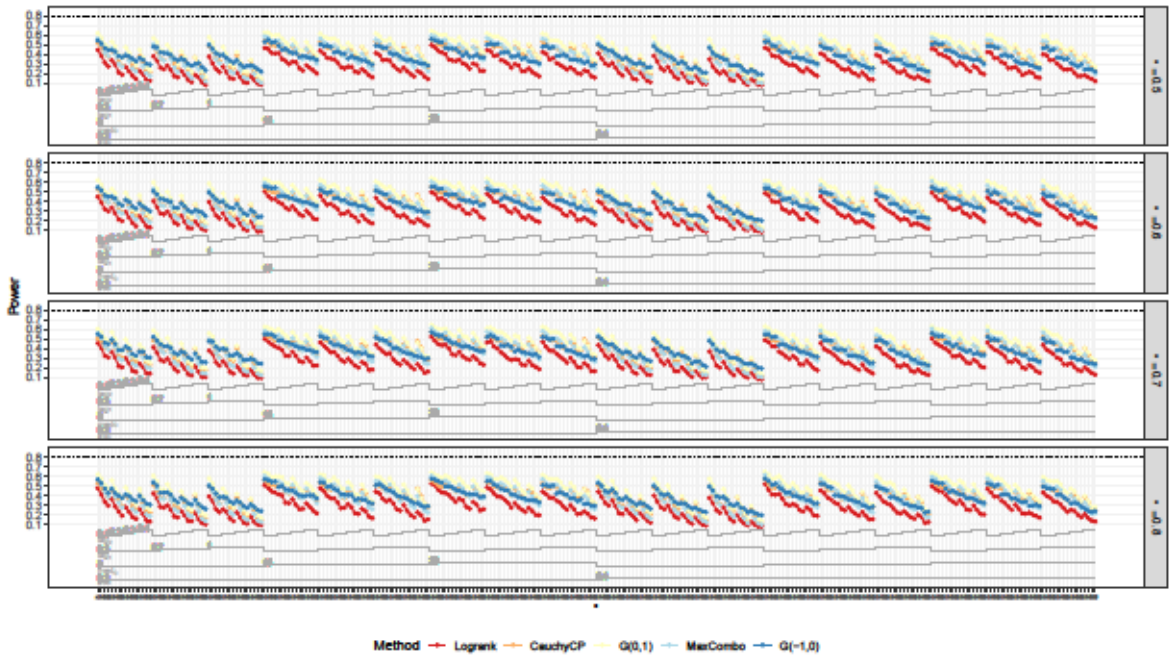

Figure 29: *Power in NPH scenarios of non-parametric methods that outperform the logrank test in all NPH scenarios for decreasing hazards together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

The power of these nonparametric methods ranges from 10.1% to 64.8% where the Fleming-Harrington test $G(0,1)$ shows best performance in a majority of scenarios. For low median survival of $\text{med}_C = 5$ months the Gray-Tsiatis test (*G(-1,0)*) scores second but is outperformed by the Cauchy changepoint model (*CauchyCP*) and the *MaxCombo* test for greater median survival.

For these methods the performance in PH scenarios is also displayed in Figure 30 and it can be seen that the power is similar to that of the logrank test and the methods show a smaller reduction in power than the parametric methods. In PH scenarios the Fleming-Harrington test *G(0,1)* expectedly performs worst with a power of 65.4%-70.1%. The other tests perform very similar and are close to the logrank test except for the Gray-Tsiatis test (*G(-1,0)*) in scenarios with low median survival of $\text{med}_C = 5$ months and a high overall study duration of $\tau = 48, 60$ months.

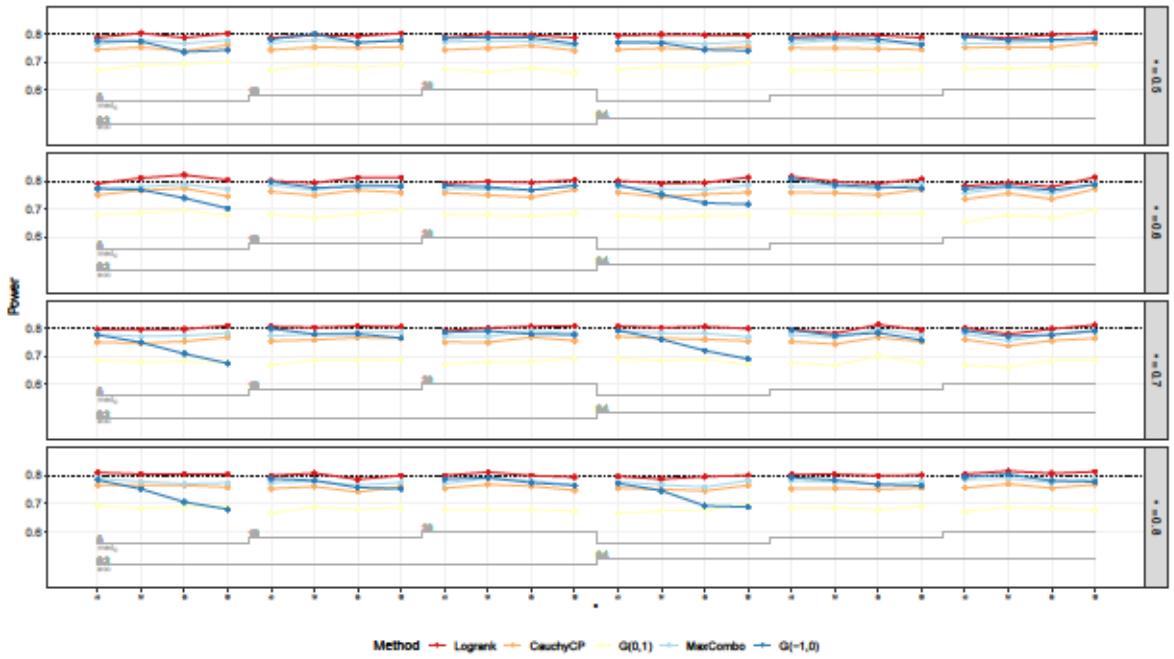

Figure 30: *Power in PH scenarios of non-parametric methods that outperform the logrank test in all NPH scenarios for decreasing hazards together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

Order of the methods: From a general overview over the different methods and a comparison of these methods to the logrank test the focus will now be on investigating the effect of the different delay patterns.

The following scatterplots show the power of the different methods arranged by the delay and changepoint used in the generalized linear lag model for the data generating process. Separate plots for the study duration τ, the median survival in the control arm med_C, the maximum treatment effect θ and the accrual proportion acc were created which resulted in 96 plots. In contrast to Figure 25 the focus does not lie on an overall comparison of the methods across all scenarios but on the impact of the delay and changepoint parameter on the relationship between the power of the methods. Therefore, the methods displayed on the x-axis are ordered by the power of the methods in the upper left panel of the plot to see if the ordering remains the same throughout the different delay and changepoint combinations. Although the order of the methods is slightly different in each plot, the quality of it is the

same and hence only the plot for an overall study duration of $\tau = 48$ months, an accrual of $0.2 \cdot \tau = 9.6$ months, and a median survival in the control group of $\text{med}_C = 15$ months is shown both for the maximum effect $\theta = 0.5$ (Figure 31) and the minimum effect $\theta = 0.8$ (Figure 32).

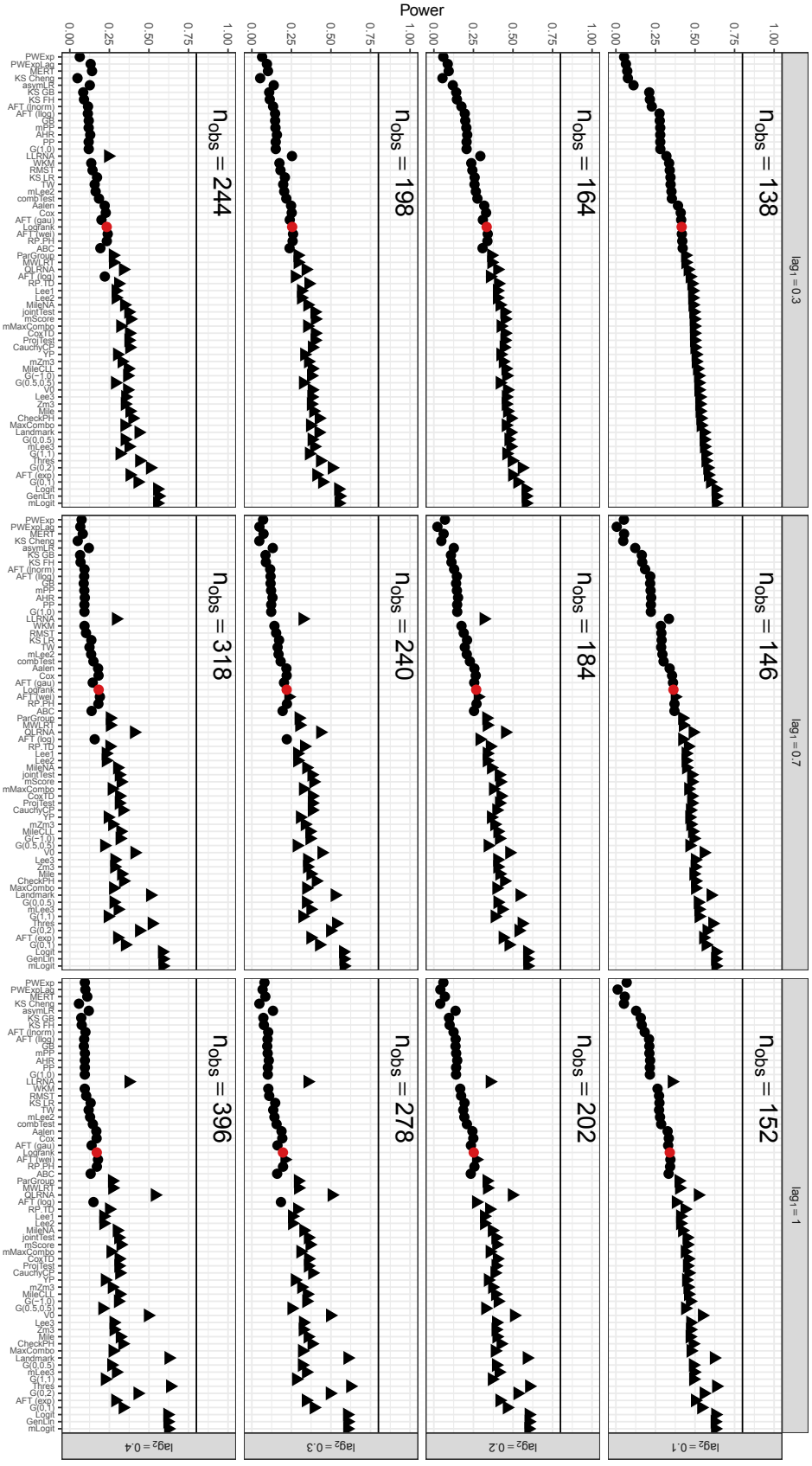

Figure 31: *Power of the different methods for a study duration of 48 months, an accrual of 9.6 months, median survival of 15 months and maximum effect of HR 0.5 when hazards are decreasing. Triangular shape indicates that the power exceeds the power of the logrank test by more than the Monte-Carlo standard error based on the evaluable datasets. Panels are arranged by delay proportion (lag_2) and changepoint proportion (lag_1). The logrank test is highlighted in red.*

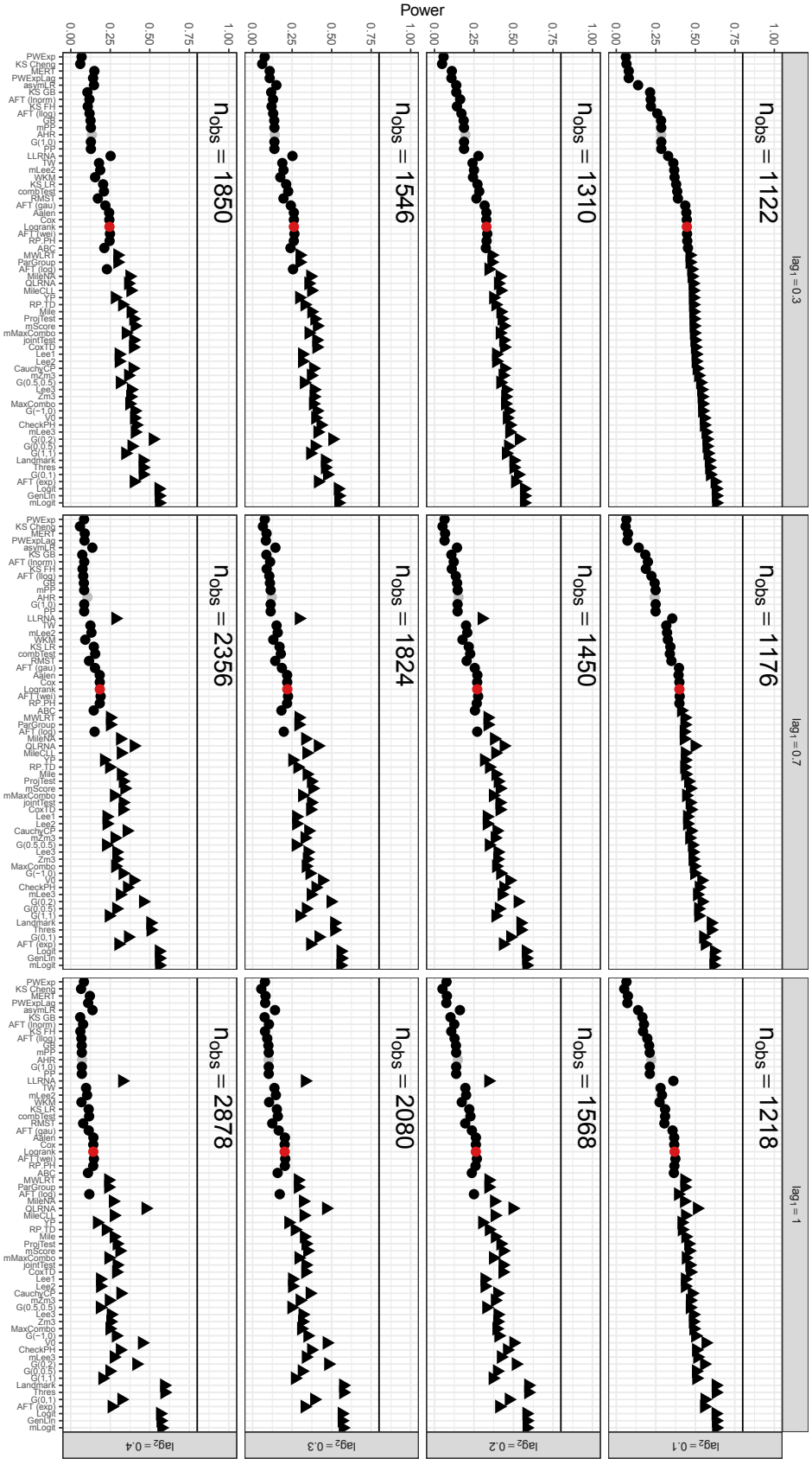

Figure 32: *Power of the different methods for a study duration of 48 months, an accrual of 9.6 months, median survival of 15 months and minimum effect of HR 0.8 when hazards are decreasing. Triangular shape indicates that the power exceeds the power of the logrank test by more than the Monte-Carlo standard error based on the evaluable datasets. Panels are arranged by delay proportion (lag_2) and changepoint proportion (lag_1). The logrank test is highlighted in red.*

As can be seen in these figures the overall order of the methods based on their power is not preserved between the different delay patterns. Of the methods that had lower power than the logrank test in the upper left panel, the linear combination of the logrank test with the Milestone survival based on the Nelson-Aalen estimator (*LLRNA*) sticks out, as its power is relatively stable or even increasing with increasing deviations from the PH assumption, so that it exceeds the power of the logrank test in more extreme scenarios. The remaining methods with lower power decrease uniformly so that order of the methods does not change much, which is completely different for the methods with higher power, which spread out much more the extremer scenarios get. As seen before the generalized linear lag model (*GenLin*) and the *Logit* models are always on top throughout all scenarios. Furthermore, the quadratic combination of the logrank test and Milestone survival (*QLRNA*), the *V0* test and the *Landmark* and *Thres* methods do not only keep their power above the power of the logrank test but do also decrease less than the other methods the bigger the deviation from the PH assumption gets.

Chance of rejection of each method: Furthermore, to quantify the impact of the delay and changepoint parameter a logistic regression model is calculated for each combination of $\text{med}_C, \text{acc}, \tau, \theta$ and for each method separately. To do this the independent variables are taken as proportions of the overall study duration τ and then transformed on a base 2 logarithm scale so that with binary rejection indicator $Y := \mathbb{1}(p - value <= 0.05)$ the model becomes

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 \log_2(\text{lag}_2 + 1) + \beta_2 \log_2(\text{lag}_2 \cdot \text{lag}_1 + 1),$$

where the independent variables were treated as continuous variables. The logarithmic transformation has the appeal that the resulting odds ratio can be interpreted as the effect per doubling of the independent variable, which is better interpretabel than an increase of 1 in case of proportions. Adding 1 before logarithmic transformation has the effect that setting the independent variables to 0 again corresponds to the PH scenario so that the intercept β_0 represents the log odds of rejection for PH scenarios.

Figure 33 shows the results of these logistic regression models with the estimated odds ratio on the y-axis and the different methods on the x-axis. The boxplots summarize the results

for the different combinations of $\text{med}_C, \text{acc}, \tau$ and θ and are arranged by the independent variable. The logrank test is again highlighted in red.

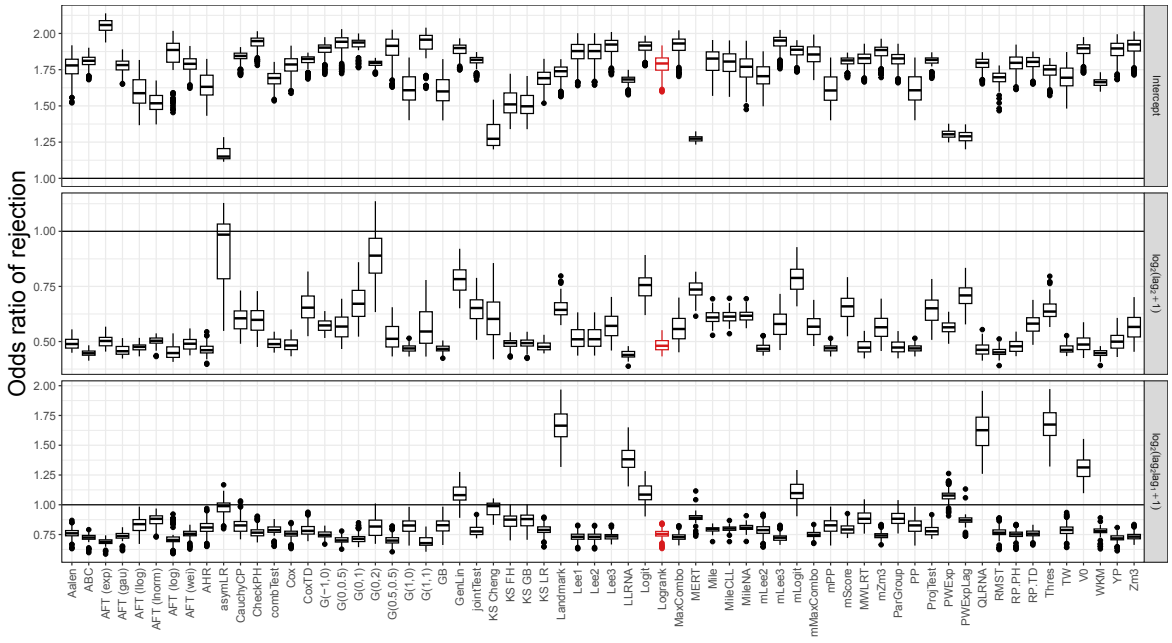

Figure 33: *Results from logistic regression models for rejection of the null hypothesis under the alternative for the single methods in decreasing hazard scenarios with independent variables delay proportion (*lag_2*) and changepoint proportion (*lag_1*). The logrank test is highlighted in red.*

The first row of Figure 33 shows the chance of rejecting the null hypothesis if no delay and changepoint are present, which corresponds to the PH scenarios. As was already observed in the section on power in the PH scenarios the asymptotic logrank test (*asymLR*), the Kolmogorov-Smirnov type test based on the weighted logrank statistic by Cheng (*KS Cheng*) and the piecewise exponential models (*PWExp, PWExpLag*) have a very low chance of rejecting the null hypothesis. This is also true for the *MERT* test, with a chance of approximately 5:4. The impact of the delay proportion is shown in the second row of Figure 33. For most methods it is similar to the impact of the delay proportion on the logrank test where a twice as high delay halves the chance of rejection. The asymptotic logrank test (*asymLR*) and the Fleming Harrington test for late difference (*G(0,2)*) seem to be less affected by the delay proportion as there are scenarios where the chance of rejection does not change. In addition the *GenLin*, the *Logit* and modified Logit (*mLogit*) method also seem to be less affected than

most methods by the delay proportion. The effect of the changepoint proportion is shown in the third row of Figure 33 and has a lesser effect on most methods than the delay proportion as the odds ratio is closer to one but for most methods below one hence reducing the chance of rejection. For *GenLin, Landmark, LLRNA, Logit, mLogit, QLRNA*, Threshold (*Thres*) and *V0*, however, the effect of the changepoint proportion is reversed leading to an increase of the chance of rejection which is highest for *Landmark*, Threshold (*Thres*) and *QLRNA*.

3.3.3 Constant hazards

Now the power of the methods is compared if the hazard in the control arm is constant. Again it was calculated in how many of the 1248 simulated scenarios each method exceeded the power of the logrank test by more than the Monte-Carlo standard error given the number of evaluable datasets. This is displayed in a bar chart in Figure 34. A green bar indicates that the method exceeds the logrank test in all NPH scenarios and a proportion of zero indicates that the method has lower power than the logrank test in all NPH scenarios.

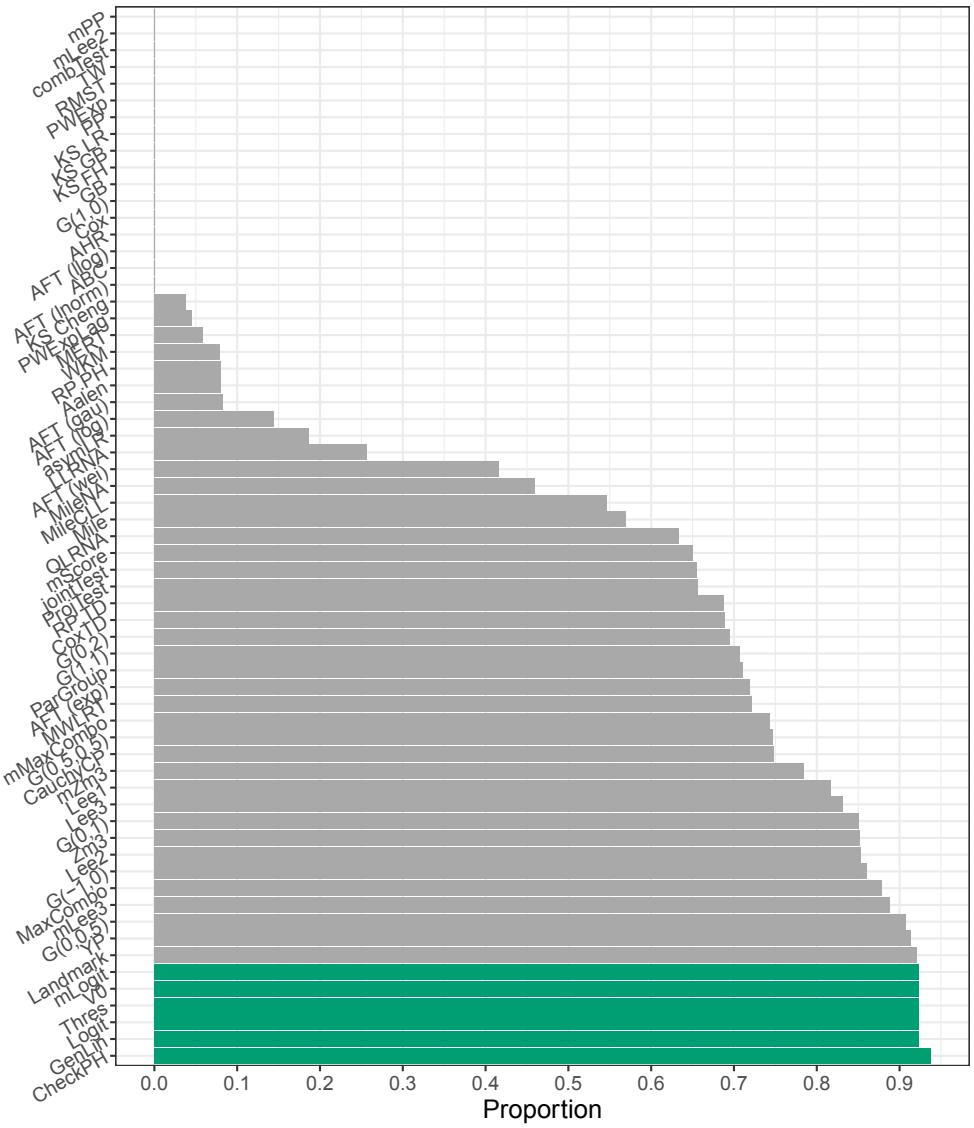

Figure 34: *Proportion of scenarios in which each method exceeds the power of the logrank test in constant hazard scenarios. Green bars indicate that method outperforms the logrank test in all NPH scenarios.*

Of the five methods which were identified in Section 3.2.2 to have increased type 1 error (*CheckPH, YP, RP.TD, AFT (log), G(-1,0)*), only *CheckPH* outperformed the logrank test in all NPH scenarios and in 18 of 96 PH scenarios. Even the Gray-Tsiatis test (*G(-1,0)*) which had an type 1 error inflation of up to 11.4% outperformed the logrank test only in 1074 (86.06%) of the 1248 scenarios. The other five methods that outperform the logrank test in all NPH scenarios are all parametric methods which control type 1 error: *GenLin, Logit, mLogit, V0* and *Thres*. As for nonparametric methods which control type 1 error, the following methods outperform the logrank test in more than 80% of the scenarios: the Fleming-Harrington tests for late differences (*G(0,0.5), G(0,1)*), the combinations tests by Lee (*Lee1, Lee2, Lee3, mLee3*), the *MaxCombo* test and the *Zm3* test.

3.3.3.1 PH scenarios

Based on the data-generating process the PH scenarios are those scenarios where no delay and hence no changepoint is present ($t_2^* = t_1^* = 0$). Considering only the scenarios with constant hazard results in 96 scenarios defined by the combination of median survival in the control arm med_C, the accrual proportion acc, the overall study duration τ and the maximum treatment effect θ. For the methods for which parameters had to be chosen the results displayed here are based on the moderate parameter, i.e. 20% of the overall study duration for the methods using the Landmark cutoff and an interval of $[0, 0.2\tau]$ for the methods using the GenLin parameter.

As a sample size calculation tailored to this test has been performed, the logrank test should achieve approximately 80% power in each scenario. Figure 35 shows all methods on the x-axis and the power of the method in all 96 scenarios summarized in a boxplot with the Logrank test highlighted in red, which - as expected - achieves the targeted power. Compared to the decreasing hazard scenarios the boxplots are for many methods much bigger and 20 of the 63 methods have an interquartile range of more than 5 percentage points. However, the asymptotic logrank test (*asymLR*), the Kolmogorov-Smirnov type test based on Cheng (*KS Cheng*), the Milestone survival (*Mile, MileNA*) and the piecewise exponential model (*PWExp*) have again very variable power with an interquartile range of more than 10 percentage points. For constant hazard this is also true for the *MERT* test.

For all these methods except the Milestone survival (*Mile, MileNA*) the median power in the PH scenarios was below 50%. Compared to the decreasing hazard scenarios the *Landmark* and Threshold (*Thres*) test improved achieving a median power of over 60%.

The best of the Kolmogorov-Smirnov type tests is again based on the logrank statistic (*KS LR*) which has a stable power of approximately 71% to 79%.

Royston-Parmar PH model (*RP.PH*) performs really good with a power of approx. 77% to 82% which is always higher than the Royston-Parmar TD model (*RP.TD*) with a power of 69% to 73%. Power of Milestone survival rate comparisons (*Mile, MileNA, MileCLL*) ranges from approx. 37% to 77%. *RMST* obtained a power of approximately 54% to 80% which is not as stable as in the decreasing hazard scenarios.

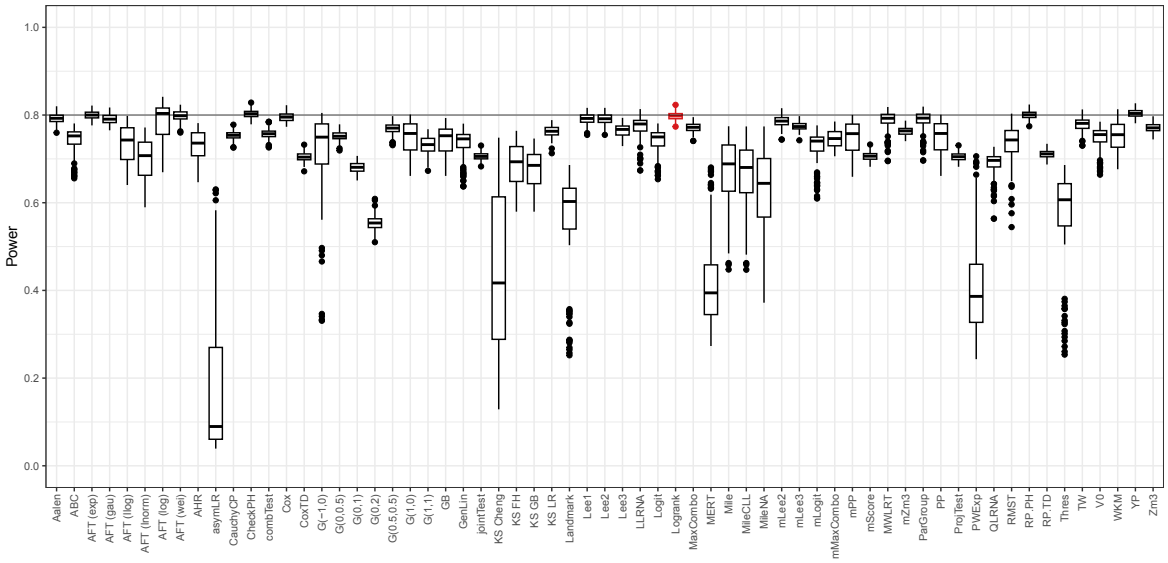

Figure 35: *Boxplot of power in PH scenarios with constant hazard for all 96 scenarios. Logrank is highlighted in red.*

For the aforementioned methods with very variable power, i.e. an interquartile range of more than 10 percentage points, a nested loop plot was created to identify the cause of this behavior (Figure 36). In this plot the power of the method is displayed on the y-axis with a reference line for the target power of 80% and the x-axis is defined by the overall study duration clustered by the remaining simulation parameters med_C and acc, whose values are displayed in the bottom part of the plot as steps. For each treatment effect θ a plot was created and

these four plots are combined in a grid. Due to the performed sample size calculation n_{obs} increases the larger the median survival in the control group and for fixed median survival it gets smaller the longer the study duration and the smaller the accrual period.

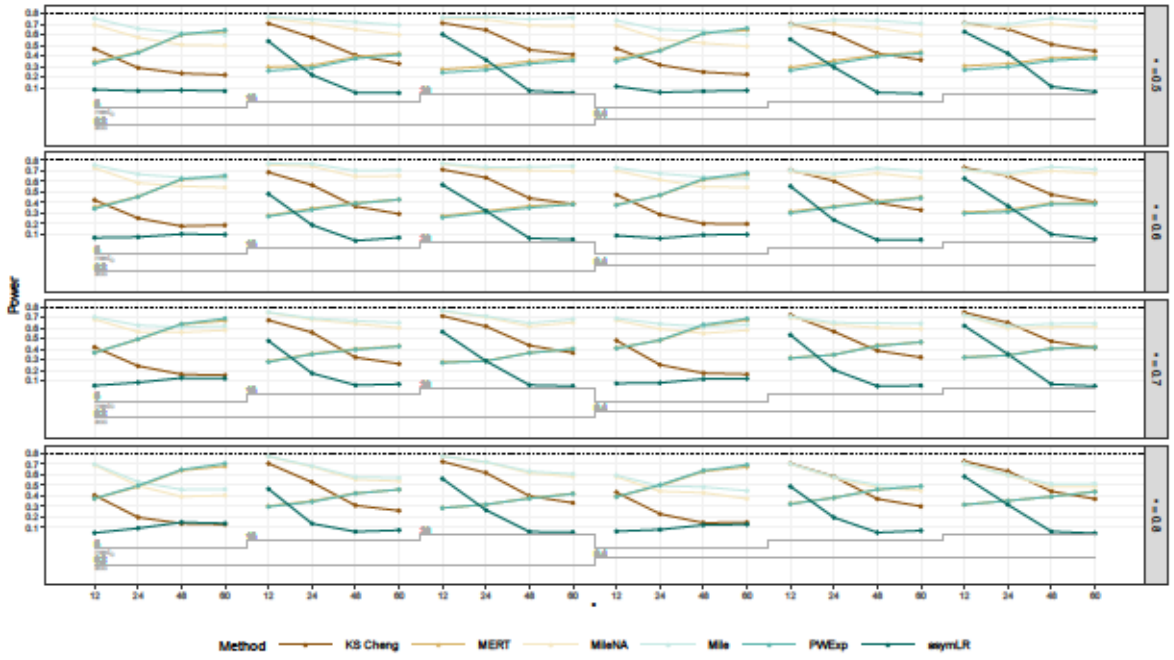

Figure 36: *Nested loop plot of power in PH scenarios with constant hazard for methods with variable power. The overall study duration τ is plotted on the x-axis and clustered by median survival in the control arm* med_C *and duration of accrual* acc. *The panels are arranged by the treatment effect θ.*

Milestone survival rate comparisons (*Mile, MileNA*) performed best in scenarios with longer median survival in the control arm. In the scenario with $\text{med}_C = 5$ months they are outperformed by the piecewise exponential model and the *MERT* test if the overall study duration is high. The variability of their power values increases with decreasing θ as explained for decreasing hazards. Again the piecewise exponential model (*PWExp*) revealed a contradictory behavior that the power increases the longer the study duration and the shorter the median survival in the control arm, which corresponds to a decrease in sample size. This could also be observed for the *MERT* test which performed very similar. The power values of these two methods were more spread out the higher the study duration τ or the smaller the median survival med_C.

The worst performance revealed the asymptotic logrank test of which the power increased with increasing sample size, but stayed below the power of all other tests except for the case of very low study duration ($\tau = 12$ months) and a median survival of more than 5 months in the control arm. The variability of the power values is driven by increasing med_C and decreasing τ and θ.

3.3.3.2 NPH scenarios

The 1152 NPH scenarios with constant hazard are uniquely defined by the combination of median survival in the control arm med_C, the accrual proportion acc, the overall study duration τ, the maximum treatment effect θ and the delay t_2^* and changepoint t_1^*. For the methods for which parameters had to be chosen the results displayed here are based on the correct specified parameter, i.e. t_1^* for the methods using the Landmark cutoff and an interval of $[t_1^*, t_2^*]$ in case of linear lag scenarios for the methods using the GenLin parameter. In case of threshold lag scenarios ($t_1^* = t_2^*$) there is no correct interval for the latter methods and the interval $[0.8 \cdot t_2^*, 1.2 \cdot t_2^*]$ was chosen.

Overall power of all methods: Figure 37 shows all methods on the x-axis and the power of the method in all NPH scenarios summarized in a boxplot with the Logrank test highlighted in red. For comparison grey boxplots are added that represent the power in PH scenarios as shown in the preceding section. It can be seen that the boxplots are much wider than in the PH scenarios with a mean interquartile range of 24.4 and a mean overall power range of 71.5 percentage points ranging from an average minimum power of 5.4% to an average maximum power of 76.9%. This is due to fact that the NPH scenarios are much more heterogeneous and differ in the extent they deviate from the PH assumption.

In contrast, the *MERT* test and the piecewise exponential models (*PWExp, PWExpLag*) have the lowest interquartile range of up to 5.5 percentage points and an overall range of up to 20.3 percentage points which is due to the overall poor power of at most 23%. Another method with a comparatively low range of 38.3 percentage points is the asymptotic logrank test (*asymLR*) ranging from 4.1% to 42.4% power. The range of all other methods is more than 60 percentage points. Again *Landmark* and Threshold test (*Thres*) as well as the Fleming-Harrington test for late difference (*G(0,2)*) achieve higher power than in the PH scenarios

as ignoring early events in the phase where the groups are equal enhances power. For all other methods the power in the NPH scenarios is reduced compared to the power in the PH scenarios but as the power range of each method is very wide the differences between NPH and PH scenarios are not that accentuated.

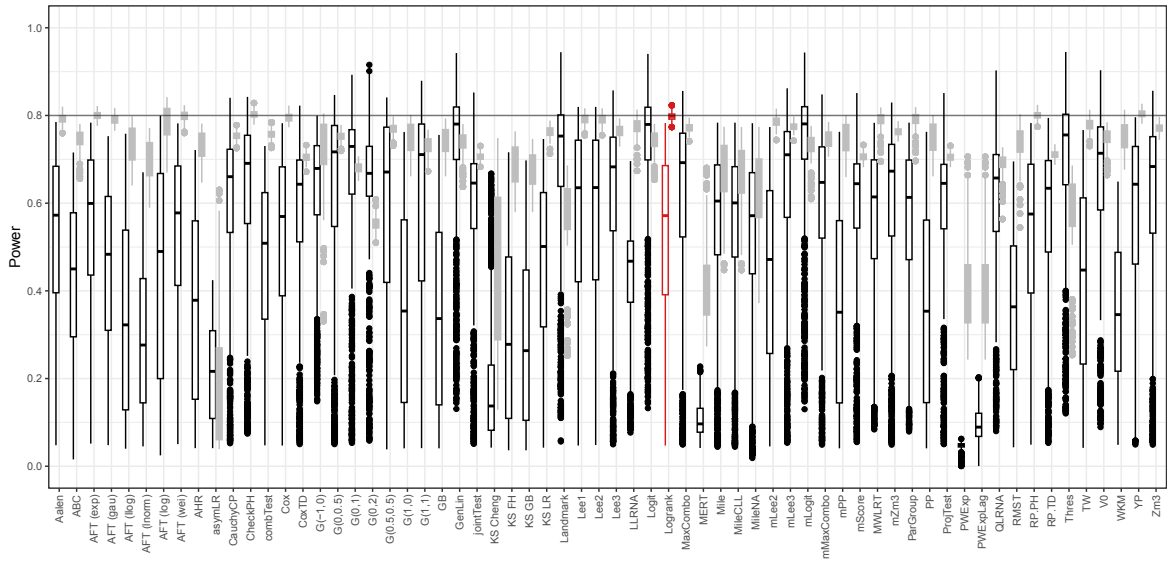

Figure 37: *Boxplot of power in NPH scenarios with constant hazard in black and for comparison the power in PH scenarios in grey. Highlighted in red is the logrank test.*

To disentangle the effect of the different delay and changepoint combinations the following Figure 38 is similar to the previous one but with the NPH scenarios arranged by delay t_2^* and changepoint t_1^*. The grey boxplots of PH scenarios are the same in each panel. As can be seen the variability of the power of the methods icnreases with increasing delay t_2^* and is very low for $\text{lag}_2 = 0.1$. Then *GenLin* model and both *Logit* methods are relatively stable with high power in each panel but have outliers with very low power which makes them appear less stable in the overall boxplot in Figure 37.

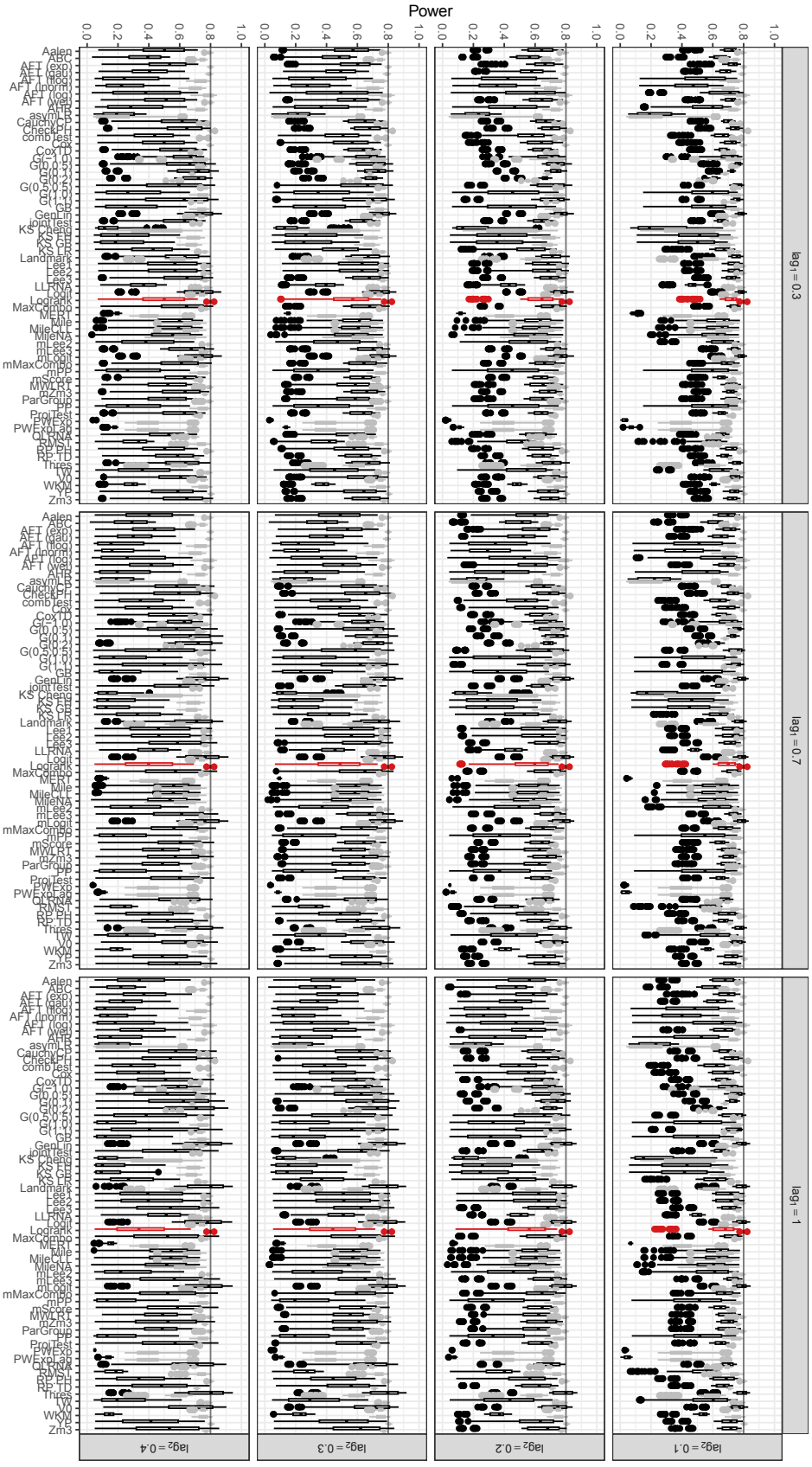

Figure 38: *Boxplot of power in NPH scenarios with constant hazard arranged by delay proportion* lag_2 *and changepoint proportion* lag_1*. For comparison the power in PH scenarios is given in grey and the same in each panel and the logrank test is highlighted in red.*

Power of single methods: The impact of the different simulation parameter on the power of the single methods is displayed in a nested loop plot. In this plot the power of the method is displayed on the y-axis and the x-axis is defined by the overall study duration τ clustered by the remaining simulation parameters med_C, acc, lag_2 and lag_1, whose values are displayed in the bottom part of the plot as steps. A reference line is given at the target power of 80%. For each maximum treatment effect θ a plot was created and these four plots are combined in a grid.

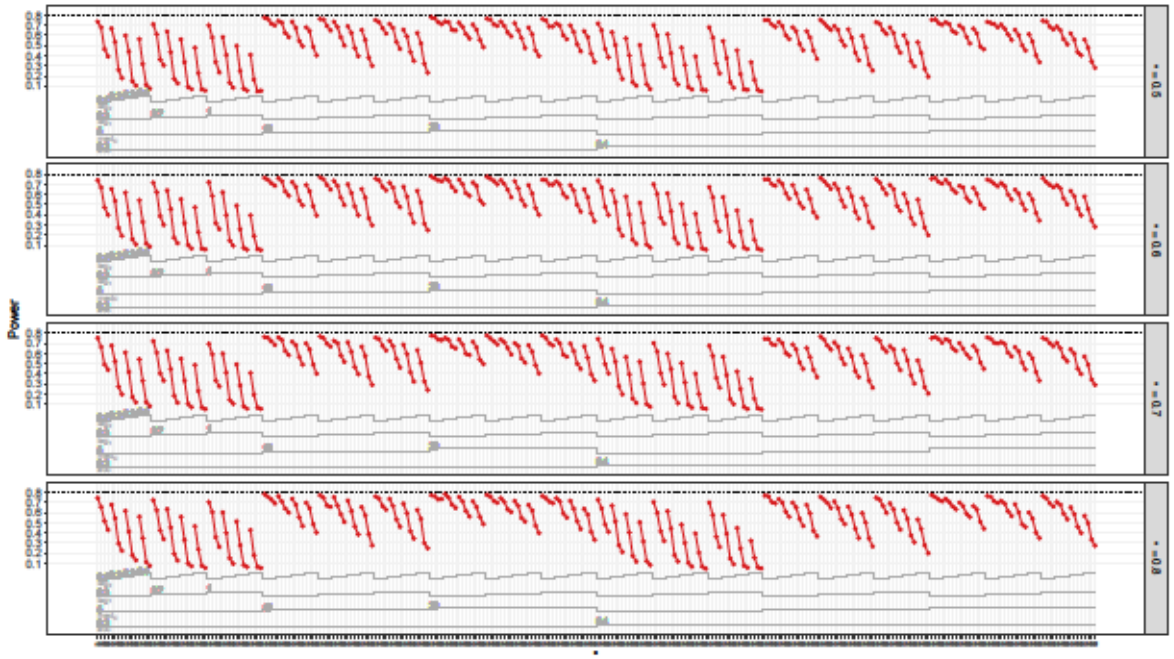

Figure 39: *Power of the logrank test in NPH scenarios for constant hazards with the overall study duration τ on the x-axis and further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

Figure 39 displays the behavior of the well established logrank test. Although the power of the logrank test has been seen to be relatively stable at 80% in PH scenarios this drops to approx. 78% to 5% in NPH scenarios. With increasing study duration and therefore decreasing sample size the power is reduced drastically in scenarios with $\text{med}_C = 5$ months but less so for scenarios with longer median survival. This reduction is further driven by the delay proportion lag_2, whereas the changepoint proportion lag_1, the median survival in the control group med_C and the accrual proportion acc play a minor role.

As has been shown at the beginning of section 3.3.3 there are five methods which control type 1 error and outperform the logrank test in all NPH scenarios. These methods are all parametric methods: the generalized linear model (*GenLin*), the logit and modified logit model (*Logit, mLogit*), the threshold model (*Thres*) and the *V0* test. The five methods are displayed against the logrank test in the following nested loop plot for NPH scenarios (Figure 40). As can be seen all methods perform very similar and their power as well as the power of the logrank test is increasing the higher the median survival in the control arm (med_C). For scenarios with an med_C of more than 5 months the power loss due to the higher overall study duration is smaller. Additionally, the power in threshold lag scenarios ($\text{lag}_1 = 1$) increases the bigger the delay lag_2. For the methods the *GenLin* and *(m)Logit* methods often perform best except for threshold lag scenarios where they are outperformed by the *Thres* test. The *V0* test has the lowest power in most of the methods but still outperforms the logrank test as mentioned before.

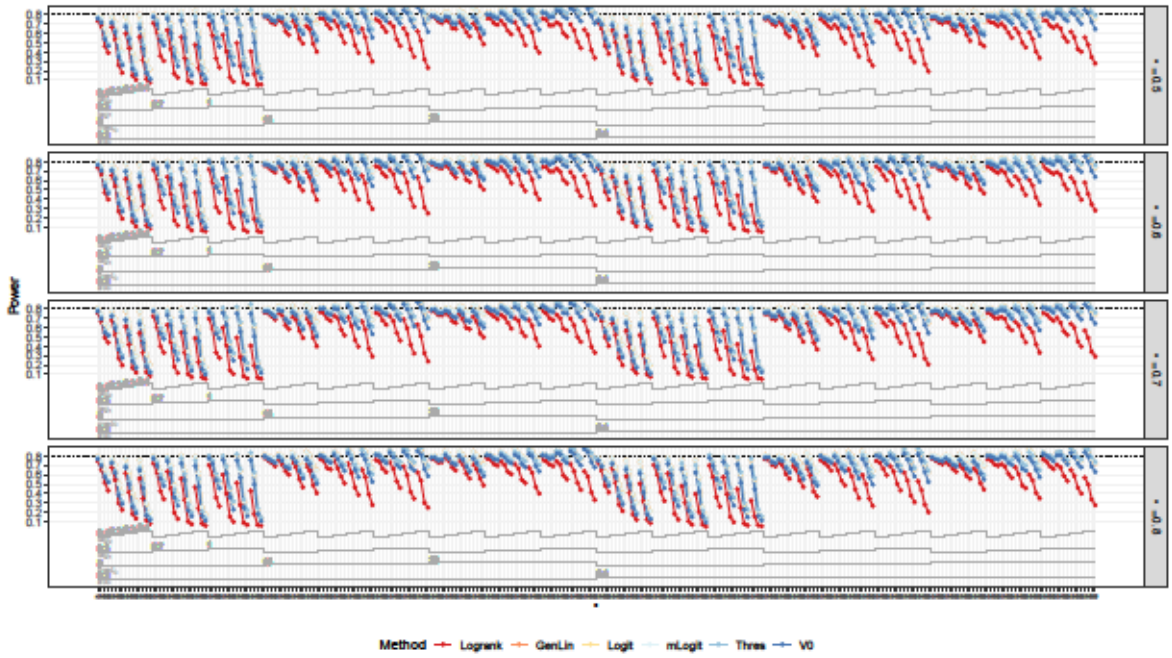

Figure 40: *Power in NPH scenarios of the parametric methods that outperform the logrank test in all NPH scenarios for constant hazards together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

Although these methods perform very well in NPH scenarios with a power between 9% and 94%, their performance in PH scenarios is poor since all of these methods ignore or downweight data based on the prespecified parameters of analysis. To see how much the power is reduced if one uses these methods although the PH assumption is not violated, their performance in PH scenarios is shown in Figure 41.

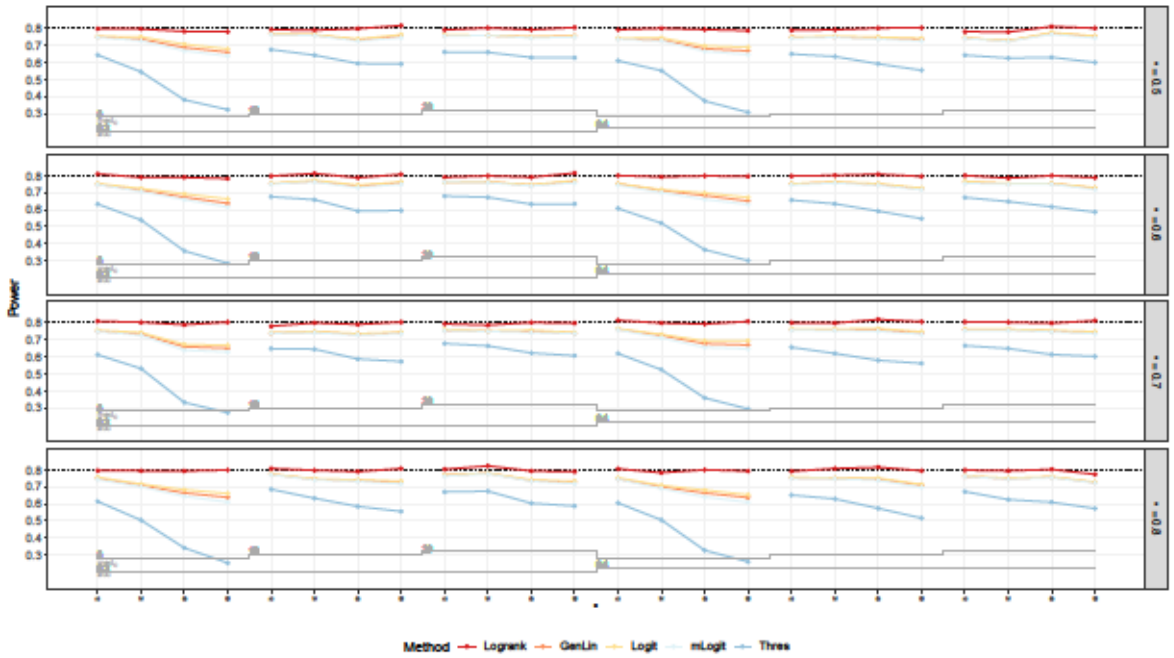

Figure 41: *Power in PH scenarios of the parametric methods that outperform the logrank test in all NPH scenarios for constant hazards together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

Of these parametric methods the *Thres* test performs worst and shows a to have heavily reduced power in PH scenarios ranging from approximately 25% to 69%. The best performing method is the *V0* test which achieves between 66% and 78% power in PH scenarios. The other parametric methods are comparable to the *V0* test and only slightly inferior, but have the disadvantage that two parameters for analysis have to be specified. It is hence remarkable that the *V0* test performs that well although the *GenLin* model is much closer to the data generating process.

Their dependence on the choice of parameters for analysis and with it their possible susceptibility to parameter misspecification can make their uasge difficult in practice and hence nonparametric methods might be more robust. As mentioned at the beginning of Section 3.3.3 none of the nonparametric methods outperformed the logrank test in all NPH scenarios with constant hazard. However, of the methods that outperform the logrank test in more than 80% of the scenarios the following were chosen for comparison: the Fleming-Harrington tests for late differences (*G(0,1)*), the combination test by Lee (*Lee2*), the *MaxCombo* test and the *Zm3* test. Their performance in NPH scenarios is given in the following plot (Figure 42).

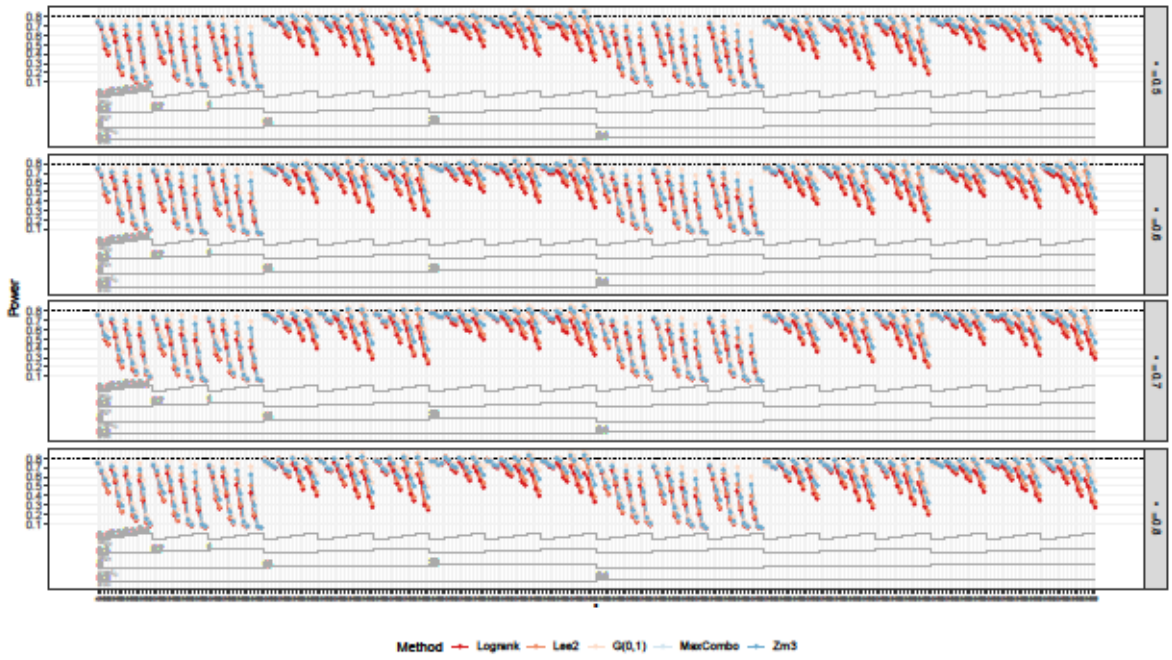

Figure 42: *Power in NPH scenarios of non-parametric methods that outperform the logrank test in 80% of NPH scenarios for constant hazard together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

In the majority of scenarios the Fleming-Harrington test *G(0,1)* achieves the highest power. The overall power of the methods ranges from 5% to 89% and of the combination tests the tests that consider more Fleming-Harrington weighted logrank statistics perform slightly better than those that use less.

For these methods the performance in PH scenarios is also displayed in Figure 43. As the *MaxCombo* test performs very similar as the *Zm3* test it is covered entirely by the line of the latter. This shows that inclusion of a further Fleming-Harrington test does not give any power benefit in PH scenarios. The Fleming-Harrington test for late differences (*G(0,1)*) expectedly performs worst as it downweights early events and achieves only a power of 65% to 71%. Quite surprisingly the *Lee2* test with a power between 75% and 82% performs better than the *MaxCombo* and *Zm3* test although the logrank statistic is, in contrast to the other two tests, not considered in the construction of the *Lee2* test.

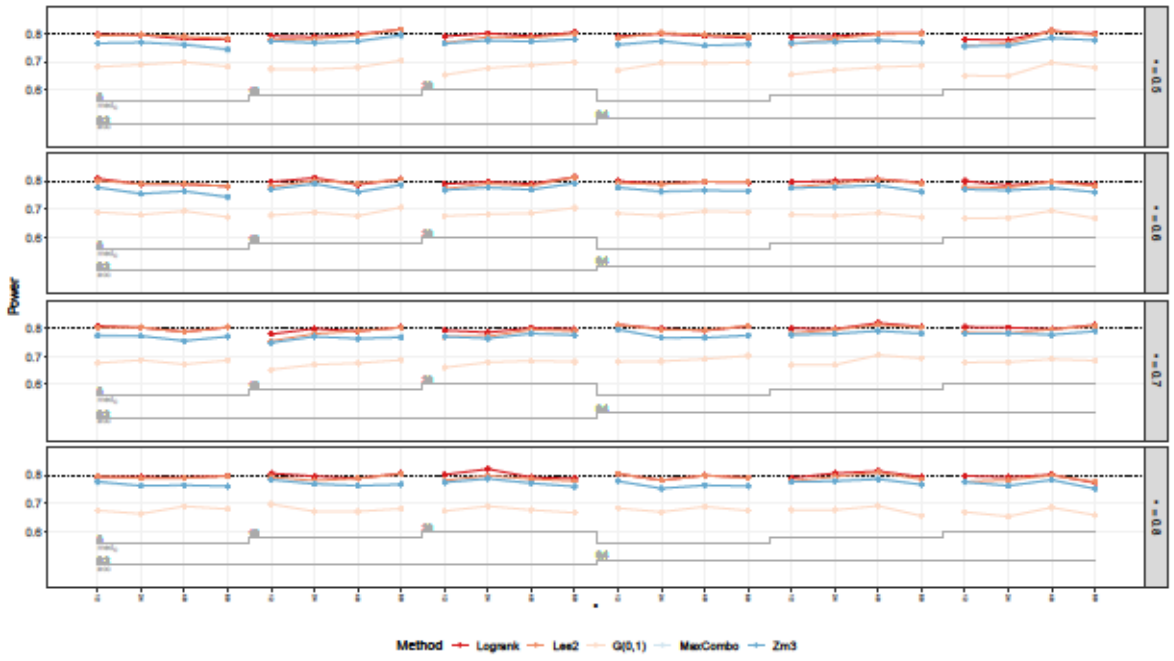

Figure 43: *Power in PH scenarios of non-parametric methods that outperform the logrank test in 80% of NPH scenarios for constant hazard together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

Order of the methods: Figures 44 and 45 show the power of the different methods but arranged by the delay t_2^* and changepoint t_1^* used in the generalized linear lag model of the data generating process. Separate plots for the study duration τ, the median survival in the control arm med_C, the maximum treatment effect θ and the accrual proportion acc were created which resulted in 96 plots. In contrast to Figure 38 the focus does not lie on an

overall comparison of the methods across all scenarios but on the impact of the delay and changepoint parameter on the relationship between the power of the methods. Therefore, the methods are displayed on the x-axis ordered by the upper left panel of the plot to see if the ordering remains the same throughout the different delay and changepoint combinations. Although the order of the methods is slightly different in each plot, the quality of it is the same and hence only the plot for an overall study duration of $\tau = 48$ months, an accrual of $\text{acc} = 0.2 \cdot \tau = 9.6$ months, and a median survival in the control group of $\text{med}_C = 15$ months is shown both for the maxmimum effect $\theta = 0.5$ (Figure 44) and the minimum effect $\theta = 0.8$ (Figure 45).

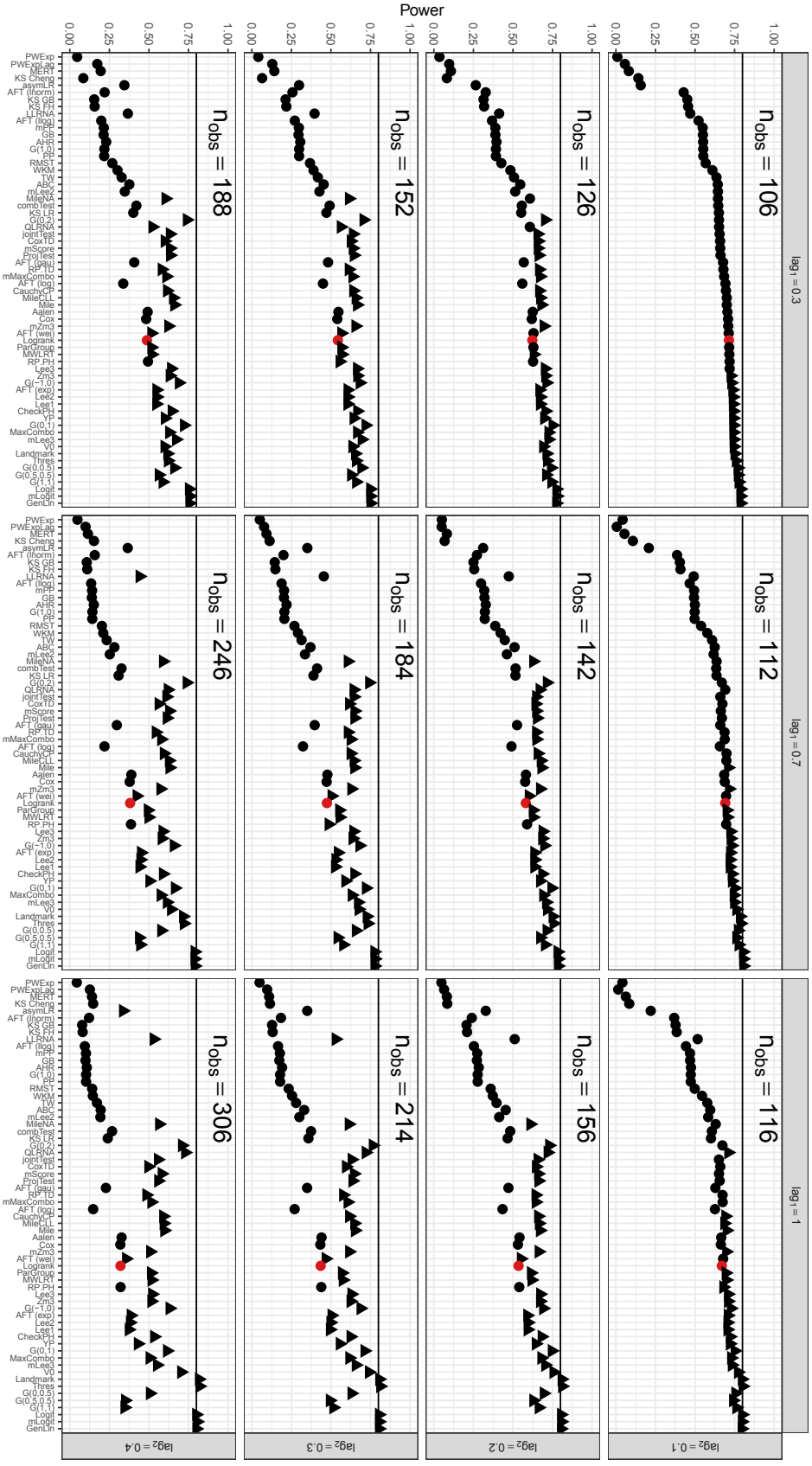

Figure 44: *Power of the different methods for a study duration of 48 months, an accrual of 9.6 months, median survival of 15 months and maximum effect of HR 0.5 when hazards are constant. Triangular shape indicates that the power exceeds the power of the logrank test by more than the Monte-Carlo standard error based on the evaluable datasets. Panels are arranged by delay proportion (lag_2) and changepoint proportion (lag_1). The logrank test is highlighted in red.*

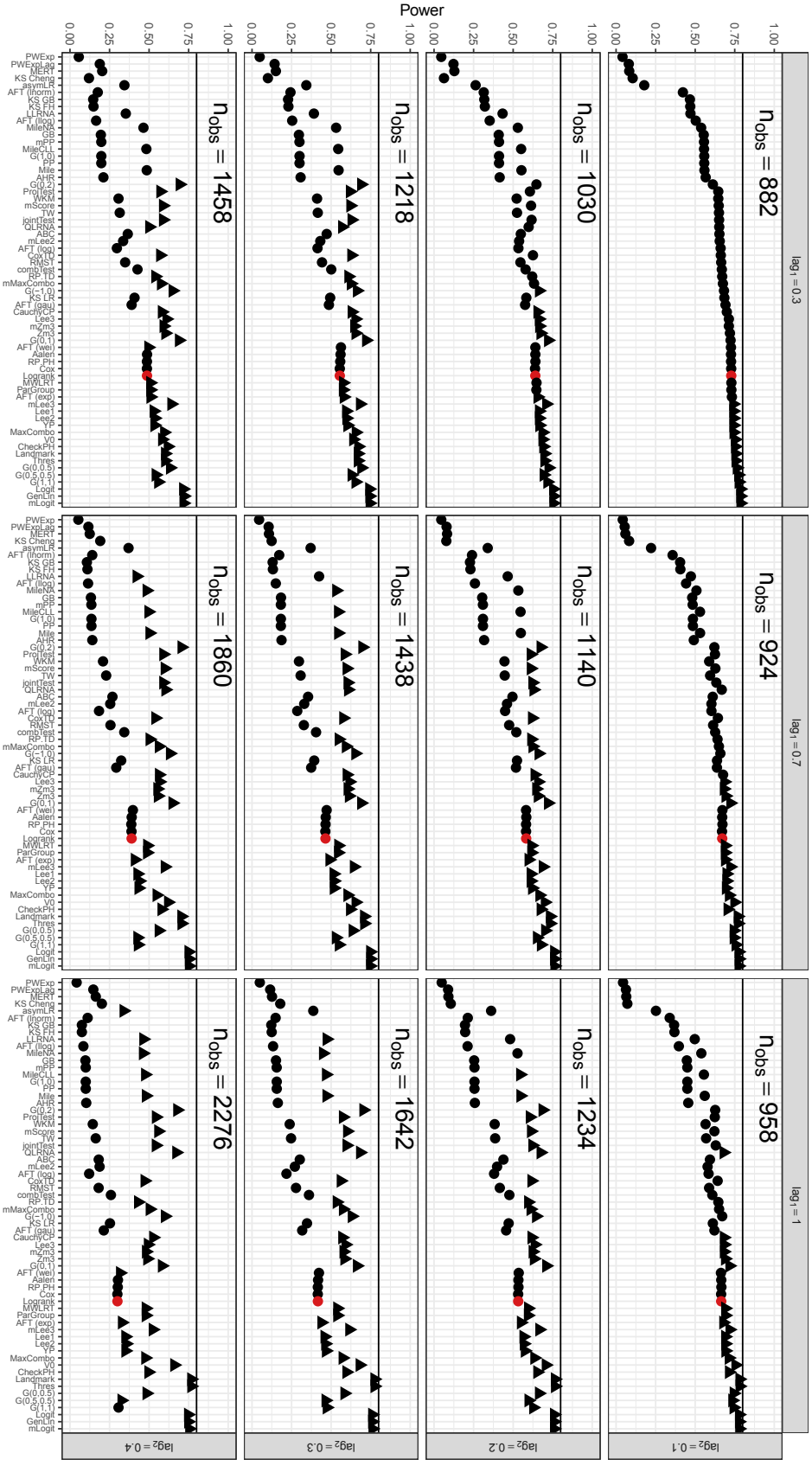

Figure 45: *Power of the different methods for a study duration of 48 months, an accrual of 9.6 months, median survival of 15 months and minimum effect of HR 0.8 when hazards are constant. Triangular shape indicates that the power exceeds the power of the logrank test by more than the Monte-Carlo standard error based on the evaluable datasets. Panels are arranged by delay proportion (lag_2) and changepoint proportion (lag_1). The logrank test is highlighted in red.*

As can be seen in these figures the overall order of the methods based on their power is not preserved between the different delay patterns. The overall number of methods that outperform the logrank test increases with increasing deviation from the PH assumption, i.e. with increasing delay and changepoint proportion. Furthermore, there are almost no methods that outperformed the logrank test in the upper left panel and dropped below the power of the logrank test in the other panels. The methods with lower power than the logrank test in the upper left panel split up into methods whose power decreases and methods whose power stays stable the stronger the deviation from the PH assumption gets.

Chance of rejection of each method: Figure 46 shows the results of the logistic regression models with the estimated odds ratio on the y-axis and the different methods on the x-axis. The boxplots summarize the results for the different combinations of $\text{med}_C, \text{acc}, \tau$ and θ and are arranged by the independent variable. The logrank test is again highlighted in red and a reference line for an odds ratio of 1 is given.

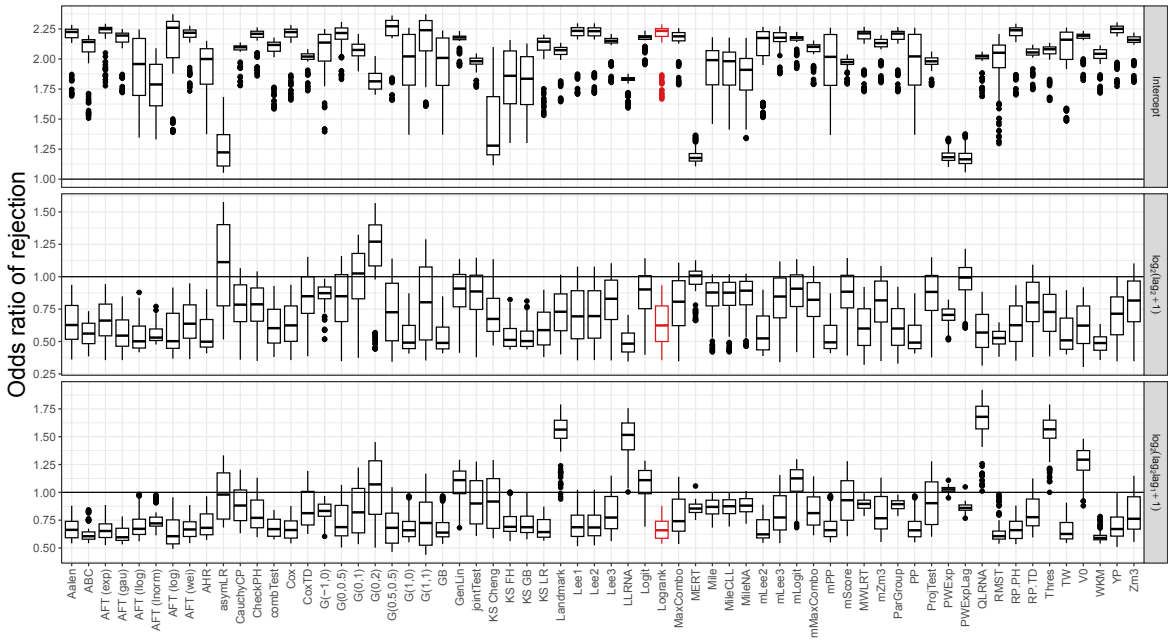

Figure 46: *Results from logistic regression models for rejection of the null hypothesis under the alternative for the single methods in constant hazard scenarios with independent variables delay proportion (lag_2) and changepoint proportion (lag_1). The logrank test is highlighted in red and for an odds ratio of 1 a reference line is given.*

The first row of Figure 46 shows the chance of rejecting the null hypothesis if no delay and changepoint are present, which corresponds to the PH scenarios. As was already observed in Section 3.3.3.1 the asymptotic logrank test (*asymLR*), the *MERT* test and the piecewise exponential models (*PWExp, PWExpLag*) have a very low chance of rejecting the null hypothesis. As already seen in Figure 35 the width of the boxplots is bigger than in decreasing hazard scenarios. The impact of the delay proportion lag_2 is shown in the second row of Figure 46. For most methods it stays below an odds ratio of one indicating that the chance of rejection is smaller the bigger the delay proportion. The asymptotic logrank test (*asymLR*) and two of the Fleming Harrington tests for late difference (*G(0,1), G(0,2)*) have an odds ratio greater than one in more than 50% of scenarios and hence benefit from increasing delay proportions. The effect of the changepoint proportion is shown in the third row of Figure 46 and has a similar effect on most methods with and odds ratio below one. For *GenLin, Landmark, LLRNA, Logit, mLogit, QLRNA*, Threshold (*Thres*) and *V0*, however, the effect of the changepoint proportion is reversed leading to an increased chance of rejection which is highest for *Landmark*, Threshold (*Thres*), *LLRNA, QLRNA* and *V0*.

3.3.4 Increasing hazards

Lastly, the power of the methods is compared if the hazard in the control arm is increasing. Again it was calculated in how many of the 1248 simulated scenarios each method exceeded the power of the logrank test by more than the Monte-Carlo standard error given the number of evaluable datasets. This is displayed in a bar chart in Figure 47. As no method exceeds the logrank test in all NPH scenarios, no bar is colored green.

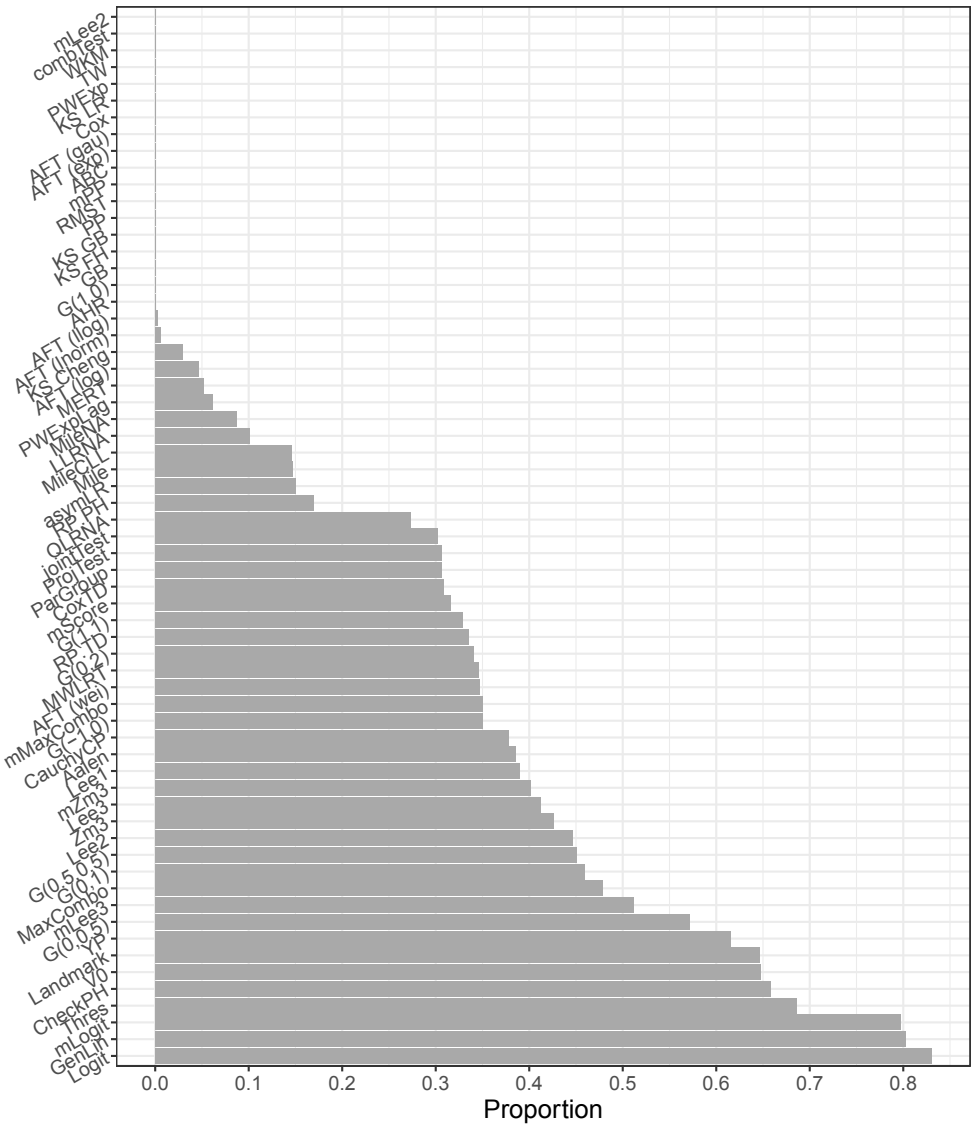

Figure 47: *Proportion of scenarios in which each method exceeds the power of the logrank test in increasing hazard scenarios.*

Of the four methods which were identified in Section 3.2.3 to have increased type 1 error in increasing hazard scenarios (*CheckPH, YP, RP.TD, AFT (log)*), *CheckPH* performed best outperforming the logrank test in 821 of the 1248 scenarios. Only the *Logit* and *GenLin* method achieved a higher power than the logrank test in more than 80% of scenarios.

3.3.4.1 PH scenarios

Figure 48 shows all methods on the x-axis and the power of the method in all 96 scenarios summarized in a boxplot with the Logrank test highlighted in red. As a sample size calculation tailored to this test has been performed, the logrank test should achieve approximately 80% power in each scenario. For the methods for which parameters had to be chosen the results displayed here are based on the moderate parameter, i.e. 20% of the overall study duration for the methods using the Landmark cutoff and an interval of $[0, 0.2\tau]$ for the methods using the GenLin parameter.

For increasing hazards the AFT model based on exponentially distributed error terms (*AFT (exp)*), the Kolmogorov-Smirnov type test based on Cheng (*KS Cheng*), the *MERT* test, the Gray-Tsiatis test (*G(-1,0)*) and the piecewise exponential model (*PWExp*) have very variable power with an interquartile range of approximately 20 percentage points.

For all these methods except the Gray-Tsiatis test (*G(-1,0)*) the median power in the PH scenarios was below 50%. The *Landmark* and Threshold (*Thres*) test had very extreme outlier with power below 25% which occurred in scenarios with very low median survival of $\text{med}_C = 5$ months and high overall study duration of $\tau = 48, 60$ months. This is due to the fact, that the parameter for analysis was chosen based on the study duration τ and is hence quite high. This is problematic in these scenarios where events occur very early on in the study as the methods then ignore too many events.

Again the Royston-Parmar PH model (*RP.PH*) performs really good with a power of approx. 78% to 82% which is always higher than the Royston-Parmar TD model (*RP.TD*) with a power of 68% to 74%. Power of Milestone survival rate comparisons (*Mile, MileNA, MileCLL*) ranges from approx. 38% to 77%. RMST obtained a power of approximately 43% to 78%.

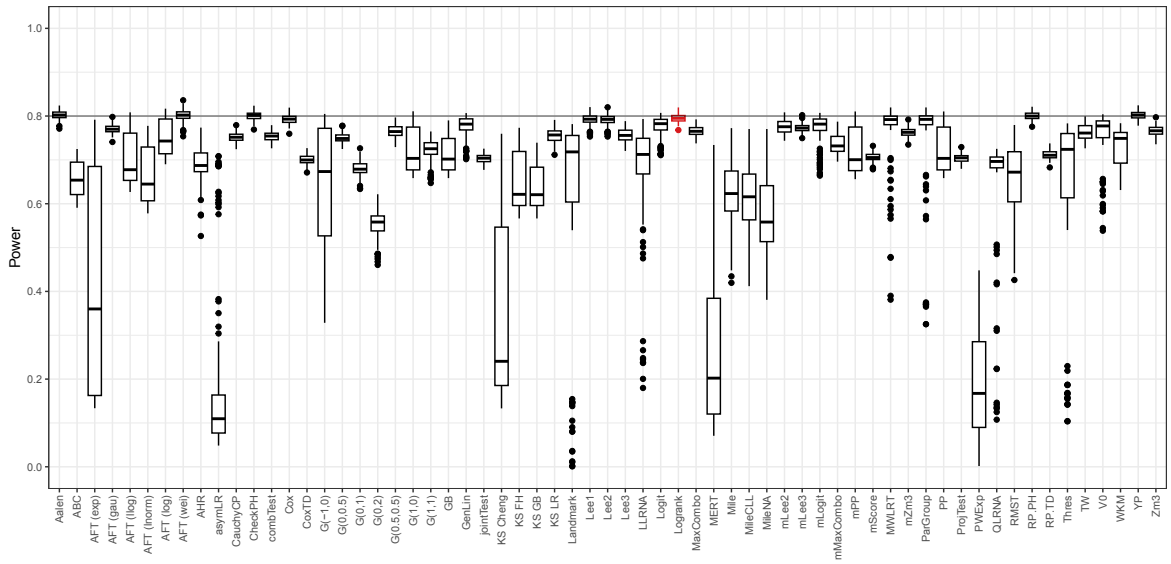

Figure 48: *Boxplot of power in PH scenarios with increasing hazard for all 96 scenarios. Logrank is highlighted in red.*

For the aforementioned methods with very variable power, i.e. an interquartile range of approximately 20 percentage points, a nested loop plot was created to identify the cause of this behavior (Figure 49).

The *MERT* test and the piecewise exponential model *PWExp* perform very similar in scenarios with high median survival in the control arm $\text{med}_C = 15, 20$ months where the power ranges from 0.2% to 37.6%. For small $\text{med}_C = 5$ months the *MERT* test is better than the *PWExp* model and the power difference increases the longer the overall study duration τ. The maximum power achieved in these scenarios is 73.4% which explaines the high variability.

The Gray-Tsiatis test (*G(-1,0)*) performs best in scenarios with higher med_C, while in the scenarios with $\text{med}_C = 5$ months it is often outperformed by the *MERT* test. In all scenarios, however, it performs better than the AFT model with exponentially distributed error terms (*AFT (exp)*) and the Kolmogorov-Smirnov type test based on the weighted logrank statistc by Cheng (*KS Cheng*). For $\text{med}_C = 5$ months scenarios the power of these three methods was relatively stable, but in the other scenarios higher overall study duration lead to an big decrease in power. These patterns are the same within each row and hence the effect of θ is negligible.

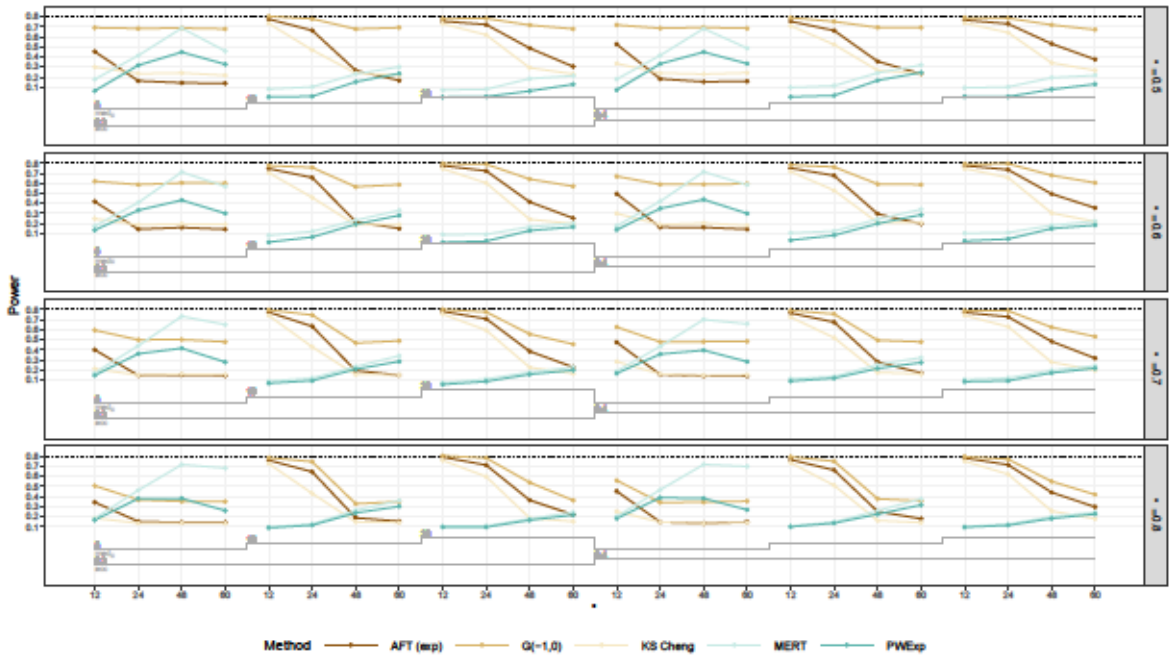

Figure 49: *Nested loop plot of power in PH scenarios with increasing hazard for methods with variable power. The overall study duration τ is plotted on the x-axis and clustered by median survival in the control arm* med_C *and duration of accrual* acc. *The panels are arranged by the treatment effect θ.*

3.3.4.2 NPH scenarios

The 1152 NPH scenarios with increasing hazard are uniquely defined by the combination of median survival in the control arm med_C, the accrual proportion acc, the overall study duration τ, the maximum treatment effect θ and the delay t_2^* and changepoint t_1^*. For the methods for which parameters had to be chosen the results displayed here are based on the correct specified parameter, i.e. t_1^* for the methods using the Landmark cutoff and an interval of $[t_1^*, t_2^*]$ in case of linear lag scenarios for the methods using the GenLin parameter. In case of threshold lag scenarios ($t_1^* = t_2^*$) there is no correct interval for the latter methods and the interval $[0.8 \cdot t_2^*, 1.2 \cdot t_2^*]$ was chosen.

Overall power of all methods: Figure 50 shows all methods on the x-axis and the power of the method in all NPH scenarios summarized in a boxplot with the Logrank test highlighted in red. For comparison grey boxplots are added that represent the power in PH scenarios as shown in the preceding section. It can be seen that the boxplots are much wider than in the PH scenarios with a mean interquartile range of 36.7 and a mean overall power range of 89

percentage points ranging from an average minimum power of 3.4% to an average maximum power of 92.3%. This is due to the fact that the NPH scenarios are much more heterogeneous and differ in the extent they deviate from the PH assumption. For 26 of 64 methods the median power is above 80%.

In contrast, the *MERT* test and the piecewise exponential models (*PWExp, PWExpLag*) have the lowest interquartile range of up to 10.2 percentage points and an overall range of up to 35.6 percentage points which is due to the overall poor power of at most 36%. These are also the methods that show a lower average power in NPH scenarios than in PH scenarios. All other methods have a range of more than 70 percentage points.

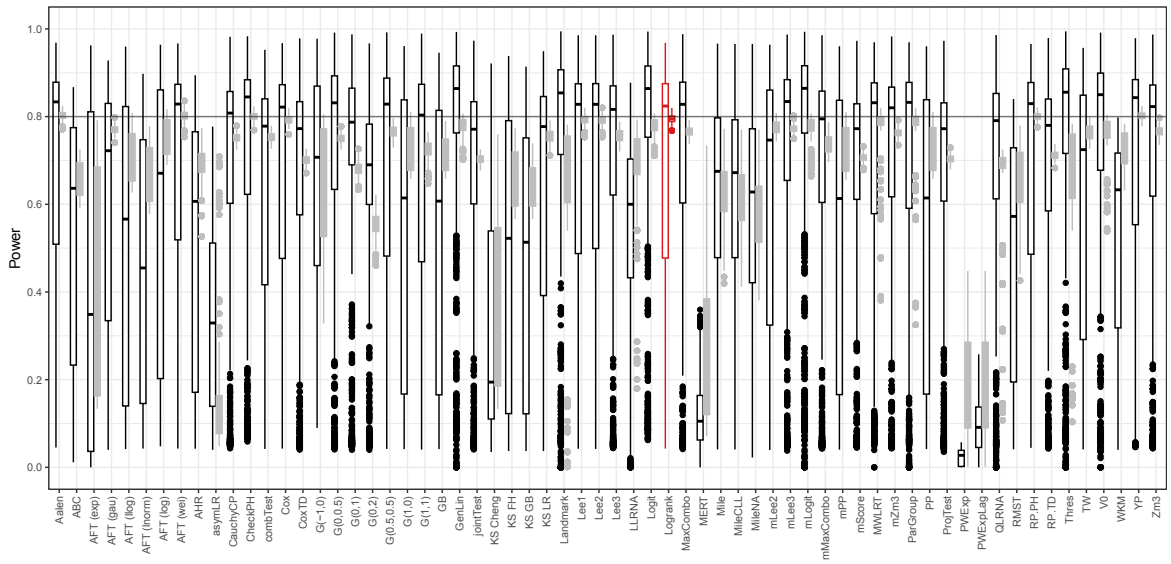

Figure 50: *Boxplot of power in NPH scenarios with increasing hazard in black and for comparison the power in PH scenarios in grey. Highlighted in red is the logrank test.*

To disentangle the effect of the different delay and changepoint combinations the following Figure 51 is similar to the previous one but with the NPH scenarios arranged by delay t_2^* and changepoint t_1^*. The grey boxplots of PH scenarios are the same in each panel.

As can be seen in this figure, the range of the power values for each method increases the greater the deviation from the PH assumption, i.e. the greater the dealy and changepoint proportions. In the upper left panel the boxplots are considerably more narrow with more outlier values. With increasing changepoint and delay the boxplots widen and the power decreases which makes the outlier more extreme. In the bottom row with greatest delay, there

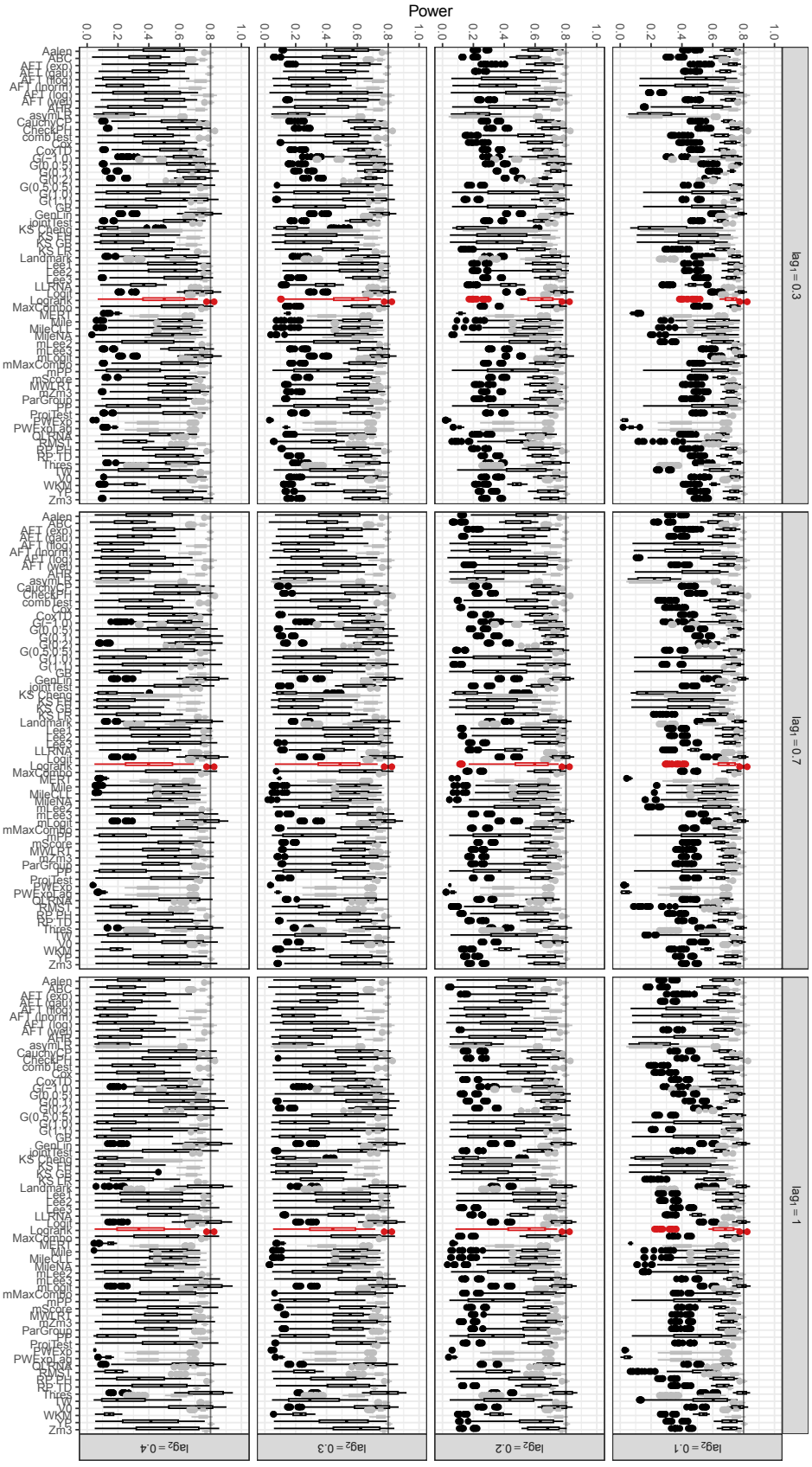

Figure 51: *Boxplot of power in NPH scenarios with increasing hazard arranged by delay proportion* lag_2 *and changepoint proportion* lag_1*. For comparison the power in PH scenarios is given in grey and the same in each panel and the logrank test is highlighted in red.*

is only a small number of methods whose power values do not spread out across such a wide range and have hence no outliers. Of these methods the Fleming-Harrington tests for detecting late differences (*G(0,1), G(0,2)*), the *GenLin* method and the *Logit* and *mLogit* methods have relatively high power values when the changepoint proportion is low ($\text{lag}_1 = 0.3$). For threshold lag models ($\text{lag}_1 = 1$) the *Landmark* model, the threshold test (*Thres*), the *V0* test and the quadratic combination of logrank and Nelson-Aalen test at fixed timepoints (*QL-RNA*) also have high power and few outliers. The *MERT* test and the piecewise exponential (*PWExp, PWExpLag*) methods have low power throughout all scenarios.

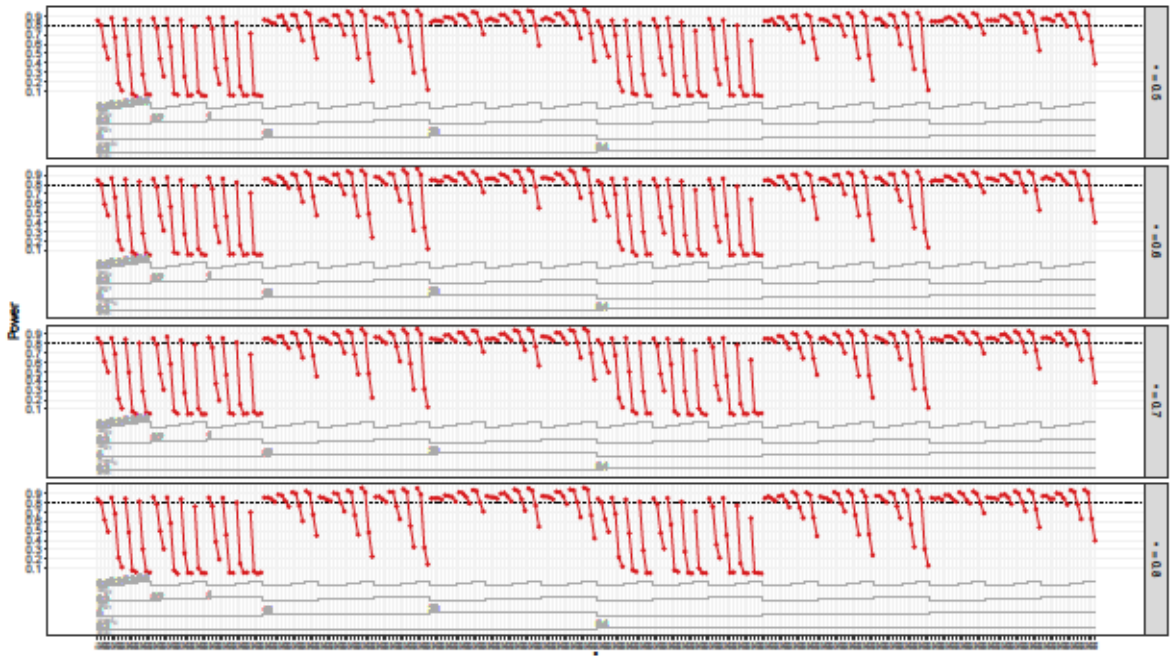

Figure 52: *Power of the logrank test in NPH scenarios with increasing hazards with the overall study duration τ on the x-axis and further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

Power of single methods: Figure 52 displays the behavior of the well established logrank test in nested loop plots arranged by the maximum treatment effect θ. In these scenarios the power of the logrank test is very variable ranging from 4.3% to 96.8%. The high variability mainly occurs for scenarios with very low median survival of $\text{med}_C = 5$ months where the power decreases drastically the longer the study duration τ with power ranging from 4.3% to 88.3%. This is remedied with increasing med_C revealing power values from 11% to 96.8% for $\text{med}_C = 15$ months and 38.4% to 96.5% for $\text{med}_C = 20$ months. The loss in power with

increasing study duration can be explained as this leads to a decreased sample size due to the sample size calculation. The other parameters delay proportion lag_2, changepoint proportion lag_1, accrual proportion acc and maximum treatment effect θ play a minor role.

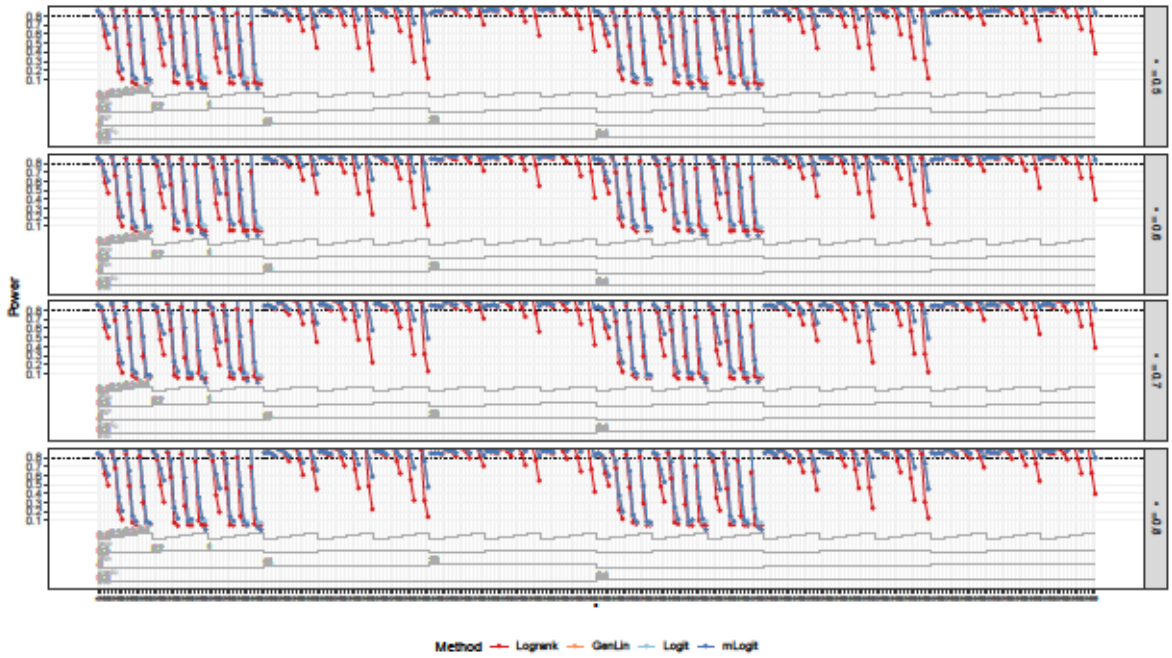

Figure 53: *Power in NPH scenarios of the parametric methods that outperform the logrank test in approx. 80% of scenarios with increasing hazards together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm med_C, the accrual proportion* acc*, the delay proportion lag_2 and changepoint proportion lag_1. The panels are arranged by the maximum treatment effect θ.*

As has been shown at the beginning of section 3.3.4 no method outperforms the logrank test in all NPH scenarios. However, three methods outperform the logrank test in approximately 80% of the cases which are: the generalized linear model (*GenLin*) and the logit and modified logit model (*Logit, mLogit*). The three methods are displayed against the logrank test in the following nested loop plot for NPH scenarios (Figure 53).

As can be seen all methods perform very similar and their power as well as the power of the logrank test is increasing the higher the median survival in the control arm (med_C). For scenarios with low overall study duration τ the power gain compared to the logrank test is very small to non-existent and gets bigger with increasing study duration. Additionally, the

power gain for high study duration is small in scenarios with low med_C and increases for higher med_C to approx. 40%. For scenarios with $\text{med}_C = 20$ months all methods have a power of at least 80% while the power of the logrank test goes down to 38.4%.

For comparison their performance in PH scenarios is shown in Figure 54.

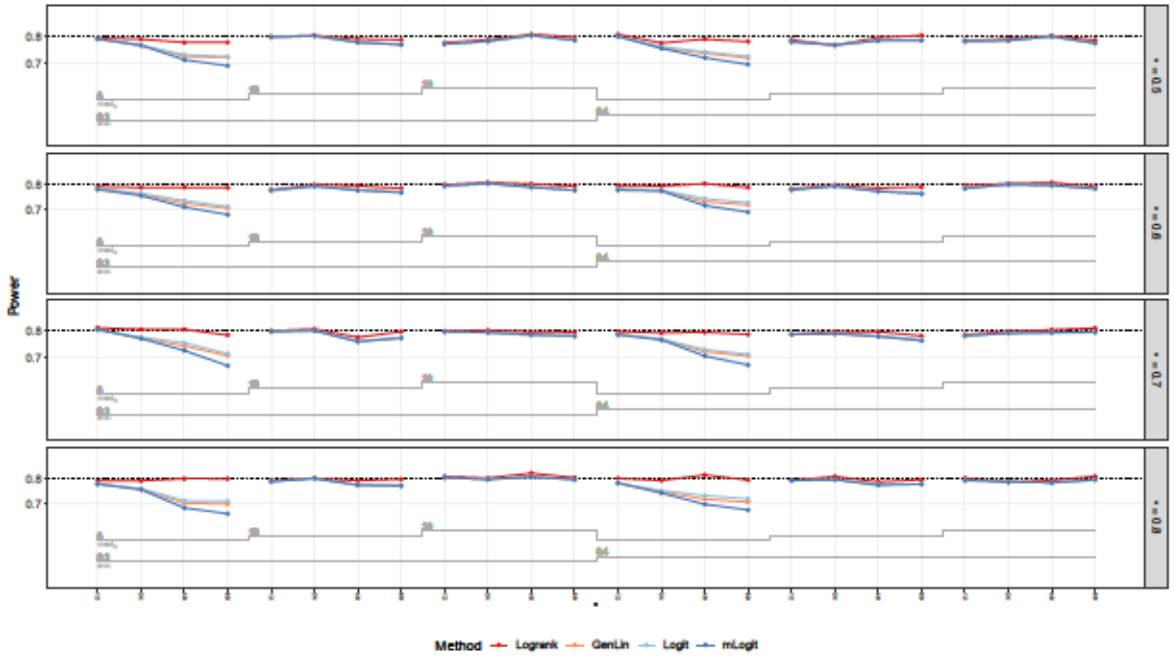

Figure 54: *Power in PH scenarios of the parametric methods that outperform the logrank test in approx. 80% of NPH scenarios with increasing hazards together with the power of the logrank test for comparison. The plots show the overall study duration τ on the x-axis and are further clustered by median survival time in the control arm* med_C*, the accrual proportion* acc*, the delay proportion* lag_2 *and changepoint proportion* lag_1*. The panels are arranged by the maximum treatment effect θ.*

The methods again perform very similar and for longer median survival in the control arm there is almost no difference to the power of the logrank test of 80%. For low median survival of $\text{med}_C = 5$ months the difference to the logrank test becomes more apparent and increases with higher study duration as well as the difference between the methods. The biggest difference is achieved at a study duration of $\tau = 60$ months and with a power of 66.4% for the modified Logit model (*mLogit*).

Although nonparametric methods are more robust to parameter misspecification, no non-parametric method outperformed the logrank test in all NPH scenarios. The best method

compared to the logrank test were the Fleming-Harrington test for late differences (*G(0,0.5)*) and the modified Lee3 test (*mLee3*), which outperformed the logrank test in 57% and 51% of the scenarios, respectively. Due to this low number of outperforming scenarios the non-parametric methods were not investigated more deeply for the increasing hazard scenarios.

Order of the methods: Figures 55 and 56 show the power of the different methods. The plots are arranged by the delay t_2^* and changepoint t_1^* used in the generalized linear lag model of the data generating process. The methods are ordered by the upper left panel of the plot to see if the ordering remains the same throughout the different delay and changepoint combinations. This is done for an overall study duration of $\tau = 48$ months, an accrual of $\text{acc} = 0.2 \cdot \tau = 9.6$ months, and a median survival in the control group of $\text{med}_C = 15$ months. Figure 55 shows the results for the maxmimum effect $\theta = 0.5$ and Figure 56 for the minimum effect $\theta = 0.8$.

As before the overall order of the methods based on their power is not preserved between the different delay patterns. Many of the methods with higher power than the logrank test in the upper left panel do not only keep their power above the power of the logrank test but do also decrease less than the other methods the bigger the deviation from the PH assumption gets. Of the methods that had lower power than the logrank test in the upper left panel, there are many with relatively stable or even increasing power such as Fleming-Harrington weighted logrank test for late differences (*G(0,1), G(0, 0.5), G(0,2)*) and regression models with time-dependent treatment effects (*CoxTD, RP.TD*) but mainly combination tests (e.g. *MaxCombo, Zm3, mZm3*). But also the Cauchy changepoint method (*CauchyCP*) performed really well. The remaining methods with lower power decrease uniformly so that order of the methods does not change much, which is completely different for the methods with higher power, which spread out much more the extremer scenarios get. As seen before the generalized linear lag model (*GenLin*) and the *Logit* models are always on top throughout all scenarios.

Chance of rejection of each method: Figure 57 shows the results of the logistic regression models with the estimated odds ratio on the y-axis and the different methods on the x-axis. The boxplots summarize the results for the different combinations of $\text{med}_C, \text{acc}, \tau$ and

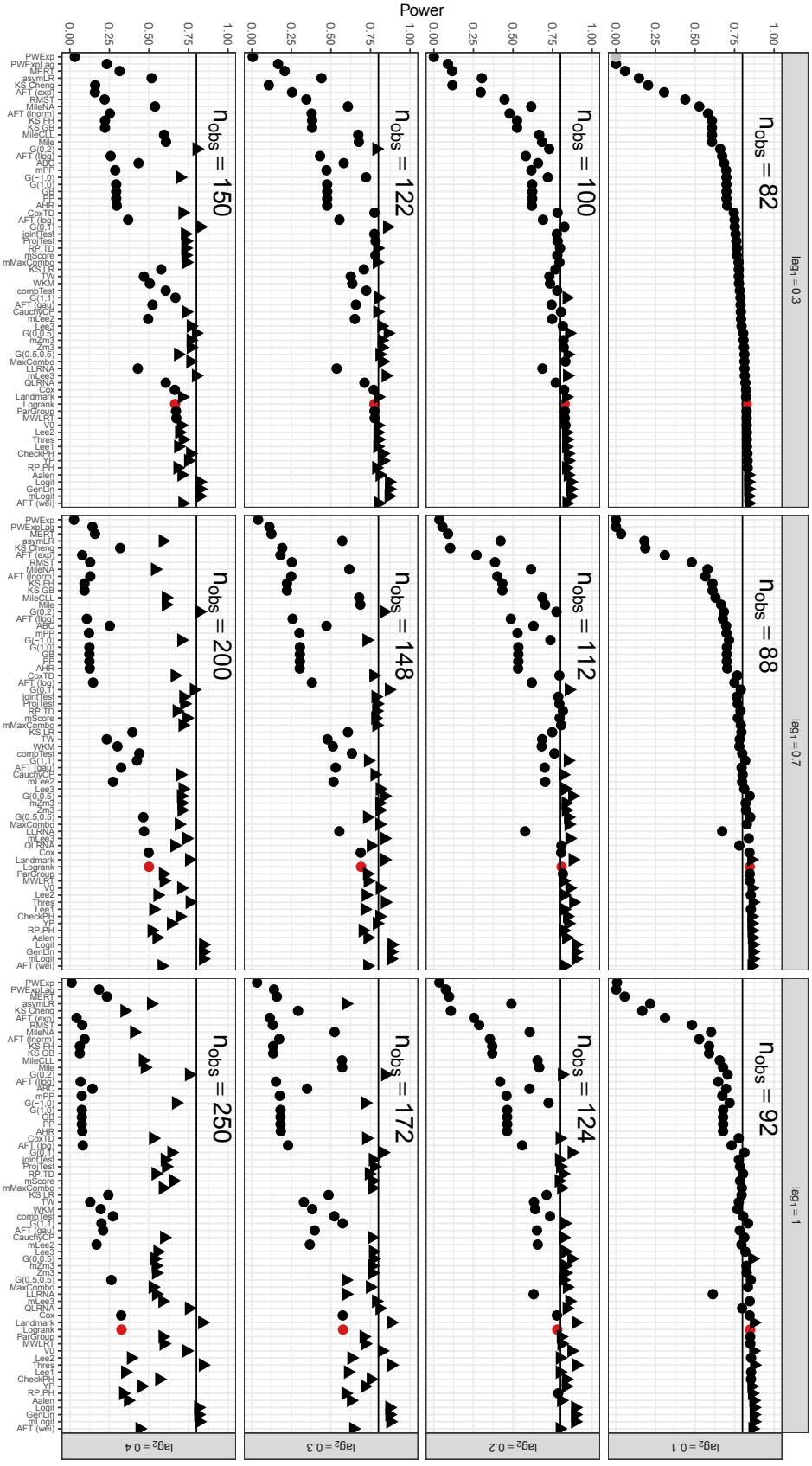

Figure 55: *Power of the different methods for a study duration of 48 months, an accrual of 9.6 months, median survival of 15 months and maximum effect of HR 0.5 when hazards are increasing. Triangular shape indicates that the power exceeds the power of the logrank test by more than the Monte-Carlo standard error based on the evaluable datasets. Panels are arranged by delay proportion (lag_2) and changepoint proportion (lag_1). The logrank test is highlighted in red.*

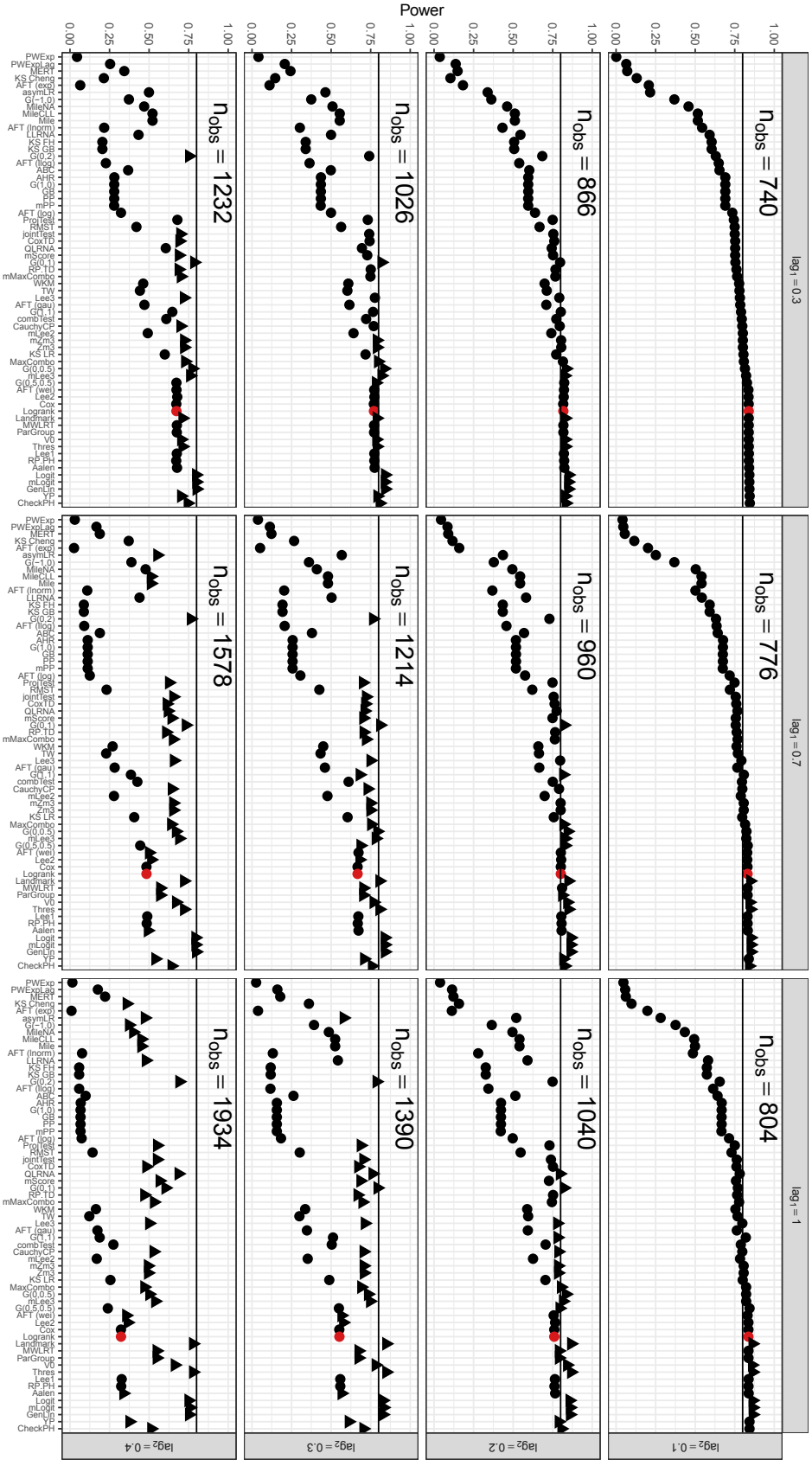

Figure 56: *Power of the different methods for a study duration of 48 months, an accrual of 9.6 months, median survival of 15 months and minimum effect of HR 0.8 when hazards are increasing. Triangular shape indicates that the power exceeds the power of the logrank test by more than the Monte-Carlo standard error based on the evaluable datasets. Panels are arranged by delay proportion (lag_2) and changepoint proportion (lag_1). The logrank test is highlighted in red.*

θ and are arranged by the independent variable. The logrank test is again highlighted in red and a reference line for an odds ratio of 1 is given.

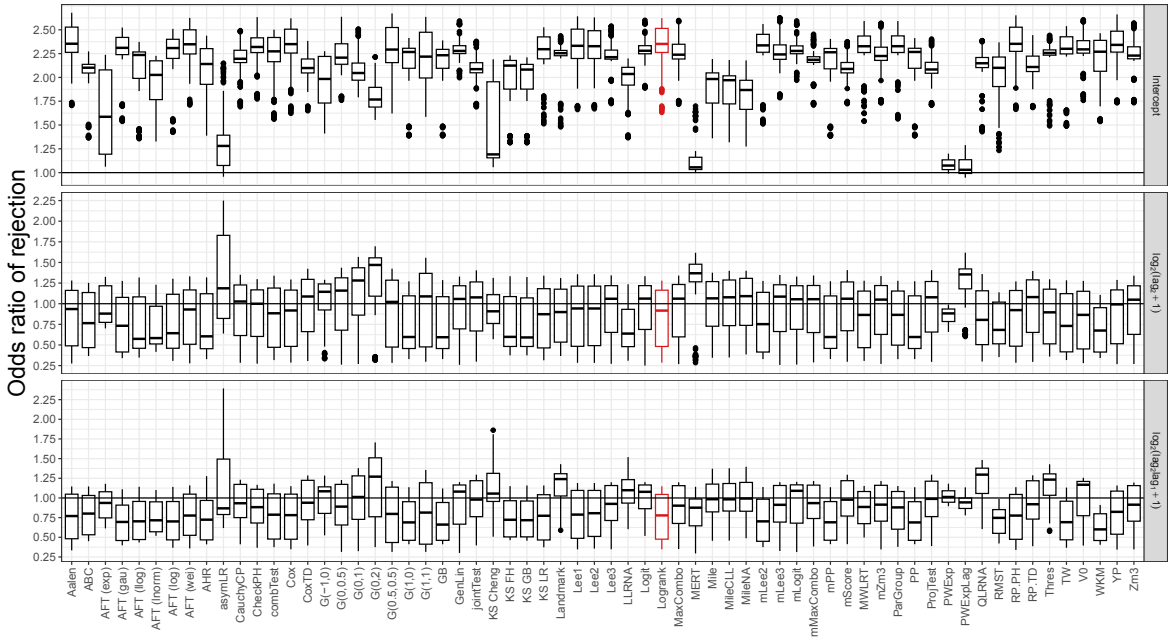

Figure 57: *Results from logistic regression models for rejection of the null hypothesis under the alternative for the single methods in increasing hazard scenarios with independent variables delay proportion (*lag_2*) and changepoint proportion (*lag_1*). The logrank test is highlighted in red and for an odds ratio of 1 a reference line is given.*

The first row of Figure 57 shows the chance of rejecting the null hypothesis if no delay and changepoint are present, which corresponds to the PH scenarios. Again the asymptotic logrank test (*asymLR*), the *MERT* test and the piecewise exponential models (*PWExp, PWExpLag*) have a very low chance of rejecting the null hypothesis.

The impact of the delay proportion lag_2 is shown in the second row of Figure 57. For almost all methods the box of the boxplot contains an odds ratio of one indicating that the chance of rejection is not influenced by lag_2 or gets even bigger the bigger the delay proportion in at least a quarter of the scenarios. The Fleming-Harrington test for late difference (*G(0,2)*), the *MERT* test and the piecewise exponential lag model (*PWExpLag*) have an odds ratio greater than one in more than 75% of scenarios and hence benefit from increasing delay proportions.

The effect of the changepoint proportion is shown in the third row of Figure 57 and has a similar effect on most methods with and odds ratio below one. For *Landmark, QLRNA* and

the Threshold (*Thres*) method the effect of the changepoint proportion is reversed in more than 75% of scenarios leading to an increased chance of rejection. Additionally, there are scenarios where the chance of rejection of the asymptotic logrank test (*asymLR*) and the Kolmogorov-Smirnov type test based on the weighted logrank test by Cheng (*KS Cheng*) is approximately doubled when the changepoint proportion is doubled.

3.4 Recommendation at planning stage

To conclude this chapter it was endeavored to give guidance on the choice of approppriate methods when a new trial is designed. To do so the results of the previous sections are weighed up and a ranking system is introduced depending on the amount of knowledge available at the time of planning the new study. For the ranking system the three methods that showed an inflated type 1 error in all scenarios, i.e. *CheckPH, YP* and *AFT (log)*, were not taken into consideration. For all other methods the ranks based on their power are calculated for each scenario contributing lower ranks to higher power and then averaged across all scenarios corresponding to the planning assumptions. For the planning assumptions the following four were considered: assuming proportional hazards (PH), assuming a delayed treatment effect without specific knowledge of the delay (NPH), assuming a delayed treatment effect with knowledge of the timepoint when the full effect is achieved (TLM), and assuming a delayed treatment effect with full knowledge of the timepoint the treatment effect increases and the timepoint when the full effect is achieved (GLLM). All of these four planning assumptions are further refined by the shape of the hazard, which corresponds to knowledge on the disease progression and it was distinguished between no previous knowledge, decreasing, constant and increasing hazard. The planning assumptions are summarized in Table 16.

Table 16: *Summary of the planning assumptions*

Assumption	Excluded methods	Simulation scenarios
No delay (PH)	Landmark, MWLRT, Thres, V0, PWExp, ParGroup, LLRNA, QLRNA, PWExpLag, MERT, (m)Logit, GenLin	all PH scenarios ($\text{lag}_2 = \text{lag}_1 = 0$)
General delay without further knowledge (NPH)	Landmark, MWLRT, Thres, V0, PWExp, ParGroup, LLRNA, QLRNA, PWExpLag, MERT, (m)Logit, GenLin	all NPH scenarios ($\text{lag}_2 \neq 0$)
Delay with known onset of delay (TLM)	PWExpLag, MERT, (m)Logit, GenLin	TLM scenarios ($\text{lag}_2 \neq 0, \text{lag}_1 = 1$)
Delay with full knowledge (GLLM)	all methods	GLLM scenarios ($\text{lag}_2 \neq 0, \text{lag}_1 \neq 1$)

The results are shown in Figures 58, 59, 60 and 61. Each figure displays the results of the PH, NPH, TLM and GLLM planning assumption and is subdivided into four panels based on the shape of the hazard. In each panel the results are summarized as barplots with red bars indicating that the method showed a type 1 error inflation in this situation and green bordered bars indicating the five methods with the lowest rank.

Assuming PH If there is no indication that the PH assumption is violated at the planning stage specification of the parameters for analysis of the Landmark and GenLin type methods is not possible and hence these methods were excluded. The ranks of the remaining methods are then averaged over all PH scenarios and the results are shown in Figure 58.

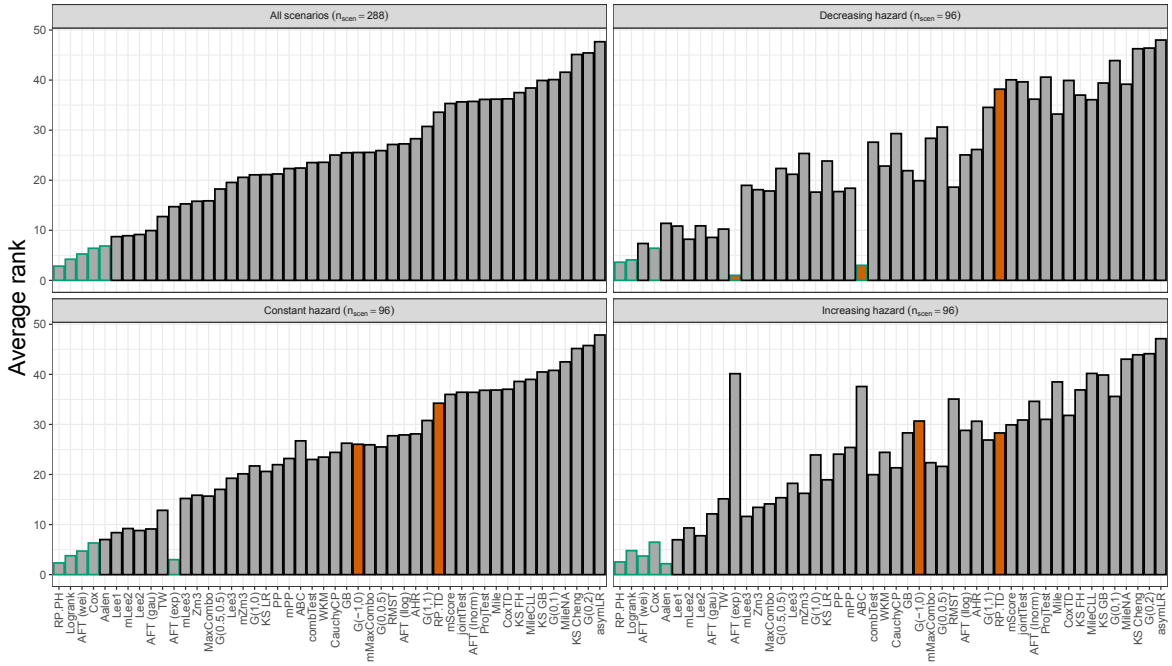

Figure 58: *Barplot of the mean rank of each method averaged over all PH scenarios and over all PH scenarios with decreasing, constant and increasing hazard separately. A red bar indicates that this method showed inflated type 1 error in this setting and green bordered bars indicate the five methods with the best rank. As methods of the Landmark and GenLin type were excluded the number of methods was 48.* n_{scen} *is the number of scenarios considered in each plot.*

As expected the *Logrank, Cox* and Royston-Parmar model for proportional hazards (*RP.PH*) are among the five best performing methods in each panel. Although the Royston-Parmar

model achieves a better average rank than the logrank test, this is within the Monte-Carlo standard error and hence the logrank test can be recommended.

Assuming NPH If, at the planning stage, there is an indication that a delay will be present but no data is available to quantify it, specification of the parameters for analysis of the Landmark and GenLin type methods is not possible and hence these methods were again excluded. The ranks of the remaining methods are then averaged over all NPH scenarios and the results are shown in Figure 59.

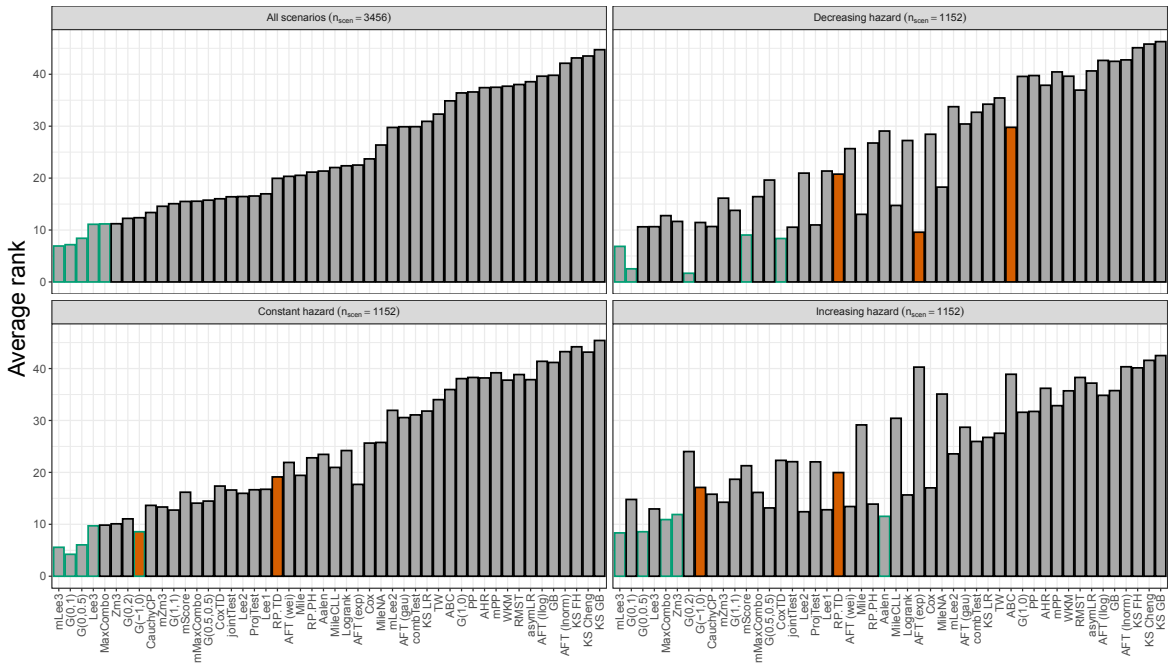

Figure 59: *Barplot of the mean rank of each method averaged over all NPH scenarios and over all NPH scenarios with decreasing, constant and increasing hazard separately. A red bar indicates that this method showed inflated type 1 error in this setting and green bordered bars indicate the 5 methods with the best rank. As methods of the Landmark and GenLin type were excluded the number of methods was 48.* n_{scen} *is the number of scenarios considered in each plot.*

The modified versatile test by Lee (*mLee3*) which is the maximum of the logrank and the Fleming-Harrington test for late differences is in all scenarios among the best five methods and the best method if no knowledge on the shape of the hazard is available. If it is known a priori that the underlying hazard is decreasing or constant it is better to choose the Fleming-Harrington test for late differences (*G(0,1)*).

Assuming TLM Assuming that it is known when the full treatment effect will set in allows specification of the parameters for analysis of the Landmark type methods but not of the GenLin type methods and hence these were excluded. The ranks of the remaining methods are then averaged over all TLM scenarios and the results are shown in Figure 60.

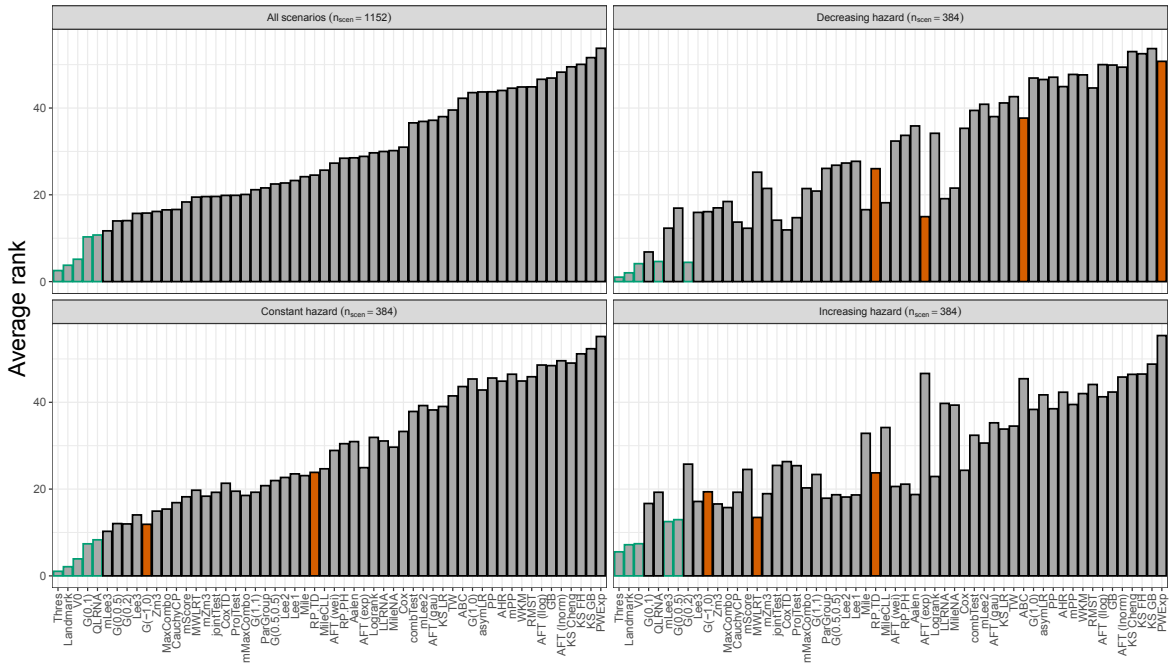

Figure 60: *Barplot of the mean rank of each method averaged over all TLM scenarios and over all TLM scenarios with decreasing, constant and increasing hazard separately. A red bar indicates that this method showed inflated type 1 error in this setting and green bordered bars indicate the 5 methods with the best rank. As methods of the Landmark and GenLin type were excluded the number of methods was 56.* n_{scen} *is the number of scenarios considered in each plot.*

In all scenarios the threshold test (*Thres*), the *Landmark* test and the *V0* test are the three best performing methods. If a method is to be used which does not require the specification of the delay, the Fleming-Harrington test for late differences can be used. However, the amount of weight placed on late timepoints should be chosen depending on the shape of the hazard, i.e. *G(0,1)* if the hazard is constant, *G(0,2)* if the hazard is decreasing, and *G(0,0.5)* if the hazard is increasing.

Assuming GLLM If the exact structure of the delay, i.e. the time when the treatment effect increases and the time when the full effect sets in, is known, all methods can be used.

The ranks are then averaged over all GLLM scenarios and the results are shown in Figure 61.

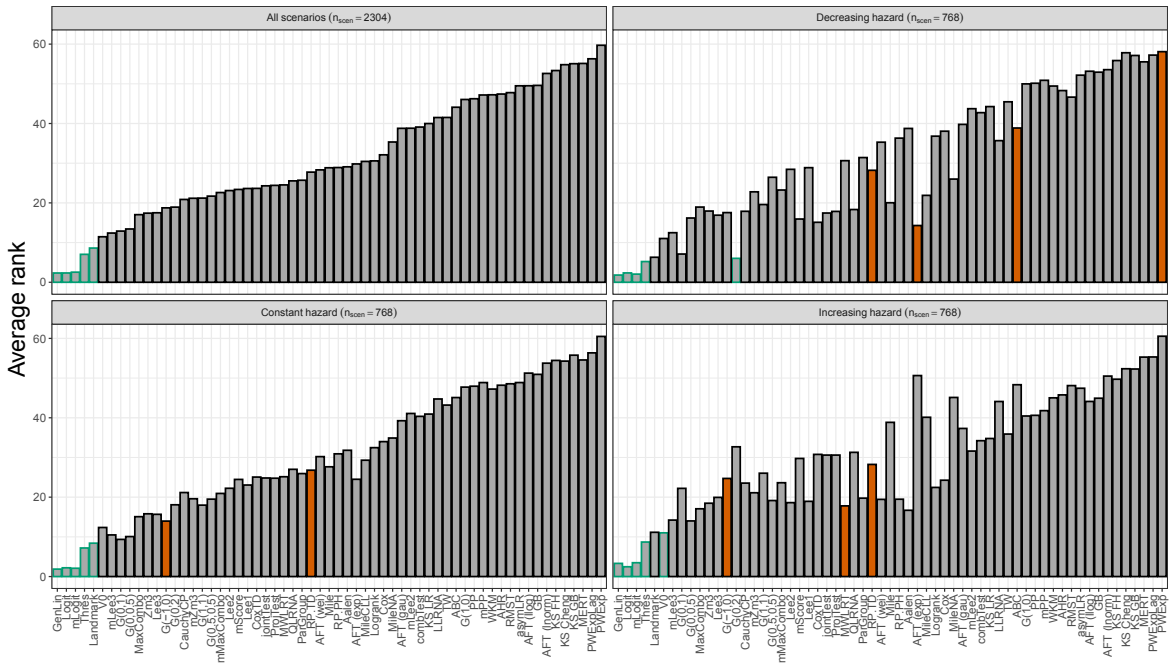

Figure 61: *Barplot of the mean rank of each method averaged over all GLLM scenarios and over all GLLM scenarios with decreasing, constant and increasing hazard separately. A red bar indicates that this method showed inflated type 1 error in this setting and green bordered bars indicate the 5 methods with the best rank. As no methods were excluded the number of methods was 61.* n_{scen} *is the number of scenarios considered in each plot.*

As expected the best performing methods are the *GenLin* test and the *Logit* and modified logit (*mLogit*) test. A good alternative presents the threshold test (*Thres*) which only requires the specification of the time when the full effect sets in for analysis.

Chapter 4

Discussion

In this chapter the contribution of this thesis to current research is presented as well as limitations and directions for future research.

4.1 Contributions to research and discussion

Handling of non-proportional hazards has become increasingly important due to recent advances in oncological immunotherapies. The choice of an appropriate statistical method to assess superiority of a new treatment is essential to the success of a clinical trial. It was shown by Peto and Peto (1972) that a rank-invariant test such as a weighted logrank test is most powerful if the weights are proportional to the logarithmic hazard ratio. This makes the standard logrank test unappealing if a delayed onset of treatment effect can be expected in advance. Hence, the selection of an appropriate statistical method should be made with careful consideration of the underlying mechanism of the oncological entity as was already stressed by many authors, e.g. Ananthakrishnan et al. (2021), Mukhopadhyay et al. (2020).

The aim of this thesis was to systematically evaluate the performance of alternatives to the commonly used logrank test, that were identified in a systematic literature search, when comparing a time-to-event endpoint between two groups in settings where the proportional hazard assumption is violated. For these methods a comprehensive overview is given and if possible R packages in which these methods are implemented as well as own code in an

online repository referenced in the Appendix to facilitate their application. Comparison of these methods was done in an extensive simulation study assessing type 1 error and power under a variety of scenarios in a concise and structured manner allowing to disentangle the impact of different aspects such as the study duration, accrual time or the extent and form of the delay.

In the systematic literature search 42 articles published before 2022 on the general topic of non-proportional hazards were identified of which only 22 considered non-proportional hazard scenarios with delayed treatment effects in their simulation study. However, in most cases this was only a single scenario among other non-proportional hazard scenarios and not a systematic approach to investigate the effect of the delay in detail. Due to this limited number of methods considered in these articles and the relatively simple structure of delayed treatment effect scenarios, this thesis can be understood to extend the work of already published simulation studies by investigating a greater number of available methods. It is found that the performance of the single methods highly depends on the constellation of simulation parameters. For type 1 error the Milestone survival rate comparison based on the Nelson-Aalen estimator was overly conservative whereas the naive procedure of checking the PH assumption, the time-dependent Royston-Parmar model and the model by Yang and Prentice showed an error inflation throughout all simulated scenarios. For accelerated failure time models control of type 1 error for different distributional assumptions depended on the shape of the underlying hazard of the data generating process, e.g. assuming a exponentially distributed error term in the AFT model controlled type 1 error for constant hazard scenarios and led to an error inflation if the hazard was decreasing or an overly conservative test if the hazard was increasing. Now throughout this thesis the best performing methods in NPH scenarios with respect to power were the generalized linear lag model (*GenLin*) and the Logit models (*Logit, mLogit*) since these models are very close to the data generating process and their parameters for analysis were chosen to correspond to the true parameters. This has also been observed by Xu et al. (2018) and Yu et al. (2021) who invented these methods. Although it was assessed that these methods are robust against misspecification of their parameters in NPH scenarios it might be desireable to use methods for which no parameters must be chosen of which the Fleming-Harrington test for late differences stood out in terms of power. This has also been observed by many other authors, e.g. Arfè et al. (2021), Flandre

and O'Quigley (2019), Jiménez (2022), Lin and León (2017), Tang (2022) and Yang et al. (2020). Mukhopadhyay et al. (2020) also observed that this test achieves very high power in delayed treatment effect scenarios but caution to use it thoughtless as its power can decrease drastically if no or another type of non-proportional hazards is observed. In this thesis the Fleming-Harrington test for late differences also performed better than the logrank test in all NPH scenarios with decreasing hazards and in most of the NPH scenarios with constant and increasing hazards if the median survival was low and the overall study duration high, but showed a power loss of 10%-15% compared to the logrank test in PH scenarios. Another quite elaborated simulation study was performed by Jachno et al. (2021) who considered decreasing, constant and increasing hazard situations but only for threshold lag scenarios with a single maximum treatment effect comparing 13 methods in their simulation study. They concluded that the best performing test was the *Zm3* test as it maintained power. The combination of weighted logrank tests of the Fleming-Harrington family was also advocated by other authors depending on which combinations were considered in the simulation study. Chen et al. (2022), Cheng and He (2021) and Ristl et al. (2021) observed that the *MaxCombo* test represents a robust variant whereas Royston and Parmar (2020) preferred their modified version of the Zm3 test, the *mZm3* test. For Lin et al. (2020) general combination tests are a useful tool if no prior knowledge is available to select optimal weigthed logrank tests. In this thesis it was also observed that these combined tests are very similar to the logrank test in PH scenarios with only marginal power loss. In NPH scenarios, however, the advantage of these methods over the logrank test depends on the shape of the underlying distribution. While for decreasing hazards the logrank test was always outperformed by the combination tests, this reduced to approx. 80% of scenarios with constant hazards and further to approx. 40% of scenarios with increasing hazards. It should be mentioned, however, that Magirr (2021) criticized the use of Fleming-Harrington family of weighted logrank tests since this choice of weights can be problematic as it downweights early events so much that early detrimental effects of the new treatment can be rewarded by these statistics. This was not observed in this thesis as all scenarios assumed equality of the survival distributions before the delay.

Furthermore, a more recently published simulation study by Klinglmüller et al. (2023), which was mentioned in the introduction but not considered in the literature search within this thesis, also investigated delayed treatment effect scenarios but considered only a constant

hazard. As a result of their simulation study the modestly weighted logrank test showed to be a viable and robust choice which was also favored by Magirr and Burman (2019) who proposed this test. This recommendation is based on the fact that, although this test achieves lower power than weighted logrank test alternatives specifically tailored to detect delayed treatment effects, such as the Fleming-Harrington test for late differences, the *MWLRT* loses less power in PH scenarios. In general this is in line with the findings of this thesis. However, for increasing and constant hazard NPH scenarios with low median survival and high overall study duration the power of the *MWLRT* dropped below the power of the Fleming-Harrington test for late differences. Furthermore, the modestly weighted logrank test showed inflated type 1 error in some scenarios with increasing hazard. Another aspect touched by Klinglmüller et al. (2023) is treatment effect estimation where they investigated confidence interval coverage for three available summary measures (AHR, median difference, and RMST difference) which they found to be close to the nominal level. The disadvantage of many methods considered in this thesis is that either no statistical measure, that is equivalent to the test decision, is available to quantify the treatment effect or, if it is, it is not easily interpretable. This was not further investigated in this thesis, but, if possible, equivalent measures were given for the methods described in Section 2.3.

The final conclusion drawn in this thesis is a recommendation on the choice of the most powerful method based on different assumptions made at the planning stage of a trial. This was done by implementing a ranking system which ranks all methods with no type 1 error inflation with respect to their power within each scenario and averages over all scenarios corresponding to the knowledge at planning stage. It showed, that in PH scenarios the logrank test, together with the Cox and Royston-Parmar PH model tests performs best and all alternative methods reduced power in PH scenarios what is in line with the result by Peto and Peto (1972). In the unlikely case that the underlying data-generating process is known a priori, the use of GenLin type methods that exploit the full knowledge is advised. More interesting, however, is the case that it is known in advance that the treatment effect will be delayed but no further knowledge on when the delayed effect sets in is available. In this case the use of the modified versatile test by Lee (*mLee3*) is a robust choice for analysis.

4.2 Limitations and directions for future research

Although the simulation study presented in this thesis extends the number of methods and scenarios of recent simulation studies, it has its limitations which can give rise to future research. One limitation that is often present in analyzing time-to-event data is that the tests considered in this simulation study were used to test the null hypothesis of equal survival distributions versus the alternative hypothesis of unequal survival distributions. This, however, cannot be translated to uniform superiority of one group over the other. This again highlights that caution is needed to analyze and interpret time-to-event data and that careful inspection of the survival distributions is necessary.

Not only the final phase of evaluating and summarizing collected trial data needs careful consideration, but also during the design and planning phase of a trial different aspects should be decided thoughtfully. The aspect of sample size calculation has been touched in this thesis, but was not investigated more deeply. Sample size calculation in PH scenarios allowed to set a benchmark for the logrank test which facilitated the comparisons to other methods. For NPH scenarios sample size calculation is more difficult and in this thesis the approach based on the naive average hazard ratio was pursued. It has been observed that this works reasonably well in scenarios with increasing hazards and longer median survival. For other scenarios different approaches such as the average hazard ratio or approaches presented by Ananthakrishnan et al. (2021) or the flexible method proposed by Lakatos (1988) could be a field of further investigation. Another aspect for future research could be to explore ways to exploit adaptive or group-sequential design tools in context of delayed treatment effects. Some authors (e.g. Chen et al. (2022), Magirr and Jiménez (2022)) have already investigated application of group-sequential designs but these only scratch the surface of possibilities. Further topics of research could be whether an optimal statistical test for final analysis based on the results of an interim analysis could be chosen or, if the test for final analysis is already specified, if certain methods increase the chance of predicting the final test decision if applied at interim analysis.

4.3 Conclusion

A strength of this thesis is that it provides an extensive overview of proposed methods to deal with NPH when analyzing time-to-event data, which were compiled through a comprehensive literature search. Not all of these methods could be considered for the simulation study, as was already explained in Section 3.1, but references are given to revise the details of these methods. For the methods included in the simulation study, detailed information is given in this thesis, and the methods are presented in a well-structured way. If available the R package in which the method is implemented is referenced. All other methods have been self-implemented and the R code is provided in an online repository referenced in the appendix which makes these methods more easily accesible for future investigation. Another strength is the comprehensive simulation study which allows to assess various aspects of oncological trials with delayed treatment effects, which hopefully facilitates the choice of an appropriate statistical test.

To conclude, this thesis contains a comprehensive investigation of the performance of alternatives to the logrank test in NPH scenarios with delayed treatment effect. It provides a detailed and structured description of these methods together with details or references of their implementation. The results of the extensive simulation study can bee seen as a basis to choose the most appropriate method with respect to type 1 error control and power when planning future trials. Furthermore, the general structure and ideas of this thesis might also build a foundation for future investigations.

Chapter 5

Summary

This thesis is motivated by recent advances in oncology, where therapies rely on the activation and augmentation of the immune system to identify and fight tumor cells. These immunotherapeutic approaches come with own characteristics and one commonly observed trait is that the observable treatment effect is delayed by the time needed to train the response of the immune system. Especially for time-to-event endpoints such as overall survival or progression-free survival this demands careful consideration with respect to the statistical evaluation as commonly used methods rely on the assumption of proportional hazards. This assumption is violated if the treatment effect is delayed and the usual methods have reduced power to detect a difference between therapies. The aim of this thesis was hence to investigate the performance of various alternatives to the commonly used logrank test in terms of type 1 error and power in this setting.

Firstly, a systematic literature search was performed to identify statistical methods that have been suggested to analyze time-to-event data and especially if these methods have been developed to handle non-proportional hazards. The methods were then compared in an extensive simulation study taking the following parameters into account: the overall study duration τ, the accrual proportion acc, the shape of the Weibull hazard k_C and median survival med_C in the control arm and the maximum treatment effect θ as well as the delay t_2^* and changepoint t_1^* of the generalized linear lag model. As performance measure type 1 error and power of the methods was assessed and the effect of the different simulation parameters

on these performance measures was examined. For methods where parameters for analysis had to be chosen the effect of misspecifying these parameters was also investigated.

Most of the methods controlled type 1 error and those that did not were already known from previous publications. With respect to power the methods that come close to the data-generating process performed best in the simulation studies and were not sensitive to parameter misspecification at least in non-proportional hazard scenarios. However, the specification of these parameters requires knowledge of the form and extent of the delay that can be expected in advance. If this is not available other alternatives such as the Fleming-Harrington test for late differences or combinations of weighted logrank statistics can be used. Their power, however, depends on the underlying distribution of the data. For proportional hazard scenarios the logrank test is the most powerful of all tests that control type 1 error.

Application of the presented methods is facilitated by referencing the R package in which the method is implemented or providing own code in an online repository referenced in the Appendix. The interpretation of the results of these methods is, however, often complicated due to the absence of interpretable summary measures of the detected treatment effect. It is also pointed out that some aspects of this thesis such as sample size calculation or application of the methods in adaptive or group-sequential designs could be worth future research.

It is concluded that this thesis provides a detailed and well structured compilation of alternative methods to the logrank test when analyzing time-to-event data in the presence of non-proportional hazards. Furthermore, the extensive simulation study together with the ranking system can help to make a substantiated choice of an appropriate method for analysis of future studies. In case of proportional hazards the logrank test remains the method of choice and in case the data-generating process of the non-proportional hazard is completely understood methods should be chosen that exploit this additional knowledge. If, however, the exact structure of the delay is unknown versatile methods that combine weighted logrank statistics should be resorted to.

Chapter 6

Zusammenfassung

Die jüngsten Fortschritte in der Onkologie zu Therapien, die auf der Aktivierung und Verstärkung des Immunsystems Tumorzellen zu erkennen und zu bekämpfen beruhen haben diese Arbeit motiviert. Eine häufig beobachtete Eigenschaft dieser immuntherapeutischen Ansätze ist, dass der beobachtbare Behandlungseffekt um die benötigte Zeit eine Immunantwort zu trainieren verzögert wird. Insbesondere bei Ereigniszeitendpunkten wie dem Gesamtüberleben oder dem progressionsfreien Überleben muss dies bei der statistischen Auswertung sorgfältig berücksichtigt werden, da die üblicherweise verwendeten Methoden auf der Annahme der proportionalen Hazards beruhen. Diese Annahme ist verletzt, wenn der Behandlungseffekt verzögert eintritt und die herkömmlichen Methoden haben eine geringere statistische Power einen Unterschied zwischen den Therapien zu erkennen. Ziel dieser Arbeit war es daher, die Performance verschiedener Alternativen zum häufig verwendeten Logrank-Test in Bezug auf den Fehler 1. Art und die Power in dieser Situation zu untersuchen.

In einer systematischen Literaturrecherche wurden zuerst statistische Methoden identifiziert, die für die Analyse von Ereigniszeitdaten vorgeschlagen wurden, und ob diese insbesondere entwickelt wurden, um nicht-proportionale Hazards umzugehen. Die Methoden wurden dann in einer umfangreichen Simulationsstudie unter Berücksichtigung der folgenden Parameter verglichen: die Gesamtstudiendauer τ, der Anteil des Rekrutierungszeitraums acc, die Form des Weibull-Hazards k_C und das mediane Überleben med_C im Kontrollarm und der maximale Behandlungseffekt θ sowie die Verzögerung t_2^* und der Zeitpunkt der Änderung t_1^* des

generalisierten linearen Lag-Modells. Als Gütemaß wurde der Fehler 1. Art und die Power der Methoden und der Einfluss der verschiedenen Simulationsparameter auf diese Gütemaße untersucht. Bei Methoden, bei denen Analyseparameter gewählt werden mussten, wurde auch die Auswirkung einer falschen Spezifikation dieser Parameter untersucht.

Die meisten Methoden kontrollierten den Fehler 1. Art, und die dies nicht taten, waren bereits aus früheren Veröffentlichungen bekannt. In Bezug auf die Power schnitten dem datengenerierenden Prozess nahekommende Methoden in der Simulationsstudie am besten ab und waren zumindest in Szenarien mit nichtproportionalen Hazards nicht anfällig gegenüber einer falschen Spezifikation der Analyseparameter. Die Spezifikation dieser Parameter erfordert jedoch die Kenntnis von Form und Ausmaß der zu erwartenden Verzögerung im Voraus. Ist dies nicht der Fall, können andere Alternativen wie der Fleming-Harrington Test für späte Unterschiede oder Kombinationen von gewichteten Logrank-Statistiken verwendet werden. Ihre Power hängt jedoch von der zugrundeliegenden Verteilung der Daten ab. Für Szenarien mit proportionalen Hazards hat der Logrank-Test die größte Power unter allen Tests, die den Typ 1 Fehler kontrollieren.

Durch einen Verweis auf das R-Paket, in dem die Methode implementiert ist, oder durch die Bereitstellung eigenen Codes im Anhang wird die Verwendung der vorgestellten Methoden erleichtert. Die Interpretation der Ergebnisse dieser Methoden ist jedoch durch das Fehlen interpretierbarer Effektmaße oft schwierig. Einige Aspekte dieser Arbeit, wie z.B. die Berechnung des Stichprobenumfangs oder die Anwendung der Methoden in adaptiven oder gruppensequentiellen Designs könnten zukünftige Forschung wert sein.

Diese Arbeit liefert eine detaillierte und gut strukturierte Übersicht alternativer Analysemethoden zum Logrank-Test, wenn Ereigniszeitdaten mit nichtproportionalen Hazards vorliegen. Darüber hinaus kann die umfangreiche Simulationsstudie zusammen mit dem Platzierungssystem helfen geeignete Methode zur Analyse zukünftiger Studien fundiert auszuwählen. Im Falle proportionaler Hazards bleibt der Logrank-Test die Methode der Wahl, und sofern der Datenerzeugungsprozess bei nichtproportionalen Hazards vollständig verstanden ist, sollten Methoden gewählt werden, die dieses zusätzliche Wissen nutzen. Falls jedoch die genaue Struktur der Verzögerung nicht bekannt ist, sollte auf vielseitige Methoden zurückgegriffen werden, die gewichtete Logrank-Statistiken kombinieren.

Chapter 7

References list

Aalen, O. O. (1975). **Statistical inference for a family of counting processes**. Institute of Mathematical Statistics, University of Copenhagen, Copenhagen.

Ananthakrishnan, R., Green, S., Previtali, A., Liu, R., Li, D., and LaValley, M. (2021). **Critical review of oncology clinical trial design under non-proportional hazards**. Crit Rev Oncol Hematol, 162:103350–103350, doi: 10.1016/j.critrevonc.2021.103350.

Arfè, A., Alexander, B., and Trippa, L. (2021). **Optimality of testing procedures for survival data in the nonproportional hazards setting**. Biometrics, 77(2):587–598, doi: 10.1111/biom.13315.

Bagdonavicius, V. B., Levuliene, R. J., Nikulin, M. S., and Zdorova-Cheminade, O. (2004). **Tests for equality of survival distributions against non-location alternatives**. Lifetime Data Anal, 10(4):445–460, doi: 10.1007/s10985-004-4777-7.

Behnisch, R. (2023). **Performance of time-to-event methods in delayed treatment effect scenarios - a simulation plan**.

Brendel, M., Janssen, A., Mayer, C.-D., and Pauly, M. (2014). **Weighted logrank permutation tests for randomly right censored life science data**. Scand Stat Theory Appl, 41(3):742–761, doi: 10.1111/sjos.12059.

Breslow, N. (1970). **A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship**. Biometrika, 57(3):579–594, doi: 10.1093/biomet/57.3.579.

Breslow, N. (1974). **Covariance analysis of censored survival data**. Biometrics, 30(1):89–99, doi: 10.2307/2529620.

Broström, G. (2020). **eha: Event history analysis**. R package version 2.8.3, `https://cran.r-project.org/package=eha`.

Buckley, J. and James, I. (1979). **Linear regression with censored data**. Biometrika, 66(3):429–436, doi: 10.1093/biomet/66.3.429.

Callegaro, A. and Spiessens, B. (2017). **Testing treatment effect in randomized clinical trials with possible nonproportional hazards**. Stat Biopharm Res, 9(2):204–211, doi: 10.1080/19466315.2016.1257436.

Campbell, H. and Dean, C. (2014). **The consequences of proportional hazards based model selection**. Stat Med, 33(6):1042–1056, doi: 10.1002/sim.6021.

Chauvel, C. and O'Quigley, J. (2014). **Tests for comparing estimated survival functions**. Biometrika, 101(3):535–552, doi: 10.1093/biomet/asu015.

Chen, Y. M., Lawrence, J., and Lee, M. L. T. (2022). **Group sequential design for randomized trials using "first hitting time" model**. Stat Med, 41(13):2375–2402, doi: 10.1002/sim.9360.

Cheng, H. and He, J. (2021). **A maximum weighted logrank test in detecting crossing hazards**.

Cox, D. R. (1972). **Regression models and life-tables**. J R Stat Soc Series B Stat Methodol, 34(2):187–220, doi: 10.1111/j.2517-6161.1972.tb00899.x.

Cox, D. R. (1975). **Partial likelihood**. Biometrika, 62(2):269–276, doi: 10.1093/biomet/62.2.269.

Ding, X. and Wu, J. (2020). **Designing cancer immunotherapy trials with delayed treatment effect using maximin efficiency robust statistics**. Pharm Stat, 19(4):424–435, doi: 10.1002/pst.2003.

Dunkler, D., Ploner, M., Schemper, M., and Heinze, G. (2018). **Weighted cox regression using the R package coxphw**. J Stat Softw, 84(2):1–26, doi: 10.18637/jss.v084.i02.

Flandre, P. and O'Quigley, J. (2019). **Comparing kaplan-meier curves with delayed treatment effects: applications in immunotherapy trials**. J R Stat Soc Ser C Appl Stat, 68(4):915–939, doi: 10.1111/rssc.12345.

Fleming, T. R. and Harrington, D. P. (1981). **A class of hypothesis tests for one and two sample censored survival data**. Commun Stat Theory Methods, 10(8):763–794, doi: 10.1080/03610928108828073.

Fleming, T. R. and Harrington, D. P. (1991). **Counting processes and survival analysis**. Wiley series in probability and mathematical statistics : Applied probability and statistics section. Wiley, New York, Chicester, Weinheim. Literaturverz. S. 401 - 412.

Fleming, T. R., Harrington, D. P., and O'sullivan, M. (1987). **Supremum versions of the log-rank and generalized wilcoxon statistics**. J Am Stat Assoc, 82(397):312–320, doi: 10.1080/01621459.1987.10478435.

Gehan, E. A. (1965). **A generalized wilcoxon test for comparing arbitrarily singly-censored samples**. Biometrika, 52(1/2):203–223, doi: 10.1093/biomet/52.1-2.203.

Grambsch, P. M. and Therneau, T. M. (1994). **Proportional hazards tests and diagnostics based on weighted residuals**. Biometrika, 81(3):515–526, doi: 10.1093/biomet/81.3.515.

Gray, R. J. and Tsiatis, A. A. (1989). **A linear rank test for use when the main interest is in differences in cure rates**. Biometrics, 45(3):899–904, doi: 10.2307/2531691.

Greenwood, M. (1926). **A report on the natural duration of cancer**. Reports on public health and medical subjects. H.M. Stationery Office, London.

He, P. and Su, Z. (2015). **A novel design for randomized immuno-oncology clinical trials with potentially delayed treatment effects**. Contemp Clin Trials Commun, 1:28–31, doi: 10.1016/j.conctc.2015.08.003.

Hodi, F. S., Chiarion-Sileni, V., Gonzalez, R., Grob, J.-J., Rutkowski, P., Cowey, C. L., Lao, C. D., Schadendorf, D., Wagstaff, J., Dummer, R., Ferrucci, P. F., Smylie, M., Hill, A., Hogg, D., Marquez-Rodas, I., Jiang, J., Rizzo, J., Larkin, J., and Wolchok, J. D. (2018). **Nivolumab plus ipilimumab or nivolumab alone versus ipilimumab alone in advanced melanoma (checkmate 067): 4-year outcomes of a multicentre, randomised, phase 3 trial**. Lancet Oncol, 19(11):1480–1492, doi: 10.1016/S1470-2045(18)30700-9.

Huang, B. and Kuan, P. (2018). **Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point**. Pharm Stat, 17(3):202–213, doi: 10.1002/pst.1846.

Jachno, K., Heritier, S., and Wolfe, R. (2021). **Impact of a non-constant baseline hazard on detection of time-dependent treatment effects: a simulation study**. BMC Med Res Methodol, 21(1):177, doi: 10.1186/s12874-021-01372-0.

Jackson, C. (2016). **flexsurv: A platform for parametric survival modeling in R**. J Stat Softw, 70(8):1–33, doi: 10.18637/jss.v070.i08.

Jiménez, J. L. (2022). **Quantifying treatment differences in confirmatory trials under non-proportional hazards**. J Appl Stat, 49(2):466–484, doi: 10.1080/02664763.2020.1815673.

Kalbfleisch, J. D. and Prentice, R. L. (1973). **Marginal likelihoods based on cox's regression and life model**. Biometrika, 60(2):267–278, doi: 10.1093/biomet/60.2.267.

Kaplan, E. L. and Meier, P. (1958). **Nonparametric estimation from incomplete observations**. J Am Stat Assoc, 53(282):457–481, doi: 10.1080/01621459.1958.10501452.

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

10.1177/1536867x1601600308.

Klein, J. P., Logan, B., Harhoff, M., and Andersen, P. K. (2007). **Analyzing survival curves at a fixed point in time**. Stat Med, 26(24):4505–4519, doi: 10.1002/sim.2864.

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

Kolonko, M. (2008). **Stochastische simulation**. Vieweg+Teubner, Wiesbaden, doi: doi:10.1007/978-3-8348-9290-4.

Koziol, J. A. (1978). **A two sample CRAMÉR-VON MISES test for randomly censored data**. Biom J, 20(6):603–608, doi: 10.1002/bimj.4710200608.

Lakatos, E. (1988). **Sample sizes based on the log-rank statistic in complex clinical trials**. Biometrics, 44(1):229–241, doi: 10.2307/2531910.

Lee, S.-H. (2007). **On the versatility of the combination of the weighted log-rank statistics**. Comput Stat Data Anal, 51(12):6557–6564, doi: 10.1016/j.csda.2007.03.006.

Letón, E. and Zuluaga, P. (2001). **Equivalence between score and weighted tests for survival curves**. Commun Stat Theory Methods, 30(4):591–608, doi: 10.1081/STA-100002138.

Lin, R. S. and León, L. F. (2017). **Estimation of treatment effects in weighted log-rank tests**. Contemp Clin Trials Commun, 8:147–155, doi: 10.1016/j.conctc.2017.09.004.

Lin, R. S., Lin, J., Roychoudhury, S., Anderson, K. M., Hu, T. L., Huang, B., Leon, L. F., Liao, J. J. Z., Liu, R., Luo, X. D., Mukhopadhyay, P., Qin, R., Tatsuoka, K., Wang, X. J., Wang, Y., Zhu, J., Chen, T. T., Iacona, R., Bhagavatheeswaran, P., Cong, J., Geraldes, M., Heinzmann, D., Huang, Y. F., Li, Z. R., Liu, H. L., Mai, Y. B., Qian, J. N., Xu, L. A., Ye, J. B., Zhao, L. P., and Cross-Pharma, N.-p. (2020). **Alternative analysis methods for time to event endpoints under non-proportional hazards: A comparative analysis**. Stat Biopharm Res, 12(2):187–198, doi: 10.1080/19466315.2019.1697738.

Lin, X. and Xu, Q. (2010). **A new method for the comparison of survival distributions**. Pharm Stat, 9(1):67–76, doi: 10.1002/pst.376.

Liu, T. T., Ditzhaus, M., and Xu, J. (2020). **A resampling-based test for two crossing survival curves**. Pharm Stat, 19(4):399–409, doi: 10.1002/pst.2000.

Magirr, D. (2021). **Non-proportional hazards in immuno-oncology: Is an old perspective needed?** Pharm Stat, 20(3):512–527, doi: 10.1002/pst.2091.

Magirr, D. (2022). **modestwlrt: Functions for implementing modestly-weighted logrank tests.** R package version 0.1.0.

Magirr, D. and Burman, C. F. (2019). **Modestly weighted logrank tests**. Stat Med, 38(20):3782–3790, doi: 10.1002/sim.8186.

Magirr, D. and Jiménez, J. L. (2022). **Design and analysis of group-sequential clinical trials based on a modestly weighted log-rank test in anticipation of a delayed separation of survival curves: A practical guidance**. Clin Trials, 19(2):201–210, doi: 10.1177/17407745211072848.

Miller, R. G. (1976). **Least squares regression with censored data**. Biometrika, 63(3):449–464, doi: 10.1093/biomet/63.3.449.

Moreau, T., Maccario, J., Lellouch, J., and Huber, C. (1992). **Weighted log rank statistics for comparing two distributions**. Biometrika, 79(1):195–198, doi: 10.1093/biomet/79.1.195.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). **Using simulation studies to evaluate statistical methods**. Stat Med, 38(11):2074–2102, doi: 10.1002/sim.8086.

Mukhopadhyay, P., Huang, W. M., Metcalfe, P., Ohrn, F., Jenner, M., and Stone, A. (2020). **Statistical and practical considerations in designing of immuno-oncology trials**. J Biopharm Stat, 30(6):1130–1146, doi: 10.1080/10543406.2020.1815035.

Nelson, W. (1969). **Hazard plotting for incomplete failure data**. Journal of Quality Technology, 1(1):27–52, doi: 10.1080/00224065.1969.11980344.

Pepe, M. S. and Fleming, T. R. (1989). **Weighted kaplan-meier statistics: A class of distance tests for censored survival data**. Biometrics, 45(2):497–507, doi: 10.2307/2531492.

Peto, R. and Peto, J. (1972). **Asymptotically efficient rank invariant test procedures**. J R Stat Soc Ser A, 135(2):185–207, doi: 10.2307/2344317.

Ristl, R., Ballarini, N. M., Götte, H., Schüler, A., Posch, M., and König, F. (2021). **Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze rcts in oncology**. Pharm Stat, 20(1):129–145, doi: 10.1002/pst.2062.

Roychoudhury, S., Anderson, K. M., Ye, J., and Mukhopadhyay, P. (2021). **Robust design and analysis of clinical trials with nonproportional hazards: A straw man guidance from a cross-pharma working group**. Stat Biopharm Res, doi: 10.1080/19466315.2021.1874507.

Royston, P. and Parmar, M. K. B. (2002). **Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects**. Stat Med, 21(15):2175–2197, doi: 10.1002/sim.1203.

Royston, P. and Parmar, M. K. B. (2014). **An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect**. Trials, 15:314, doi: 10.1186/1745-6215-15-314.

Royston, P. and Parmar, M. K. B. (2016). **Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated**. BMC Med Res Methodol, 16:16, doi: 10.1186/s12874-016-0110-x.

Royston, P. and Parmar, M. K. B. (2020). **A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome**. Trials, 21(1):315, doi: 10.1186/s13063-020-4153-2.

Scheike, T. H. and Martinussen, T. (2006). **Dynamic regression models for survival data**. Springer, NY.

Schemper, M. (1992). **Cox analysis of survival data with non-proportional hazard functions**. Statistician, 41(4):455–465, doi: 10.2307/2349009.

Schemper, M., Wakounig, S., and Heinze, G. (2009). **The estimation of average hazard ratios by weighted cox regression**. Stat Med, 28(19):2473–2489, doi: 10.1002/sim.3623.

Schoenfeld, D. (1982). **Partial residuals for the proportional hazards regression model**. Biometrika, 69(1):239–241, doi: 10.1093/biomet/69.1.239.

Sooriyarachchi, M. R. and Whitehead, J. (1998). **A method for sequential analysis of survival data with nonproportional hazards**. Biometrics, 54(3):1072–84, doi: 10.2307/2533858.

Sposto, R., Stablein, D., and Carter-Campbell, S. (1997). **A partially grouped logrank test**. Stat Med, 16(6):695–704, doi: 10.1002/(sici)1097-0258(19970330)16:6<695::Aid-sim436>3.3.Co;2-3.

Stablein, D. M. and Koutrouvelis, I. A. (1985). **A two-sample test sensitive to crossing hazards in uncensored and singly censored data**. Biometrics, 41(3):643–652, doi: 10.2307/2531284.

Tang, Y. Q. (2022). **Complex survival trial design by the product integration method**. Stat Med, 41(4):798–814, doi: 10.1002/sim.9256.

Tarone, R. E. and Ware, J. (1977). **On distribution-free tests for equality of survival distributions**. Biometrika, 64(1):156–160, doi: 10.1093/biomet/64.1.156.

Therneau, T. M. (2023). **survival: A package for survival analysis in r**. R package version 3.5.7, `https://CRAN.R-project.org/package=survival`.

Tsiatis, A. A. (1990). **Estimating regression parameters using linear rank tests for censored data**. Ann Stat, 18(1):354–372, doi: 10.1214/aos/1176347504.

Uno, H., Tian, L., Horiguchi, M., Cronin, A., Battioui, C., and Bell, J. (2022). **survrm2: Comparing restricted mean survival time**. R package version 1.0-4.

van Houwelingen, H. C. and Stijnen, T. (2020). **Cox regression model**. In: Handbook of survival analysis, Eds Klein, J., van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H., 1st edition, CRC Press, London, p. 5-25.

Wang, Y., Wu, H., Anderson, K., Roychoudhury, S., Hu, T., and Liu, H. (2017). **nphsim: Non proportional hazards sample size and simulation**. R package version 0.1.1.9000.

Wassmer, G. and Pahlke, F. (2023). **rpact: Confirmatory adaptive clinical trial design and analysis**. R package version 3.3.4, `https://CRAN.R-project.org/package=rpact`.

Wu, Y.-L., Lu, S., Cheng, Y., Zhou, C., Wang, J., Mok, T., Zhang, L., Tu, H.-Y., Wu, L., Feng, J., Zhang, Y., Luft, A. V., Zhou, J., Ma, Z., Lu, Y., Hu, C., Shi, Y., Baudelet, C., Cai, J., and Chang, J. (2019). **Nivolumab versus docetaxel in a predominantly chinese patient population with previously treated advanced nsclc: Checkmate 078 randomized phase 3 clinical trial**. J Thorac Oncol, 14(5):867–875, doi: 10.1016/j.jtho.2019.01.006.

Xu, Z. Z., Park, Y., Zhen, B. G., and Zhu, B. (2018). **Designing cancer immunotherapy trials with random treatment time-lag effect**. Stat Med, 37(30):4589–4609, doi: 10.1002/sim.7937.

Xu, Z. Z., Zhen, B. G., Park, Y., and Zhu, B. (2017). **Designing therapeutic cancer vaccine trials with delayed treatment effect**. Stat Med, 36(4):592–605, doi: 10.1002/sim.7157.

Yang, M., Hua, Z. W., Xue, L., and Hu, M. X. (2020). **Z(max) test for delayed effect in immuno-oncology clinical trials**. J Biopharm Stat, 30(2):244–266, doi: 10.1080/10543406.2019.1632873.

Yang, S. and Prentice, R. (2005). **Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data**. Biometrika, 92(1):1–17, doi: 10.1093/biomet/92.1.1.

Yang, S. and Prentice, R. (2010). **Improved logrank-type tests for survival data using adaptive weights**. Biometrics, 66(1):30–8, doi: 10.1111/j.1541-0420.2009.01243.x.

Yang, S. and Prentice, R. L. (2011). **Estimation of the 2-sample hazard ratio function using a semiparametric model**. Biostatistics, 12(2), doi: 10.1093/biostatistics/kxq061.

Yang, S. and Zhao, Y. (2007). **Testing treatment effect by combining weighted log-rank tests and using empirical likelihood**. Stat Probab Lett, 77(12):1385–1393, doi: 10.1016/j.spl.2007.03.025.

Ye, T. and Yu, M. (2018). **A robust approach to sample size calculation in cancer immunotherapy trials with delayed treatment effect**. Biometrics, 74(4):1292–1300, doi: 10.1111/biom.12916.

Yu, C., Huang, X., Hui, N. A., and He, P. L. (2021). **A weighted log-rank test and associated effect estimator for cancer trials with delayed treatment effect**. Pharm Stat, 20(3):528–550, doi: 10.1002/pst.2092.

Zhang, H. (2022). **Cauchycp: Powerful test for survival data under non-proportional hazards**. R package version 0.1.1.

Zhang, H., Li, Q., Mehrotra, D. V., and Shen, J. (2021). **Cauchycp: A powerful test under non-proportional hazards using cauchy combination of change-point cox regressions**. Stat Methods Med Res, 30(11):2447–2458, doi: 10.1177/09622802211037076.

Zhang, Y. and Zhang, Z. (2020). **The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications**. Cell Mol Immunol, 17(8):807–821, doi: 10.1038/s41423-020-0488-6.

Zucker, D. M. and Lakatos, E. (1990). **Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment**. Biometrika, 77(4):853–864, doi: 10.1093/biomet/77.4.853.

Appendix

Implementations in R

To ensure reproducibility and make the implemented methods easily accessible R program code that was used in this thesis was uploaded to an online repository on github and can be accessed at `https://github.com/BeRouven/Methods-T2E-DelayedTreatment`. The repository contains the following R-Code-Scripts:

- Costume functions: The first script "0_CostumeFunctions_DataGeneration.R" contains the implementation of the inverse cumulative hazard function Λ_E^{-1} needed to generate the survival times by applying the inversion method as outlined in Section 2.2.3. The second script "0_CostumeFunctions_Analysis.R" contains the function "weightedLR" to calculate all weighted logrank statistics and, if needed, their correlation for combinations of these statistics as described in 2.3.3. The function "MethodsSurvival" is used to apply these methods together with methods from publicly available R packages.
- Data Generation: These scripts conduct the data generation for the type 1 error ("1_DataGeneration_T1E.R") and power ("1_DataGeneration_Power.R") scenarios.
- Data Analysis: The scripts "2_DataAnalysis_T1E.R" and "2_DataAnalysis_Power.R" conduct the data analysis of the generated data for all scenarios. As this calculation is computationally very expensive, the analysis was parallelized and multiple cores were used to reduce runtime.

To reproduce the results from this thesis the data need to be generated first which for type 1 error assessment can be done by simply running the "1_DataGeneration_T1E.R" script. To generate the data for the power assessment the functions defined in the script "0_CostumeFunctions_DataGeneration.R" need to be loaded first and then the "1_DataGeneration_Power.R" script can be run. To analyze the data the analysis function from the "0_CostumeFunctions_Analysis.R" script need to be loaded before the "2_DataAnalysis_T1E.R" and "2_DataAnalysis_Power.R" scripts can be run.

Acknowledgments

First and foremost I would like to thank my supervisor Prof. Dr. Meinhard Kieser for giving me the opportunity to write this thesis at the Institute of Medical Biometry. I am very grateful that he gave me the chance to contribute to this relevant topic, for his constructive suggestions and his continued support that made a succesful completion of this thesis possible.

In addition I thank my current and former colleagues at the Institute of Medical Biometry who accompanied me on this journey and helped me grow so much. They all created such a friendly and supportive atmosphere, and always gave me moral support and valuable feedback. Special thanks goes to Dr. Marietta Kirchner for all the time she invested to read and discuss my work and for sharing her experience with me and to Christopher Büsch for all the joint in-depth discussions and constructive critique.

I also would like to thank my family, my family-in-law and my friends for their support during the last seven years. A final and special thanks goes to my husband Christian, who always believed in me, endured my bad mood when things did not work as planned and also motivated me when it got difficult. Without all your encouragement and support I would not have come to this point and I am deeply grateful and thank you all very much.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema „Comparison of methods to analyze time-to-event endpoints when treatment effect is delayed“handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift