Received: 8 February 2019 | Revised: 24 September 2019 | Accepted: 25 November 2019

DOI: 10.1111/biom.13196

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Sample size and power for the weighted log-rank test and Kaplan-Meier based tests with allowance for nonproportional hazards

Godwin Yung[1] | **Yi Liu**[2]

[1]Takeda Pharmaceuticals, 35 Landsdowne St, Cambridge, Massachusetts

[2]Nektar Therapeutics, 455 Mission Bay Boulevard South, San Francisco, California

Correspondence
Godwin Yung, Takeda Pharmaceuticals, Cambridge, MA 02139.
Email: godwin.yung@takeda.com

Abstract

Asymptotic distributions under alternative hypotheses and their corresponding sample size and power equations are derived for nonparametric test statistics commonly used to compare two survival curves. Test statistics include the weighted log-rank test and the Wald test for difference in (or ratio of) Kaplan-Meier survival probability, percentile survival, and restricted mean survival time. Accrual, survival, and loss to follow-up are allowed to follow any arbitrary continuous distribution. We show that Schoenfeld's equation—often used by practitioners to calculate the required number of events for the unweighted log-rank test—can be inaccurate even when the proportional hazards (PH) assumption holds. In fact, it can mislead one to believe that 1:1 is the optimal randomization ratio (RR), when actually power can be gained by assigning more patients to the active arm. Meaningful improvements to Schoenfeld's equation are made. The present theory should be useful in designing clinical trials, particularly in immuno-oncology where nonproportional hazards are frequently encountered. We illustrate the application of our theory with an example exploring optimal RR under PH and a second example examining the impact of delayed treatment effect. A companion R package `npsurvSS` is available for download on CRAN.

KEYWORDS

asymptotic theory, clinical trial, randomization ratio, survival analysis

1 | INTRODUCTION

The log-rank test is the most commonly used test to design and analyze two-arm clinical trials with a time-to-event endpoint. Investigators typically begin by defining end of study by the number of events d achieved, where d is calculated from the equation

$$d = \frac{(z_{\alpha/2} + z_{\beta})^2}{p_0 p_1 \theta^2}; \quad (1)$$

$z_{\alpha/2}$ and z_{β} are the upper $\alpha/2$ and β percentiles of the standard normal distribution; p_0 and p_1 are the proportions of patients assigned to the control and active arm; and θ is the logarithm of the hazard ratio quantifying the relative treatment benefit (Schoenfeld, 1981). Then, sample size n is determined taking into account the accrual distribution, survival distribution, loss to follow-up distribution, and study duration (Collett, 2015).

The calculation of d and n based on Schoenfeld (1981) has been referred to as the "back-of-the-envelope" approach (Pak *et al.*, 2017). In this paper, we refer to it as the d-n method

because of the order in which d and n are calculated. The d-n method has the nice property that, as long as θ is correct, power $1-\beta$ stays the same even if the assumed accrual, survival, and loss to follow-up distributions are misspecified. However, it has two conditions that are limiting. It applies only under the assumption of proportional hazards (PH) and when the unweighted log-rank test (or a traditional test based on the related Cox model) is used for the primary hypothesis testing. If either condition is not met, the d-n method can be misleading. How to calculate sample size and power in a more general setting is the focus of our paper.

This problem has become increasingly relevant with the advent of novel immunotherapies. Consider nivolumab (OPDIVO; Bristol-Myers Squibb), an anti-PD-1 monoclonal antibody that harnesses a patient's own immune system to fight cancer. Unlike traditional chemotherapy, the overall survival benefit of nivolumab may not become apparent until several months after initial treatment. This has resulted in trials exhibiting various patterns of nonproportional hazards (NPH) including delayed separation, curves that cross, and long-term survival with "cured" patients (eg, see trials CheckMate 057, 066, and 141 in nivolumab's FDA label).

When the PH assumption is violated, the log-rank test may be less efficient. The estimated hazard ratio (HR) will also be difficult to interpret. Thus, concurrent with advances in immunotherapy, there has been an increased interest in using alternatives for hypothesis testing and quantifying treatment benefit. Uno *et al.* (2014) proposed to use nonparametric tests that have obvious, corresponding measures of efficacy. This includes the difference in milestone survival probability, survival percentile, and restricted mean survival time (RMST) based on the Kaplan-Meier (KM) estimate. Many statisticians have echoed the same recommendation (Chen, 2015; Hoering *et al.*, 2016; Pak *et al.*, 2017). There are also others who advocate for the weighted log-rank test, which weighs periods of the survival curve differently than the ordinary, unweighted log-rank test (Fine, 2007).

Several papers have proposed analytical solutions for sample size and power calculation under NPH. However, most of these solutions apply only to the weighted log-rank test under particular patterns of NPH, for example, piecewise exponential survival (Barthel *et al.*, 2006; Luo *et al.*, 2019), PH cure model (Wang *et al.*, 2011), or delayed treatment effect model (Hasegawa, 2014). It is unclear how these solutions will translate under alternative assumptions (eg, Weibull survival) or when other nonparametric tests are used. Royston and Parmar (2013) derived sample size and power equations for the KM RMST. However, in the presence of censoring, parameters in their equations need to be estimated via Monte Carlo simulations, which can be computationally intensive to perform and lead to Monte Carlo error.

The aim of this paper is to propose a unified framework for calculating sample size and power that can be efficiently applied to diverse patterns of survival—patterns under PH or NPH—and to any of the aforementioned nonparametric statistical tests: difference in KM survival probability, percentile, RMST, and the weighted or unweighted log-rank test. Specifically, we consider the practical setting where patients are assigned to one of two treatments via block randomization and the study is ended when either a prespecified duration has elapsed or a prespecified number of events has been achieved. Accrual, survival, and loss to follow-up are allowed to follow any arbitrary, continuous distribution. Asymptotic theory is used to derive analytical equations, rendering simulations unnecessary.

In the course of our work, we use an improved estimate for the large-sample distribution of the weighted log-rank test that is not widely recognized in literature. Motivated by this improvement, we propose a new method to calculate sample size and event size. We call this method the n-d method because, unlike the d-n method, n is calculated before d. The n-d method can be applied under NPH and to nonparametric tests beyond the unweighted log-rank test. Even when the PH assumption is satisfied and the unweighted log-rank test is used for hypothesis testing, it results in more accurate calculations of sample size and power than the d-n method. Differences in precision are especially pronounced when HR is small (similar to results in Barthel *et al.*, 2006) or when randomization ratio (RR) is imbalanced.

The rest of this paper is organized as follows. In Sections 2.1–2.3, we describe the notation and derive sample size and power equations for time-driven trials. In Section 2.4, we discuss how these results can be translated to event-driven trials by using the n-d method. Section 2.5 gives an overview of our companion R package `npsurvSS`. In Section 3.1, we use our theory to explore RR and its impact on power of the unweighted log-rank test under PH. Contrary to conventional thinking, 1:1 RR does not maximize power. In Section 3.2, we illustrate the application of our theory by evaluating the impact that delayed treatment effect has on power in the CheckMate 141 (CM-141) trial. We conclude with some discussions in Section 4.

2 | METHODS

2.1 | Study setting and notation

Suppose $n_0 = np_0$ patients are randomized to the control arm and $n_1 = np_1$ to the active arm. Let X_i denote the treatment indicator (0 = control, 1 = active), R_i the time of study entry, L_i the time to loss of follow-up, and T_i the time to event, $i = 1, \ldots, n$. We assume that R_i has continuous cdf $H(\cdot)$; $L_i|X_i = j$ has continuous cdf $G_j(\cdot) = 1 - \overline{G}_j(\cdot)$; and $T_i|X_i = j$ is independent of C_i and has continuous cdf $F_j(\cdot) = 1 - \overline{F}_j(\cdot)$, pdf $f_j(\cdot)$, and hazard function $\lambda_j(\cdot)$. Patients are accrued over a

period τ_a and followed for an additional period τ_f, so that the total study duration is $\tau = \tau_a + \tau_f$ and time to censoring due to either loss of follow-up or administrative censoring is given by $C_i = L_i \wedge (\tau - R_i)$, where $\wedge$ denotes taking the minimum. At the end of the trial, we observe for each patient the triplet (X_i, U_i, δ_i) where $U_i = T_i \wedge C_i$ and $\delta_i = 1(U_i = T_i)$.

Define the counting processes $N_i(t) = 1(U_i \le t, \delta_i = 1)$ and $Y_i(t) = 1(U_i \ge t)$. Conditional on $X_i = j$, $N_i(t)$ and $Y_i(t)$ have expectations $\nu_j(t) = \int_0^t f_j(s)\overline{G}_j(s)H(\tau_a \wedge (\tau - s))ds$ and $\pi_j(t) = \overline{F}_j(t)\overline{G}_j(t)H(\tau_a \wedge (\tau - t))$. They intuitively represent the probability that a patient in arm j will experience an event within time t, or survive beyond time t after entering the study. The term $H(\tau_a \wedge (\tau - t))$ reflects the fact that if a patient is at risk at time t, she must have been accrued at or before time $\tau_a \wedge (\tau - t)$. We denote the unconditional expectations of $N_i(t)$ and $Y_i(t)$ by $\nu(t) = p_0\nu_0(t) + p_1\nu_1(t)$ and $\pi(t) = p_0\pi_0(t) + p_1\pi_1(t)$. Evaluated at $t = \tau$, $\nu(\tau)$ is the probability a patient will experience an event during his or her participation in the trial.

2.2 | Sample size and power

In Section 2.3, test statistics Z will be proposed of the form such that, for large n, Z follows approximately a normal distribution with mean $\sqrt{n}\Delta/\sigma$ and variance $\tilde{\sigma}^2/\sigma^2$. That is, $Z \dot{\sim} N(\sqrt{n}\Delta/\sigma, \tilde{\sigma}^2/\sigma^2)$, where $\sigma, \tilde{\sigma} > 0$, and Δ is equal to 0 under the null hypothesis of no treatment benefit and less than 0 under the alternative hypothesis of beneficial treatment. Power and sample size for a one-sided $\alpha/2$-level test (or approximate two-sided α-level test) can therefore be calculated using the equations

$$1 - \beta = 1 - \Phi\left(\frac{\sigma z_{\alpha/2} + \sqrt{n}\Delta}{\tilde{\sigma}}\right) \tag{2}$$

$$n = \frac{(\sigma z_{\alpha/2} + \tilde{\sigma} z_{\beta})^2}{\Delta^2}, \tag{3}$$

with $\Phi(\cdot)$ being the cdf of the standard normal distribution. Besides the weighted log-rank statistic, Z will be a Wald statistic with Δ corresponding to some measure of treatment efficacy (eg, difference in survival probability at 12 months) and $\tilde{\sigma} = \sigma$.

2.3 | Tests for comparing two survival curves

We now present various tests in survival analysis and their approximate, large-sample distributions. Detailed derivations can be found in Web Appendix A. Readers interested in a technical treatment of counting processes, martingale theory, and their application to survival data are referred to Fleming and Harrington (1991).

2.3.1 | KM survival probability and percentile

One approach to differentiating two survival curves is to compare their survival probabilities at some time t, for example, by the difference $\Delta = \overline{F}_0(t) - \overline{F}_1(t)$. The null hypothesis $H_0 : \Delta = 0$ can be tested using the Wald statistic

$$Z_{\overline{F}} = \frac{\hat{\overline{F}}_0(t) - \hat{\overline{F}}_1(t)}{\sqrt{\widehat{\mathrm{Var}}(\hat{\overline{F}}_0(t)) + \widehat{\mathrm{Var}}(\hat{\overline{F}}_1(t))}},$$

where $\hat{\overline{F}}_j(t)$ is the KM estimate and $\widehat{\mathrm{Var}}(\hat{\overline{F}}_j(t))$ is a consistent estimate for its asymptotic variance (see Greenwood's formula, for example). Having incorporated censoring and accrual into the counting processes $N_i(t)$ and $Y_i(t)$, the martingale central limit theorem can be used to show that $\sqrt{n_j}(\hat{\overline{F}}_j(t) - \overline{F}_j(t)) \xrightarrow{d} N(0, \sigma_j^2)$, where $\xrightarrow{d}$ represents convergence in distribution and $\sigma_j^2 = \overline{F}_j(t)^2 \int_0^t \lambda_j(s)/\pi_j(s)ds$. Thus, $Z_{\overline{F}} \dot{\sim} N(\sqrt{n}\Delta/\sigma, 1)$ with $\sigma^2 = \sigma_0^2/p_0 + \sigma_1^2/p_1$.

In practice, the given time t, or milestone t as it is often referred to, should be prespecified to avoid selection bias. A common choice is a fixed clinically meaningful duration of follow-up, for example, $t = 2$ years or $t = \tau_f$, the minimum period that every enrolled patient has the opportunity to be followed for. Other choices include the minimax (minimum of the two arms maximum) event time and the minimax observed time (event or censoring). Cumulative distribution functions for both random variables are available in Web Appendix A.1 to help facilitate choosing an appropriate milestone. For fixed t, investigators should check during the design stage that both survival curves will be estimable at time t. For random variable t, investigators should check that likely values of t are clinically meaningful.

For other measures comparing two survival probabilities (eg, ratio of probabilities or difference in complementary log-log), the delta method may be applied to propose a different Wald test statistic and to approximate its large-sample distribution. Likewise, the functional delta method (Andersen *et al.*, 1993) may be applied for measures comparing two KM survival percentiles such as the KM medians (Web Appendix A.2).

Fleming and Harrington understood in 1991 that accrual and censoring can be explicitly incorporated while deriving the asymptotic variance of a KM-based statistic; in their Example 6.3.1, the authors provide σ_j^2 under uniform accrual and exponential survival. However, for some reason this understanding has not translated into practice. Recent studies exploring power of KM-based tests or proposing software for

sample size calculation are all based on simulations. Part of the goal for this paper is to bridge the gap between literature and implementation of methods.

2.3.2 | KM RMST

The RMST is the mean survival time in patients followed up to time t. It is also simply the area under the survival curve up to t: $RMST_j(t) = E(T \wedge t) = \int_0^t \overline{F}_j(s)ds$. As such, the KM RMST is a robust, nonparametric summary of the survival curve that takes into account its temporal profile. Given t, let $\Delta = RMST_0(t) - RMST_1(t)$. The null hypothesis $H_0 : \Delta = 0$ can be tested using the Wald statistic

$$Z_{RMST} = \frac{\widehat{RMST}_0(t) - \widehat{RMST}_1(t)}{\sqrt{\widehat{\text{Var}}(\widehat{RMST}_0(t)) + \widehat{\text{Var}}(\widehat{RMST}_1(t))}},$$

where $\widehat{RMST}_j(t) = \int_0^t \hat{\overline{F}}_j(s)ds$ and $\widehat{\text{Var}}(\widehat{RMST}_j(t))$ can be obtained using perturbation (Zhao *et al.*, 2016). Following arguments similar to Theorem 1 in Sander (1975), $\sqrt{n_j}(\widehat{RMST}_j(t) - RMST_j(t)) \xrightarrow{d} N(0, \sigma_j^2)$, where $\sigma_j^2 = \int_0^t [\int_{s_1}^t \overline{F}_j(s_2)ds_2]^2 \lambda_j(s_1)/\pi_j(s_1)ds_1$. Thus, $Z_{RMST} \dot{\sim} N(\sqrt{n}\Delta/\sigma, 1)$ with $\sigma^2 = \sigma_0^2/p_0 + \sigma_1^2/p_1$. Unlike in Royston and Parmer (2013), this closed-form approximation accommodates general distributions of accrual and loss to follow-up.

2.3.3 | Weighted log-rank test

A popular approach in nonparametric literature is to compare entire survival curves between two groups. In this case, the null hypothesis is written as $H_0 : \overline{F}_0(\cdot) = \overline{F}_1(\cdot)$, where two survival curves are the same at every time point. The weighted log-rank statistic is proposed as a test statistic with the following form:

$$Z_{WLR} = \frac{U}{\sqrt{V}} = \frac{\sum_k a(t_k)\left(X_{(k)} - \frac{\sum Y_1(t_k)}{\sum Y_0(t_k)+\sum Y_1(t_k)}\right)}{\sqrt{\sum_k a(t_k)^2 \frac{\sum Y_0(t_k)\cdot\sum Y_1(t_k)}{(\sum Y_0(t_k)+\sum Y_1(t_k))^2}}},$$

where $a(\cdot)$ is a predictable weight function, $\{t_k\}_k$ is the complete set of event times in the trial, $X_{(k)}$ is the assigned treatment for the patient who fails at t_k, and $\sum Y_j(t_k) = \sum_{i:X_i=j} Y_i(t_k)$ is the number of patients in arm j at risk at time t_k. Letting $w(t) = \lim_{n\to\infty} a(t)$ denote the limit of $a(t)$, the Gehan-Breslow, Tarone-Ware, and Fleming-Harrington tests have $w(t)$ equal to $\pi(t)$, $\sqrt{\pi(t)}$, and $(\sum_j p_j \overline{F}_j(t))^p(1 - \sum_j p_j \overline{F}_j(t))^q$, respectively. When $a(\cdot) = 1$, Z_{WLR} reduces to the ordinary, unweighted log-rank statistic.

Although the weighted log-rank statistic does not correspond to any well-known or readily interpretable measure of treatment efficacy, the numerator can be rewritten as

$$U = \sum_k a(t_k)\frac{\sum Y_0(t_k)\sum Y_1(t_k)}{\sum Y_0(t_k) + \sum Y_1(t_k)}\left(\frac{X_{(k)}}{\sum Y_1(t_k)} - \frac{1 - X_{(k)}}{\sum Y_0(t_k)}\right),$$

which in discrete time is a weighted sum of the difference in estimated hazards. As shown in Web Appendix A.3, $\sqrt{n}(U/n - \Delta) \xrightarrow{d} N(0, \tilde{\sigma}_b^2)$, where $\tilde{\sigma}_b^2$ is some constant and

$$\Delta = \int_0^\tau w(s)\frac{p_0\pi_0(s)p_1\pi_1(s)}{\pi(s)}\{\lambda_1(s) - \lambda_0(s)\}ds.$$

Note that Δ can be similarly interpreted as a weighted average of the difference in hazards. By the Law of Large Numbers, $V/n \xrightarrow{p} \sigma^2$, where $\sigma^2 = \int_0^\tau w(s)^2 \frac{p_0\pi_0(s)p_1\pi_1(s)}{\pi(s)^2} v'(s)ds$ and $v'(s) = H(\tau_a \wedge (\tau - s))\sum_j p_j f_j(s)\overline{G}_j(s)$, the derivative of $v(\cdot)$ evaluated at s. Denoting the centered weighted log-rank statistic by $Z_{CWLR} = \sqrt{n}(U/n - \Delta)/\sqrt{V/n}$, it follows by Slutsky's theorem that $Z_{CWLR} \xrightarrow{d} N(0, \tilde{\sigma}_b^2/\sigma^2)$ and $Z_{WLR} \dot{\sim} N(\mu, \tilde{\sigma}_b^2/\sigma^2)$, where $\mu = \sqrt{n}\Delta/\sigma$.

Luo *et al.* (2019) also previously derived the asymptotic variance for U. However, their variance estimate—denoted here by $\tilde{\sigma}_s^2$—assumes patients are assigned treatment via simple randomization, that is, X_is are sampled as Bernoulli random variables with probability p_1. By contrast, $\tilde{\sigma}_b^2$ is derived under block randomization, that is, exactly $n_1 = n * p_1$ are randomized to the treatment arm while the rest are randomized to the control arm. We show in Web Appendix A.3 that $\tilde{\sigma}_b^2 \leq \tilde{\sigma}_s^2$. In Table 1, we see that $\tilde{\sigma}_b^2/\sigma^2$ is decidedly smaller than $\tilde{\sigma}_s^2/\sigma^2$ whenever HR is small and RR is imbalanced. For other more practical scenarios, however, the two variances lead to similar sample size and power calculations.

Under local alternatives, that is, $\sup_{t\leq\tau}|\log\{\lambda_1(t)/\lambda_0(t)\}| = O(n^{-1/2})$, $\tilde{\sigma}_b^2/\sigma^2$ and $\tilde{\sigma}_s^2/\sigma^2$ equal $1 + o(n^{-1/2})$. Thus, an alternative approximation for the large-sample distribution of Z_{WLR} is $N(\mu, 1)$. Previous works have similarly assumed local alternatives to simplify calculations (Schoenfeld, 1981; Barthel *et al.*, 2006; Hasegawa, 2014). One might jump to the conclusion that, under the more practical assumption of fixed alternatives (eg, PH), $N(\mu, 1)$ leads to less accurate power calculations than $N(\mu, \tilde{\sigma}_b^2/\sigma^2)$ and $N(\mu, \tilde{\sigma}_s^2/\sigma^2)$. Indeed, Luo *et al.* (2019) make this very conclusion, citing that $Z_{WLR} \xrightarrow{d} N(\mu, \tilde{\sigma}_b^2/\sigma^2)$. However, convergence in distribution for Z_{WLR} itself requires the assumption of local alternatives (see proof in Web Appendix A.3). Under fixed alternatives, $\tilde{\sigma}_b^2/\sigma^2$, $\tilde{\sigma}_s^2/\sigma^2$, and 1 may all serve as approximations for the large-sample variance of Z_{WLR}, but none of them are the limiting variance of Z_{WLR}. Consequently, there is no guarantee that one is always more accurate than the others.

TABLE 1 Comparing $\tilde{\sigma}_b^2/\sigma^2$ and $\tilde{\sigma}_s^2/\sigma^2$ to the empirical variance of Z_{CWLR} and Z_{WLR} from 10 000 time-driven clinical trials simulated under various randomization ratios (RR), randomization schemes (RS), and hazard ratios (HR)

RR	RS	HR	$\tilde{\sigma}_b^2/\sigma^2$	$\tilde{\sigma}_s^2/\sigma^2$	Var $[Z_{CWLR}]$	Var $[Z_{WLR}]$
1:2	Block	0.80	1.004	–	1.012	0.981
	Simple	0.80	–	1.008	1.004	0.970
	Block	0.67	0.978	–	0.975	0.946
	Simple	0.67	–	0.987	0.979	0.945
	Block	0.33	0.782	–	0.771	0.807
	Simple	0.33	–	0.813	0.818	0.855
1:1	Block	0.80	0.980	–	1.005	1.025
	Simple	0.80	–	0.980	0.975	0.993
	Block	0.67	0.945	–	0.943	0.996
	Simple	0.67	–	0.945	0.945	0.998
	Block	0.33	0.777	–	0.767	0.941
	Simple	0.33	–	0.788	0.800	0.969
2:1	Block	0.80	0.960	–	0.965	1.030
	Simple	0.80	–	0.960	0.967	1.028
	Block	0.67	0.921	–	0.927	1.057
	Simple	0.67	–	0.940	0.948	1.061
	Block	0.33	0.797	–	0.811	1.166
	Simple	0.33	–	0.956	0.985	1.182

Note. It is assumed that 9000 patients are enrolled uniformly over 14 months, assigned to active and control arm 1:2, 1:1, or 2:1 via block or simple randomization, and followed for an additional 11 months; that control patients have exponential survival with median 6 months; and that, ignoring the risk of death, all patients are subject to loss of follow-up at an exponential rate of ~ 0.0004, equivalent to 1% every 25 months.

Table 1 illustrates this subtle point. We see that $\tilde{\sigma}_b^2/\sigma^2$ and $\tilde{\sigma}_s^2/\sigma^2$ are necessarily good estimates for the large-sample, empirical variance of Z_{CWLR} from trials simulated under PH and block or simple randomization. However, they are not as good of an estimate for the empirical variance of Z_{WLR}, which is more relevant for accurate sample size and power calculations. In fact, under 1:1 or 2:1 RR, the variance of Z_{WLR} is better estimated by 1.

Using a Taylor series expansion, $\{\lambda_1(s) - \lambda_0(s)\}\pi(s)$ in Δ may be replaced by $\log\{\frac{\lambda_1(s)}{\lambda_0(s)}\}v'(s)$, resulting in $\Delta_{GS} = \int_0^\tau w(s)\log\{\frac{\lambda_1(s)}{\lambda_0(s)}\}\frac{p_0\pi_0(s)p_1\pi_1(s)}{\pi(s)^2}v'(s)ds$. ($\Delta_{GS}$ generalizes Equation (2) in Schoenfeld (1981) by accommodating for administrative censoring. It will not be discussed further due to poor performance.) If in addition $a(\cdot) = w(\cdot) = 1$, $\log\{\lambda_1(s)/\lambda_0(s)\} = \theta$, $G_j(s) = G(s)$, and $\pi_j(s)$ is replaced by $\pi(s)$—its limit under local alternatives, then $\Delta_{ED} = \theta p_0 p_1 v(\tau)$ and $\sigma^2 = p_0 p_1 v(\tau)$. Plugging $N(\mu_{ED}, 1) \equiv N(\sqrt{n}\Delta_{ED}/\sigma, 1)$ into Equation (3) and setting $d = \lceil nv(\tau)\rceil$, one arrives at Equation (1). Schoenfeld's equation therefore depends on the normal approximation $N(\mu_{ED}, 1)$ derived for the unweighted log-rank test under PH assumption. Its popularity has led to the perception that the number of events, not patients, is the effective sample size in survival studies.

TABLE 2 Comparing μ and μ_{ED} to the empirical mean of Z_{WLR} from 10 000 time-driven clinical trials employing block randomization and simulated under various randomization ratios (RR) and hazard ratios (HR)

RR	HR	μ (%-error)	μ_{ED} (%-error)	$E[Z_{WLR}]$
1:2	0.80	−9.015 (0.16%)	−9.133 (1.15%)	−9.029
	0.67	−15.979 (0.03%)	−16.398 (2.58%)	−15.985
	0.33	−38.341 (0.02%)	−42.169 (10.01%)	−38.332
1:1	0.80	−9.622 (0.07%)	−9.625 (0.04%)	−9.629
	0.67	−17.155 (0.01%)	−17.173 (0.09%)	−17.157
	0.33	−42.280 (0.03%)	−42.832 (1.33%)	−42.269
2:1	0.80	−9.132 (0.01%)	−9.016 (1.28%)	−9.132
	0.67	−16.394 (0.02%)	−15.981 (2.54%)	−16.397
	0.33	−41.834 (0.02%)	−38.514 (7.96%)	−41.842

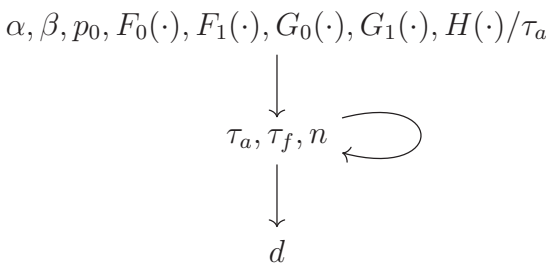
Note. Study assumptions are identical to those in Table 1. Percent-error is calculated as $|\text{estimate} - \text{empirical}|/\text{empirical} \times 100$.

In Table 2, we compare μ and μ_{ED} to the empirical mean of the unweighted log-rank statistic under PH. μ is notably more accurate. Differences in precision between the means are most evident when HR is small or RR is imbalanced. Together with observations from Table 1, we might anticipate $N(\mu, 1)$ to be a good approximation for practical use. In Section 3, we perform extensive simulations under PH and NPH to compare power calculations based on $N(\mu, 1)$, $N(\mu, \tilde{\sigma}_b^2/\sigma^2)$, and $N(\mu_{ED}, 1)$.

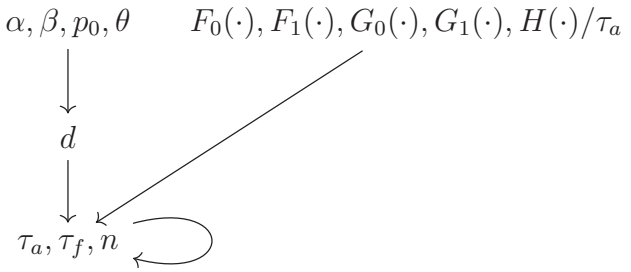
2.4 | Event-driven trials

Thus far, we have assumed that final analysis for the clinical trial is triggered at $\tau = \tau_a + \tau_f$ (unit of time) after the first patient is enrolled. τ_a and τ_f are known and we calculate the required sample size n or power $1 - \beta$ given the other. Of course, different values of τ_a and τ_f can be tried in practice, this process repeated until a desirable balance is struck between accrual rate, trial duration, sample size, and power. In this case of a time-driven analysis, the number of events at the end of the trial is not fixed, but can be estimated by the expected number of events $d = \lceil nv(\tau)\rceil$. Should an event-driven analysis be desired, the required number of events can be set as $d = \lceil nv(\tau)\rceil$. Study duration then becomes a random variable whose expectation can be approximated by τ. Even though d and τ are typically prespecified in clinical trial protocols, we point out this nuance, that depending on whether the study is time-driven or event-driven, exactly one of these two parameters will be fixed and the other will be an estimate of the eventual value.

We call the above sequence of calculating sample and event size the n-d method because, contrary to the d-n method, n is calculated before d. The n-d method has been previously applied in the context of the weighted/unweighted log-rank test (Hasegawa, 2014; Luo *et al.*, 2019). Here, we formalize the method and extend its application to other nonparametric tests. Schematically, the n-d method is as follows:



As mentioned in the introduction, current practice for sample size determination is to take a reverse approach by calculating d before n. Specifically, the d-n method uses Equation (1) to calculate d. Given $F_0(\cdot)$, $F_1(\cdot)$, $G_0(\cdot)$, $G_1(\cdot)$ and the scaled accrual distribution $H(\cdot)/\tau_a$, a desirable combination of τ_a, τ_f, and n is then chosen satisfying $\lfloor n\nu(\tau) \rfloor = d$. Schematically, the d-n method can be depicted as follows:



But again, sample size calculation based on Equation (1) translates to assuming PH and using $N(\mu_{ED}, 1)$ to approximate the large-sample distribution of the unweighted log-rank statistic. If $N(\mu, 1)$ or $N(\mu, \tilde{\sigma}_b^2/\sigma^2)$ is used instead, or if any other nonparametric test (eg, weighted log-rank test, difference in RMST) is used under PH or NPH, then it is no longer possible to use the d-n method for sample size and power calculation. In any of these cases, the n-d method should be used. We show in Section 3.1 that even when the d-n method is applicable, using the n-d method with $N(\mu, 1)$ results in more precise sample size and power calculations.

2.5 | Implementation

Results in Section 2.3–2.4 allow efficient sample size and power calculations for common nonparametric tests of survival under flexible, continuous distributions of accrual, survival, and loss to follow-up. To make these results accessible to practitioners, we have created an R package `npsurvSS` that incorporates the following distributions:

- accrual—piecewise uniform, truncated exponential (Lachin and Foulkes, 1986);
- survival—Weibull, piecewise exponential, mixture cure model (Berkson and Gage, 1952);
- loss to follow-up—Weibull.

These distributions cover a wide range of scenarios, including increasing/decreasing accrual, delayed treatment effect, crossing survival curves, and populations with cured and uncured patients. Details of each distribution can be found in Web Appendix B. For the weighted log-rank test, practitioners may choose to use the approximating distributions $N(\mu, 1)$, $N(\mu, \tilde{\sigma}_b^2/\sigma^2)$, $N(\mu, \tilde{\sigma}_s^2/\sigma^2)$, or $N(\mu_{GS}, 1) \equiv N(\sqrt{n}\Delta_{GS}/\sigma, 1)$. For the unweighted log-rank test under PH, practitioners have the additional option of using Schoenfeld's $N(\mu_{ED}, 1)$.

Besides providing functions that calculate sample size and power, `npsurvSS` can be used to calculate expected study duration or number of events depending on whether the trial is event-driven or time-driven. It also facilitates visualization of assumed distributions, selection of milestone t for KM survival or RMST, and simulation of two-arm clinical trials.

Several other sample size programs are currently available for nonparametric tests comparing two survival curves. A list of programs and their capabilities is provided in Web Table 1. We found that most programs provide only calculations for the unweighted or weighted log-rank test. Of these programs, the `PWEALL` package in R by Luo *et al.* (2019) and the ART module in Stata by Barthel *et al.* (2006) are the most advanced. Both allow users to specify nonuniform accrual, piecewise exponential survival, and flexible loss to follow-up. Where they differ is in the distributions they use to approximate the large-sample distribution of Z_{WLR}. `PWEALL` gives the option of using $N(\mu, 1)$, $N(\mu, \tilde{\sigma}_s^2/\sigma^2)$, or $N(\mu_{ED}, 1)$. ART's so-called "local alternative" option uses $N(\mu_{ART}, 1) \equiv N(\sqrt{n}\Delta_{ART}/\sigma, 1)$, where the variance of 1 is derived under local alternatives and Δ_{ART} is another approximation for the mean of U/n derived under fixed alternatives (Web Appendix A.3). ART also has a "distant alternative" option which approximates the asymptotic variance of U by performing simulations.

The R package `ssrmst` provides calculations for the KM RMST, but only under uniform accrual, Weibull survival, and no loss to follow-up. Its calculations are also simulation-based and therefore computationally intensive. We are not aware of any software offering sample size and power calculations for the KM survival or KM percentile. This is likely due to the fact that trials in the past rarely, if ever, used these test statistics

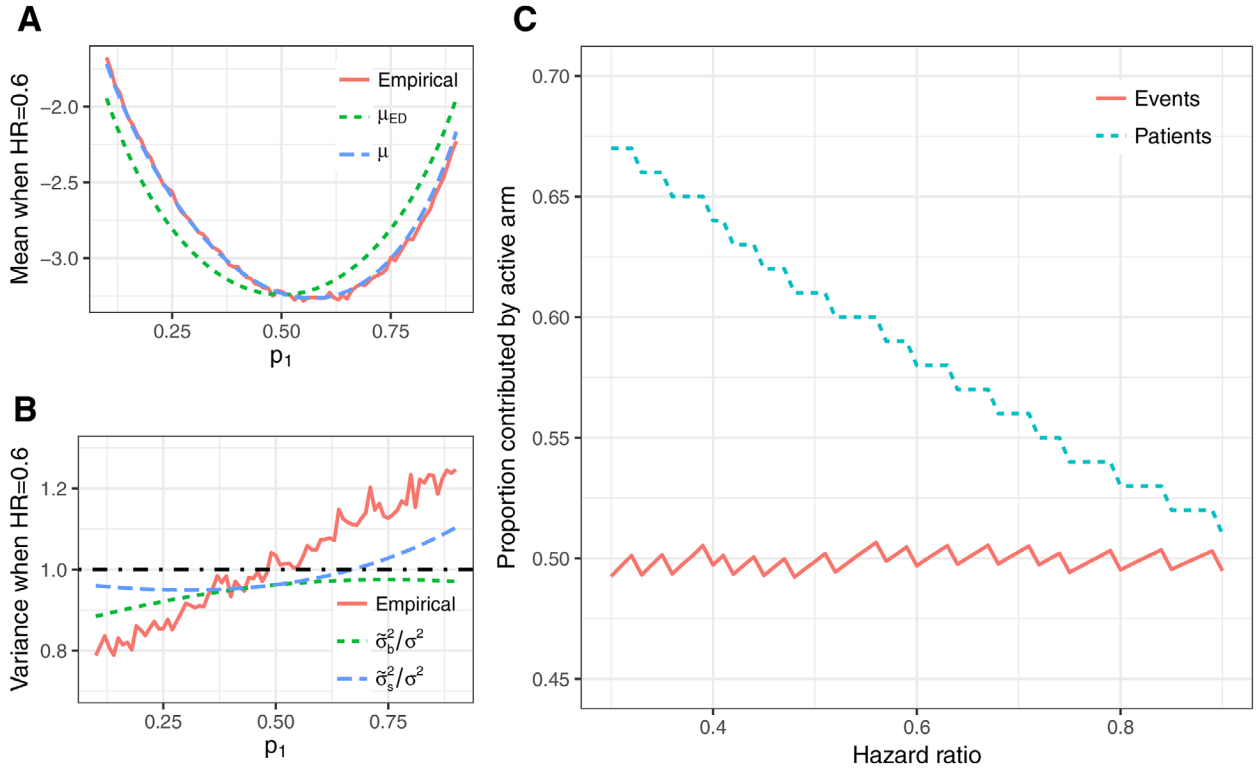


FIGURE 1 Search for the proportion p_1 of patients that should be randomized to the active arm in order to maximize power in event-driven trials under proportional hazards. *Note*. Trial assumptions including number of patients, patient accrual, loss to follow-up, and control median survival are detailed in Web Appendix C. (A–B) Empirical versus estimated mean and variance of the unweighted log-rank statistic, for the trial with HR = 0.6; (C) expected proportion of events and patients contributed by the active arm when the optimal p_1 based on log-rank approximation $N(\mu, 1)$ is adopted. Results suggest that more patients should receive the beneficial treatment so that number of events at trial completion is balanced across arms. HR, hazard ratio. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

for formal comparison. Thus, `npsurvSS` is the most flexible, robust, and efficient option to date.

3 | EXAMPLES

3.1 | Optimal RR under PH

It is well known for continuous endpoints that equal allocation optimizes statistical power (Meinert, 1986). The same would seem to be true for survival endpoints under PH: the unweighted log-rank test is the most efficient test in this case and its power according to Equation (1) is maximized when $p_1 = 0.5$. However, as mentioned earlier, Equation (1) approximates the log-rank distribution using $N(\mu_{ED}, 1)$, which is derived under a number of assumptions. The n-d method allows us to use alternative approximations for improved sample size and power calculations. We now compare accuracy between various power calculations while exploring the question of optimal RR under PH.

Consider a series of event-driven clinical trials with HR = 0.30, 0.31, …, 0.89, 0.90. To make each trial realistic, we calculate the required number of events d using Equation (1) with $p_1 = 0.5$, $\alpha = 0.05$, and $\beta = 0.1$. We define the ratio d/n of events to patients, the control arm median m, and accrual duration τ_a as increasing functions of HR taking on values between 0.6–0.7, 0.5–2, and 1–4 years, respectively (Web Appendix C). As a result of these parameters, minimum duration of follow-up under equal allocation is $\geq m$ and the total study duration is between 1.5 and 5.5 years. The impact of RR is assessed by varying p_1 while fixing d, n, m, and τ_a. We compare power calculated using the n-d to the d-n method, that is, $N(\mu, 1)$, $N(\mu, \tilde{\sigma}_b^2/\sigma^2)$, or $N(\mu_{ED}, 1)$ to approximate the distribution of the log-rank statistic.

Figure 1A,B shows the empirical and estimated log-rank mean and variance for the trial design with HR = 0.6. While the empirical mean closely follows μ, the empirical variance can be under- or over-estimated by $\tilde{\sigma}_b^2/\sigma^2$. If one uses instead the simpler variance of 1, then maximizing power under $N(\mu, 1)$ is equivalent to minimizing μ. In this case, it would suggest randomizing 151 of the total 264 patients to the active arm ($p_1 = 0.57$).

Figure 1C looks across all the designs and provides the proportion of patients and expected events in the active arm when the optimal RR based on $N(\mu, 1)$ is adopted, keeping other design parameters d, n, m, and τ_a fixed. Results suggest that

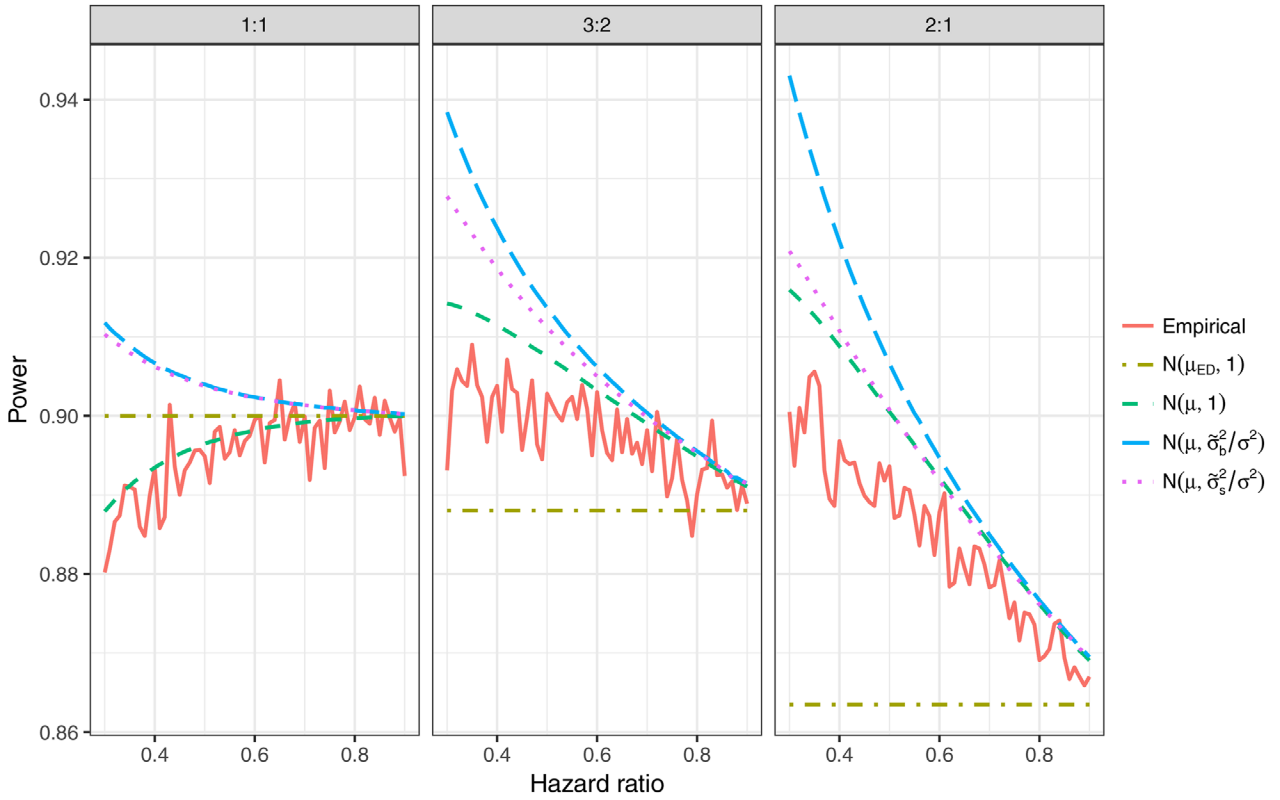


FIGURE 2 Empirical versus estimated power of the unweighted log-rank test, under the same setup as Figure 1 but with randomization ratio restricted to 1:1, 3:2, and 2:1 in favor of the active arm. *Note*. Power estimates are calculated using one of four normal approximations for the log-rank distribution. 3:2 and 2:1 may be attractive options whenever the true hazard ratio is less than 0.7 and 0.5, respectively. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

power is maximized when more patients are randomized to the active arm so that expected events is balanced across arms.

In practice, the optimal RR might be a decimal number that is difficult to implement. A conventional ratio such as 3:2 and 2:1 might be preferable for operational convenience. Figure 2 compares power under 1:1, 3:2, and 2:1. $N(\mu, 1)$ consistently leads to the most accurate power calculations, though there is still room for improvement. It is worth noting the difference between $N(\mu, 1)$ and $N(\mu_{ED}, 1)$. $N(\mu_{ED}, 1)$ suggests that randomizing patients 3:2 and 2:1 decreases power by 1.2% and 3.7% for all HRs, whereas $N(\mu, 1)$ suggests that the power loss is much less. For highly effective treatments, there may even be power gain.

We explored additional scenarios by varying n, m, and τ_a (not shown). Based on our results, we encourage investigators to consider 3:2 and 2:1 when the design HR is less than 0.7 and 0.5, because power will be similar to or higher than 1:1. Of course, other considerations need to be factored into the decision-making process. For example, in order to achieve 90% power in the above design with HR = 0.6, randomizing the 264 patients 3:2 instead of 1:1 would extend trial duration from ~41 months to ~42 months. Investigators need to consider the potential trade-offs between statistical power, operational convenience, and study timeline.

We also explored some scenarios to compare `npsurvSS` with ART (Web Table 2). Power calculated using $N(\mu, 1)$ was slightly more accurate than $N(\mu_{ART}, 1)$ and always within 1% of the empirical power. Perhaps more importantly, $N(\mu, 1)$ and $N(\mu_{ART}, 1)$ consistently led to more accurate power calculations than $N(\mu, \tilde{\sigma}_b^2/\sigma^2)$ and Barthel's distant alternative approach. Thus, contrary to previous thought, estimating the asymptotic variance of U/n under nonlocal alternatives does not necessarily improve power calculations.

3.2 | Pressure testing CheckMate 141

CM-141 was a randomized phase 3 trial comparing nivolumab to standard therapy in patients with head and neck squamous cell carcinoma (Ferris *et al.*, 2016). According to its protocol, 278 deaths would be needed to achieve 90% power under HR = 0.67 and using the log-rank test with two-sided $\alpha = 0.05$. Assuming uniform accrual, no loss to follow-up, and exponential survival with 9- and 6-month medians, the trial planned to enroll 360 patients (2:1 RR) and projected 25-month study duration (14 accrual, 11 follow-up).

CM-141 eventually observed median survivals close to the protocol assumptions: 7.5 and 5.1 months. However, the

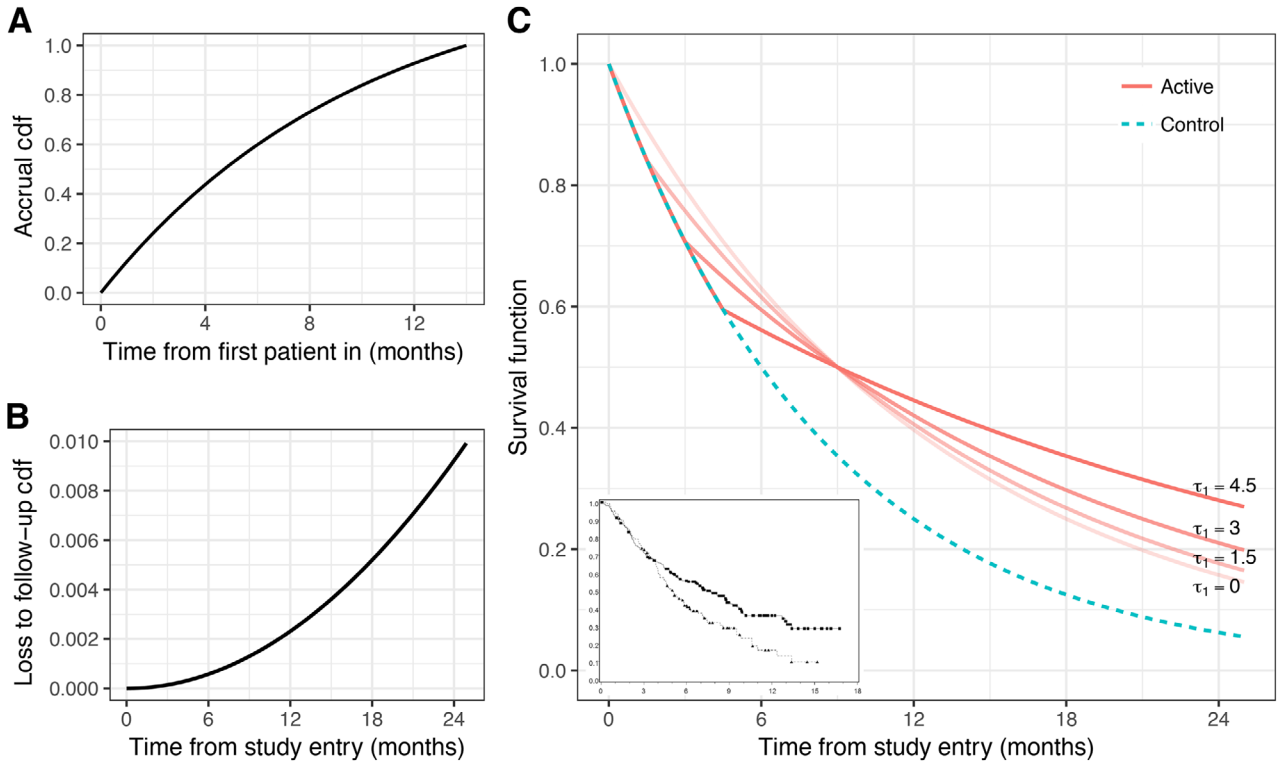


FIGURE 3 Example of a trial with delayed treatment effect, motivated by CheckMate 141. (A) Accrual following a truncated exponential distribution; (B) loss to follow-up following a Weibull distribution; (C) survival distributions with various degrees of delay in treatment effect (control = exponential with 6-month median; active = piecewise-exponential with 9-month median and change point at $\tau_1 \in \{0, 1.5, 3, 4.5\}$ months). Superimposed on the bottom-left corner are the observed overall survival curves from CheckMate 141 (Ferris *et al.*, 2016). cdf, cumulative distribution function. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

assumption of exponential survival distributions was notably violated: the survival curves overlap in the early months and separate only after 4 months (Figure 3C). We find these results interesting because whether from literature search or clinical experience, investigators often have a sense of the median survival. On the other hand, exponential survival is assumed for sake of simplicity and without strong prior evidence. This assumption may not be appropriate in trials comparing drugs with different mechanisms. Therefore, we might ask ourselves, "Under the original protocol design, what would have been the impact on power and timeline if delayed treatment effect was anticipated?"

To answer that question, suppose the same assumptions as CM-141 hold except that survival in the active arm follows a two-piece exponential distribution with change point τ_1 to characterize the delayed effect. Also, accrual and censoring are assumed to follow more flexible distributions (Figure 3). Figure 4A shows the power for the unweighted log-rank test (and other nonparametric tests for completeness) as a function of τ_1 when final analysis is conducted at 278 events. The corresponding study duration is depicted in Figure 4B.

The protocol-specified unweighted log-rank test is not always the most powerful, but its power is always above 90%. In fact, power increases as the change point is delayed further. This might seem surprising. We typically think delayed separation results in a loss of power. This would be true if the relative effect after separation was fixed. But since the median survivals were fixed at 9 and 6 months, HR after τ_1 decreases as τ_1 increases, counteracting the lengthened period of no treatment benefit. For instance, when $\tau_1 = 3$, HR after the change point is equal to 0.333. This is substantially smaller than the HR 0.667 when $\tau_1 = 0$.

Comparing all tests, the Fleming-Harrington weighted log-rank test with $p = 1$ and $q = 1$ is the most powerful test when $\tau_1 \geq 0.5$. Difference in 11-month survival has moderately less power (0–15%) than the unweighted log-rank test. Difference in 11-month RMST and difference in median survival suffer even greater losses of power (14–80%). The impact on timeline is minimal when $\tau_1 \leq 4$. When $\tau_1 > 4$, a time-driven analysis at 25 months may be preferable. The resulting power curves shown in Figure 4C are similar to those in Figure 4A except for the Gehan-Breslow weighted log-rank test. Although the Gehan-Breslow test is less powerful than other weighted log-rank tests because it upweights earlier periods of the survival curve, it stands to benefit more from an event-driven trial where study duration is prolonged and where there is a clear distinction of treatment benefit in the tail.

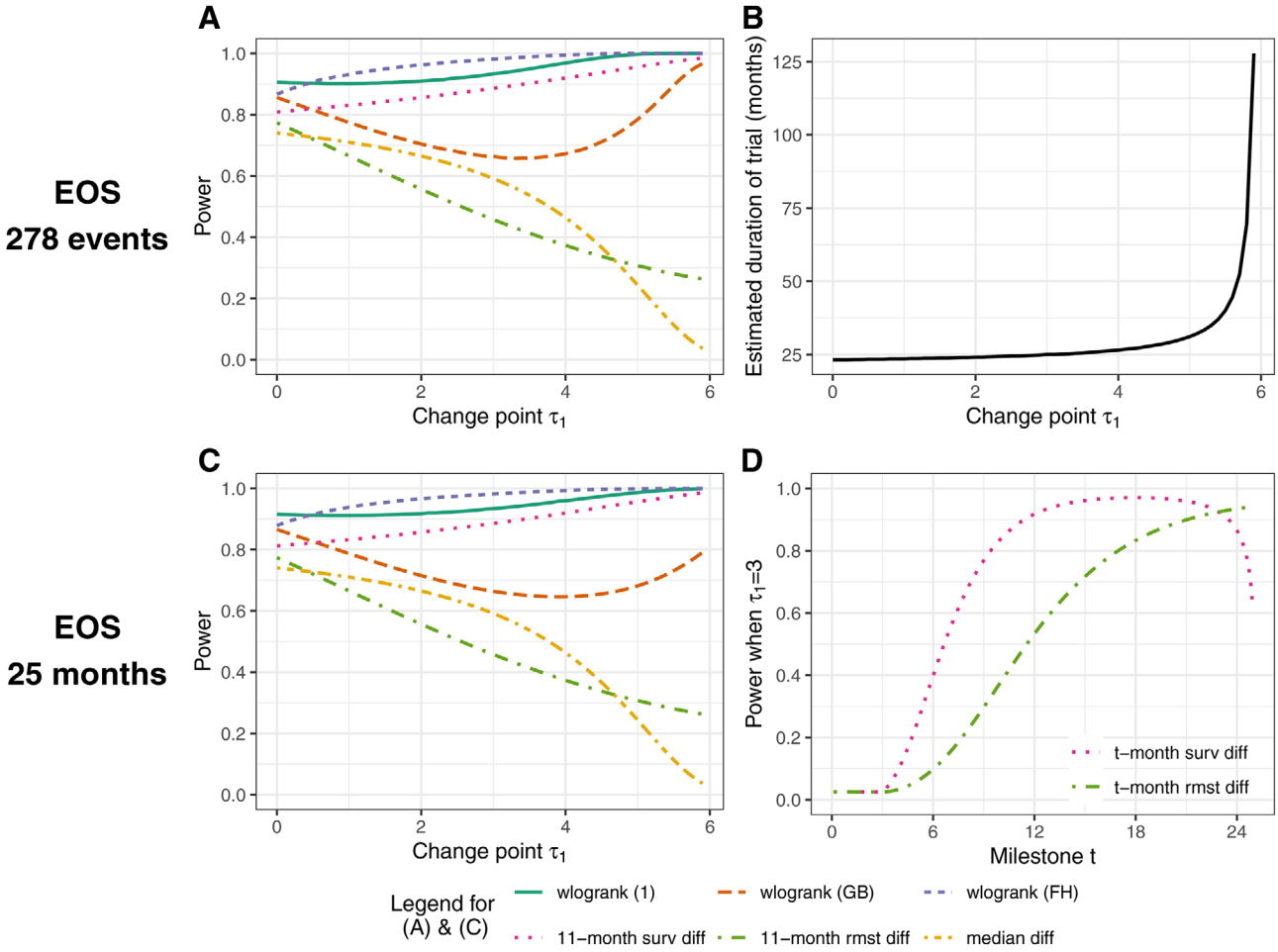


FIGURE 4 Power for various nonparametric tests in the example trial of Figure 3. *Note*. It is assumed throughout that 240 and 120 patients are enrolled to the active and control arm, respectively. Subparts (A) and (B) apply when the end of study (EOS) is triggered by 278 events. Subparts (C) and (D) apply when EOS is instead triggered at 25 months after the first-patient-in. (A) Power in the event-driven trial, under increasing delays in treatment effect and based on the unweighted log-rank test "wlogrank (1)," log-rank test with Gehan-Breslow weights "wlogrank (GB)," log-rank test with Fleming-Harrington weights $p = 1$ and $q = 1$ "wlogrank (FH)," difference in 11-month survival probability "11-month surv diff," difference in 11-month RMST "11-month rmst diff," or difference in median survival "median diff." (B) Estimated duration of the event-driven trial. (C) Power as in (A), but in the time-driven trial. (D) Power in the time-driven trial when $\tau_1 = 3$, specifically for the difference in t-month survival and RMST evaluated at various milestones t. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

Figure 4D suggests that using an 11-month milestone for RMST analysis throws away substantial information of clinical and statistical relevance. A milestone at a later time—where some but not all patients are followed up to—can increase power greatly.

Simulations were performed to compare power calculations in Figure 4 to the empirical power (Web Appendix D). In general, calculations were accurate within 1–2%, with the exception of the difference in KM median survival which can be unstable due to the small number of patients at the tail end of a survival curve. Considering the half-width of the 95% confidence interval is about 1.3% for an estimate of 90% power based on 2000 simulations, our results suggest that calculations based on asymptotic theory provide adequate accuracy for practical application. Lastly, weighted log-rank calculations in Figure 4 were based on the approximation $N(\mu, 1)$. Using $N(\mu, \tilde{\sigma}_b^2/\sigma^2)$ did not increase accuracy in this example.

4 | DISCUSSION

We have proposed sample size and power equations for common nonparametric tests in survival analysis. For KM-based tests, equations or software under such flexible and practical settings were previously unavailable. For the weighted log-rank test, we derived and compared multiple normal approximations for its large-sample distribution. Schoenfeld's approximation $N(\mu_{ED}, 1)$ works well for the unweighted test when randomization is balanced and HR is a constant >0.5.

However, if either condition is not met, then it can lead to inaccurate sample size and power calculations. We saw in Section 3.1 that $N(\mu_{ED}, 1)$ wrongfully concludes that 1:1 maximizes power, when in fact power can be gained by randomizing more patients to the active arm. $N(\mu, \tilde{\sigma}_b^2/\sigma^2)$ offers a viable alternative under block randomization that can also be applied for general, weighted log-rank tests under NPH. But the effort required to calculate $\tilde{\sigma}_b^2/\sigma^2$ does not guarantee better results compared to using the traditional variance of 1. Based on extensive simulations under PH and NPH, our final recommendation to practitioners is to use the simpler—yet still flexible and accurate—approximation $N(\mu, 1)$.

Our companion software `npsurvSS` provides a comprehensive tool for trial design. We imagine that it can be used as follows: (1) "Compare and Select." `npsurvSS` can be used to compare potential study designs. Designs can differ, for example, in their RR, duration of accrual and follow-up, or primary statistical test. (2) "Pressure test." Once a design is chosen, it can be evaluated under plausible deviations from design assumptions to ensure that power is adequately robust, for example, deviations with respect to the accrual, censoring, or survival distributions (Section 3.2). (3) "Simulate." Finally, if utmost precision is desired, then simulations can be performed to fine-tune estimates of sample size and power.

Some questions remain regarding the practical use of the present theory. What is the minimum sample size required for Equations (2)–(3) to be accurate? In Section 3, calculations based on the approximation $N(\mu, 1)$ for the unweighted log-rank test consistently estimated power within 1–2% with as few as 50 patients. But what about KM-based tests? While we do not recommend using large-sample theory to perform calculations for difference in percentile, simulation results suggest that the rule of thumb "$n_0 \times \pi_0(t) \geq 5$ and $n_1 \times \pi_1(t) \geq 5$" (motivated by the rule "$np \geq 5$ and $n(1-p) \geq 5$" for the normal approximation to binomial data) may be appropriate for difference in t-month survival (Web Table 5).

Should trials be event-driven, time-driven, or both? For the unweighted log-rank test, event-driven trials are robust under PH, but can be under- or over-powered under NPH. Time-driven trials may be more suitable for comparing differences in t-month survival or RMST. However, investigators should keep in mind that power may be greater at some milestone t larger than the minimum duration of follow-up τ_f, even though not all patients will have the opportunity to be followed up to the milestone. Analytical results in Web Appendix A can be useful for choosing an appropriate t.

Finally, this work considered single-stage trials, but how might the present theory be applied to the increasingly popular group sequential or adaptive trials? More generally, what tests and study designs can be used to address challenges arising from NPH, while keeping connected to clinically meaningful effect estimates? These are all questions that warrant further research and that, if answered, can improve future practice and trial design.

ACKNOWLEDGMENTS

We would like to thank a co-editor, associate editor, and two anonymous referees for constructive comments, which improved our manuscript.

ORCID

Godwin Yung https://orcid.org/0000-0001-8032-0458

REFERENCES

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York, NY: Springer.

Barthel, F.M.-S., Babiker, A., Royston, P. and Parmar, M.K.B. (2006) Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and crossover. *Statistics in Medicine*, 25, 2521–2542.

Berkson, J. and Gage, R.P. (1952) Survival cure for cancer patients following treatment. *Journal of the American Statistical Association*, 47, 501–515.

Chen, T.T. (2015) Milestone survival: a potential intermediate endpoint for immune checkpoint inhibitors. *Journal of National Cancer Institute*, 107. Available at: https://doi.org/10.1093/jnci/djv156.

Collett, D. (2015) *Modelling Survival Data in Medical Research*. Boca Raton, FL: Chapman & Hall.

Ferris, R.L., Blumenschein Jr, G., Fayette, J., Guigay, J., Colevas, A.D., Licitra, L., *et al.* (2016) Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *New England Journal of Medicine*, 375, 1856–1867.

Fine, G.D. (2007) Consequences of delayed treatment effects on analysis of time-to-event end points. *Therapeutic Innovation & Regulatory Science*, 41, 535–539.

Fleming, T.R. and Harrington, D.P. (1991) *Counting Processes and Survival Analysis*. Hoboken, NJ: John Wiley & Sons.

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, 128–135.

Hoering, A., Durie, B., Wang, H. and Crowley, J. (2016) End points and statistical considerations in immuno-oncology trials: impact on multiple myeloma. *Future Oncology*, 13, 1181–1193.

Lachin, J.M. and Foulkes, M.A. (1986) Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, 42, 507–519.

Luo, X., Mao, X., Chen, X., Qiu, J., Bai, S. and Quan, H. (2019) Design and monitoring of survival trials in complex scenarios. *Statistics in Medicine*, 38, 192–209.

Meinert, C.L. (1986) *Clinical Trials Designs, Conduct, and Analysis*. New York, NY: Oxford University Press.

Pak, K., Uno, H., Kim, D.H., Tian, L., Kane, R.C., Takeuchi, M., *et al.* (2017) Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. *JAMA Oncology*, 3, 1692–1696.

Royston, P. and Parmar, M.K.B. (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, 152.

Sander, J.M. (1975) *Asymptotic normality of linear combinations of functions of order statistics with censored data*. Stanford, CA: Division of Biostatistics, Stanford University. Technical report number: 8.

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

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., et al. (2014) Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32, 2380–2385.

Wang, S., Zhang, J. and Lu, W. (2011) Sample size calculation for the proportional hazards cure model. *Statistics in Medicine*, 31, 3959–3971.

Zhao, L., Clagget, B., Tian, L., Uno, H., Pfeffer, M.A., Solomon, S.D., et al. (2016) On the restricted mean survival time curve in survival analysis. *Biometrics*, 72, 215–221.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2–3 are available with this paper at the Biometrics website on Wiley Online Library. The R package `npsurvSS` is available for download on CRAN and contains 3 vignettes illustrating its use.

How to cite this article: Yung G, Liu Y. Sample size and power for the weighted log-rank test and Kaplan-Meier based tests with allowance for nonproportional hazards. *Biometrics*. 2020;76:939–950. https://doi.org/10.1111/biom.13196