Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

Sample Size Calculation Under Nonproportional Hazards Using Average Hazard Ratios

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

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

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

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

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

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

ABSTRACT

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

1 | Background

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

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

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

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

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

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

2 | Methods

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

2.1 | The AHR and a Corresponding Test Statistic

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 | Asymptotic Sample Size Calculation

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

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

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

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

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

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

2.3 | Simulation-Based Sample Size Calculation

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

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

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

3 | Simulation Study

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

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

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

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

3.1 | Scenarios and Parameter Values

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

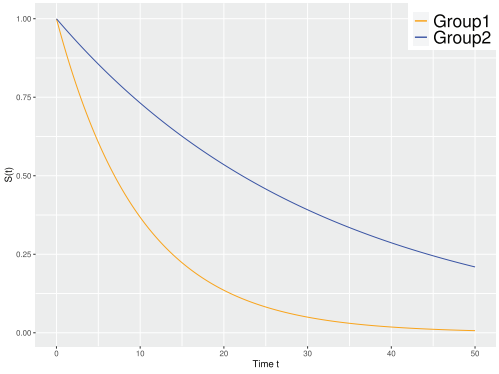
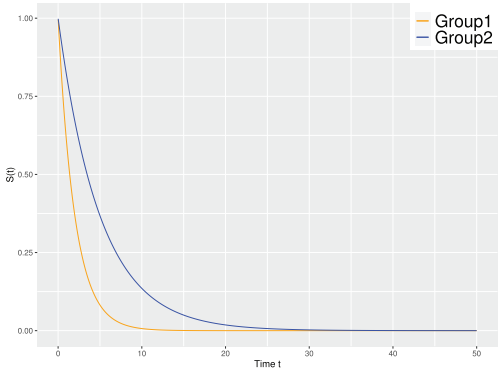
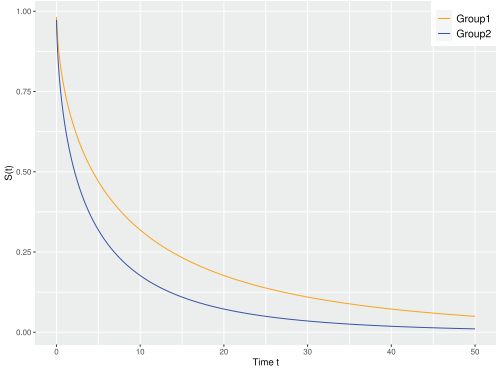
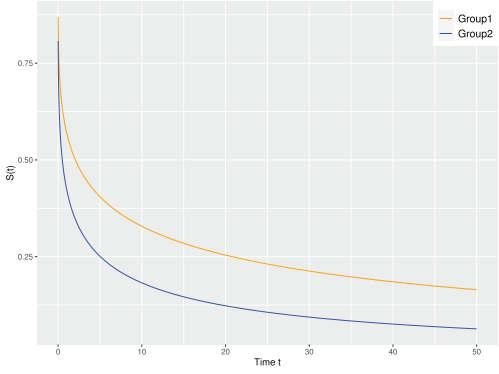
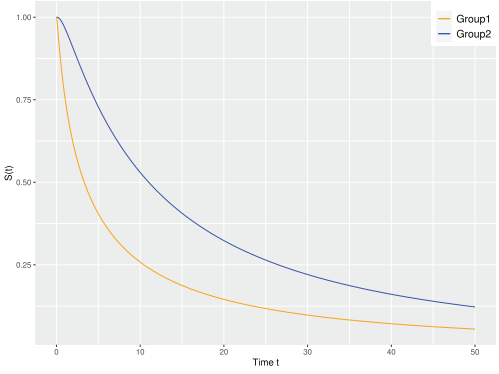
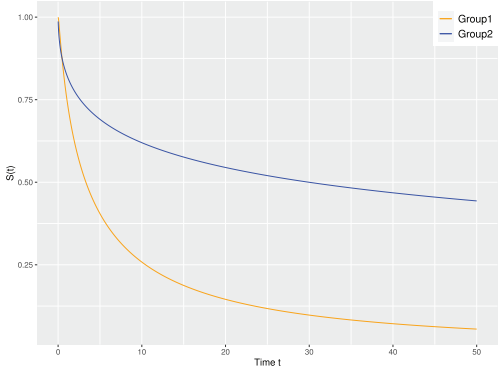
3.2 | Results

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

3.2.1 | G1: Comparison of Different Sample Size Calculation Approaches

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

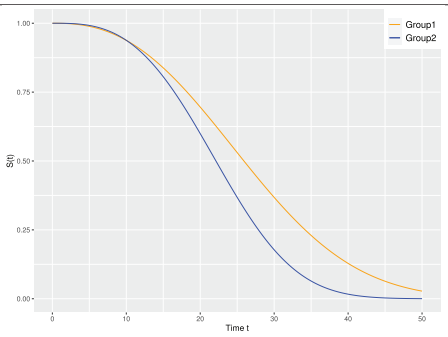
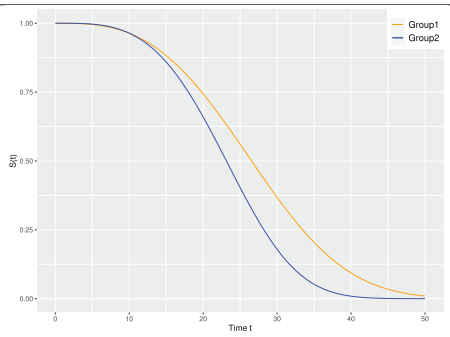
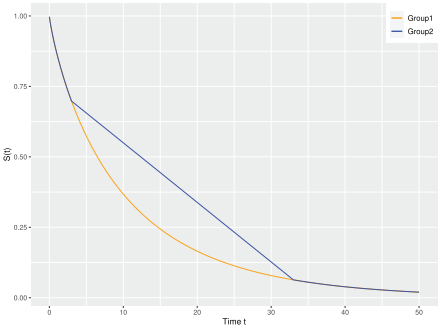
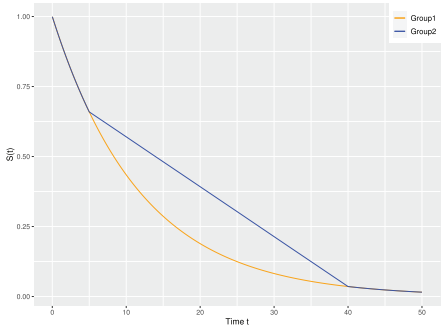
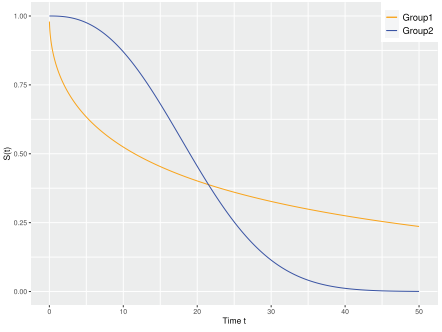
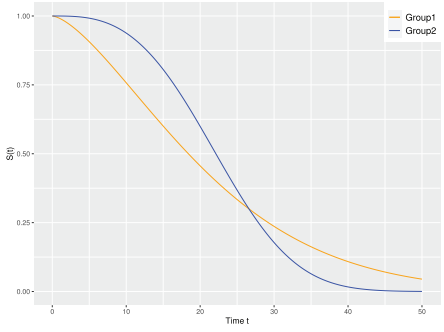
TABLE 1 | Simulation scenarios and the alternative distribution considered. The hazard ratio (HR) given is the mean based on 100 datasets with $n_0 = n_1 = 150$. The alternative scenario (alt) was chosen to have the same HR as the initial scenario.

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

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

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

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

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

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

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

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

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

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

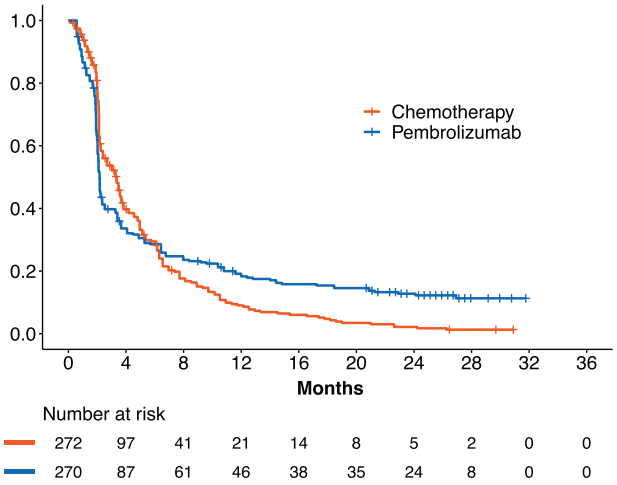


FIGURE 1 | Reconstructed individual patient data based on the UC study by Fradet et al. (2019).

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

4 | Illustrative Data Example

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

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

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

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

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

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

Significant values are given in bold.

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

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

5 | Discussion

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

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

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

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

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

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

Acknowledgments

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

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

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

Open Research Badges

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Royston, P., and M. K. Parmar. 2011. "The Use of Restricted Mean Survival Time to Estimate the Treatment Effect in Randomized Clinical Trials When the Proportional Hazards Assumption is in Doubt." *Statistics in Medicine* 30, no. 19: 2409–2421.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Supporting Information

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