

MAIN PAPER

Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology

Robin Ristl¹  | Nicolás M Ballarini¹  | Heiko Götte²  | Armin Schüller²  |
Martin Posch¹  | Franz König¹ 

¹Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria

²Merck Healthcare KGaA, Darmstadt, Germany

Correspondence

Franz König, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria.

Email: franz.koenig@meduniwien.ac.at

Funding information

Merck Healthcare KGaA

SUMMARY

In the analysis of survival times, the logrank test and the Cox model have been established as key tools, which do not require specific distributional assumptions. Under the assumption of proportional hazards, they are efficient and their results can be interpreted unambiguously. However, delayed treatment effects, disease progression, treatment switchers or the presence of subgroups with differential treatment effects may challenge the assumption of proportional hazards. In practice, weighted logrank tests emphasizing either early, intermediate or late event times via an appropriate weighting function may be used to accommodate for an expected pattern of non-proportionality. We model these sources of non-proportional hazards via a mixture of survival functions with piecewise constant hazard. The model is then applied to study the power of unweighted and weighted log-rank tests, as well as maximum tests allowing different time dependent weights. Simulation results suggest a robust performance of maximum tests across different scenarios, with little loss in power compared to the most powerful among the considered weighting schemes and huge power gain compared to unfavorable weights. The actual sources of non-proportional hazards are not obvious from resulting populationwise survival functions, highlighting the importance of detailed simulations in the planning phase of a trial when assuming non-proportional hazards. We provide the required tools in a software package, allowing to model data generating processes under complex non-proportional hazard scenarios, to simulate data from these models and to perform the weighted logrank tests.

1 | INTRODUCTION

Comparing survival distributions based on censored data is in general challenging since conclusions are sought about distributions which are observed incompletely. In medical statistics, and in particular in oncology, the logrank test and the Cox

R.R. and N.B. received funding from Merck Healthcare KGaA Healthcare KGaA connected to this work.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Pharmaceutical Statistics* published by John Wiley & Sons Ltd.

proportional hazards model have been established as key tools. Under the assumption of proportional hazards, they are efficient and their results can be interpreted unambiguously. The proportional hazards assumption implies that the benefit gained through active treatment over control is the same at each time-point. Even though this assumption may be reasonable in a homogeneous population and over a limited time span, it is increasingly challenged in more complex settings.¹

Here we investigate four different sources of non-proportional hazards which are frequently encountered in oncology trials with survival endpoints. We address (1) the effect of delayed onset of treatment action, which has been described particularly for the new class of immuno-oncology drugs.² We take into account (2) the possibility of reduced survival probabilities after a disease progression event. This is a source of non-proportional hazards if, for example, the active treatment prolongs the time to progression in addition to reducing the hazard rate in each disease state. We further study (3) the impact of differential effect of the treatment in different subgroups. This is of particular interest if stronger effects are expected in patients who are positive for a certain biomarker. For example, in metastatic colorectal cancer, targeted therapy inhibiting the epidermal growth factor receptor is effective only against tumors that are free from mutations in the KRAS or NRAS genes.³ Subgroup effects have also been described in settings where a fraction of patients is permanently cured (see for example, the examples in Reference 4).

Finally (4), in an actual trial patients may at some time-point receive medication that is different from the planned medication in their trial arm, for various reasons. One focus of the paper is on treatment switching where patients from the control group may receive the potentially active treatment after they have experienced disease progression, or in an alternative scenario, patients from either group may receive an effective follow-up medication after disease progression. For example, in a randomized controlled trial in metastatic non-small-cell lung cancer, 33% of patients in the control group received the treatment group medication after disease progression.⁵ In the standard intention-to-treat analysis non-proportional hazards arise since the observed treatment effect is reduced with time as a consequence of treatment switching. Alternative methods to analyze data in presence of treatment switching been studied extensively. In particular, accelerated failure time models accounting for the time under control and active treatment have been proposed.^{6,7} Our focus is on investigating the effect of switching in an intention-to-treat analysis.

When comparing two hazard functions which are not proportional, the power of the logrank test can be restored by weighting observed events. Optimal weights would correspond to the actual event time specific log hazard-ratio,⁸ which is unknown. In practice, weighted logrank tests emphasizing either early, intermediate or late event times via an appropriate weighting function may be used to accommodate for an expected pattern of non-proportionality.

In this manuscript, we focus on the Fleming-Harrington $\rho - \gamma$ family of weighted logrank tests.⁹ Since the choice of a powerful weighting function is subject to prior assumptions, we further consider the properties of maximum-type tests⁹⁻¹¹ that combine a set of differently weighted logrank test.

To study the operating characteristics of these tests under the aforementioned sources of non-proportional hazards, we propose a data generating model based on mixtures of survival distributions that are defined via piecewise constant hazards. Similar methods have been applied in the literature to study the sole impact of delayed onset,¹² of a subpopulation of cured patients^{4,13} and the combined effect of delayed onset, disease progression and treatment switching, though not subgroups of patients.¹⁴ However, subgroups with different response characteristics are an important element to be considered in trials for targeted therapies such as immuno-oncology drugs. Even in studies with restrictive inclusion criteria, the study population may consist of previously unknown subpopulations.

Furthermore, the recruitment scheme and trial duration have an impact on the distribution of observed event times. Under non-proportional hazards, the power of hypothesis tests is also depending on this distribution, see for example, Reference 13. Consequently, we also study the influence of different recruitment rates under the considered non-proportional hazards scenarios.

In the manuscript, we aim to provide guidance on the methods for the primary analysis under differently complex non-proportional hazard settings. To this end we further provide software as an R package labeled `nph`¹⁵ which allows one to model data generating processes under complex non-proportional hazard scenarios, to simulate data from these models and to analyze datasets using the weighted logrank tests we studied.

The remainder of the manuscript is structured as follows. In Section 2 the proposed data generating model is defined. In Section 3 the weighted logrank tests are described. In Section 4 we present a case study based on an actual trial.⁵ In Section 5 the power of weighted logrank and maximum tests depending on different sources of non-proportional hazards and their extents is investigated systematically by simulation. Details on R package `nph`¹⁵ providing functions to simulate and analyze data under non-proportional hazards are given Section 6. We conclude with a discussion in Section 7. Additional information on the simulation scenarios and results for further simulation scenarios are provided in the online supplemental material.

2 | MODELING NON-PROPORTIONAL HAZARDS

Let subscripts $i \in \{ctr, trt\}$ denote the control and treatment group in a randomized controlled clinical trial with survival as main endpoint. For a patient in group i , let the survival time $T \in [0, \infty)$ be a continuous non-negative random variable with distribution function $F_i(t)$, survival function $S_i(t) = 1 - F_i(t)$ and density $f_i(t)$. The hazard function is defined as $\lambda_i(t) = \lim_{h \downarrow 0} P(T \in (t, t + h] | T > t)/h = f_i(t)/S_i(t)$. The cumulative hazard is defined as $\Lambda_i(t) = \int_0^t \lambda_i(s) ds$. The survival function can be expressed as $S_i(t) = \exp(-\Lambda_i(t))$.

For traditional study planning, the survival functions are often modeled via constant hazards over time, resulting in exponentially distributed survival times. Such a model may fail to address the complexity of actual trial data and in particular, does not cover non-proportional hazards. In this section, we propose a model that is still simple and easy to apply but allows for increasingly complex hazard functions and in particular addresses several sources of non-proportional hazards. Figure 1 illustrates the different states and rates assumed in this model.

2.1 | Piecewise constant hazards and delayed onset of treatment effect

The hazard rates $\lambda_i(t)$, $i \in \{trt, ctr\}$ may be modeled as piecewise constant functions to provide an intuitively simple model with non-proportional hazards that still allows for complexity when needed.

Define k time intervals $[t_{j-1}, t_j)$, $j = 1, \dots, k$ with $0 = t_0 < t_1 < t_k = \infty$ and constant hazards λ_{ij} . The resulting hazard function in group i is $\lambda_i(t) = \sum_{j=1}^k \lambda_{ij} 1_{t \in [t_{j-1}, t_j)}$, the cumulative hazard function is $\Lambda_i(t) = \int_0^t \lambda_i(s) ds = \sum_{j=1}^k ((t_j - t_{j-1}) \lambda_{ij} 1_{t > t_j} + (t - t_{j-1}) \lambda_{ij} 1_{t \in [t_{j-1}, t_j)})$ and we obtain the survival function as $S_i(t) = e^{-\Lambda_i(t)}$.

In particular, we apply the piecewise constant hazard approach to model the effect of delayed onset of treatment action. Assume that the treatment has an effect on the hazard rate only after a certain time span t_{onset} from initiation of the treatment. The hazard rate in the treatment group is modeled as $\lambda_{trt}(t) = \lambda_{preonset} 1_{t < t_{onset}} + \lambda_{postonset} 1_{t \geq t_{onset}}$ whereas the hazard rate in the control group is constantly $\lambda_{ctr}(t) = \lambda_{preonset}$. The resulting survival curves are identical until time t_{onset} and are separating afterward (assuming $\lambda_{postonset}$ is different from $\lambda_{preonset}$).

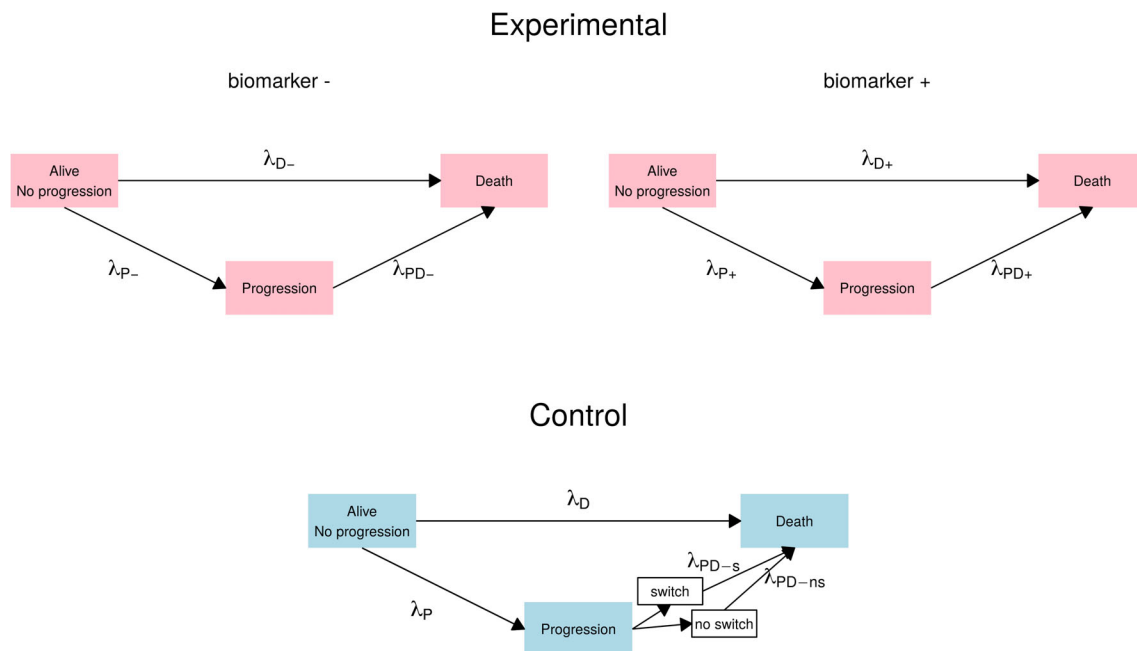


FIGURE 1 Multi-state representation of the modeled sources of non-proportional hazards. Different states are represented by boxes, transitions are represented by arrows. The transition rates are indicated next to the respective arrows. All transition rates may be piecewise constant functions of time. For simplicity of representation, biomarker dependent subgroups are shown only in the experimental group and treatments switching is only indicated in the control group. However the proposed framework allows for both, subgroups and treatment switching, in each group

A similar model has been used by Hasegawa¹² to provide a sample size formula for weighted logrank tests under the scenario of a delayed treatment effect. While we focus on models used for study planning, piecewise constant hazard models may also be estimated given observed data, see for example, References 16,17.

2.2 | Changing hazards after disease progression

The piecewise constant hazard functions defined above allow for changing hazards at defined time-points. However, the hazard rate may also change at random time-points, in particular after progression of disease (PD). To model this aspect, we assume that the time to disease progression T_{PD} is governed by a separate process with hazard function $\eta_i(t)$, $i \in \{trt, ctr\}$, which does not depend on the hazard function for death $\lambda_i(t)$. $\eta_i(t)$, too, may be modeled as piecewise constant or, for simplicity, as constant over time. Define $\lambda_{i,prePD}(t)$ and $\lambda_{i,postPD}(t)$ as the hazard functions for death before and after disease progression. Conditional on $T_{PD} = s$, the hazard function for death in group i is $\lambda_i(t | T_{PD} = s) = \lambda_{i,prePD}(t)I_{t \leq s} + \lambda_{i,postPD}(t)I_{t > s}$ and the conditional survival function is $S_i(t | T_{PD} = s) = \exp\left(-\int_0^t \lambda_i(t | T_{PD} = s) ds\right)$. The unconditional survival function results from integration over all possible progression times as $S_i(t) = \int_0^t S_i(ts) dP(T_{PD} = s)$. Note that, even though $\eta_i(t)$ does not depend on $\lambda_i(t)$, progression free survival and overall survival times will be correlated since progression free survival time is $\min(T, T_{PD})$.

2.3 | Biomarker subgroups

A further element in our model are subgroups in the patient population that may exhibit different hazard functions. Patients who are positive with respect to certain biomarkers may show a better response to treatment. For modeling differential subgroup effects, we assume the number and relative sizes of subgroups are known, however in the analysis we regard the presence of subgroups as an unknown aspect. Thus, we are interested in the marginal survival functions over the full population. Given m subgroups with relative sizes p_1, \dots, p_m and subgroup-specific survival functions $S_{i,l}(t)$, the marginal survival function is the mixture $S_i(t) = \sum_{l=1}^m p_l S_{i,l}(t)$. Note that the respective hazard function is not a linear combination of the subgroup-specific hazard functions. It may be calculated by the general relation $\lambda_i(t) = -\frac{dS_i(t)}{dt} \frac{1}{S_i(t)}$.

Other authors have focused on the presence of a subgroup of long-term survivors or completely cured patients and their implications in study planning,^{4,13} sample size reassessment¹⁸ and data analysis.¹⁹

2.4 | Treatment switching after progression

The final aspect in our model is the possibility that the study medication is changed, for example, to a further line treatment, due to a disease progression event. Such treatment switching may occur with a certain probability only. We address the probabilistic aspect by defining two subpopulations, nested within any subpopulation as defined previously, of patients that will switch medication after disease progression and patients that will not switch with survival functions $S_{l,switch}(t)$ and $S_{l,noswitch}(t)$. Denote the relative sizes of these two subpopulations as $p_{l,switch}$ and $p_{l,noswitch}$, such that $p_{l,switch} + p_{l,noswitch} = p_l$ and $p_{l,switch}/p_l$ corresponds to the probability for switching in subgroup l . The marginal survival function, covering subpopulations and switching, is $S_i(t) = \sum_{l=1}^m \sum_{s \in \{switch, noswitch\}} p_{l,s} S_{i,l,s}(t)$. This approach is a specific application of a scenario with changing hazards after disease progression as considered in Section 2.2.

2.5 | Sampling from the modeled survival function

Once the marginal survival functions $S_i(t)$, $i \in \{ctr, trt\}$ are derived, we may draw random samples via the inverse c.d.f. method. That is, draw a uniform random number $U \sim Unif(0, 1)$ and calculate the random survival time as $S_i^{-1}(U)$. In the numeric calculations we regard t as discrete with smallest increments of 1 day, which facilitates the computation of the inverse function.

Note that sampling n observations from the overall marginal distribution of the population is equivalent to first sampling the subgroup status of n patients and then sampling the conditional survival times given the subgroups.

3 | HYPOTHESIS TESTING

We consider the setting of randomized clinical trials comparing an experimental treatment to a control treatment. The aim of such trials typically is to show superiority of the experimental treatment over control. Under a proportional hazards assumption the treatment effect is conveniently expressed as hazard ratio which facilitates formulating an appropriate null hypothesis and deciding on a testing procedure. Typically, to avoid any further assumptions on the underlying distribution of survival times, the logrank test or, equivalently, a test based on the Cox proportional hazard model is applied, which is an optimal strategy in this context.²⁰

Under non-proportional hazards (in combination with censored data) deciding on a relevant parametrisation of the testing problem is less straight forward. In this paper we focus on the general one-sided null hypothesis of (weighted) logrank tests, which is $H_0 : \lambda_{ctr}(t) \leq \lambda_{trt}(t), \forall t \geq 0$. Rejecting H_0 means there is a treatment benefit at least in some time interval. See Section 7 for further discussion on this approach and on alternative effect measures.

3.1 | (Weighted) logrank tests

The logrank test is frequently applied to test the one-sided null hypothesis $H_0 : \lambda_{ctr}(t) \leq \lambda_{trt}(t), \forall t \geq 0$. Note that H_0 implies $S_{ctr}(t) \geq S_{trt}(t), \forall t \geq 0$. The reverse implication $S_{ctr}(t) \geq S_{trt}(t), \forall t \geq 0 \Rightarrow \lambda_{ctr}(t) \leq \lambda_{trt}(t), \forall t \geq 0$ is true under proportional hazards but does not generally hold under non-proportional hazards (see⁸).

For a given sample, let \mathcal{D} be the set of unique event times. For a time-point $t \in \mathcal{D}$, let $n_{t,ctr}$ and $n_{t,trt}$ be the number of patients at risk in the control and treatment group and let $d_{t,ctr}$ and $d_{t,trt}$ be the respective number of events. The expected number of events in the control group is calculated under the least favorable configuration in H_0 , $\lambda_{ctr}(t) = \lambda_{trt}(t)$, as $e_{t,ctr} = (d_{t,ctr} + d_{t,trt}) \frac{n_{t0}}{n_{t0} + n_{t1}}$. The conditional variance of $d_{t,ctr}$ is calculated from a hypergeometric distribution as $var(d_{t,ctr}) = \frac{n_{t0}n_{t1}(d_{t0} + d_{t1})(n_{t0} + n_{t1} - d_{t0} - d_{t1})}{(n_{t0} + n_{t1})^2(n_{t0} + n_{t1} - 1)}$. Further define a weighting function $w(t)$. The weighted logrank test statistic for a comparison of two groups is

$$z = \sum_{t \in \mathcal{D}} w(t)(d_{t,ctr} - e_{t,ctr}) / \sqrt{\sum_{t \in \mathcal{D}} w(t)^2 var(d_{t,ctr})}$$

Under the least favorable configuration in H_0 , the test statistic is asymptotically standard normally distributed and large values of z are in favor of the alternative.

Under proportional hazards and for local alternatives (ie, the hazard ratio approaching 1), the unweighted logrank test is the most powerful rank-invariant test for H_0 .²⁰ With arbitrary hazard functions, optimal weights for testing H_0 would correspond to the true log-hazard ratio $w(t) = \log(\lambda_{ctr}(t)/\lambda_{trt}(t))$ when $\lambda_{ctr}(t) > \lambda_{trt}(t)$ and $w(t) = 0$ otherwise.⁸ The true hazard function is of course unknown. However, by defining particular weight functions $w(t)$ the effect size at early or late event times may receive greater influence on the test statistic.

In this paper we consider particular weights in the Fleming-Harrington $\rho - \gamma$ family⁹ $w(t) = \hat{S}(t)^\rho (1 - \hat{S}(t))^\gamma$. Here, $\hat{S}(t) = \prod_{s \in \mathcal{D}: s \leq t} 1 - \frac{d_{t,ctr} + d_{t,trt}}{n_{t,ctr} + n_{t,trt}}$ is the pooled sample Kaplan-Meier estimator. Weights $\rho = 0, \gamma = 0$ correspond to the standard logrank test with constant weights $w(t) = 1$. Choosing $\rho = 0, \gamma = 1$ puts more weight on late events, $\rho = 1, \gamma = 0$ puts more weight on early events and $\rho = 1, \gamma = 1$ puts most weight on events at intermediate time points.

3.2 | Maximum logrank test

Depending on the true unknown alternative, differently weighted logrank tests may differ substantially in power. In general hypothesis testing problems, instead of deciding a-priori for a single test, a viable strategy is to combine different tests for the same hypothesis via the mean or maximum of their individual test statistics.²¹ In many applied problems, including weighted logrank tests, the asymptotic joint distribution of test statistics is multivariate normal, which

facilitates the derivation of an approximate distribution of a combination statistic. In this context Tarone¹⁰ studied the distribution of the maximum of the unweighted logrank test statistic and the generalized Wilcoxon statistic which uses as weights the total number at risk at each event time. Lee¹¹ compared the operating characteristics of combination tests based on the maximum or an average of $\rho = 0, \gamma = 1$ and $\rho = 1, \gamma = 0$ weighted logrank test statistics. The simulation results of Lee suggest that the maximum was the more robust combination statistic in terms of preserving power across various scenarios. Li et al²² studied the performance of the same statistics and several other test statistics, which are based on integrated differences of estimates of the survival functions or cumulative hazard functions, under scenarios of crossing survival functions. The robustness of the maximum test was confirmed also in this setting.

Karrison²³ describes the implementation of a maximum-type logrank test based on different $\rho - \gamma$ weights for the software Stata. Recently, the idea to apply maximum-type combination tests to analyze survival data with non-proportional hazards is gaining increasing attention.²⁴

To perform a maximum-type combination test, a set of r different weight functions $w_1(t), \dots, w_r(t)$ is specified and the correspondingly weighted logrank statistics z_1, \dots, z_r are calculated. The maximum test statistic is $z_{max} = \max_{i=1, \dots, r} z_i$. If at least one of the selected weight functions results in high power, we may expect a large value of z_{max} . Under the least favorable configuration in H_0 , approximately $(Z_1, Z_r) \sim N_r(\mathbf{0}, \Sigma)$. The P -value of the maximum test, $P_{H_0}(Z_{max} > z_{max}) = 1 - P(Z_1 \leq z_{max}, Z_r \leq z_{max})$, is calculated based on this multivariate normal approximation via numeric integration.

This approach automatically corrects for multiple testing with different weights and does so efficiently since the correlation between the different tests is incorporated in Σ . For actual calculations, Σ is replaced by an estimate.

Note that $cov(w_i(t)d_{t,ctr}, w_j(t)d_{t,ctr}) = w_i(t)w_j(t)var(d_{t,ctr})$, at least approximately assuming weights are converging in probability to a non-random function. Thus the i, j th element of Σ is estimated as

$$\hat{cov}(Z_i, Z_j) = \frac{\sum_{t \in \mathcal{D}} w_i(t)w_j(t)var(d_{t,ctr})}{\sqrt{\sum_{t \in \mathcal{D}} w_i^2(t)var(d_{t,ctr}) \sum_{t \in \mathcal{D}} w_j^2(t)var(d_{t,ctr})}}$$

Since the individual tests are typically highly correlated, the required multiplicity adjustment can be moderate while there is a good chance to gain from the most powerful among the included weighted test.

4 | CASE STUDY

To investigate increasingly complex examples for a trial under non-proportional hazards we devised five case study scenarios. As example for an actual trial which shared several characteristics of these scenarios see Reference 5. For easier interpretation, piecewise constant hazard functions are described by the median event time corresponding to the respective hazard rate in each time interval, that is, the presented median event time for interval i equals $\log(2)/\lambda_i$.

4.1 | Scenario A

As reference, we assume a scenario with proportional hazards, with the hazard for death corresponding to median survival times of 20 months and 12 months for the treatment and control arm, respectively.

4.2 | Scenario B

For the treatment arm we assume a hazard for death before PD corresponding to a median survival time of 24 months, a hazard for death after PD corresponding to a median survival time of 16 months, and a hazard for PD corresponding to 12 months median time to progression. For the control arm, the respective hazards correspond to median times of 22, 7 and 7 months, for the hazard of death before and after PD and for the hazard of PD, respectively.

4.3 | Scenario C

In this scenario we assume a delayed effect. During the first 2 months, the treatment group has the same hazards as the control group. After the first 2 months, the hazards used in scenario B for the treatment group are used.

4.4 | Scenario D

We consider delayed effect and the subgroup of female patients (prev 50%) with an additional benefit. During the first 2 months, the treatment group has the same hazards as the control group (irrespective of subgroups). After 2 months, and for the subgroup of male patients, we keep the true underlying parameters as in Scenario B. For the subgroup of female patients, we assume no disease progression and a hazard for death corresponding to a median time of 24 months is used.

4.5 | Scenario E

Here we consider treatment switching after disease progression for patients under the control treatment. Building on the previous scenarios, now we assume that 1/3 of the patients in the control group switch to the treatment group after progression. Note that for these patients, we also account for female/male subgroups with a prevalence of 50% each.

4.6 | Scenario F

Finally, we study a scenario where treatment switching to an effective follow-up medication occurs in both groups after disease progression. Similar to scenarios A and B, we assume that before PD the hazard for death corresponds to a median of 20 and 12 months, in the treatment and control group respectively, and the rates for PD correspond to median times of 12 and 7 months. After progression, patients in either group are switched to an effective follow-up medication with 75% probability, and the subsequent hazard for death corresponds to 18 months. For patients who are not switched we assume subsequent median survival times of 12 and 7 months.

We present the theoretical survival, hazard and hazard ratio functions over time in Figure 2. An interesting aspect is that the survival curves are similar across the scenarios, even if we compare the case where the proportionality assumption holds with the cases where it does not. In an actual trial, the investigator will only examine a single realization from the theoretical survival function and, therefore, it may not be easy to identify non-proportionality from the observed curves.

Table 1 shows the power for the weighted logrank test and the maximum test across the four scenarios under design assumptions similar to Reference 5. We assumed that patients are recruited over 1 year with a constant rate of 616 patients per year and are randomized between the treatment and control arm with a 2:1 allocation ratio.

The study ends when 190 events have been observed. This value corresponds to 80% of the number of events observed in Reference 5 and was chosen such that the power of the considered tests in our case study is approximately between 80% and 90%.

Censoring was included only in terms of administrative censoring at the end of the study. The one-sided level of significance was 0.025. The power was calculated from 50 000 simulation runs. As expected, the usual logrank test ($\rho = \gamma = 0$) outperforms the others in Scenario A since the proportional hazards assumption holds. However, in Scenarios C, D and E, using the weighted logrank test with stronger weights on late events provides higher power and in Scenario F stronger weights on early events provides higher power. In this sense, the maximum logrank test performs the best, since its power is among the highest in all cases.

5 | SIMULATION STUDIES

In this section we evaluate the effect of the different sources of non-proportional hazards on the power of unweighted logrank tests, the $\rho - \gamma$ weighted logrank tests with $(\rho, \gamma) \in \{(0, 1), (1, 1), (1, 0)\}$, and two maximum-type tests. The first

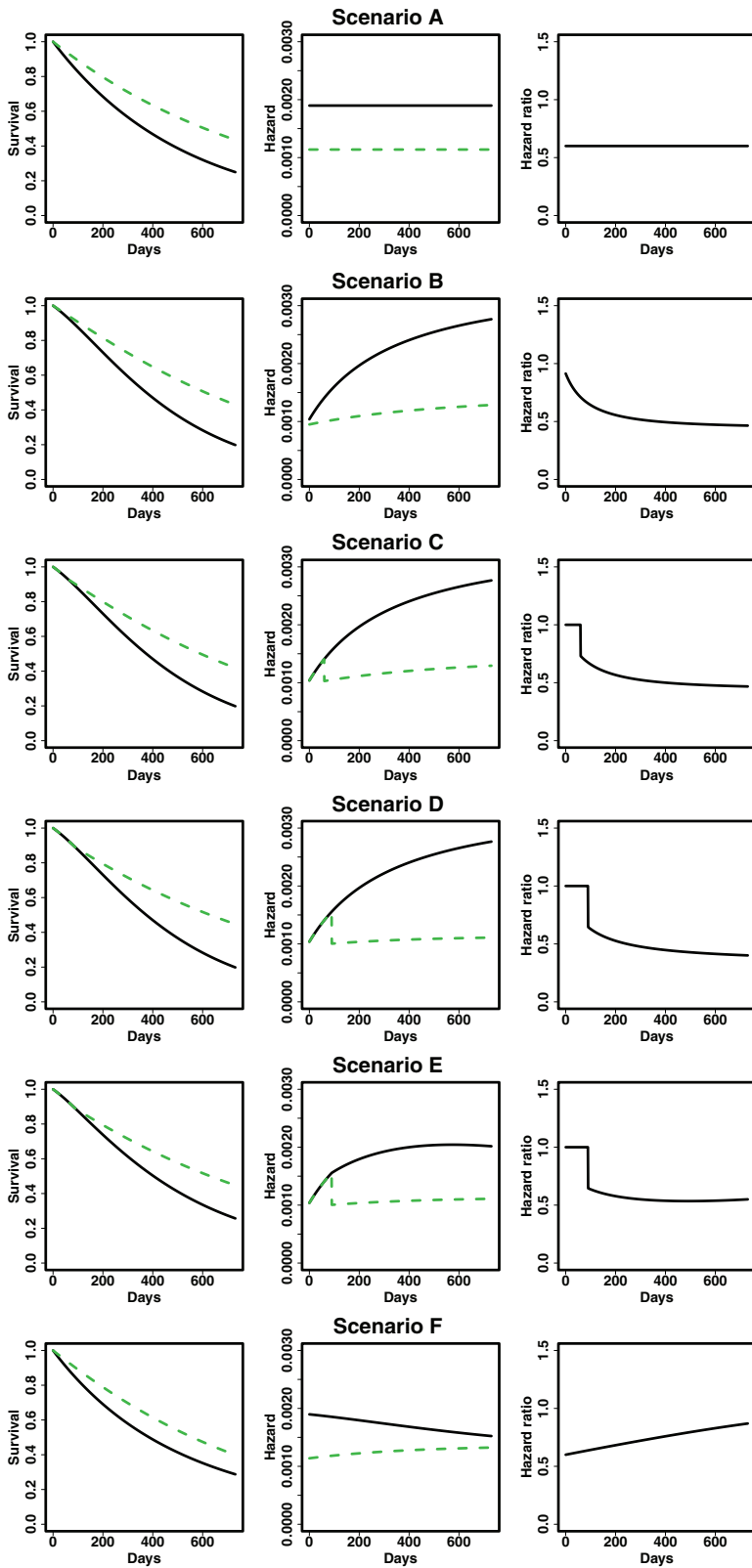


FIGURE 2 Survival function, hazard function and hazard ratio for treatment (dash green line) and control (solid black line) arms as described in Scenarios A-F

maximum test combines all four logrank tests, the second one combines the unweighted logrank test and the test with weights $\rho = 0, \gamma = 1$, which puts more weight in late events and should be beneficial under delayed onset.

We perform simulations for three study designs corresponding to slow, intermediate and fast recruitment. In all simulation scenarios patients are randomly assigned to treatment or control with equal probabilities. The recruitment periods are 2 years, 1 year and half a year for the scenarios of slow, intermediate and fast recruitment, with constant

TABLE 1 Power of weighted logrank tests in the studied scenarios for weighted logrank tests with $(\rho, \gamma) \in \{(0, 0), (0, 1), (1, 1), (1, 0)\}$, and a maximum test combining all four logrank tests (denoted by Max log-rank 4). The weighted logrank tests with $(\rho = 0, \gamma = 0)$ corresponds to the conventional logrank test

| Test | ρ | γ | Scenario | | | | | |
|------------------|--------|----------|----------|------|------|------|------|------|
| | | | A | B | C | D | E | F |
| Max logrank 4 | | | 91.8 | 92.4 | 86.1 | 89.3 | 77.1 | 78.1 |
| Weighted logrank | 0 | 0 | 93.0 | 90.9 | 78.8 | 78.1 | 65.3 | 80.9 |
| Weighted logrank | 0 | 1 | 82.4 | 91.1 | 87.3 | 91.3 | 79.8 | 57.5 |
| Weighted logrank | 1 | 1 | 85.9 | 92.7 | 88.3 | 91.4 | 80.6 | 63.8 |
| Weighted logrank | 1 | 0 | 92.6 | 87.8 | 71.8 | 69.3 | 57.0 | 81.7 |

recruitment rates of 100, 300 and 800 patients per year, respectively. In the simulations, trials stop for data analysis when a total number of 130 events is reached. (Under proportional hazards this number of events would correspond to approximately 80% power of the logrank test with a hazard ratio of 0.6 at one-sided level of significance of 0.025.) To focus on the effects of non-proportional hazards we did not consider random censoring due to drop outs in the simulation. However, the R package `nph`¹⁵ we provide does allow for additional random censoring.

Data are simulated from survival functions that were defined using the methods described in Section 2 and with parameters detailed below in this section. All tests are performed at a one-sided level of significance of 0.025. For each scenario we performed 50 000 simulation runs.

Details on the survival functions and hazard functions in the considered scenarios can be found in Supplemental material S1.

5.1 | Effect of subgroup prevalence

Here we assume a subgroup of patients has increased benefit from treatment. We modeled patients under control have a constant hazard corresponding to median survival of 11 months, and patients under treatment to have constant hazards with 18 months median survival if they do not belong to the subgroup and 30 months median survival if they do. The prevalence of the subgroup was varied in (0,0.2,0.4,0.6). Note that these are population prevalences, with observed sample proportions deviating according to random sampling.

In a second set of scenarios we added a delay of 100 days before onset of the treatment effect, meaning that in the first 100 days the hazard rate under treatment is the same as under control (for both subgroups) and afterward corresponds to the values of the previous scenario.

Visualizations of the resulting survival and hazard functions are included in the online supplemental material. Figure 3 shows the resulting power curves. Without delayed onset the extent of non-proportionality of hazard rates is limited (see Supplement), hence the logrank test is superior to the other tests. Also there is little dependence on the recruitment scheme. In scenarios with delayed effect, however, there is a strong non-proportionality and, both, putting too much weight on early events as well as recruiting too fast and hence observing mainly early events results in reduced power. Both maximum tests perform similar and the drop in power when using either maximum test compared to the best single weighted logrank test is small.

5.2 | Effect of hazard ratio in subgroup

Here we assume a setup identical to Section 5.1, except that the subgroup prevalence is held constant at 40% whereas the extra benefit in the subgroup is varying in terms of a multiplicative factor of the base hazard ratio in {1,0.72,0.6,0.45}, which corresponds to median survival times in the subgroup under treatment of {18,25,30,40}. As before, we considered an additional set of scenarios with delayed onset of treatment effect of 100 days. Resulting survival and hazard functions are found in the supplement and resulting power curves are shown in Figure 4. The observed pattern of power values is similar to that of the scenarios with varying subgroup prevalence, with increasing treatment benefit in the subgroup having a similar effect as increasing the subgroup prevalence.

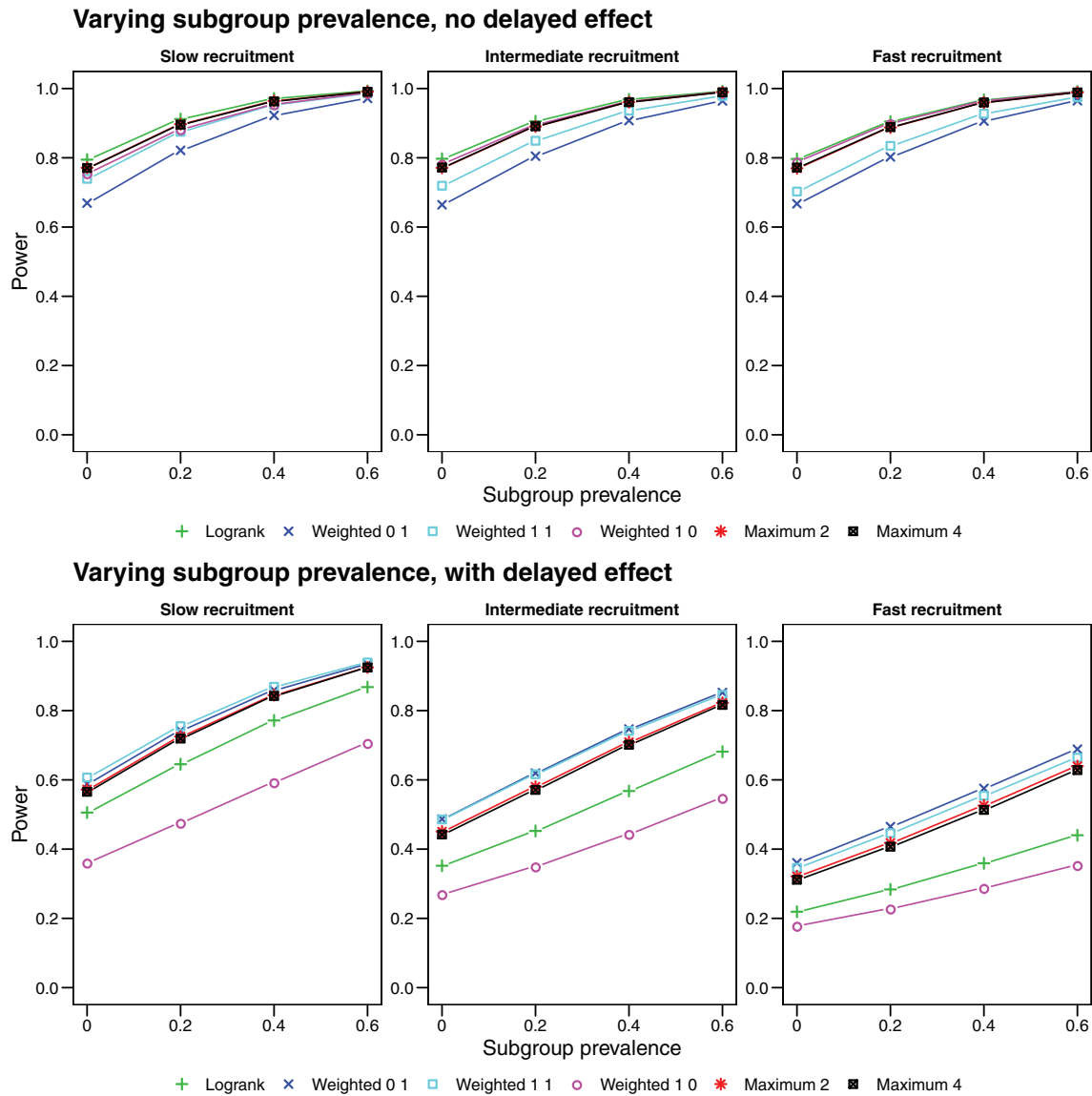


FIGURE 3 Power as a function of subgroup prevalence. Upper row: Scenarios without delayed onset of treatment effect. Lower row: Scenarios with delayed onset of treatment effect after 100 days. We show the power for the logrank test and three weighted logrank tests with $(\rho, \gamma) \in \{(0, 1), (1, 1), (1, 0)\}$ and maximum tests (i) combining the logrank and $(\rho = 0, \gamma = 1)$ (denoted by Maximum 2) and (ii) combining all four weighted logrank tests (denoted by Maximum 4)

5.3 | Effect of treatment switching

We again start from a setup with hazard rates as in Section 5.1 and the subgroup prevalence held constant at 40%. We now add the possibility to switch from control to active treatment after disease progression. Progression is modeled as an independent process with constant hazard and median time to progression of 5 months (fast progression scenarios) or 9 months (slow progression scenarios). In the treatment group progression has no further effect. In the control, groups patients will switch to the treatment group with a probability varying between scenarios in $\{0, 0.2, 0.4, 0.6\}$. After switching, former control patients will have the same hazards as a patient in the treatment group. The beneficial subgroup effect will come into effect for a switcher with a probability of 40%, since belonging to the subgroup is modeled as an intrinsic property of each patient. To calculate the power we assume an intention-to-treat analysis, that is, each patient is counted according to the initial randomization and with the full available observation time, regardless of potential subsequent switching.

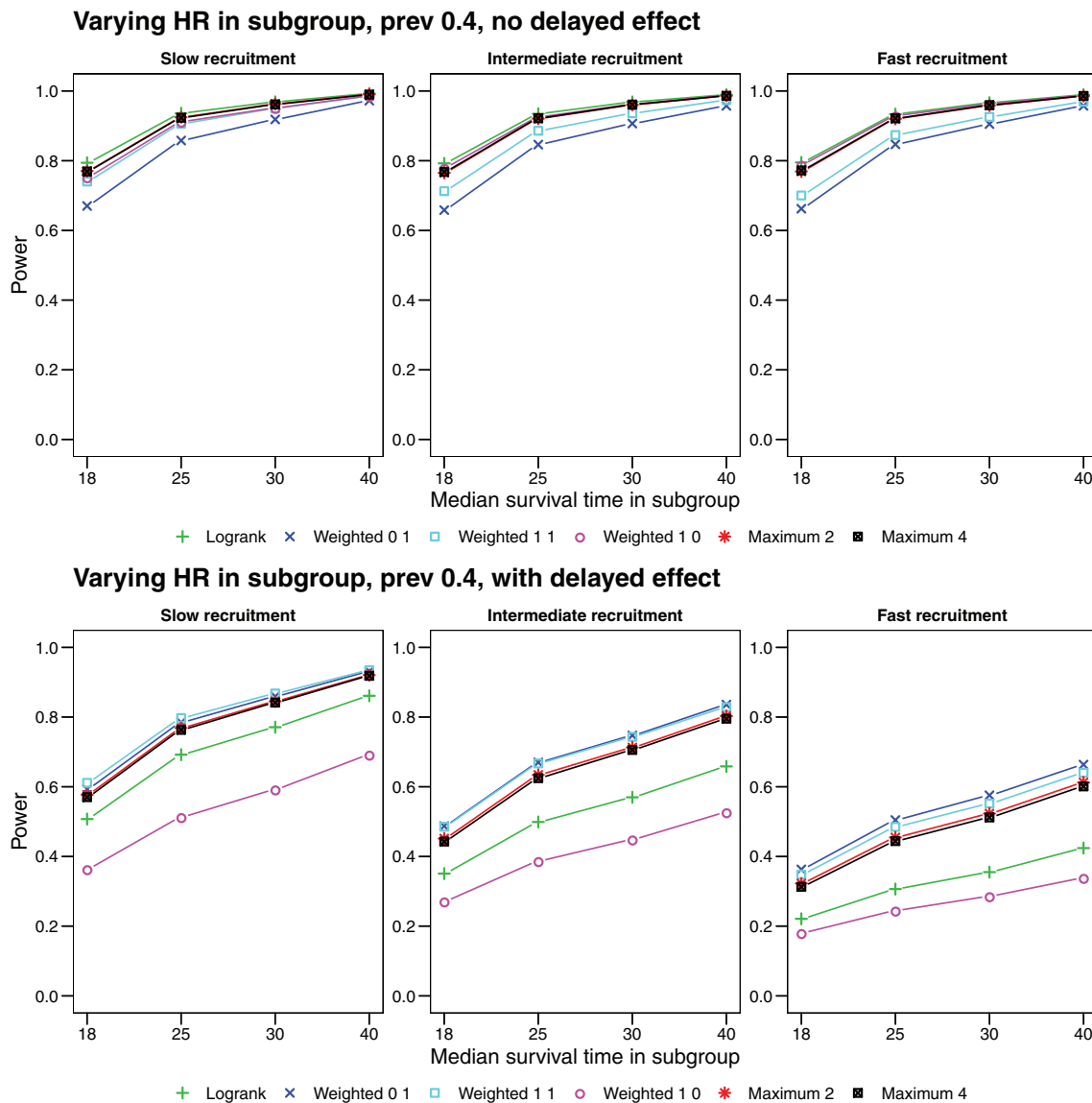


FIGURE 4 Power as a function of the additional hazard ratio in a subgroup with 40% prevalence. Upper row: Scenarios without delayed onset of treatment effect. Lower row: Scenarios with delayed onset of treatment effect after 100 days. We show the power for the logrank test and three weighted logrank tests with $(\rho, \gamma) \in \{(0, 1), (1, 1), (1, 0)\}$ and maximum tests (i) combining the logrank and $(\rho = 0, \gamma = 1)$ (denoted by Maximum 2) and (ii) combining all four weighted logrank tests (denoted by Maximum 4)

As before, survival and hazard functions are shown in the supplement. Resulting power curves are shown in Figure 5. As the possibility of switching decreases the hazard ratio with time, tests with more weight on early events have more power and the loss in power when using less optimal weights can be substantial. As seen before, the maximum tests are able to recover most of the power of the best test they include. The maximum test that includes only the $\rho = 0, \gamma = 0$ and the $\rho = 0, \gamma = 1$ weighted tests is less powerful than the maximum test that includes all four considered weighted logrank tests. However, the difference is only up to few percentage points, since the unweighted logrank test still has considerably power if the effect is mostly present at early event times. When switching is the main cause of non-proportional hazards, fast recruitment, and hence observing more early events, results in larger power values for all considered test, since in this scenario the observed treatment effect is then strongest at early event times. Similarly, the reduction in power with increasing switching probability is less pronounced in a fast recruitment regimen.

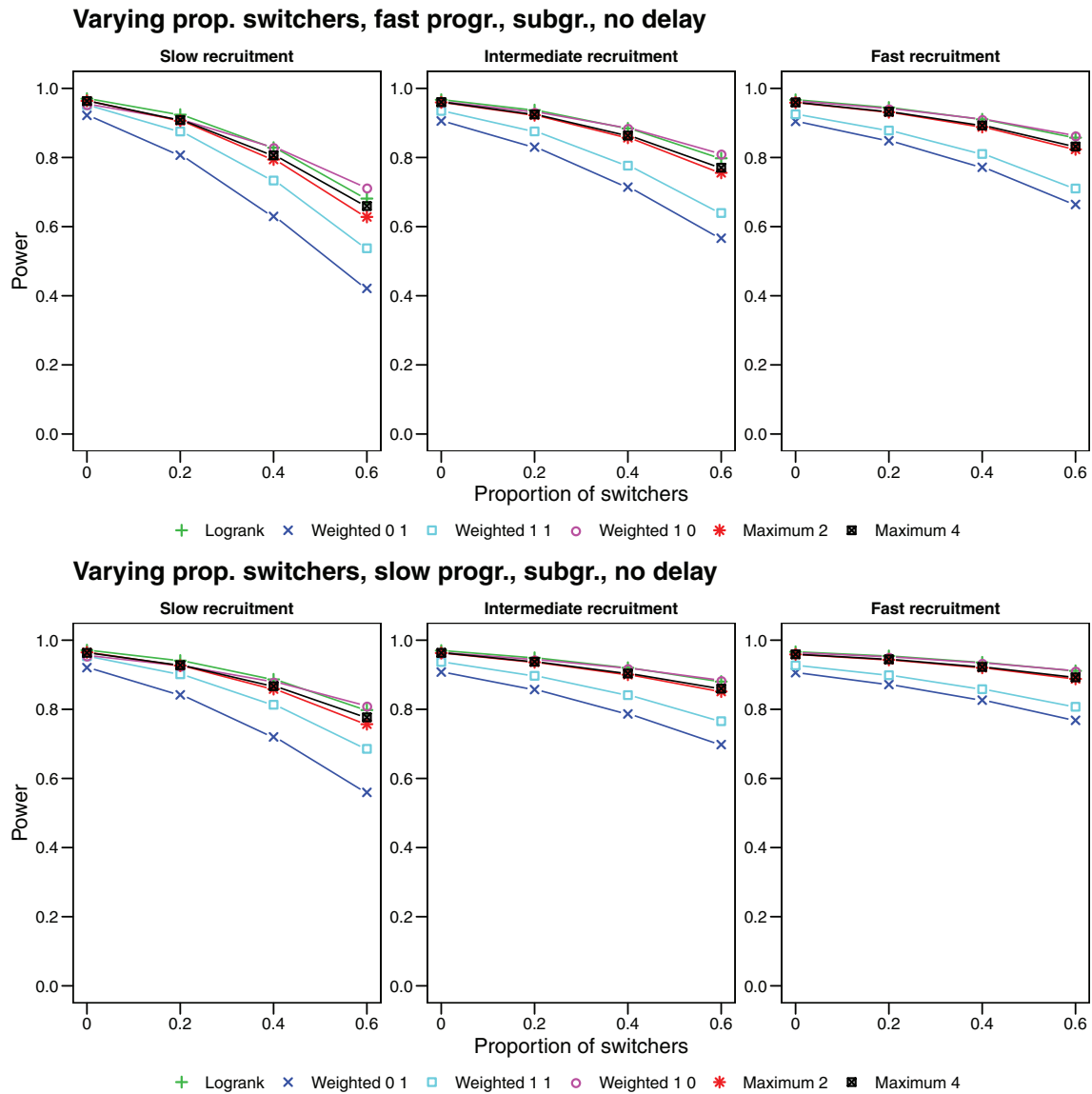


FIGURE 5 Power as a function of the probability to switch from control to treatment after disease progression. Upper row: Fast progression scenarios with a median time to progression of 5 months. Lower row: Slow progression scenarios with a median time to progression of 9 months. We show the power for the logrank test and three weighted logrank tests with $(\rho, \gamma) \in \{(0, 1), (1, 1), (1, 0)\}$ and maximum tests (i) combining the logrank and $(\rho = 0, \gamma = 1)$ (denoted by Maximum 2) and (ii) combining all four weighted logrank tests (denoted by Maximum 4)

5.4 | Effect of disease progression

We finally study the impact of altered hazard rates after disease progression. This results in non-proportional hazards for death if disease progression is occurring at a different rate in the two groups, or if the between-group hazard ratio is different before and after progression. Here we consider the former case. We assume constant hazard rates before and after progression, respectively, corresponding to median survival times of 18.33 and 15 months under treatment and 11 and 9 months under control. Hence the hazard ratio is 0.6, both, before and after progression. The rate of progression is constant with median time to progression of 5 months in the control group and a value in {5,7,9,11} months under treatment. Slower progression under treatment results in an initial decrease (stronger effect) in the hazard ratio over approximately 200 days and subsequent increase, however the extent of this non-proportionality is rather small (see Supplement). Consequently, the logrank test is most powerful for this setting, closely followed by maximum tests and the $\rho = 1, \gamma = 0$ weighted test. The recruitment scheme does not have a notable impact in these tests, however, the power of other weighted tests, which emphasize late events, are affected by too fast recruitment. As expected, a stronger

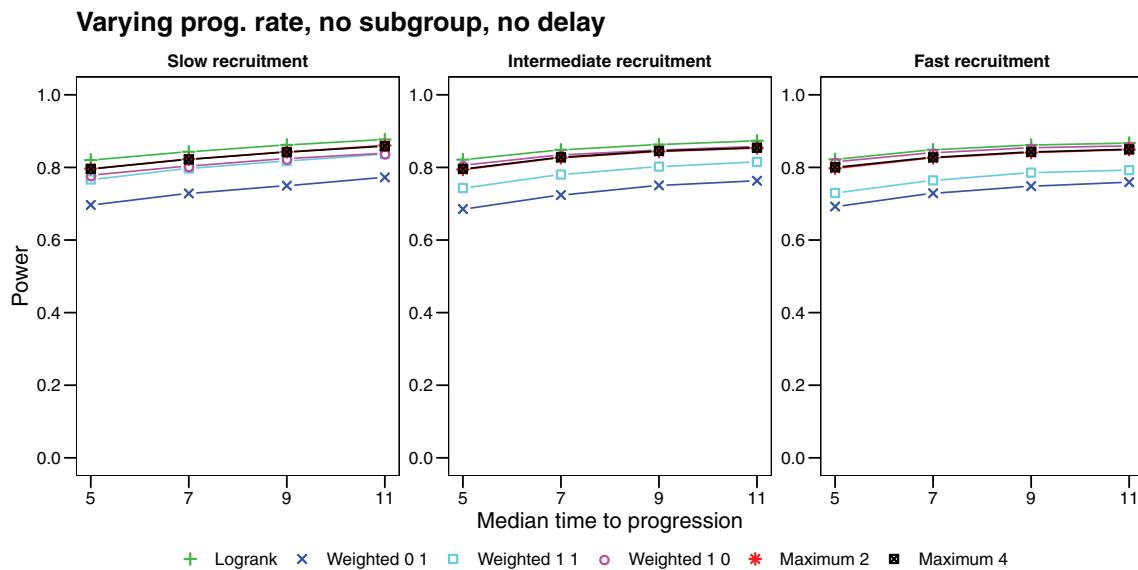


FIGURE 6 Power as a function of the median time to disease progression when hazards increase after progression in both groups. We show the power for the logrank test and three weighted logrank tests with $(\rho, \gamma) \in \{(0, 1), (1, 1), (1, 0)\}$ and maximum tests (i) combining the logrank and $(\rho = 0, \gamma = 1)$ (denoted by Maximum 2) and (ii) combining all four weighted logrank tests (denoted by Maximum 4)

treatment effect with respect to delaying progression increases the power in all scenarios. Figure 6 shows the detailed power curves.

5.5 | Summary of simulation results

The extent of non-proportionality caused by different sources of non-proportional hazard varies strongly. In the studied scenarios, delayed onset and the possibility of treatment switching resulted in the strongest deviations from proportional hazards. However, under more extreme scenarios also the presence of subgroups with different response to treatment and progression effects have the potential for more pronounced non-proportionality. Delayed onset and treatment switching have opposing effects with respect to the time points of strongest observable effects, hence weighting early vs late events has opposing effects on power with these two sources on non-proportionality. See Table 2 for a concise overview of the conclusions drawn from the simulation studies and literature. In line with previous simulation studies,^{11,22,23} the maximum test has the potential to safeguard against misspecification of a single weighed test. Interestingly, the price of combining two or four tests in the maximum test, in terms of reduced power compared to the best test single test in the simulation, was almost identical. This can be attributed to the strong correlation between the included tests and supports the idea of using the maximum tests as robust variant in the analysis of trials where the pattern of non-proportional hazards cannot be predicted well.

6 | SOFTWARE IMPLEMENTATION

In Section 5 we studied the effect of different sources of non-proportional hazards separately, while Section 4 provided a case study of a clinical trial in which at least two sources were present. Naturally, different diseases and trial characteristics will exhibit different patterns in terms of non-proportionality mechanisms. We provide the R package `nph`¹⁵ which is available at the CRAN repository so that researchers can explore different scenarios using the theory presented in this paper. The package includes functions to model survival distributions in terms of piecewise constant hazards and with differential effects due to disease progression and subgroups, to simulate data from the specified distributions, and to perform the weighted logrank tests and maximum tests described in Section 3.1.

Functions for modeling/setting the underlying survival model include:

TABLE 2 Summary of the effect of considered population characteristics and design aspects on the proportionality of hazards and the power of between group comparisons of survival times

| Aspect | Effect on non-proportional hazards or on power | Key message |
|------------------------------------|--|---|
| Delayed onset of treatment effect | Treatment effect increases with time, since the treatment is effective only with some delay. | Reducing the weight of early events increases power. Zero weight for early events may be considered as most extreme case. ¹⁹ |
| Disease progression | Treatment effect changes with time if active treatment delays disease progression or the treatment effect is different before and after progression. | Optimal weights depend on treatment effect on progression and hazard ratios before and after progression. The effects can be explored by simulation in the proposed piecewise constant hazards framework. |
| Biomarker subgroups | Treatment effect is increasing with time, when a biomarker positive subgroup with particularly strong treatment benefit is included, since this subpopulation is enriched with time. | More weight on late events increases power. As an alternative to weighted logrank test, parametric cure rate models may be considered. ⁴ |
| Treatment switch after progression | Treatment effect is decreasing with time, if patients may switch to a common medication after disease progression. | More weight on early events may increase power. Accelerated failure time models accounting for different failure rates before and after treatment switch have been proposed as parametric alternative. ^{6,7} |
| Recruitment scheme | Fast recruitment with short follow up results in larger power if the treatment effect decreases with time. Slow recruitment with long follow up results in larger power if the treatment effect increases with time. | Under non-proportional hazards, the power of logrank tests also depends on the timing of observed events. This in turn depends on recruitment speed and total study duration. The effect of different recruitment schemes should be considered during study planning. |

- `pchaz`: Calculate survival functions defined through piecewise constant hazards
- `subpop_pchaz`: Calculate survival functions defined through piecewise constant hazards, allowing for a change in the piecewise constant hazard regimen after a random progression time
- `pop_pchaz`: Calculate survival functions for a mixture of subpopulations. In each subpopulation the survival function is defined through piecewise constant hazards and these are allowed to change after a random progression time.

Functions for generating simulated datasets given a specified survival model are:

- `sample_fun`: Sample survival times based on study settings
- `sample_conditional_fun`: Sample conditional survival times based on study settings and given observed interim data

Functions for performing statistical tests:

- `logrank.test`: Weighted logrank test
- `logrank.maxtest`: Maximum logrank test

Plotting functions:

- `plot.mixpch`: Generic plot function for `mixpch` objects, which result from `pchaz`, `subpop_pchaz` and `pop_pchaz`
- `plot_diagram`: Creates the diagram for the model, similar to Figure 1
- `plot_shhr`: Plot of survival, hazard and hazard ratio of two groups as function of time

Additionally, we provide in the supplementary material (S2 file) a document with basic usage instructions for the package and outlining further more complex examples of non-proportional hazards along with the R code to model these situations.

7 | DISCUSSION

The proportional hazards assumption may be violated due to different inherent features of a study design, patient population or mode of treatment action. The modeling approach pursued in this paper allows one to study the effect of different sources of non-proportional hazards on the survival and hazard functions. It also allows for easy simulation of study data and comparison of different testing strategies under complex non-proportional hazard scenarios.

Under proportional hazards the standard unweighted logrank test corresponds to the optimal test in the family of weighted logrank tests, whereas under non-proportional hazards other weighting schemes may result in larger power. If the shape of the hazards function for an actual trial is well understood, for example, through modeling the assumed data generating process, and well known from available data, a suitable weighting function may be selected. For example, Liu et al²⁵ proposed nearly optimal weights under a specific cure rate model with delayed onset. In many cases, however, different hazard functions may be considered reasonable. In this situation the combination of several weighting functions in terms of the maximum test may be preferable. The power of this test is typically closely below that of the best included individual test, providing a robust method. Other combination functions with similar properties may be explored though, for example, the combination of individual tests via the harmonic mean of their P -values.²⁶

Group sequential methods have been proposed as another approach to tackle the planning uncertainties under unknown patterns of non-proportional hazards. Jimenez et al²⁷ describe sample size reassessment methods in the setting of the delayed onset and comparing two groups with $\rho - \gamma$ weighted logrank test. They propose to plan the sample size under a model of delayed onset, for example, for a $\rho = 0, \gamma = 1$ weighted test. If, however, the assumption about the delay was not correct, the sample size might be too small and in this case sample size reassessment based on interim is proposed to increase the sample size (and power) if necessary. Korn and Freidlin²⁸ study the loss in power when futility stopping rules at interim are applied in the setting of delayed onset. To reduce the influence of the suspected delayed onset phase, they propose to modify stopping rules by the additional condition that a reasonable fraction (eg, at least two-thirds) of the events observed at interim have occurred at least a sufficient time (eg, 3 months) after randomization. In the presence of a potential subgroup of long term survivors, the planned number of events may be reached later than anticipated. In this context, Chen⁴ proposed to select and fit a parametric model, choosing from a set of distributions such as lognormal and Weibull, with blinded interim data and thereon predict the remaining study duration.

We focused on testing the null hypothesis of ordered hazard functions. Here rejecting H_0 using a (weighted) logrank test first of all means that at least for t in some time interval $\lambda_{trt}(t) < \lambda_{ctr}(t)$. In the clinical trial setting this means the experimental treatment has some beneficial effect at least in some time interval. However, which time intervals are actually observed and how strongly their information contributes to the test decision depends on the censoring distribution and the applied weights (see Reference 8). A definite assessment of the treatment benefit should therefore include considerations on possible detrimental effects in time intervals which were not observed or which received less weight. In some cases subject matter knowledge on the safety of the treatment may even allow for the assumption $\lambda_{trt}(t) \leq \lambda_{ctr}(t)$ for all t .

However the survival of two (or more) groups may also be compared in terms of other parameters. A simple strategy is so the called landmark or milestone analysis in which the values of the survival functions for a pre-specified time point, for example, one-year survival, are compared between groups. However, selecting a fixed time point in the planning phase is difficult and an inappropriate choice could lead to power loss. In addition, this approach suffers from information loss as only one point of the estimated survival curve is considered. Logan and Mo¹⁹ combine both approaches. They argue that under non-proportional hazards and under potential crossing of survival curves, long term survival is a relevant outcome. They propose a (group sequential) test for the intersection null hypothesis of equal survival function at a pre-specified time point t_0 and equal hazard functions after t_0 . Alternative effect measures which we did not consider here include the difference in restricted mean survival times²⁹ and the average hazard ratio.

A well-known alternative, the difference in restricted mean survival times²⁹ of two groups may be used as effect measure. Tests based on this parameter can be more powerful than the unweighted logrank test, however depending on the actual hazard functions and the chosen time point t_{restr} up to which the restricted means are calculated.³⁰ The interpretation of this parameter is entirely focused on the hazards before t_{restr} , as the difference in restricted mean survival times corresponds to the expected life time gained up to the time t_{restr} . Note that under crossing survival functions, the probability to survive longer than time t_{restr} may be larger in one group while the restricted mean survival time up to time t_{restr} is larger in the other group.

Also the general non-parametric effect measure $P(T_{trt} > T_{ctr})$ may be utilized, which is referred to under various names such as relative effect or probabilistic index. In this context T_{trt} and T_{ctr} are the event times of two randomly chosen subjects under treatment and control, respectively. Under proportional hazards, the hazard ratio equals $\frac{P(T_{trt} > T_{ctr})}{1 - P(T_{trt} > T_{ctr})}$, see for example, Reference 31. For non-proportional hazard settings, several authors propose to use as effect measure an average hazard ratio³²⁻³⁵ defined as $\frac{\int \lambda_{trt}/(\lambda_{trt} + \lambda_{ctr})dW(t)}{\int \lambda_{ctr}/(\lambda_{trt} + \lambda_{ctr})dW(t)}$, with $W(t)$ a survival function that serves as weighting function.

When choosing $W(t) = S_{ctr}(t)S_{trt}(t)$, the average hazard ratio conveniently equals $\frac{P(T_{trt} > T_{ctr})}{1 - P(T_{trt} > T_{ctr})}$. However other choices of $W(t)$ may result in larger power depending on the true hazard functions. With censored data, the average hazard ratio needs to be computed in a truncated fashion up to some time t_{trun} , which is within the time span of observed events. In this case it can be interpreted to compare the survival times T_{trt} and T_{ctr} conditional on at least one of the two event times being smaller than t_{trun} .

Explicit modeling of non-proportional hazards may be considered, see for example, the parametric approach in Reference 4 discussed above. However, to allow for unambiguous interpretation of observed effects in a parametric analysis model, the model must be predefined and should be supported by subsequent model diagnostics.

There is not a unique answer on how to address the impact of different disturbances of the proportional hazards assumptions. As key finding from our investigation with respect to choosing a hypothesis test we conclude that the maximum test has the potential to safeguard against deviations from optimal assumptions and is robust with respect to the number of subsumed tests. With respect to identification of sources non-proportionality we conclude that inspection of estimated survival curves, for example, from previous trials, does in general not reveal underlying sources of non-proportional hazards. In part as consequence of the difficulty to empirically identify these sources, we conclude that simulations are essential to understand the operational characteristics of a planned trial under all relevant deviations from the proportional hazards assumption.


The modeling approach, the hypothesis tests and the according software we described provide a comprehensive set of tools to address these challenges. While we focused on weighted logrank test for statistical inference on hazard functions, the set of method can readily be extended with analysis methods based on further parameters of interest.

DATA AVAILABILITY STATEMENT

"We provide the R package *nph* [15] which is available at the CRAN repository so that researchers can explore different scenarios using the theory presented in this paper." [See Section 6]

ORCID

Robin Ristl  <https://orcid.org/0000-0002-4163-9236>

Nicolás M Ballarini  <https://orcid.org/0000-0002-3432-8931>

Heiko Götte  <https://orcid.org/0000-0002-6305-4466>

Armin Schüller  <https://orcid.org/0000-0003-3833-5757>

Martin Posch  <https://orcid.org/0000-0001-8499-8573>

Franz König  <https://orcid.org/0000-0002-6893-3304>

REFERENCES

1. Finke LH, Wentworth K, Blumenstein B, Rudolph NS, Levitsky H, Hoos A. Lessons from randomized phase iii studies with active cancer immunotherapies—outcomes from the 2006 meeting of the Cancer Vaccine Consortium (CVC). *Vaccine*. 2007;25:B97-B109.
2. Anagnostou V, Yarchoan M, Hansen AR, Wang H, Verde F, Sharon E, Collyar D, Chow LQ, Forde PM. Immuno-oncology trial endpoints: capturing clinically meaningful activity; 2017.
3. Chan DLH, Segelov E, Wong RSH, et al. Epidermal growth factor receptor (egfr) inhibitors for metastatic colorectal cancer. *Cochrane Database Syst Rev*. 2017;(6). <https://doi.org/10.1002/14651858.CD007047.pub2>.
4. Chen T-T. Predicting analysis times in randomized clinical trials with cancer immunotherapy. *BMC Med Res Methodol*. 2016;16:12.
5. Gandhi L, Rodriguez-Abreu D, Gadgeel S, et al. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *N Engl J Med*. 2018;378:2078-2092.
6. Latimer NR, Abrams KR, Lambert PC, et al. Adjusting for treatment switching in randomised controlled trials – a simulation study and a simplified two-stage method. *Stat Methods Med Res*. 2017;26(2):724-751.
7. Latimer NR, White IR, Abrams KR, Siebert U. Causal inference for long-term survival in randomised trials with treatment switching: should re-censoring be applied when estimating counterfactual survival times? *Stat Methods Med Res*. 2018;28(8):2475-2493.
8. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*. 1981;68(1):316-319.

9. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Vol 169. Hoboken, New Jersey: John Wiley & Sons; 2011.
10. Tarone RE. On the distribution of the maximum of the logrank statistic and the modified wilcoxon statistic. *Biometrics*. 1981;37:79-85.
11. Lee S-H. On the versatility of the combination of the weighted log-rank statistics. *Comput Stat Data Anal*. 2007;51(12):6557-6564.
12. Hasegawa T. Sample size determination for the weighted log-rank test with the Fleming–Harrington class of weights in cancer vaccine studies. *Pharm Stat*. 2014;13(2):128-135.
13. Hasegawa T. Group sequential monitoring based on the weighted log-rank test statistic with the Fleming–Harrington class of weights in cancer vaccine studies. *Pharm Stat*. 2016;15(5):412-419.
14. Luo X, Mao X, Chen X, Qiu J, Bai S, Quan H. Design and monitoring of survival trials in complex scenarios. *Stat Med*. 2019;38(2):192-209.
15. Ristl R, Ballarini N. *nph: Planning and Analysing Survival Studies under Non-Proportional Hazards*. 2020. R package version 2.0. <https://CRAN.R-project.org/package=nph>.
16. Kitchin J, Langberg NA, Proschan F. *A New Method for Estimating Life Distributions From Incomplete Data*. Tallahassee, Florida: Technical report. Florida State Univ Tallahassee Dept of Statistics; 1980:1-20.
17. Malla G, Mukerjee H. A new piecewise exponential estimator of a survival function. *Sta Prob Lett*. 2010;80(23–24):1911-1917.
18. Chen T-T. Statistical issues and challenges in immuno-oncology. *J Immunother Cancer*. 2013;1(1):18.
19. Logan BR, Mo S. Group sequential tests for long-term survival comparisons. *Lifetime Data Anal*. 2015;21(2):218-240.
20. Peto R. Rank tests of maximal power against lehmann-type alternatives. *Biometrika*. 1972;59(2):472-475.
21. Yang S, Hsu L, Zhao L. Combining asymptotically normal tests: case studies in comparison of two groups. *J Stat Plan Infer*. 2005;133(1):139-158.
22. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One*. 2015;10(1):1-18.
23. Karrison TG et al. Versatile tests for comparing survival curves based on weighted log-rank statistics. *Stata J*. 2016;16(3):678-690.
24. The National Press Club, Washington, DC. Public workshop: Oncology clinical trials in the presence of non-proportional hazards; 2018. <https://healthpolicy.duke.edu/events/public-workshop-oncology-clinical-trials-presence-non-proportional-hazards>. Accessed January 30, 2019.
25. Liu S, Chu C, Rong A. Weighted log-rank test for time-to-event data in immunotherapy trials with random delayed treatment effect and cure rate. *Pharm Stat*. 2018;17(5):541-554.
26. Wilson DJ. The harmonic mean p-value for combining dependent tests. *Proc Natl Acad Sci*. 2019;116(4):1195-1200.
27. Jiménez JL, Stalbovskaya V, Jones B. Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects. *Pharm Stat*. 2018;18(3):287-303.
28. Korn EL, Freidlin B. Interim futility monitoring assessing immune therapies with a potentially delayed treatment effect. *J Clin Oncol*. 2018;36(23):2444-2449.
29. Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Epidemiol Infect*. 1949;47(2):188-189.
30. Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13(1):152.
31. De Neve J, Gerds TA. On the interpretation of the hazard ratio in Cox regression. *Biometr J*. 2020;62(3):742-750. <https://onlinelibrary.wiley.com/doi/full/10.1002/bimj.201800255>.
32. Kalbfleisch JD, Prentice RL. Estimation of the average hazard ratio. *Biometrika*. 1981;68(1):105-112.
33. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted cox regression. *Stat Med*. 2009;28(19):2473-2489.
34. Brückner M, Brannath W. Sequential tests for non-proportional hazards data. *Lifetime Data Anal*. 2017;23(3):339-352.
35. Rauch G, Brannath W, Brueckner M, Kieser M. The average hazard ratio—a good effect measure for time-to-event endpoints when the proportional hazard assumption is violated? *Methods Inf Med*. 2018;57(03):89-100.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Ristl R, Ballarini NM, Götte H, Schüler A, Posch M, König F. Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology. *Pharmaceutical Statistics*. 2021;20:129–145. <https://doi.org/10.1002/pst.2062>