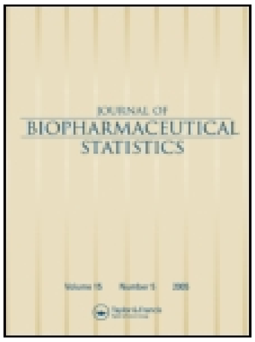

JOURNAL OF
BIOPHARMACEUTICAL
STATISTICS

Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

Some Issues of Sample Size Calculation for Time-to-Event Endpoints Using the Freedman and Schoenfeld Formulas

Ulrich R. Abel, Katrin Jensen, Irini Karapanagiotou-Schenkel & Meinhard Kieser

To cite this article: Ulrich R. Abel, Katrin Jensen, Irini Karapanagiotou-Schenkel & Meinhard Kieser (2015) Some Issues of Sample Size Calculation for Time-to-Event Endpoints Using the Freedman and Schoenfeld Formulas, Journal of Biopharmaceutical Statistics, 25:6, 1285-1311, DOI: 10.1080/10543406.2014.1000546

To link to this article: https://doi.org/10.1080/10543406.2014.1000546

Published online: 30 Jul 2015.

Journal of Biopharmaceutical Statistics, 25: 1285–1311, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2014.1000546

SOME ISSUES OF SAMPLE SIZE CALCULATION FOR TIME-TO-EVENT ENDPOINTS USING THE FREEDMAN AND SCHOENFELD FORMULAS

Ulrich R. Abel[1], Katrin Jensen[2], Irini Karapanagiotou-Schenkel[1], and Meinhard Kieser[2]
[1]*National Center for Tumor Diseases Heidelberg, Heidelberg, Germany*
[2]*Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany*

This article deals with seven special issues related to the assumptions, applicability, and practical use of formulas for calculating power or sample size, respectively, for comparative clinical trials with time-to-event endpoints, with particular focus on the well-known Freedman and Schoenfeld methods. All problems addressed are illustrated by numerical examples, and recommendations are given on how to deal with them in the planning of clinical trials.

Key Words: Controlled clinical trial; Freedman formula; Sample size; Schoenfeld formula; Time-to-event data.

1. INTRODUCTION

In the past 40 years, with the rise of Phase III clinical trials, a vast amount of research has been done on methods for calculating power and sample size, respectively, for comparative clinical trials with time-to-event endpoints (see the overviews of Oellrich et al., 1997; Rogon, 2009; Ryan, 2013). These methods cover a wide range of assumptions for the planning of clinical trials and are of varying complexity. In this article, we focus on the methods and formulas devised by Freedman (1982) and Schoenfeld (1981, 1983), which have particularly simple assumptions and are arguably the tools most widely used in practice. The use of these standard sample size methods is deceptively clear and straightforward, but, as we will show, if care is not taken to avoid several issues and pitfalls, large errors can be incurred.

In Section 2, the assumptions and notation used throughout the article are introduced, the Freedman and the Schoenfeld method are briefly reviewed, and the simulation studies we performed are described. Section 3 gives the results of our investigations that consider the following seven topics: (1) consequences of small deviations from the distributional assumptions, (2) consequences of failing to account for mixed populations with a cured subgroup, (3) consequences of using approximations of unknown survival distributions, (4) consequences of using a special implementation of Freedmans formula,

Received March 10, 2014; Accepted October 17, 2014
Address correspondence to Katrin Jensen, Institute of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg, Germany; E-mail: jensen@imbi.uni-heidelberg.de

(5) comparison of the properties and performance characteristics of the Freedman and the Schoenfeld formula, (6) considerations on measuring the treatment effect by looking at the survival difference at a particular time point, and (7) need to take random variations and systematic changes of the accrual into account. We conclude with a brief discussion in Section 4.

2. METHODS

2.1. Assumptions

- comparison of two groups (A vs. B) using the log-rank test (two-sided testing, significance level α; in our examples we use $\alpha = 5\%$)
- one analysis only, taking place at the end of the planned follow-up period
- equal group sizes (Note that some considerations regarding the power of the log-rank test for unbalanced designs can be found in Hsieh, 1992).
- uniform patient accrual in a given time interval $[0, a]$
- no "dropouts", i.e., all patients are observed until the time of analysis or an event (whichever occurs first).

Remarks

1. Uniform accrual during a fixed time interval $[0, a]$ is an assumption often made in the literature (including Schoenfeld, 1981, 1983; see the overviews by Oellrich et al., 1997; Rogon, 2009). One alternative sometimes considered (and accommodated by our simulation programs) is to model patient accrual as a Poisson process with expected duration a. In some instances we will compare our results with those obtained when using a Poisson process for modeling accrual.
2. A fixed follow-up period, resulting in type I censoring, reflects the practice of clinical trial protocols as we have encountered them in our professional work. Also, it is an assumption widely adopted in methodological papers (including Freedman, 1982; Schoenfeld, 1981, 1983; see Oellrich et al., 1997; Rogon, 2009). One Reviewer pointed out that since the power of the log-rank statistic is not determined by the sample size but by the number of events observed, the time of analysis should be determined by the number of events observed. Such designs have been studied, e.g., by Case and Morgan (2001) and Xiong et al. (2003). We incorporated this event-oriented follow-up (type II censoring) in our methodological framework. Some results obtained with this alternative methodology will be given and discussed.

2.2. Notation

a	duration of accrual
f	duration of follow-up after the recruitment of the last recruited study patient
N, n	total sample size, sample size per group
$S(t)$, $S_A(t)$, $S_B(t)$	survivor function (of the total study cohort, standard therapy group A and experimental therapy group B, respectively)
$h(t)$, $h_A(t)$, $h_B(t)$	hazard functions (of the total study cohort, standard therapy group A and experimental therapy group B, respectively)

P_e, $P_{e,A}$, $P_{e,B}$ probability for an event during the observation time of the study (of the total study cohort, under standard therapy group A, under experimental therapy group B, respectively)

t_1, t_2 time points of the individual follow-up at which non-event rates (used for expressing the treatment effect to be detected in the study) are specified

HR hazard ratio

Months (m) or years (y) will be used as time units.

2.3. Brief Recapitulation of the Freedman and Schoenfeld Methods

Both the Freedman and Schoenfeld approaches to calculating the required sample size in a time-to-event setting comparing two samples consist of two parts:

1. An "event formula" determining the number N_e of events required in the trial in order to detect the effect postulated under the alternative hypothesis H_1. N_e depends on the size of this effect as well as on the type I and type II errors, α and β. The event formulas derived by Freedman and Schoenfeld are both based on the proportional hazards (PH) assumption for the time-to-event, but apart from that do not have any distributional assumptions regarding the event times.
2. A formula calculating the probability P_e (calculated under H_1) that a randomly selected study patient experiences an event during the observation time of the study. P_e depends on the accrual time a, the duration of follow-up f, as well as on assumptions regarding the distribution of event times.

The required total number of patients is then given by

$$N = N_e / P_e.$$

Let Θ be the *HR* of group A vs. B. Under the PH model $\Theta = \log(S_A(t))/\log(S_B(t))$ for arbitrary t. Let z_γ be the γ-quantile of the standard normal distribution, and let $z = z_{1-\alpha/2} + z_{1-\beta}$. Then, for the special case of 1:1 allocation considered here, the event formulas are as follows:

Freedman (1982):

$$N_e = \frac{z^2(1+\Theta)^2}{(1-\Theta)^2} \tag{1}$$

Schoenfeld (1981):

$$N_e = \frac{4z^2}{\log^2(\Theta)} \tag{2}$$

As for P_e, first note that

$$P_e = \frac{P_{e,A} + P_{e,B}}{2}. \tag{3}$$

All formulas for $P_{e,g}$, g = A, B, are based on the simple equation

$$P_{e,g} = \int_0^a \mathrm{Pr}_g(e|entry\,at\,time\,t)\mathrm{Pr}_g(entry\,at\,t)\,dt \tag{4}$$

for $a > 0$, and $P_{e,g} = S_g(f)$ for $a = 0$. For uniform accrual, and after some rearrangement, (4) becomes the following (see Collett, 2003, p. 308):

$$P_{e,g} = 1 - \frac{1}{a}\int_f^{a+f} S_g(u)\,du. \tag{5}$$

Equation (5) can either be solved in closed form or calculated by means of numeric integration. In the context of the Freedman or Schoenfeld approaches, it appears logical to require that the PH assumption, which was the basis for determining the required number of events, also be satisfied when calculating $P_{e,g}$. Given that under the PH model $S_A(t) = S_B(t)^{\Theta}$, this is tantamount to focusing on Lehmann alternatives families only.

There are several examples of common parametric families satisfying the proportional hazards assumptions, including the exponential distribution, the Weibull subfamilies W_p characterized by the hazard function $h(t) = \lambda p(\lambda t)^{p-1}$ with fixed p, and the Gompertz subfamilies G_γ given by hazard functions $h(t) = \lambda\exp(\gamma t)$, γ fixed. Among these, the exponential distribution, where (5) transforms to the well-known formula

$$P_{e,g} = 1 - \exp\{-\lambda_g(f + a)\}\frac{\exp(\lambda_g a) - 1}{\lambda_g a} \tag{6}$$

(e.g., Bernstein and Lagakos, 1978; Schoenfeld and Richter, 1982), is by far the most popular one, and literature on power or sample size calculation under the PH assumption for parametric models other than the exponential is sparse (e.g., Cantor, 1992; Lakatos and Lan, 1992; Heo et al., 1998; Jiang et al., 2012).

In what follows we will address a number of issues related to the use of sample size or power calculations using the framework described above. Broadly speaking, all points are concerned with deviations from the assumptions regarding, e.g., the survival distributions, the validity of approximations, or the parameter values used when applying the formulas. We want to emphasize that some of the points raised here are not specific to the Schoenfeld and Freedman method, but are rather issues that these methods have in common with many other power or sample size calculations for time-to-event endpoints.

It should be noted that while software packages or add-ons (e.g., STATA) have been developed allowing the user to simulate survival analysis under varying assumptions (Ryan, 2013), these tools do not address the issues dealt with in the present article.

2.4. Simulation Studies

We used computer simulations of clinical trials satisfying the above-mentioned assumptions. All programs were written in R code. The R workspace and program descriptions are available on request from the corresponding author.

The simulation programs are modular and include the following functions:

Param: This calculates the parameters of the Weibull, log-normal, log-logistic, and Gompertz survival distribution from points $(t_1, rate_1)$, $(t_2, rate_2)$ on the survival curve, as well as the parameters of the two exponential distributions passing through the first and second data point, respectively. For the special case of Gompertz distributions, *Param* invokes two functions (*Gompertz* and *Gamma)* supplying the numerical solution of the equation for the parameters (see Appendix A).

Recruit: This produces recruitment times for n patients, either uniform in $[0, a]$ or following a homogeneous Poisson process with rate n/a.

Sample: This generates a data frame containing a random sample of censored survival data for two groups A, B of identical size. *Sample* accommodates a positive cure rate, which is assumed to be identical in both groups. Follow-up optionally for a fixed duration f or up until a given number of events has been observed. It invokes functions *Recruit* and *SurvSample*.

SimPower: This is a simulation program for estimating the power of the log-rank test (two-tailed) comparing two patient groups of identical size. *SimPower* invokes *Param* and *Sample*.

SimPowerMix: This is a simulation program for calculating the power of the log-rank test (two-tailed) comparing two patient groups of identical size, assuming a cure rate model. This assumes exponential survival of uncured patients and either zero hazard or exponential survival for the cured patients. *SimPowerMix* invokes *Sample*.

SurvSample: This generates random survival times for n patients following one of the five distributions mentioned above.

3. RESULTS

3.1. A Small Leak Will Sink a Great Ship: Impact of Deviations from Distributional Assumptions

Except for the special case of positive cure fractions (see the next paragraph), only a few papers have addressed the question of how sample size or power calculations along the lines described above are affected if the distributional assumptions are violated (e.g., Barthel et al., 2006; Heo et al., 1998; Lakatos and Lan, 1992). Given the multitude of design parameters of a trial and the vast number of different ways to formulate the true and the assumed distributions, an exhaustive study of this question is not feasible. Our aim is more modest, namely, to make the reader aware of non-robustness by showing that a severe power loss may occur even in situations where deviations from the assumed distribution (here: the exponential) are very small and practically undetectable over a large range of follow-up times. This point will be highlighted by two examples which were devised by making the following assumptions:

1. The survival distribution in at least one of the groups is not exponential, but selected from one of four two-parameter families (Weibull, log-normal, log-logistic, Gompertz) which are sometimes used for modeling survival, in particular in clinical trials of cancer patients (e.g., Allan, 1978; Cantor, 1992; Moghimi-Dehkordi et al., 2008; Wang et al., 2010; Hayat et al., 2010).
2. The distributions are, however, "similar" to the exponential used for the sample size calculation. This is enforced by requiring that (a) the survival rates $S(t_1)$ be correctly specified (i.e., the survivor functions of the assumed exponential and the true distributions coincide at t_1); and (b) at a second time point $t_2 > t_1$, the survival rates in the non-exponential group(s) differ only very slightly from the values implied by the exponential distribution.

This particular approach of formulating alternative distributions via survival rates was chosen because in our experience it constitutes the easiest way (especially for clinicians) to express and visualize differences in survival. Formulas for calculating the parameters of the distributions from two data points $S(t_1)$ and $S(t_2)$ are given in Appendix A. The true power was calculated by means of computer simulations (100,000 trials generated for each scenario). The exponential distribution was used as the benchmark; i.e., sample sizes were set to achieve 80% or 90% power for the exponential distribution with parameter $\lambda = -\log(S(t_1))/t_1$ and using the Schoenfeld formulas (2) and (6).

3.1.1. Example 1 (Deviation from Exponentiality in Both Groups). The following values were used for the sample size calculation based on the exponential distribution: $a = 36m$, $f = 24\text{m}$, $t_1 = 12\text{m}$, $S_A(t_1) = 60\%$, $S_B(t_1) = 75\%$. The corresponding parameters of the exponential distribution are $\lambda_A = 0.0426$ and $\lambda_B = 0.024$. Thus, according to (2) and (6), 67 (89) patients per group are needed to achieve a power of at least 80% (90%). At $t_2 = 24\text{m}$ the exponential distributions yield $S_A(t_2) = 36.0\%$ and $S_B(t_2) = 56.25\%$, respectively. Let S' denote the survivor functions of any of the four alternative survival distributions. $S'(t_1)$, $S'(t_2)$ were stipulated as follows: $S'_A(12\text{m}) = S_A(12\text{m})$, $S'_B(12\text{m}) = S_B(12\text{m})$, $S'_A(24\text{m}) = 38\%$, $S'_B(24\text{m}) = 54\%$, implying that $S_.(t_2)$ and $S_.'(t_2)$ do not differ by more than about 2 percentage points. Fig. 1 compares the corresponding survivor functions with the exponential up to $t = 48\text{m}$. The power values obtained from computer simulations (100,000 runs) are shown in Table 1a.

The apparent power loss may be reason for concern, especially given the similarity of the distributions, in particular in case of the Weibull model. In fact, we submit that (except perhaps for the Gompertz distribution) few clinicians would be able to tell if the true, unknown survivor function is exponential rather than one of the alternatives shown here. Of particular concern is the Weibull example where the survivor curve of group A up to 48m was very nearly exponential, whereas the exponentiality assumption for group B would be slightly optimistic if the Weibull distribution were, in fact, the correct one. Note that optimism regarding new treatments may be the rule rather than an exception.

The Gompertz distribution is special in that the corresponding survivor function of group A is improper, with $S'_A(t) \to \exp(\lambda_A/\gamma_A) > 0$ for $t \to \infty$, owing to the fact that the specifications of Example 1 lead to a negative parameter value γ_A of the Gompertz hazard function $h(t) = \lambda \cdot \exp(\gamma t)$. As a result, groups A and B have crossing survivor curves, the crossing occurring at about $t = 68\text{m}$. This explains the markedly lower power.

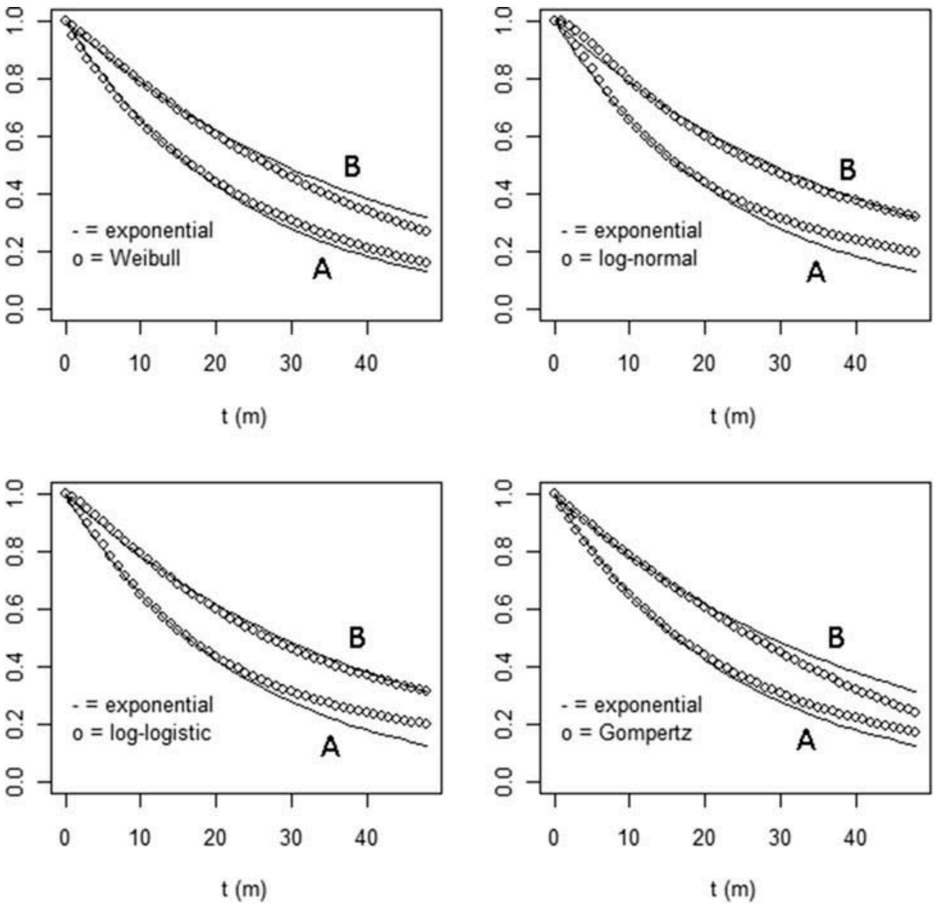

Figure 1 Survival curves as specified in Example 1.

One may object that Example 1 assumes a misspecification of both survivor curves whereas in reality previous data for the standard therapy group A are often available. The next example deals with this situation.

3.1.2. Example 2 (Deviation from Exponentiality Only in Experimental Group B). Let a, f, t_1, t_2, $S_A(t_1)$, $S_B(t_1)$ be as in Example 1. Survival in group A is assumed to have been correctly specified as exponential. Further assume that $S'_B(24\text{m}) = 51\%$, i.e., roughly a 5 percentage point difference with respect to the rate derived from the exponential.

We look at long-term survival up to 100 months. While in case of the Weibull and Gompertz distribution the deviations from the exponential are more obvious (results not shown), this is not the case if S'_B follows a log-normal or log-logistic distribution: As Fig. 2 shows the differences are very slight and, again, over most part of the follow-up the misspecification is in the "optimistic" direction. Power as calculated by computer simulations (100,000 runs each) is shown in Table 1b.

Table 1 Results of power calculations[1] for the parameter constellations considered in Examples 1 and 2

	Power calculated from computer simulations	
	Target power 80%	Target power 90%
a. Example 1		
Survival distribution in groups A and B		
Exponential	80.1%	89.9%
Weibull	53.7%	67.4%
Log-normal	54.9%	66.1%
Log-logistic	52.3%	63.9%
Gompertz	47.0%	57.7%
b. Example 2		
Survival distribution in group B		
Exponential	80.1%	90.0%
Log-normal	63.2%	75.0%
Log-logistic	61.9%	74.0%

Note. [1]100,000 simulation runs per scenario.

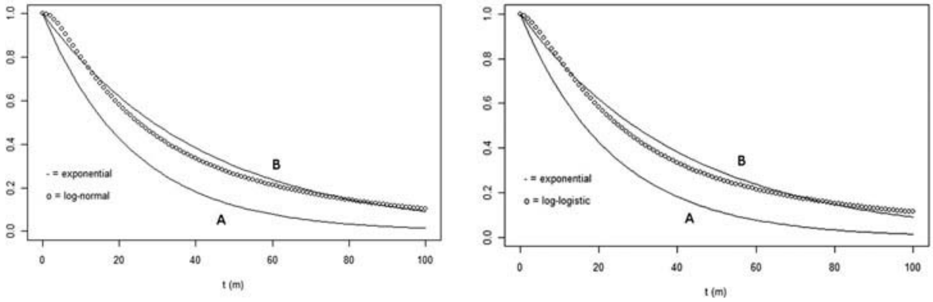

Figure 2 Survival curves as specified in Example 2.

Again, the decrease of power, though smaller than in Example 1, is still considerable. It may be of interest that a similar power loss was found in computer simulations carried out by Hoos et al. (2010), where survival in the standard group was assumed to be exponential, while the experimental group was assumed to show a "delayed separation" from the standard group, with a *HR* of 1.0 in the first 3 months, and then decreasing linearly to become 0.7 after 6 months.

Two supplementary observations may be of interest: (1) Modelling accrual as a Poisson process instead of using a uniform distribution did not materially change the results. For example, the power values for log-logistic distributions in Example 1 for 80% and 90% target power were found to be 51.7% and 63.8%, respectively (100,000 simulation runs), compared to 52.3% and 63.9% shown in Table 1a. (2) We also examined the effect of a change of follow-up type, setting the end of follow-up at the *k*-th event, where *k* is the number of events required by the formulas. Generally, the differences were fairly small. Thus, using the Schoenfeld formula as in Table 1a and b, and target power 90% (i.e., 89 patients per group, and 128 events required according to the formula), the power values obtained from the simulations with type II censoring were 89.7%, 66.1%, 67.4%, 63.7%, and 58.4% in case of Table 1a, and 89.7%, 73.8%, and 72.9% in case of Table 1b.

What are the lessons to be learned? Clearly, if exponential distributions are assumed, the validity of this assumption appears to be quite critical for the power and sample sizes. Given that even slight deviations—implying nearly undistinguishable survivor curves over a rather long follow-up period—may cause a severe power loss, it is advisable to calculate different scenarios involving alternative distributions when planning clinical trials. The standard approach based on formulas (2) and (6) should only be used as an initial guide.

3.2. Beware of Mixed Populations

A situation that arises quite often in clinical trials is that only a part of a patient population has a risk for an event, while the remainder can be considered as "cured". For example, in clinical trials of adjuvant cancer therapy, the patient population splits into two parts, say, $\prod^-$, $\prod^+$, with notoriously different time-to-event distributions (regardless of whether survival or disease-free survival is chosen as the time-to-event endpoint), namely patients whose tumors were entirely resected ($\prod^-$; these patients can be considered cured by surgery alone), and those ($\prod^+$) who, after surgery, have undetected residual tumor mass, e.g., in form of micro-metastases. Only the latter actually have a risk of relapse, and their risk of death is dramatically increased. In fact, the survival in this subpopulation is often modeled using an exponential distribution. Obviously, only the subpopulation with residual tumor mass can benefit from adjuvant therapy.

Starting with the early work of Berkson and Gage (1952), cure rates models have been studied extensively in biostatistical literature (Maller and Zhou, 1996; Othus et al, 2012; Wang et al., 2012, with further references). In principle, there are two ways to model this situation: Model type 1 (M1): This model uses event-time distributions with improper survivor functions $S(t)$, i.e., survivor functions with $S(t) \to c > 0$ for $t \to \infty$, as was the case for the negative Gompertz distribution obtained in Example 1 or for classes of hazard functions considered by Sposto and Sather (1985), where $h(t)$ is set equal to zero after a given "cure time". The quantity c may be interpreted as the proportion of patients having zero risk. Note that if the model is based on continuous parametric hazard functions, the cure rate is implied by the model parameters. Model type 2 (M2): This approach—arguably more common than M1 and the one we consider here—is a mixture model, where survival in the uncured fraction is modeled by specifying the survivor function $S^+(t)$ (or the hazard functions $\lambda^+(t)$), and the cure rate c is used in an explicit way to obtain the distribution in the total patient populations, e.g.,

$$S(t) = (1-c)S^+(t) + c\,. \tag{7}$$

Sample sizes for clinical trials accommodating a cure rate have been studied by several authors (e.g., Sposto and Sather, 1985; Cantor, 1991, 1992; Halpern and Brown, 1987, 1993). Recently, Wang et al. (2012) developed formulas for calculating sample sizes for the PH cure model, where the PH model was assumed to hold for the uncured fractions. Their analysis allowed for different cure rates in the treatment groups and arbitrary survival function of the censoring times, i.e., flexible accrual rate functions. In their numerical examples Wang et al. studied, among other things, the question of how sample size and/or power is affected if the conventional Schoenfeld approach is applied to the total population, assuming a zero cure rate when in reality some patients are cured. The sample sizes were calculated using the following assumptions and parameter values: duration of accrual $a = 2y$, duration of follow-up $f = 4y$, exponential survival with parameter $\lambda_A = 1/2$ for uncured patients in the control group, and $\lambda_B = 1/3$, i.e., a hazard ratio of $\Theta = 3/2$. The probability of an event in the control group, $P_{e,A}$, was used instead of formula (3) to calculate P_e (S. Wang, personal communication). One major result was that for small differences of the cure rates (and in particular for identical cure rates) in groups A and B the sample sizes from the PH model in the constellations considered by Wang et al. (2012) were much smaller than those from the PH cure model. This result is important because it shows that ignoring cure rates in the study planning may lead to a severe underpowering of the study.

The comparison of the PH cure model vs. the PH model made by Wang et al. (2012) was based on the concept that in both cases the same *HR* was to be detected in the uncured patients (which in case of the PH model was assumed to be the total population). While this is a valid approach, it has one limitation, namely, it implies that, if the PH cure model is the correct one, an investigator ignoring the cure rate entirely misspecifies the survival curves in both groups. This follows from the easily established fact that if survival is exponential in the uncured patients, the survival curve for the same group with or without a cure rate cannot have any point in common for $t > 0$.

In this article, we take a different approach to the consequences of neglecting cure fractions. We assume that the investigator, based on his experience or literature, correctly specifies at least one point of the survivor curve of the standard treatment, say $S_A(t_1)$, and that he requires a given percentage point difference between groups A and B at t_1 (say, 15% or 20%) to be detected in the trial. Furthermore, we assume that he is unaware of, or fails to account for, the presence of a cure fraction. Specifically, we want to address the question of what happens if the Freedman or Schoenfeld formulas are used indiscriminately, without taking cure rates into account. This is an approach we encountered occasionally when reviewing clinical trial protocols.

First observe that if the PH assumption for the treatment vs. the control group is valid in the subpopulations of uncured patients and if survivor functions in these subpopulations are proper ones, then, as is easy to prove, the PH assumption cannot be valid for all patients (see Appendix B, Proposition 1). Thus, in this situation the application of the Schoenfeld and Freedman event formulas to the total population can never be strictly justified.

The *HR* in the uncured subgroup equals $\{\log(S_A(t_1) - c_A) - \log(1 - c_A)\}/\{\log(S_B(t_1) - c_B) - \log(1 - c_B)\}$ and is thus unequal to the value $\log(S_A(t_1))/\log(S_B(t_1))$ calculated for the total population (if the PH model is erroneously being assumed for all patients). As a result, our approach leads to different *HR*s in the total population vs. the uncured subgroup (the latter being lower than the former, especially with higher cure rates). In this respect our analysis differs from that carried out by Wang et al. (2012).

We examine the effect of ignoring the cure fractions when using the standard Schoenfeld approach with an assumed exponential survival, where the parameters of the exponential distribution are calculated from $S_A(t_1)$ and $S_B(t_1)$, respectively. For simplicity, we assume identical cure fractions in groups A and B, i.e., $c_A = c_B = c$.

However, we allow the population $\prod^{-}$ to have a (small) risk described by a survivor function $S^{-}(t)$, i.e.,

$$S_g(t) = (1-c)S_g^{+}(t) + cS^{-}(t), \tag{8}$$

$g = A, B$. Unless events are excluded that are not related to the disease this appears to be a realistic and even a necessary assumption, especially in case of long-term trials and when death is counted as an event. $S^{-}(t)$ is also modeled by an exponential distribution.

Table 2 illustrates the impact of ignoring the cure rate for some constellations. The sample sizes were calculated to achieve a target power of 80% or 90% assuming a zero cure rate. The results given for $c > 0$ were based on computer simulations with 100,000 trials per constellation. As can be seen, the true power for the mixed population—obtained with the parameters of the simulations—may be dramatically lower than the desired target value, but in some situations (those characterized by lower event numbers) it may also be higher. This confirms findings of Sposto and Sather (1985), who compared the exponential and Weibull distributions with cure models of type M1 for some combinations of accrual rates and follow-up times, though the extent of the power loss shown in Table 2 is far greater than that found

Table 2 Impact on power of ignoring a cure fraction: Some examples [1]

						Power (%)		
	t_1(y)	$S_A(t_1)$	$S_B(t_1)$	$S^{-}(t_1)$	n[2]	c = 0%	c = 10%	c = 30%
a. Target power = 80%								
a = 3y, f = 4y	1	30	45	100	94	80.0	41.2	–[3]
	1			80		80.0	54.8	32.9
	3			100	114	80.3	69.8	–[3]
	3			80		80.3	74.3	61.3
	1	40	55	100	89	80.3	42.7	19.8
	1			80		80.3	59.8	31.8
	3			100	118	80.2	72.9	53.1
	3			80		80.2	76.4	64.7
a = 2y, f = 2y	4	40	60	100	114	80.1	81.1	90.8
	4			80		80.1	81.1	85.9
b. Target power = 90%								
a = 3y, f = 4y	1	30	45	100	126	90.1	51.7	–[3]
	1			80		90.1	67.0	41.7
	3			100	152	90.2	81.7	–[3]
	3			80		90.2	85.6	74.2
	1	40	55	100	119	90.2	52.9	24.6
	1			80		90.2	72.4	40.2
	3			100	158	90.2	84.3	65.3
	3			80		90.2	87.2	77.0
a = 2y, f = 2y	4	40	60	100	153	90.2	91.4	96.9
	4			80		90.2	91.0	94.1

Notes. [1]All survival rates expressed in (%); exponential survival assumed in all groups; power calculations for c > 0 based on 100,000 simulation runs; [2]Patient number per group calculated using the Schoenfeld formulas (1), (6) with α = 5%, and power = 80% or 90%, assuming a zero cure rate; [3]Inadmissible constellation ($c \geq S_A(t_1)/S^{-}(t_1)$).

by Sposto and Sather. By contrast, the approach taken by Wang et al. (2012) invariably resulted in a power loss from ignoring cure fractions. One may also note that according to our results the deviations from the aspired 80% or 90% power seem to be mitigated in cases where the "cured" group itself has a positive risk.

Interestingly, type II censoring—i.e., more specifically, follow-up until the number of events required by the formulas (ignoring the cure rate) has been observed—uniformly resulted in lower power values. For example, the power obtained with this type of follow-up for the scenario of Table 1a, first line, $c = 10\%$, was 39.1% compared to 41.2% obtained with fixed duration of follow-up (100,000 simulation runs). This may appear paradoxical, considering that a fixed follow-up leads to lower observed event numbers when ignoring the cure rate. Thus in the example given above, the number of events required assuming a zero cure rate is $N_e = 187$, but a fixed follow-up of 4 years resulted in a mean number of only 168.5 ($SD = 4.2$) observed events (100,000 simulation runs). However, this phenomenon becomes plausible when looking more closely into the details of the log-rank statistic. Owing to the positive cure rate, at later follow-up and event times both groups will continue to contain a relatively high number of patients at risk. As a result, in the higher risk group, say, A, the terms (observed – expected) will remain non-negligible, even as the events in this group are depleted with increasing follow-up time. However, with longer follow-up these contributions will be negative, i.e., they will *lower* the sum over the terms (observed – expected) in group A, because events will increasingly come from less event-depleted lower-risk group B. This explains the power loss with extended follow-up in cure rate models.

Summarizing, the failure to account for a true positive cure rate in a subgroup of patients when planning a trial may result in a very serious error. In particular, the conventional Schoenfeld or Freedman approach is inadequate when planning a trial where the cure rate is > 0, as in the adjuvant treatment of cancer. As a reviewer pointed out, while the log-rank test is asymptotically efficient under the proportional hazards model (Peto and Peto, 1972), it may not be the optimal statistic when this assumption does not hold. In particular, in the case of cure rate models and mixture models, where proportional hazards can rarely be assumed (see Appendix B, Proposition 1), the use of the log-rank test is questionable and alternative methods should be applied for hypothesis test and sample size calculation.

3.3. Probability P_e: Can Approximations of Unknown Survival Distributions Be Used?

Whenever the distribution of the event-times is unknown—and this may be the rule rather than the exception—the integral (5) must be replaced with an approximation. Schoenfeld (1983) suggested to use Simpson's formula for one of the groups and then to calculate P_e for the second group taking into account the proportional hazards assumption. If rates are estimated for the standard treatment group, then in our notation this results in the following formulas:

$$P_{e,A} \approx 1 - \frac{1}{6}\{S_A(f) + 4S_A(f + 0.5a) + S_A(f + a)\} \tag{9a}$$

$$P_{e,B} \approx 1 - (1 - P_{e,A})^{1/\Theta}. \tag{9b}$$

Note that (9b) is a simplification that does not exactly reflect proportional hazards. While, as stated above, in case of proportional hazards the equation $\Theta = \log(S_A(t))/\log(S_B(t))$ and thus $S_B(t) = (S_A(t))^{1/\Theta}$ holds for all t, the analogous formula obtained by replacing $S_A(t)$ with $1 - P_{e,A}$ and $S_B(t)$ with $1 - P_{e,B}$ is generally not correct. (It does hold in the special case $a = 0$, for then $S_g(f) = 1 - P_{e,g}$, $g = A,\ B$.)

This poses a dilemma which is perhaps sometimes overlooked: The use of survival rates that do not satisfy the proportional hazards assumption, in combination with an event formula that relies on exactly this assumption, appears somewhat inconsistent. This approach necessitates justification, which can be done by demonstrating (using computer simulations with distributional models for event-times) that the formulas are unlikely to induce a large error in the particular setting to be studied. In fact, it appears that the approximation implied by (9a,b), when $S_A(t)$ is correctly specified at $t = f$, $f + 0.5a, f + a$ and the PH assumption holds, is usually fairly good.

For example, for $a = f = 24$m and using the exact survival rates implied by exponential distributions with parameters $\lambda_A = 0.03$ and $\lambda_B = 0.02$ (resp.), (9a,b) yield $P_{e,A} = 65.30\%$ and $P_{e,B} = 50.62\%$, i.e., $P_e = 57.96\%$, compared to the exact values of $P_{e,A} = 65.30\%$, $P_{e,B} = 50.86\%$, $P_e = 58.08\%$ calculated from formula (6).

Schoenfeld (1983) and Collett (2003, p. 308ff) discussed a variant of this approach consisting of using Simpson's rule for both groups A and B and expressing P_e in terms of three average rates:

$$P_e = 1 - \frac{1}{6}\left\{\bar{S}(f) + 4\bar{S}(f + 0.5a) + \bar{S}(f + a)\right\} \tag{10}$$

where $\bar{S}(t) = \{S_A(t) + S_B(t)\}/2$. They suggested to calculate $S_B(t)$ at $t = f$, $f + a/2$, $f + a$, using $S_B(t) = (S_A(t))^{1/\Theta}$ for all t, with $\Theta = \log(S_A(t_1))/\log(S_B(t_1))$. This approach is superior to (9a,b) in two respects: First, it maintains proportionality and second it generally has an even smaller numerical inaccuracy. For example, for the example above, (10) yields an estimate of $P_e = 58.08\%$, coinciding up to four digits with the exact value.

While this approach is satisfactory when applicable, it still has the problem that it relies on survival rates for the standard treatment group A at three time points. This may be a problem, unless either data from a previous trial or cohort study in a very similar setting are available (this was the case in the example given by Collett) or distributional assumptions for the event times are made. However, in the latter case, the approximation introduced by Simpson's formula would be unnecessary. Also, of course, Collett's approach (like any other approximation to the true distributions) does not solve the problem addressed above that—even if survival in group A were perfectly known—a small deviation from the distributional assumption for experimental group B may invalidate the power calculation.

3.4. The Special Case of Freedman's P_e Formula: Do not Use It

In his article, Freedman (1982) made no distributional assumptions for calculating P_e. Instead he used the following equation (transformed to our notation):

$$P_e = \{2 - S_A(t_1) - S_B(t_1)\}/2, \tag{11}$$

where, as he wrote, $S_A(t_1)$, $S_B(t_1)$ are "survival rates … at some chosen time point" (p. 126). Clearly, this formulation is generally not correct. Rather, $P_{e,g} = 1 - S_g(t_1),\ g = A, B,$ can only hold true (independent of distributional assumptions) if $t_1 = f$ and $a = 0$. If individual follow-up data beyond $t = f$ were excluded (an assumption made by Freedman which, however, not necessarily reflects the reality of randomized trials) then the duration of accrual may be positive; however, it is still required that $t_1 = f$.

Obviously, the condition $t_1 = f$, i.e., the condition that the treatment effect must be stipulated at the end of the individual follow-up time, severely limits the applicability of this approach. For the more realistic case where $a > 0$ and the condition that individual follow-up data beyond $t = f$ are excluded is dropped, Freedman proposed to use the approximation

$$P_{e,g} \approx 1 - S_g(f + a/2). \tag{12}$$

In view of formula (5), (12) becomes an exact equality if $S_g(t)$ is linear in the interval $[f, a + f]$. However, it is easy to prove that linearity of the survivor function (even in a small interval) is not compatible with the PH assumption (see Appendix B, Proposition 2), thus making the combination of (12) with the event formula (which requires the PH property) questionable from a logical viewpoint.

There are further issues with the approximation (12). First, in some cases it may lead to an undesirable overestimation of power. If the survivor functions are convex, as is the case for exponentially distributed event times, $1 - S_g(f + a/2)$ overestimates the right-hand side of formula (5), i.e., P_e, and thus leads to an underestimation of the necessary sample sizes. With long accrual and short follow-up, the underestimation implied by formula (12) may no longer be negligible. Thus, e.g., for exponential survival times with $\lambda_A = 0.03818$ and $\lambda_B = 0.02128$, corresponding to 24-month survival rates of 40% and 60%, respectively, and assuming $a = 48$m, $f = 12$m, the exact formula (6) yields $P_e = 61.2\%$, whereas Freedman's P_e formula with $t_1 = f + a/2 = 36$m and using the corresponding rates of the exponential distributions at t_1, gives $P_e = 64.1\%$, thus implying a 4.5% underestimation of the sample size.

Another important point to note is that any deviation of t_1 from $f + a/2$ may grossly distort the sample size calculation. As Freedman (1982) wrote:

> To estimate the required number of patients under this policy, it would be *wrong* to enter the table with event-free rates appropriate to the minimum follow-up time. The number of patients needed would thereby be over-estimated, often seriously. Instead, as an approximate device, the table should be entered with event-free rates appropriate to the average follow-up time. (p. 123)

We stress this point because in our professional lives we have encountered several instances where this cautionary remark had been neglected. For example, in a recent oncologic trial the sample size was to be calculated using the following assumptions: $a = 6\text{y}, f = 2\text{y}, S_A(2\text{y}) = 47.5\%, S_B(2\text{y}) = 57.5\%, \alpha = 5\%,$ power 90%. Based on the Freedman approach, and erroneously using $t_1 = 2$y, the sample size was determined to be 511 per group (as yielded by the software nquery®). Clearly, this number was far too large, given that it essentially reflected the number of events to be expected had every patient been observed for two years. In reality, the approximate mean observation time in this study was $f + a/2 = 5$y. Assuming exponential survival, the Schoenfeld formula for this trial results in a mere 313 patients per group. The planning was corrected accordingly.

The consequence of this restriction for t_1 is that, except in situations where the effect to be detected in the trial is expressed in terms of $S_A\ (f + a/2)$, $S_B\ (f + a/2)$, the Freedman approach for calculating P_e is inappropriate. On the other hand, using $t_1 = f + a/2$ is often impractical. We will reveal the weakness of this method using the example of a long-term trial with long accrual and high event rates, a situation which is typical for studies in rare, severe diseases. In this case, $t_1 = f + a/2$ will be large and $S_g(t_1)$ will be small. But then the specification of the differences in event rates at t_1 may be clinically uninteresting or impossible to define. Thus, if in the above example the hazard rates had been constant but twice those implied by the data given above, the Freedman method would have required the clinician to specify the assumed 5-year survival rates as 6.2% in group B vs. 2.4% in group A. Few clinicians would be willing to do so.

Note that the problem cannot be circumvented by transforming $S_A\ (f + a/2)$ and $S_B\ (f + a/2)$ to a different effect measure, say, the median event times in groups A and B, or rates $S_A(t_1^{'})$, $S_B(t_1^{'})$ at $t_1^{'} \neq f + a/2$, because in order to do so a distributional assumption for the event times is needed. However, if such an assumption is made, then an exact calculation of P_e as in formula (6) or by numerical integration of formula (5) becomes possible. But then, this is no longer the Freedman approach for calculating P_e.

3.5. Event Formulas: Using Freedman or Schoenfeld?

The Freedman and Schoenfeld event formulas were not the first published ones. Earlier, several formulas had been developed for the special case of exponential distributions (Pasternack and Gilbert, 1971; George and Desu, 1974; Rubinstein et al., 1981). Interestingly, the formula given by George and Desu (1974) is identical to the Schoenfeld formula.

While the asymptotics involved in their mathematical derivation ensure that both the Schoenfeld and Freedman formulas are nearly correct for high event numbers N_e, the question which formula is more accurate for small N_e is nontrivial and important. It is not easy to decide this on theoretical grounds, because no upper bounds for the imprecision inherent to the asymptotics were given allowing a comparison for finite samples.

Both formulas are invariant against an exchange of the groups, i.e., against the transformation $\Theta \to 1/\Theta$. Since, for $\Theta > 1$, $\log(\Theta) > 2(\Theta - 1)/(\Theta + 1)$, the Schoenfeld formula leads to smaller required number of events and thus to smaller sample sizes if the same formula for P_e is used. Given that both formulas have exactly the same assumption, viz. proportional hazards, the Schoenfeld event formula would always be preferable if parsimony of sample sizes were the sole criterion. The ratio $N_{e,Schoenfeld}\big/N_{e,Freedman}$ increases monotonically as Θ approaches 1 (see Appendix B, Proposition 3), and it may be considerably lower than 1 for low or high values of Θ. For example, for $S_A(t_1) = 70\%$, $S_B(t_1) = 90\%$, i.e., a HR of $\Theta = 3.4$, the number of events calculated using the Schoenfeld formula is about 20% lower than that resulting from the Freedman formula. This difference may not be ignored.

To answer the question which formula is more accurate computer simulations are needed where the formula-based sample sizes (or power) are compared to those obtained

Table 3 Accuracy of the Schoenfeld and Freedman formulas: True vs. predicted power in selected situations

	a	f	t_1	Power(n_S) [(2)] (%)	Power(n_F) [(2)] (%)
a. Target power = 80%[(1)]					
$S_A(t_1)$ = 40%	24	12	12	79.9	82.2
$S_B(t_1)$ = 60%	24	24	24	79.7	82.1
	36	12	12	80.2	81.8
	36	36	36	79.8	82.0
$S_A(t_1)$ = 70%	24	12	12	76.1	85.3
$S_B(t_1)$ = 90%	24	24	24	75.5	84.9
	36	12	12	76.9	86.2
	36	36	36	75.2	84.8
b. Target power = 90%[(1)]					
$S_A(t_1)$ = 40%	24	12	12	89.9	91.3
$S_B(t_1)$ = 60%	24	24	24	89.9	91.6
	36	12	12	90.0	91.6
	36	36	36	90.1	91.5
$S_A(t_1)$ = 70%	24	12	12	87.8	93.8
$S_B(t_1)$ = 90%	24	24	24	87.1	93.4
	36	12	12	87.9	94.1
	36	36	36	86.8	93.5

Notes. [(1)]Sample sizes n_F and n_S calculated using the Freedman and Schoenfeld formulas (respectively) to achieve a target power of 80% or 90%; [(2)]Power (n_S), Power (n_F) = power obtained in computer simulations (100,000 simulation runs for each situation) using sample sizes n_S and n_F, respectively.

in simulations. To this end, it is necessary that the PH assumption be satisfied and that P_e be calculated to a very high precision—either using an exact formula or a numerical integration of formula (5).

While a comparison of the Rubinstein and the Freedman formulas has already been done (Lakatos and Lan, 1992), we are not aware of any published comparison of the Schoenfeld and Feedman formulas. Table 3 shows the results of computer simulations for two different sets of situations, the first one with $S_A(t_1) = 40\%$ vs. $S_B(t_1) = 60\%$, where HR is moderate (1.79) and formulas (1) and (2) yield event numbers of $N_e = 91.9$ and 97.2 (respectively) for power 80%, and 123.1 vs. 130.0 for power 90%; the second one with $S_A(t_1) = 70\%$ vs. $S_B(t_1) = 90\%$, implying a much higher HR of 3.39 and lower event numbers ($N_e = 21.1$ vs. 26.5 for power 80%, and 28.3 vs. 35.5 for power 90%). The true power was determined using computer simulations assuming exponential survival distributions. For each set-up characterized by a combination of a, f, t_1, the sample sizes n_S and n_F used were those which, according to the Schoenfeld and Freedman formulas, were required to reach 80% or 90% power ($\alpha = 5\%$).

While in the first situation the deviation from the predicted power is almost negligible, this is no longer the case when N_e is low. Here, with the parameters of the simulations, the Schoenfeld formula leads to an overestimation of power, while the Freedman formula underrates the power by almost the same absolute amount. Note that a potential underestimation of the sample size by the Schoenfeld formula was already mentioned by Collett (2003, p. 301) who advises "judicious rounding" of the obtained sample sizes. The direction of the deviation is in accordance with findings of Freedman (1982), Rubinstein et al. (1981) and Lakatos and Lan (1992), if the small difference between the formulas given by Schoenfeld and Rubinstein et al. (1981) is neglected. As

Table 3 shows, the deviation may easily reach a magnitude that makes the use of the formulas questionable. It should be noted that the result is not specific to the exponential distribution. We also performed power calculations for the case of Weibull distributions with fixed shape parameters (either $p = 0.5$ or $p = 2$) and sample sizes determined by solving equation (5) numerically. The deviations of the true power from the predicted 80% or 90% value were almost identical to those shown in Table 3 (results not shown). It thus appears that in cases where N_e is small, the reliance on asymptotics inherent in the event formulas is dubious and computer simulations are warranted.

A reviewer correctly pointed to the fact that the Freedman (1982) formula for the total required number of events is derived from the properties of the log-rank statistic, but it is based on the simplifying assumption that the proportion of patients at risk, that are on the experimental group, stays constant at .5. Freedman admits that this assumption is generally violated and, thus, leads to an over-estimate in the required number of events, which is negligible for small target hazard ratios (e.g., $HR = 1.5$), but not for large target hazard ratios (e.g., $HR = 3$). This explains some of the concerns with use of formula (1) expressed above.

As in section 3.1, following-up the patients until the event numbers required according to the Schoenfeld and Freedman formulas (respectively) have been attained did not substantially change the findings: While there was no marked and consistent change in the underestimation of power with the use of Freedman's formula, the over-estimation of power with the Schoenfeld approach was slightly mitigated, but essentially remained in place; e.g., the true power for the constellation shown in the last line of Table 3 (as estimated from 100,000 simulation runs) changed to 87.9% compared to 86.8% with fixed duration of follow-up.

One reviewer observed that the formula derived by Rubinstein et al. assuming exponential survival may often be more exact (when exponential survival can be assumed) than the Schoenfeld or Freedman approach, and partly resolve the issues addressed in Table 3. In fact, as can be easily shown, the event and patient numbers determined by this approach are invariably higher than those obtained with the Schoenfeld formula. They are more similar to those resulting from the Freedman formula—mostly somewhat smaller, but in some instances they may even exceed the Freedman sample sizes. Thus, for $a = f = t_1 = 24$, $S_A(t_1) = 70\%$, $S_B(t_1) = 90\%$, and 90% power (see Table 3b, line 6) the sample sizes per group (with expected events calculated exactly, i.e., by means of formula (6)) were as follows: Schoenfeld $n = 64$, Rubinstein et al. $n = 66$.

3.6. How Critical Is the Choice of t_1?

In our experience, while most clinicians usually enter the discussion on sample size planning with a fairly concrete idea of a percentage point difference in survival rates they wish to demonstrate (mostly 10%, 15%, or 20%), they often show an amazing flexibility as regards the timing of the difference (e.g., 15% difference in either 2-year or 3-year survival). To a certain degree this is understandable. After all, the effect to be detected in a trial always contains a discretionary element, given that it constitutes a compromise between what is desirable (viz., the detection of even small effects as long as they are clinically relevant) and what is feasible in view of the various constraints. Obviously, the power will change if the survival rates are specified at a different time point while maintaining the same absolute difference d, even if the survival of group A is assumed to remain unchanged. Though easily comprehensible for biostatisticians, the importance of this phenomenon is perhaps sometimes underrated.

Consider, e.g., a clinical trial with duration of accrual $a = 36$m, subsequent follow-up $f = 24$m, $S_A(12\text{m}) = 60\%$, $S_B(12\text{m}) = 80\%$. Then, assuming exponential survival, $n = 34$ patients per group will suffice to yield a power of $\geq 80\%$ according to the Schoenfeld formula ($\alpha = 5\%$). If survival in group A is assumed to be unchanged but power for the same 20 percentage point difference is calculated using $t_1 = 24$m or $= 36$m, the resulting change in the survival curve of B reduces the power to 50.9% and 49.8%, respectively; and, to maintain power at $\geq 80\%$, the sample size would roughly have to be doubled (to 68 and 70 per group, respectively). This considerable power loss may be counterintuitive, given that an increase in long-term survival (i.e., an increase of the 36-month survival rate from 21.6% to 41.6%) may be thought to represent a larger, clinically more important, effect—implying a higher rather than a lower power—than a 20 percentage point increase of short-term survival.

The phenomenon becomes more plausible when considering the point $t_{\max}$ where the exponential distributions defined by $S_A(t_1)$ and $S_B(t_1)$, attain their maximum absolute difference $d_{\max} = \max\{S_B(t) - S_A(t)\}$. If λ_A, λ_B, denote the parameter of these curves, then this point is given by $t_{\max} = (\log(\lambda_B) - \log(\lambda_A))/(\lambda_B - \lambda_A)$. It can be shown that, given λ_A and $d = S_B(t_1) - S_A(t_1)$, $t_{\max} > t_1$ if $S_A(t_1) > \exp(-1)$. Thus, in the aforementioned example, $t_{\max} = 34.5m$, and $d_{\max} = 29.6$ percentage points. Interestingly, the inverse implication does not hold: While it can be shown that if d is sufficiently small $S_A(t_1) < \exp(-1)$ implies $t_{\max} < t_1$, this is no longer generally true for higher values of d (see Appendix B, Proposition 4).

For the parameters specified above, Fig. 3 gives a more comprehensive impression of how power depends on the choice of t_1 when survival in group A as well as the difference

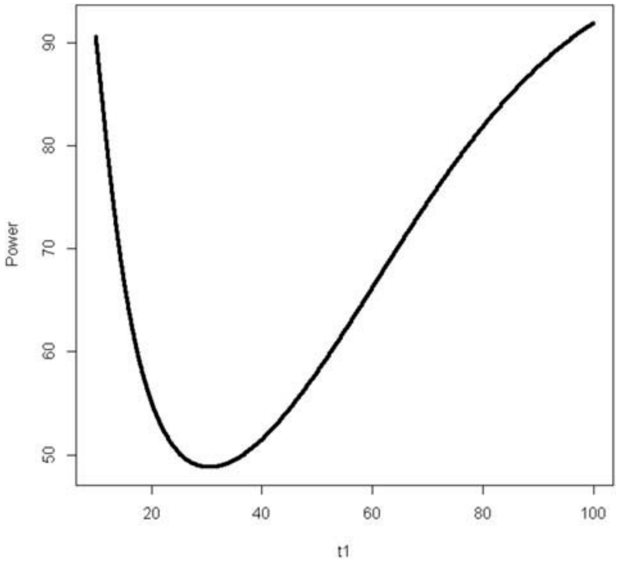

Figure 3 Power vs. t_1 (Assumptions: a = 36 months duration of accrual, f = 24 months duration of follow-up, 60% survival rate of standard therapy group A at 12 months, exponential survival, 20 percentage points difference in survival rates at time point t_1).

$S_B(t_1) - S_A(t_1)$ is kept unchanged. While there is no closed formula for calculating the minimum $t_{\min}$ of this curve, it can easily be determined numerically up to a high precision using a simple computer program (for the example considered in Fig. 3: $t_{\min} \approx 29.0$m). The relationship between power and t_1 is quite striking. Note that the peculiar shape of this curve seems to be retained, irrespective of the parameters chosen in the analysis.

Summarizing, a treatment effect is not well measured or well characterized by looking at the survival difference at a particular point in time, although it is, on occasion, characterized in such a way so as to appeal to the intuition of investigators. However, this approach includes a high risk of being misleading. For example, the survival difference of 75% vs. 95% ($HR = 5.6$), at a particular time, is very large, while the difference of 40% vs. 60% ($HR = 1.8$) is more modest. As the power of the log-rank statistic is determined by the target hazard ratio, sample size calculations based on survival differences at a single time point are not appropriate. Rather, examples for survival differences at particular time points may be used as a starting point for capturing the treatment effect in terms of the hazard ratio for which the sample size is determined. If there is some room for maneuver with respect to the timing of a rate difference, then we recommend to select a time point t_1 which is not in the vicinity of $t_{\min}$ and, if possible, is smaller than $t_{\min}$.

3.7. Do not Neglect Random Variations and Systematic Changes in Accrual Rates

Power and sample size depend on the length of accrual. If accrual is regarded as a stochastic process (often, it is modeled as a homogeneous or inhomogeneous Poisson process), then the duration of accrual a is not fixed but a random variable, whose expected value $E(a)$ is then substituted for a in the use of the sample size formulas. However, inherent in this common approach is the risk that, conditional on the true accrual process, the power may differ from the assumed value. In particular, if $a < E(a)$, the true power will be lower than calculated because fewer events will be observed. By slightly increasing the sample size over that obtained for $E(a)$, it is possible to safeguard against random variations of accrual (Abel et al., 2012).

While deviations of a from $E(a)$ are generally small, increases in the true accrual rate—implying a drop in $E(a)$—especially marked and sudden ones as may happen if further centers are added to those participating in a multicenter trial are more critical.

Assume, for example, that a trial is planned to detect an increase in 36m survival rates from 50% to 70% ($\alpha = 5\%$) with power 80% and assuming exponential survival, $f = 24$m, and an accrual as a Poisson process with an expected rate of 90 patients per year. Then the expected duration of accrual is $E(a) = 2$y and the necessary sample size is 90 per group. If, due to the addition of further centers, the accrual rate is higher than expected, e.g., 120 or even 180 patients per year, then maintaining the patient number of 90 per group will result in a power of only 77.7% and 74.7%, respectively. Conversely, if the power is maintained at 80%, the necessary sample sizes are 190 and 202, corresponding to an expected duration of accrual of $E(a) = 19.0$m and 13.5m, respectively.

In view of such a drop in power a provision in the protocol that sample sizes be recalculated once the assumed accrual rate is found to be too low may be advisable. However, this does not seem to be part of current routine study planning.

4. CONCLUSION

We have attempted to sharpen the reader's awareness for some issues involved in sample size and power calculations for time-to-event endpoints using the Freedman and Schoenfeld approaches. Great caution is required both in the application of the methods themselves and in the use of standard software. Inaccuracies and even grave errors are easily committed and seem to be rather commonplace. When planning trials, it is advisable to use the formula-based approaches only for a first orientation and to always examine various alternative scenarios using computer simulations.

Our investigations have shown that the accuracy of sample size calculations critically depends on the correctness of the assumptions. Providing good guesses for the true parameter values in the planning phase of a trial is generally an extremely difficult task, which is further boosted the more parameters come into play. The most reliable estimates for a current clinical trial situation can be obtained from the study itself. Consequently, internal pilot study designs and adaptive sample size reestimation based on a subsample gathered mid-cores have recently attracted much attention. Most of the methodological research on this topic has focused on normally distributed and binary outcomes. However, there are also methods available for time-to-event data. From a regulatory viewpoint, methods for sample size reestimation are preferred that keep the blinding of the treatment allocation. Such internal pilot study designs for survival data were, for example, proposed by Hade et al. (2010) for updating the assumed overall failure rate and by Ingel and Jahn-Eimermacher (2014) for reestimating the sample size inflation factor that takes into account the heterogeneity in the analysis of recurrent event data. For the situation that an interim analysis is to be implemented in a clinical trial to allow for early stopping, Shen and Cai (2003) (stopping for futility only), Wassmer (2006), and Jahn-Eimermacher and Ingel (2009) provided methods that enable also using the interim information on the treatment effect to adjust the sample size while controlling the type I error rate. In view of the results provided in this article, application of such methods is recommended whenever logistics allows for.

APPENDIX A

Formulas Used in This Article for Calculating Parameters of the Survival Distributions from Two Survival Rates

The parameters of the Weibull, log-normal, log-logistic, and Gompertz distribution can be calculated from two points $(t_i,\ s_i = S(t_i),\ s_i \neq 0,\ 1\ (i = 1,\ 2))$, of the survivor curves using the formulas given below. These formulas can be derived using simple arithmetics. Their validation is straightforward owing to the fact that the resulting survival distributions must reproduce the data points used as an input. Adopting the notation used by Kalbfleisch and Prentice (1980)) we have:

a. Weibull, with $S(t) = \exp(-(\lambda t)^p)$:

$$\lambda = \frac{c_1 \log(t_2) - c_2 \log(t_1)}{c_2 - c_1},$$

$$p = \frac{c_1}{\log(\lambda) + \log(t_1)},$$

where

$$c_i = \log(-\log(s_i)),\ i = 1,\ 2.$$

b. Log-normal, with $S(t) = 1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)$:

$$\mu = \log(t_2) - \frac{(\log(t_2) - \log(t_1))g(s_1)}{g(s_2) - g(s_1)},$$

$$\sigma = \frac{\log(t_2) - \log(t_1)}{g(s_2) - g(s_1)},$$

where Φ denotes the *cdf* of a standard normal variate and $g(s_i) = \Phi^{-1}(1 - s_i)$, $i = 1,\ 2$.

c. Log-logistic, with $S(t) = \frac{1}{1+(\lambda t)^p}$:

$$\lambda = \begin{cases} \exp\left(\frac{h \cdot \log(t_1) - \log(t_2)}{1-h}\right) & \textit{if } s_1 \neq 0.5 \\ \frac{1}{t_1} & \textit{if } s_1 = 0.5, \end{cases}$$

$$p = \frac{h'}{\log\left(\frac{t_2}{t_1}\right)},$$

where

$$h = \log\left(\frac{1 - s_2}{s_2}\right) / \log\left(\frac{1 - s_1}{s_1}\right),$$

$$h' = \log\left(\frac{1 - s_2}{s_2}\right) - \log\left(\frac{1 - s_1}{s_1}\right).$$

d. Gompertz, with $S(t) = \exp\left\{\frac{\lambda}{\gamma}(1 - \exp(\gamma t))\right\}$:

γ is the nontrivial (i.e., nonzero) solution of

$$\log(s_1)\exp(\gamma t_2) - \log(s_2)\exp(\gamma t_2) - \log\left(\frac{s_1}{s_2}\right) = 0 \qquad (*),$$

$$\lambda = \frac{\gamma \cdot \log(s_1)}{1 - \exp(\gamma t_1)}.$$

Remark: (*) has exactly one nonzero solution, which must be determined numerically.

APPENDIX B

Proofs of Propositions

Proposition 1 Let $S_g(t)$ follow the cure model $S_g(t) \equiv (1-c)S_g^+(t) + cS^-(t)$ with $0 < c < 1$, and assume that $S_g^+/S^-(t) \to 0$ for t$\to\infty$ $(g = A, B)$. Then, if $S_A \neq S_B$, $S_A(t)$ and $S_B(t)$ do not follow a proportional hazards model.

Proof: Assume that the proportional hazards model holds true for $S_A(t)$ and $S_B(t)$, i.e., $S_A(t) \equiv S_B(t)^\Theta$ for some $\Theta \neq 1$. Then, for $t\to\infty$:

$$\frac{S_A(t)}{S_B(t)} = \frac{(1-c)\frac{S_A^+(t)}{S^-(t)} + c}{(1-c)\frac{S_B^+(t)}{S^-(t)} + c} \to 1.$$

On the other hand,

$$\frac{S_A(t)}{S_B(t)} = [S_B(t)]^{\Theta-1} = \left\{S^-(t)\left[(1-c)\frac{S_B^+(t)}{S^-(t)} + c\right]\right\}^{\Theta-1}$$

$$\to [cS^-(t)]^{\Theta-1} \text{as } t\to\infty.$$

If S^- is a proper survivor function (approaching 0 as $t\to\infty$) then

$$[cS^-(t)]^{\Theta-1} \to \begin{cases} 0 & if\ \Theta > 1 \\ \infty & if\ \Theta < 1 \end{cases}$$

for $t\to\infty$, whereas if S^- is an improper survivor function (approaching c' with $0 < c' \leq 1$ as $t\to\infty$), then

$$[cS^-(t)]^{\Theta-1} \to [cc']^{\Theta-1} \neq 1 \text{ for } t \to \infty.$$

In both cases we have a contradiction, showing that the initial assumption was false.

Proposition 2 If the survivor functions $S_g(t)(g = A, B)$ are linear in the interval $[f, a+f]$ then

a. $P_{g,e} = 1 - S_g(f + a/2)$.

b. Unless $S_A(t)$ and $S_B(t)$ are identical, they do not satisfy the proportional hazards model.

Proof: Assume that, for $t \in [f, a+f]$, $S_g(t)$ is linear, e.g., $S_g(t) = \alpha_g + \beta_g \cdot t$.

a. For convenience we drop the subscript g. We have

$$P_e = 1 - \frac{1}{a}\int_f^{a+f} S(t)dt = 1 - \frac{1}{a}\int_f^{a+f} (\alpha + \beta t)dt = 1 - \left(\alpha + \beta\left(\frac{a}{2} + f\right)\right).$$

However this is equal to1 $- S(f + a/2)$, as is easily verified.

b. Assume that $S_A(t)$ and $S_B(t)$ satisfy the proportional hazards model, i.e.,

$$\frac{\log(S_A(t))}{\log(S_B(t))} = \Theta$$

(with $\Theta \neq 1$) for all $t > 0$. Then, for all $t \in (f, a+f)$,

$$\log(\alpha_A + \beta_A t) = \Theta \cdot \log(\alpha_B + \beta_B t).$$

Taking first derivatives on both sides and rearranging terms, we have

$$\frac{\alpha_B}{\beta_A} + t \equiv \frac{\alpha_B}{\Theta\beta_B} + \frac{t}{\Theta}$$

for all $t \in (f, a+f)$. Since this is an identity of two straight lines which is valid for all $t \in (f, a+f)$, it follows that the slopes must be identical, i.e., $\Theta = 1$, in conflict with the initial assumption.

Proposition 3 Let $r(\Theta) = \frac{N_{e,Schoenfeld}(\Theta)}{N_{e,Freedman}(\Theta)}$ denote the ratio of the event numbers calculated by means of the Schoenfeld and Freedman formulas, defined for a hazard ratio $\Theta \neq 1$ for probabilities α, β of type 1 and type 2 errors. Then $r(\Theta)$ is a monotonically increasing function in $(0, 1)$, and a monotonically decreasing function in $(1, \infty)$, with $\underset{\Theta \to 1}{lim}(r(\Theta)) = 1$.

Proof: Both $N_{e,Schoenfeld}(\Theta)$ and $N_{e,Freedman}(\Theta)$ are invariant against the transformation $\Theta \to \Theta' = 1/\Theta$. Therefore, $r(\Theta') = r(\Theta)$. Let

$$s(\Theta) = \frac{2(\Theta - 1)}{(\Theta + 1)\log(\Theta)}$$

Then $r(\Theta) = s^2(\Theta)$, and since $s(\Theta) > 0$ for $0 < \Theta < 1$ it suffices to show that $s(\Theta)$ is a monotonically increasing function in the interval (0,1), with $\lim_{\Theta \to 1}(s(\Theta)) = 1$. The first derivative is given by

$$s'(\Theta) = \frac{2}{(\Theta + 1)\log(\Theta)} \cdot \left\{ \frac{2}{\Theta + 1} - \frac{(\Theta - 1)}{\Theta \cdot \log(\Theta)} \right\}.$$

Arranging terms, we find that

$$s'(\Theta) > 0 \Leftrightarrow f(\Theta) := 2\Theta\log(\Theta) - \Theta^2 + 1 > 0. \tag{B1}$$

Now, $f(1) = 0$, and because of $\log(\Theta) < \Theta - 1$ for all Θ we have

$$f'(\Theta) = 2(1 + \log(\Theta) - \Theta) < 0.$$

Thus, $f(\Theta)$ is a monotonically decreasing function with $f(\Theta) > 0$ in (0,1). By virtue of (B1), it follows that $s(\Theta)$ is a monotonically increasing function in (0, 1).

As for the assertion $\lim_{\Theta \to 1}(s(\Theta)) = 1$, it follows from the Taylor series expansion (valid for $0 < \Theta < 1$)

$$\log(\Theta) = \log(1 + (\Theta - 1)) = (\Theta - 1) \cdot \left\{1 - \frac{\Theta - 1}{2} + \frac{(\Theta - 1)^2}{3} - \ldots\right\}$$

which implies

$$\frac{\log(\Theta)}{\Theta - 1} \rightarrow 1 \; as \; \Theta \rightarrow 1$$

and thus $s(\Theta) \rightarrow 1$ for $\Theta \rightarrow 1$.

Proposition 4 Let $S_A(t) = \exp(-\lambda_A t)$, $S_B(t) = \exp(-\lambda_B t)$, and assume that for given t_1 and $d > 0$ $S_B(t_1) - S_A(t_1) = d$. Let t_{max} be the value of t where the difference $d(t) = S_B(t) - S_A(t)$ attains its maximum. Then

a. If $S_A(t_1) > \exp(-1)$ then $t_{max} > t_1$.

b. If $S_A(t_1) < \exp(-1)$ and d sufficiently small, then $t_{max} < t_1$.

Remarks:

a. $S_A(t_1) > \exp(-1)$is equivalent to $t_1 < 1/\lambda_A$.

b. In Proposition 4b, the requirement that d be sufficiently small cannot be dropped. That is, for higher values of d, $S_A(t_1) < \exp(-1)$ does not imply $t_{max} < t_1$. Example: $t_1 = 12m$, $S_A(t_1) = 0.3$, $d\!: = S_B(t_1) - S_A(t_1) = 0.5$. Then $S_A(t_1) < \exp(-1)$, but $t_{max} = 20.62 > t_1$.

Proof of Proposition 4:

a. Assume that $S_A(t_1) > \exp(-1)$. Setting $d'(t) = 0$ and noting that

$$\lambda_g = \frac{-\log(S_g(t_1))}{t_1}$$

($g = A, B$) we have

$$t_{max} = t_1 \cdot \left\{\frac{\log(-\log(S_A(t_1)) - \log(-\log(S_B(t_1))}{\log(S_B(t_1)) - \log(S_A(t_1))}\right\}. \tag{B2}$$

For convenience we write: $x := S_A(t_1)$. To prove the proposition we have to show that

$$1 < \frac{\log(-\log(x)) - \log(-\log(x + d))}{\log(x + d) - \log(x)}. \tag{B3}$$

Rearranging and observing the monotonicity of the logarithm, (B3) is seen to be equivalent to

$$\frac{x}{x+d} > \frac{\log(x+d)}{\log(x)}$$

or

$$\log\left(1+\frac{d}{x}\right) > -\log(x)\frac{d}{x+d}. \tag{B4}$$

Now, for $y > -1$ the logarithm satisfies the (well-known and easily proven) inequality

$$\log(1+y) \geq \frac{y}{1+y}$$

with strict inequality for $y > 0$. Thus,

$$\log\left(1+\frac{d}{x}\right) > \frac{\frac{d}{x}}{1+\frac{d}{x}} = \frac{d}{x+d}$$

and (B4) follows from the fact that, by assumption, $x = S_A(t_1) > \exp(-1)$, or $-\log(x) < 1$.

b. Using the same argument and notation as in part a) of the proof, (see formulas (2)-(4)) we have to show that if $S_A(t_1) < \exp(-1)$, then for d sufficiently small,

$$\frac{x}{x+d} < \frac{\log(x+d)}{\log(x)}$$

or, equivalently,

$$d \cdot \log(x) + (x+d)\log\left(1+\frac{d}{x}\right) < 0. \tag{B5}$$

Now, because of $d/x < 1$ we have

$$\log\left(1+\frac{d}{x}\right) = \frac{d}{x} - \left(\frac{d}{x}\right)^2 \cdot \frac{1}{2} + \left(\frac{d}{x}\right)^3 \cdot \frac{1}{3} - \ldots < \frac{d}{x} - \left(\frac{d}{x}\right)^2 \cdot \frac{1}{2}.$$

Thus, in order to prove (B5) it suffices to show that

$$d \cdot \log(x) + (x+d)\left(\frac{d}{x} - \left(\frac{d}{x}\right)^2 \cdot \frac{1}{2}\right) < 0. \tag{B6}$$

Rearranging terms, (B6) is found to be equivalent to

$$\log(x) + 1 + \frac{d}{2x}\left(1 - \frac{d}{x}\right) < 0. \tag{B7}$$

However, since, by assumption, $x = S_A(t_1) < \exp(-1)$, i.e., $\log(x) < -1$, the left-hand side of (B7) will be <0 as $d \to 0$, thus proving the assertion.

ACKNOWLEDGEMENTS

The authors would like to thank the associate editor and the reviewers for their review and very helpful comments that gave us the opportunity to improve the article considerably.

REFERENCES

Abel, U., Jensen, K., Karapanagiotou-Schenkel, I. (2012). Sample sizes for time-to-event endpoints: Should you insure against chance variations in accrual? *Contemporary Clinical Trials* 33:456–458.

Allan, E. (1978). Breast cancer: The error of the exponential. *European Journal of Cancer* 14:1389–1393.

Barthel, F. M.-S., Babiker, A., Royston, P., 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 cross-over. *Statistics in Medicine* 25:2521–2542.

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

Bernstein, D., Lagakos, S. W. (1978). Sample size and power determination for stratified clinical trials. *Journal of Statistical Computation and Simulation* 8:65–73.

Cantor, A. B. (1991). Power estimation for rank tests using censored data: Conditional and unconditional. *Controlled Clinical Trials* 12:462–473.

Cantor, A. B. (1992). Sample size calculations for the log rank test: A Gompertz model approach. *Journal of Clinical Epidemiology* 45:1131–1136.

Case, L. D., Morgan, T. M. (2001). Duration of accrual and follow-up for two-stage clinical trials. *Lifetime Data Analysis* 7:21–37.

Collett, D. (2003). *Modelling Survival Data in Medical Research* (2nd ed.). London: Chapman & Hall.

Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine* 1:121–129.

George, S. L., Desu, M. M. (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases* 28:15–24.

Hade, E. M., Jarjoura, D., Wie, L. (2010). Sample size re-estimation in a breast cancer trial. *Clinical Trials* 7:219–226.

Halpern, J., Brown, Jr., B. (1987). Designing clinical trials with arbitrary specification of survival functions and for the log rank or generalized Wilcoxon test. *Controlled Clinical Trials* 8:177–189.

Halpern, J., Brown, Jr., B. (1993). A computer program for designing clinical trials with arbitrary survival curves and group sequential testing. *Controlled Clinical Trials* 14: 109–122.

Hayat, E. A., Suner, A., Uyar, Ö., Dursun, Ö., Orman, M. N. Kitapcioglu, G. (2010). Comparison of five survival models: Breast cancer registry data from Ege University Cancer Research Center. *Turkiye Klinikleri Journal of Medical Sciences* 30:1665–1674.

Heo, M., Faith, M. S., Allison, D. B. (1998). Power and sample size for survival analysis under the Weibull distribution when the whole life span is of interest. *Mechanism of Ageing and Development* 102:45–53.

Hoos, A., Eggermont, A. M. M., Janetzky, S., Hodi, F. S, Ibrahim, R., Anderson, A., Humphrey, R., Blumenstein, B., Old, L., Wolchok, J. (2010). Improved endpoints for cancer immunotherapy trials. *Journal of the National Cancer Institute* 102(18):1388–1397.

Hsieh, F. Y. (1992). Comparing sample size formulae for trials with unbalanced allocation using the logrank test. *Statistics in Medicine* 11:1091–1098.

Ingel, K., Jahn-Eimermacher, A. (2014). Sample-size calculation and reestimation for a semiparametric analysis of recurrent event data taking robust standard errors into account. *Biometrical Journal* (epublished ahead of print). DOI:10.1002/bimj.201300090.

Jahn-Eimermacher, A., Ingel, K. (2009). Adaptive trial design: A general methodology for censored time to event data. *Contemporary Clinical Trials* 30:171–177.

Jiang, Z., Wang, L., Li, C., Xia, J., Jia, H. (2012). A practical simulation method to calculate sample size of group sequential trials for time-to-event data under exponential and Weibull distribution. *PLOS One* 7(9):e44013. DOI:10.1371/journal.one.0044013.

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

Lakatos, E., Lan, K. K. G. (1992). A comparison of sample size methods for the logrank statistic. *Statistics in Medicine* 11:179–191.

Maller, R. A., Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Chichester, UK: Wiley.

Moghimi-Dehkordi, B., Safaee, A., Pourhoseingholi, M. A., Fatemi, R., Tabeie, Z., Zali, M. R. (2008). Statistical comparison of survival models for analysis of cancer data. *Asian Pacific Journal of Cancer Prevention* 9:417–420.

Oellrich, S., Freischläger, F., Benner, A., Kieser, M. (1997). Sample size determination on survival time data—A review. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* 2:64–85.

Othus, M., Crowley, J. J., Barlogie, B. (2012). Cure-rate survival models in clinical trials. In Crowley, J., Hoering, A. (eds.); *Handbook of Statistics in Clinical Oncology* (3rd ed.) pp. 325–337. Boca Raton, FL: CRC Press.

Pasternack, B. S., Gilbert, H. S. (1971). Planning the duration of long-term survival time studies designed for accrual by cohorts. *Journal of Chronic Diseases* 24:681–700.

Peto, R., Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A* 135:185–206.

Rogon, C. (2009). *Fallzahlplanung bei Daten mit Survivalendpunkt*. Diploma Thesis. München: Department of Statistics, University of München.

Rubinstein, L. V., Gail, M. H., Santner, T. J. (1981). Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases* 34:469–479.

Ryan, T. P. (2013). *Sample Size Determination and Power*. Hoboken, NJ: Wiley.

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

Schoenfeld, D. A., Richter, J. R. (1982). Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* 38:163–170.

Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics* 39:499–503.

Shen, Y., Cai, J. (2003). Sample size reestimation for clinical trials with censored survival data. *Journal of the American Statistical Association* 98:418–426.

Sposto, R., Sather, H. N. (1985). Determining the duration of comparative clinical trials while allowing for cure. *Journal of Chronic Diseases* 38:683–690.

Wang, S. J., Kalpathy-Cramer, J., Kim, J. S., Fuller, C. D., Thomas C. R. (2010). Parametric survival models for predicting the benefit of adjuvant chemoradiotherapy in gallbladder cancer. *AMIA Annual Symposium Proceedings* 2010:847–851.

Wang, S., Zhang, J., Lu, W. (2012). Sample size calculation for the proportional hazards cure model. *Statistics in Medicine* 31:3959–3971.

Wassmer, G. (2006). Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal* 48:714–729.

Xiong, C., Yan, Y., Ji, M. (2003). Sample sizes for comparing means of two lifetime distributions with type II censored data: Applications in an aging intervention study. *Controlled Clinical Trials* 24:283–293.