Received: 16 February 2018 | Revised: 28 October 2018 | Accepted: 14 December 2018

DOI: 10.1002/pst.1927

MAIN PAPER

WILEY

Blinded sample size reestimation in event-driven clinical trials: Methods and an application in multiple sclerosis

Tim Friede[1] | Harald Pohlmann[2] | Heinz Schmidli[3]

[1]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[2]Biostatistics NSO Franchise, Novartis Pharma AG, Basel, Switzerland

[3]Statistical Methodology, Novartis Pharma AG, Basel, Switzerland

Correspondence
Professor Tim Friede, Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany.
E-mail: tim.friede@med.uni-goettingen.de

Motivated by a recently completed trial in secondary progressive multiple sclerosis, we developed blinded sample size reestimation procedures for clinical trials with time-to-event endpoint and assessed their properties in simulation studies. Assuming independent right-censoring and proportional hazards for the two treatment groups, we considered event-driven designs with fixed number of events, which guarantees the power to be at a desired level under a certain alternative. We develop reestimation procedures based on parametric models and show that these maintain the expected duration of the trial at a target length in flexible follow-up designs across a range of nuisance parameter values by adjusting the number of patients recruited into the trial based on blinded nuisance parameter estimates. Furthermore, we provide convincing evidence from a simulation study that such procedures proposed do not inflate the type I error rate in any practically relevant way, thereby satisfying the standards set by relevant international guidelines. Inspired by practical application of these procedures, we outline a number of extensions including methods for extrapolating the observed survival curve beyond the interim time point, application of reestimation procedures to interval censored data, and situations in which a confirmation of event is required leading to a certain lag time.

KEYWORDS
adaptive design, internal pilot study, multiple sclerosis, parametric models, sample size

1 | INTRODUCTION

The idea of using data from the first stage of a two-stage design to estimate a nuisance parameter and to adjust the sample size of the second stage accordingly in order to maintain the power of a hypothesis test based on all data goes back a long way and was described by Stein in the 1940s.[1] In contrast to Stein's proposal that used the variance estimate from the first stage data in the test statistic of the final analysis, it is common practice to use the data of both design stages to estimate the nuisance parameters in the final analysis. In their seminal paper, Wittes and Brittain[2] referred to the latter as the internal pilot study design. The advantages and disadvantages of both approaches were, for example, investigated by Proschan and Wittes,[3] considering sample size reestimation based on unblinded nuisance parameter estimates. Already, from the early paper by Wittes and Brittain,[1] it was clear that the approach using naively a nuisance parameter estimate based on data aggregated across both design stages inflates the type I error rate above the nominal level and various ways of adjusting for this in the final analysis have been proposed over the years.[4-8]

Blinded procedures, which use interim data pooled across treatment code and therefore do not require breaking the treatment code at interim, were found to be equally effective in adjusting the sample size,[9] but with much smaller,

usually practically irrelevant inflations of the type I error rate.[10] These results extend to analyses with baseline adjustments.[11] A number of reviews on nuisance parameter based sample size reestimation are available.[12-16] A comprehensive comparison of blinded and unblinded procedures for variance estimation with continuous data is given in the article by Friede and Kieser.[17] Friede and Miller proposed a blinded continuous monitoring of the variance.[18] With the same expected sample size, the main advantage of several sample size reviews or continuous monitoring over a design with a single review is the reduced variance in the resulting sample size. More complex designs such as longitudinal studies, cluster randomized trials, and three-arm trials with active and placebo controls were considered by Zucker and Denne,[19] Lake et al[20] and Mütze and Friede,[21] respectively. Also, the utility of incorporating external data, eg. summarizing information from several studies on the variance in a meta-analysis[22] with blinded or unblinded data from the ongoing trial in a sample size review using Bayesian methods, was considered.[23-25]

In the ICH E9 guideline on statistical principles, it is stated that the "steps taken to preserve blindness and consequences, if any, for the type I error [...] should be explained".[26] The emphasis of the type I error rate control and maintenance of blinding, which might be interpreted as maintenance of trial integrity, is also reflected in the guidance documents on adaptive designs of the European Medicines Agency (EMA)[27] and Food and Drug Administration (FDA).[28,29] The Committee for Medicinal Products for Human Use (CHMP) reflection paper on adaptive designs[27] reads: "Whenever possible, methods for blinded sample size reassessment [...] that properly control the type I error should be used." This actually places more emphasis on blinded methods than the earlier ICH guideline. A FDA draft guidance included a statement of encouragement to use blinded sample size reestimation (BSSR) methods. In Section V.B, it was stated that "[s]ample size adjustment using blinded methods to maintain desired study power should generally be considered for most studies".[28] More recently, the FDA issued a guidance document on adaptive designs for clinical trials assessing medical devices with similarly encouraging comments.[29]

For a relatively long time, the focus in the field of nuisance parameter-based sample size reestimation has mainly been on continuous and binary endpoints with some more recent interest in recurrent events.[30-37] Also, procedures for blinded continuous monitoring for trials with recurrent event endpoints have very recently been considered.[38] In the context of group-sequential designs, the utility of BSSR procedures for time-to-event endpoints was investigated by Whitehead and colleagues.[39,40] Hade et al[41] McClure et al[42] and Todd et al[43] considered BSSR procedures for time-to-event endpoints with fixed and flexible follow-up, but did not consider specifically event-driven designs. Building on previous work by Lakatos[44,45] and Shih,[46] proposing the use of Markov models for the design of complex survival trials, Cook[47] considers interim analyses with blinded and unblinded data.

Motivated by a recently completed clinical trial in multiple sclerosis (MS), we consider in this paper BSSR procedures for time-to-event endpoints assuming proportional hazards for the two treatment groups. There are mainly two classes of designs for trials with time-to-event outcomes, namely fixed follow-up and flexible follow-up. In the former, all patients are followed up for the same amount of time (or until the event occurs) whereas with the latter patients are followed up until the study closes (or the event occurs), resulting in varying (flexible) follow-up times per patient. As in the example study described below, study close is often defined by reaching a certain number of events, which ensures a power at a desired level under a certain alternative. However, the study duration is not controlled with such a procedure, and may be much longer than anticipated at the planning stage, if for example assumptions on event rates have been too optimistic. Much longer study durations are undesirable, especially in areas with high unmet medical need. Hence, we will consider here event-driven designs where the objective is to complete the study within a given time frame. As we will see, in flexible follow-up designs with fixed number of events and therefore guaranteed power, the expected duration of the trial can be kept at the desired length by adjusting the sample size based on parametric models within the class of proportional hazards models.

The manuscript is organized as follows: The motivating example study in secondary progressive MS is outlined in Section 2. After introducing notation and statistical models in Section 3, we propose BSSR procedures in Section 4. In Section 5, their properties are assessed in a simulation study based on the example study. In Section 6, we draw attention to a few issues that might arise when applying these procedures in practice and sketch out suitable extensions to overcome these. We close with a brief discussion in Section 7.

2 | AN EXAMPLE STUDY IN MS

MS is a chronic neurological disease associated with irreversible progression of physical disability (see eg, the paper by Reich et al[48] for a recent review). It is the most common neurological disorder in younger adults affecting up to 2.5

million people worldwide.[49] While for the earlier stages (relapsing-remitting MS), a variety of treatment options are available now,[50] there is a high unmet medical need for the later stages of the disease, namely secondary progressive MS (SPMS), although a number of attempts have been undertaken over the past decades to develop an efficacious treatment for SPMS.[51] Siponimod (BAF312) is a selective sphingosine 1-phosphate receptor modulator, which is a similar drug class as fingolimod. Siponimod was recently evaluated in SPMS patients in a large phase 3 trial,[52] which is registered in ClinicalTrials.gov under NCT01665144.

In brief, the trial is a multicenter, double-blind, parallel-group, placebo-controlled variable treatment duration study evaluating the efficacy and safety of siponimod in patients with SPMS. The primary objective of the study is to demonstrate the efficacy of siponimod relative to placebo in delaying the time to 3-month confirmed disability progression where disability progression is defined as an increase in the expanded disability status scale (EDSS[53];) of 1 (0.5) point in patients with baseline score of less than or equal to 5.0 (larger 5.0). The original study protocol specified that the study is stopped when the required number of events (patients with confirmed disability progression) is observed. It was estimated that the study duration is 42 months, which, if possible, should not be prolonged.

The addendum to the ICH E9 guideline[54] highlighted the importance of clearly defining the estimand for each clinical trial. Although the SPMS trial was planned well before the release of the ICH document, we provide here information on the primary estimand, which includes details such as the study population, the outcome variable, and a strategy to deal with postrandomization events, so-called intercurrent events. The population consists of SPMS patients with eligibility criteria listed in the paper by Kappos et al.[52] The variable is the time to 3-month confirmed disability progression. A treatment-policy strategy is used for the intercurrent event of treatment discontinuation. The hazard ratio is used as a population-level summary for comparing siponimod with placebo.

For the sample size calculation, several unknown nuisance parameters have to be specified (eg, placebo event rate, enrolment and study discontinuation rate), as described in more detail in Section 4 below. However, knowledge about these nuisance parameters for studies in a SPMS population is limited and variable as a review of the literature demonstrates. For instance, Table 1 gives some examples of 2-year probabilities for disease progression observed on placebo in previous SPMS trials. Furthermore, a very recent systematic review of placebo-controlled trials in progressive MS identified a trend towards lower progression probabilities over the past three decades and shifts in trial populations with recent trials including older and more disabled patients.[61] In a phase 3 randomized placebo-controlled trial, assessing the efficacy and safety of fingolimod in relapsing MS (RMS) a 2-year probability of 25% was reported for the placebo group.[62] In systematic reviews of randomized placebo-controlled trials in relapsing MS, changes in MS trial populations[63] and in disease progression rates[64] have been reported. It is not unlikely that similar trends are also true for SPMS trial populations adding to the uncertainty in planning future trials.

TABLE 1 Examples of 2-year probabilities for disease progression observed on placebo in previous SPMS trials

Study	Population	Confirmation of progression	2-year probability for disease progression	Proportions of patients terminating follow-up prematurely
European Study Group on interferon beta-1b in Secondary progressive MS, 1998[55]	SPMS	3 months	50%	27%
SPECTRIMS, 2001[56]	SPMS	3 months	50%	18% over 3-year follow-up
The North American Study Group on interferon beta-1b in secondary progressive MS, 2004[57]	SPMS	6 months	25%	11% over 3-year follow-up
Pöhlau et al, 2007[58]	SPMS	3 months	67%	52%
MAESTRO-01 (Freedman et al, 2011)[59]	SPMS with HLA haplotype DR2+ or DR4+	3 months	30%	18% over 2-year follow-up
MAESTRO-01 (Freedman et al, 2011)[59]	SPMS with HLA haplotype DR2-/DR4-	3 months	40%	
CUPID (Zajicek et al, 2013)[60]	PMS	6 months	28%	16%

Abbreviations: MS, multiple sclerosis; PMS, progressive multiple sclerosis, ie, primary and secondary progressive multiple sclerosis; SPMS, secondary progressive multiple sclerosis.

With regard to study discontinuation, another nuisance parameter for the sample size calculation, Table 1 gives the proportions of patients terminating follow-up prematurely observed in some earlier SPMS trials. In the figolimod trial in RMS a total of 1033 of the 1272 patients completed the study resulting in study discontinuations of just under 20% over 2 years.[62]

Assumptions on recruitment are based on many factors, including past experience and actual competition for patients. In this study, it was expected that a total of 1530 patients (1020 verum, 510 placebo) need to be recruited, with an estimated enrolment period of 20 months. For the purpose of illustration, we assume that recruitment starts relatively slow with in total nine patients (six verum, three control) per month and increases linearly every month to 90 patients (60 verum, 30 control) in month 10. Thereafter, it is assumed to be fairly stable with 102 patients (68 verum, 34 control) per month in the months 11 to 15 and 105 patients (70 verum, 35 control) per month in the months 16 to 20. This results in a recruitment of 1530 patients (1020 verum, 510 control) over 20 months.

Assuming a reduction of 30% in the hazard rate for confirmed disease progressions by siponimod (hazard ratio of 0.7), a total of 374 events is required to achieve a power of 90% for a one-sided test at 2.5% significance level with a 2:1 treatment allocation (verum: placebo); see Section 4 for details on the calculation. This number of events is expected when recruiting a total of 1530 patients (1020 verum, 510 placebo) as outlined above with a maximum follow-up of 39 months plus 3 months for confirmation (resulting in a study duration of 42 months), assuming a proportion of patients with confirmed disease progression at 24 months of 30% in the placebo group and a 2-year study discontinuation probability of 20% across the two groups. For the event process and the study discontinuation process, independent exponential distributions are assumed; see Section 4 for more details.

The consequences of misspecification of the disease progression rate in the planning on study duration is illustrated in Figure 1. Given the outcomes of the more recent trials, a progression rate of 30% appears to be a reasonable assumption. As can be seen from the figure, however, the trial would be considerably longer if the progression rate was below 30%. To better control the study duration, we consider here BSSR procedures to make the trial more robust against misspecifications of nuisance parameters regarding progression, study discontinuation, and recruitment in the planning phase. With the help of a blinded sample, size reestimation procedure, the prestudy assumptions regarding progression, study discontinuation, and recruitment are assessed in a blinded review and, if necessary, the number of patients to be enrolled is increased, prior to completion of enrolment, under the constraint of a maximum study duration.

3 | NOTATION AND STATISTICAL MODEL

We consider a randomized controlled clinical trial with two treatment groups $i = 1, 2$ and n_i patients per group in a fixed sample size design. In a design with internal pilot study, we denote the sample size for stage $s = 1, 2$ and group i by n_{si}.

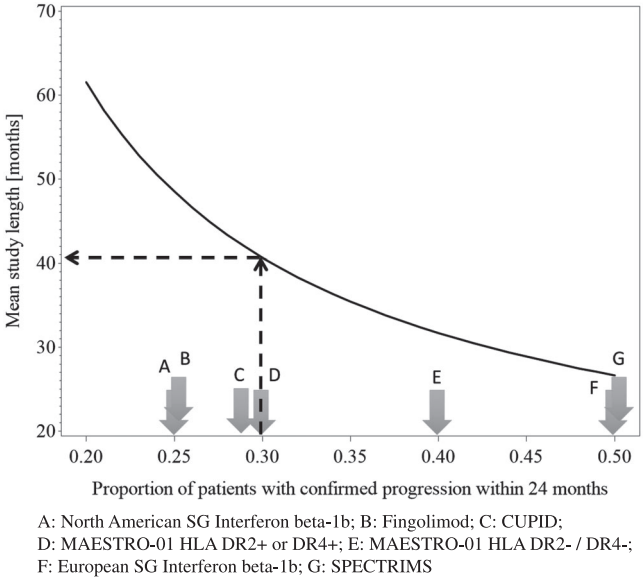

FIGURE 1 Consequences of misspecifying the proportion of patients with confirmed disease progression at 24 months in the placebo group on study duration. The siponimod trial was planned assuming a proportion of 30% in the placebo group achieving a duration of 42 months (39 months follow-up plus 3 months for confirmation). The corresponding proportions observed in the placebo control groups of some previous trials are indicated by arrows

The time to event and time to censoring (eg, administrative reasons or study discontinuation) for patient j in group i are denoted by T_{ij} and C_{ij}, respectively. The observation time, ie, the minimum of the time to event and the time to censoring, and event indicator of subject j in group i are denoted by t_{ij} and δ_{ij}, respectively, with $\delta_{ij} = 0$ indicating a censored observation and $\delta_{ij} = 1$ an event.

In what follows below, we assume independent right censoring and the same censoring process across treatment groups resulting in the following likelihood

$$\prod_{i,\,j} f_i(t_{ij})^{\delta_{ij}}\, S_i(t_{ij})^{1-\delta_{ij}}\, g(t_{ij})^{1-\delta_{ij}}\, G(t_{ij})^{\delta_{ij}},$$

with density functions f and g, and survival functions S and G for the time-to-event and time-to-censoring process, respectively. Since we are primarily interested in the time-to-event process and not the time-to-censoring process, the likelihood for inference on the time-to-event process reduces to

$$L = \prod_{i,\,j} f_i(t_{ij})^{\delta_{ij}}\, S_i(t_{ij})^{1-\delta_{ij}}.$$

Popular parametric models for time-to-event endpoints include the exponential, piecewise exponential, and Weibull distributions. The former assumes constant hazards and is commonly used in planning clinical trials since often only crude estimates of average event rates over follow-up are available in the design stage. Although the methodology is introduced more generally, we assume in large parts of the manuscript, in particular in the simulation study, that the time-to-event and time-to-censoring processes follow exponential distributions. Hence, the survival function for the event processes in group i is given by $S_i(t) = e^{-\lambda_i t}$ with constant hazard function λ_i. It follows that the density function is $f_i(t) = \lambda_i e^{-\lambda_i t}$. The survival function for the censoring process is given by $G(t) = e^{-\gamma t}$ with constant hazard function γ. The density function is then given by $g(t) = \gamma e^{-\gamma t}$.

Extending the model, we consider now piecewise exponential and Weibull distributions for the event and censoring processes. The piecewise exponential model extends the exponential model by assuming constant hazards over certain time intervals. The Weibull model is a parametric alternative to the semi-parametric Cox model, since it belongs to the same class of proportional hazards models.[65] The Weibull model belongs also to the class of accelerated failure time models, which offers an alternative interpretation. The survival functions for the event and censoring process are then given by $S_i(t) = \exp(-\lambda_i t^{\sigma})$ and $G(t) = \exp(-\gamma t^{\kappa})$, respectively, where σ and κ are the shape parameters. The (nonconstant) hazard functions are $\sigma\, t^{\sigma-1}\, \lambda_i$ and $\kappa\, t^{\kappa-1}\, \gamma$, respectively. Please note that the shape parameter σ of the event process is assumed to be constant across treatment groups. This is a common default in standard software (eg, PROC LIFEREG in SAS).

Here, we consider event-driven trials that are completed once a prespecified number of events are observed. With censoring, the required number of events might never be reached and consequently statistics such as the expected trial duration are infinite. In practice, this would be dealt with by imposing a maximum trial duration. When we refer to characteristics such as the mean duration, we implicitly assume that all trials close in finite time by either reaching the required number of events or a maximum trial duration.

4 | PROPOSED BLINDED SAMPLE SIZE REESTIMATION PROCEDURE

Interest is in testing the null hypothesis that the hazard ratio is equal to 1, ie, $H_0 : \theta = 1$. In the exponential and Weibull models discussed above, the hazard ratio is $\theta = \lambda_2/\lambda_1$. In proportional hazards models generally, the required total number of events d across both treatment groups to yield a power $1 - \beta$ testing the null hypothesis H_0 at a one-sided significance level α can be approximated by the so-called Schoenfeld formula,[66] which is given by

$$d = \frac{(1+k)^2}{k}\,\frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\log\left(\theta^*\right)^2},$$

where the hazard ratio θ^* denotes the assumed effect under the alternative hypothesis and patients are randomized using an k:1 allocation ratio of treatment group 1 to treatment group 2. If a two-sided test at level α is used, then $z_{1-\alpha}$ is replaced by $z_{1-\alpha/2}$ in the formula above.

Denoting the number of events in treatment group i by D_i and the number of patients recruited into group i in month l by r_{il} with $l = 1, \ldots, R$ and $R < L$ the expected number of events in treatment group i with administrative censoring after month L (end of study) using the general parametric notation is given by

$$\mathrm{E(D_i)} = \sum_{\mathrm{l=1}}^{\mathrm{R}} \mathrm{r_{il}\, P\big(T_{ij} < C_{ij}, T_{ij} < L\big)} = \sum_{\mathrm{l=1}}^{\mathrm{R}} \mathrm{r_{il}} \int_0^{\mathrm{L}} \mathrm{f_i(t)\, G_i(t)\, dt}.$$

With independent exponential event and censoring times, notation as introduced above and administrative censoring after month L the integral $\int_0^L f_i(t)\, G_i(t)\, dt$ is given by $\lambda_i/(\lambda_i + \gamma)\, \left(1 - e^{-(\lambda_i+\gamma)(L-l)}\right)$, as the probability $P(T_{ij} < C_{ij})$ is given by $\int_0^\infty \lambda_i e^{-\lambda_i t} e^{-\gamma t}\, dt,$ which amounts to $\lambda_i/(\lambda_i + \gamma)$. Hence, the expression for the expected number of events in group i simplifies to

$$\mathrm{E(D_i)} = \sum_{\mathrm{l=1}}^{\mathrm{R}} \mathrm{r_{il}} \frac{\lambda_\mathrm{i}}{\lambda_\mathrm{i} + \gamma} \left(1 - \mathrm{e}^{-(\lambda_\mathrm{i}+\gamma)(\mathrm{L-l})}\right).$$

A simpler representation (not considering the accrual process) and its extension to piecewise exponential hazards can be found in Appendices 1 and 2 of the paper by Whitehead[39], respectively. Relaxing the assumption of exponential event and censoring times, we now assume that the event and censoring times follow Weibull distributions. The group-specific event rates and the common shape parameter of the event process are denoted by λ_i and σ, respectively. The censoring process is assumed to be the same across treatment groups with event rate $\gamma$and shape parameter κ. The integral $\int_0^L f_i(t)\, G_i(t)\, dt$ has under these assumptions no explicit solution and (one-dimensional) numerical integration is necessary to compute the expected number of events. If Weibull event times and exponential censoring times are assumed, however, an explicit solution is available and the expected number of events in group i is given by

$$\mathrm{E(D_i)} = \sum_{\mathrm{l=1}}^{\mathrm{R}} \mathrm{r_{il}} \frac{\lambda_\mathrm{i}}{\lambda_\mathrm{i} + \gamma} \left(1 - \mathrm{e}^{-(\lambda_\mathrm{i}+\gamma)(\mathrm{L-l})^\sigma}\right).$$

In a blinded review including data of $n_{1+} = n_{11} + n_{12}$ patients, parametric models for the time-to-event and time-to-censoring processes are fitted to the blinded observations pooled across the treatment groups. The time-to-censoring process is assumed to be identical across the treatment groups and can therefore be estimated from the pooled sample. As the time-to-event processes differ between the treatment groups under the alternative, we borrow here an idea from Whitehead et al.[40] Under the assumed hazard ratio θ^*, the time to event process estimated from the combined sample is split into treatment-specific estimates. Specifically, with exponential event times the overall hazard rate $\overline{\lambda}$, ie, the weighted average $\overline{\lambda} = k/(k+1)\, \lambda_1 + 1/(k+1)\lambda_2$ of the two group specific hazard rates λ_i can be estimated from the blinded data pooled across treatment groups. Assuming a certain hazard ratio θ^* the hazard rate of the control group can be written as

$$\lambda_2 = \frac{1 + \mathrm{k}}{1 + \mathrm{k}/\theta^*}\, \overline{\lambda}.$$

From this representation of λ_2 and $\lambda_1 = \lambda_2/\theta^*$, it is apparent that the event rates λ_1 and λ_2 can be estimated in a blinded fashion. If Weibull distributions are assumed for the event times, a Weibull distribution is fitted to the blinded data pooled across both treatment groups. The estimate of the common shape parameter σ is readily available whereas the overall event rate is split into group specific ones as with the exponential distributions above.

1. The blinded sample size reestimation proceeds in the following steps: Parametric distributions are fitted for the time to event and time to censoring processes using the blinded data pooled across treatment groups. Furthermore, figures on past and predicted future recruitment are gathered.
2. Based on these estimates, the trial design is re-evaluated with regard to the expected number of events given the originally planned sample size.
3. If the expected number is at least the required number d, no change to the design is necessary. The algorithm stops here.
4. If the expected number is smaller than d, the design needs to be changed to maintain the chances to finish the study within the targeted timeframe of L months. The sample size is increased so that the expected number of events

within L months is equal to the required number of events. In practice, we propose to iterate the sample size starting with the originally planned one and prolonging recruitment by a certain time interval, say 1 month. For every sample size, the expected number of events is evaluated. The iterations continue until the expected number of events is at least as large as d.

In the following, we refer to this procedure as the BSSR design. The operating characteristics of this procedure are assessed in the next section.

5 | SIMULATION STUDY

In this section, we describe a simulation study that was carried out to assess the operating characteristics of the design with BSSR in comparison with the fixed sample size design. Here, we use the so-called clinical scenario evaluation (CSE) framework[67,68] to structure the report of the simulation study. The CSE framework consists of three components, namely assumptions, design options, and metrics. The assumptions and design options define the simulation scenarios. The metrics are the measures used to choose a design. In the following, we describe all three components for the simulation study conducted to assess the operating characteristics of the blinded reestimation procedure.

5.1 | Setup of the simulation study

Assumptions: The recruitment is assumed to follow the pattern described in Section 2, resulting in the recruitment of 1530 patients (1020 verum, 510 control) over 20 months. In any additional month of recruitment, 102 patients (68 verum, 34 control) are included into the trial. The simulation study considers independent exponentially distributed event and censoring times. With this assumption of exponential event times, the rates of the control group are chosen so that the 24 month probability of confirmed disease progression takes values between 20% to 30%. The alternative of interest is 0.7 in terms of a hazard ratio of verum to control, but other hazard ratios ranging from 0.5 to 1 are also considered. Assuming an exponential distribution for the time to withdrawal, the study discontinuation rate was fixed at a 24 month probability of 20% across both treatment arms. For the ease of computation, the likelihood ratio test based on exponential event times is used; the nominal significance level is 5% two-sided.

Design options: The following designs are considered: (a) the fixed sample size design and (b) the design with BSSR. In the fixed sample size design, the total sample size is 1530 patients. This is the sample size calculated to achieve 374 events in a 39 month study (plus 3 months for event confirmation) with the recruitment as described above, a 24 month event probability of 30% in the placebo group and a hazard ratio of 0.7. In the BSSR designs the minimum total sample size is 1530 patients, ie, the required size of a fixed design under the planning assumptions. The blinded review is carried out at month 18, by which time, 1320 patients are recruited into the study. In the designs with BSSR, the expected number of events are recalculated based on blinded estimates of hazard rates for the time to confirmed disease progression and the time to study discontinuation (assuming exponential distributions). Based on these estimates, it is then considered to extend recruitment in discrete steps by one or more months so that a total of 374 events can be expected within 39 months (plus 3 months for event confirmation). Two choices of maximum total sample sizes, ie, 1836 and 2142 patients, are considered to study the dependence of some characteristics on this design parameter. The total sample sizes correspond to extending the recruitment by up to 3 months and up to 6 months, respectively, under the assumed recruitment rate of 102 patients per month (68 verum, 34 control) for any additional month of recruitment.

Metrics: As operating characteristics, we are studying the rejection probability (ie, type I error rate under the null hypothesis and power under the alternative), the mean trial duration, and the mean sample size.

5.2 | Simulation results

Table 2 gives the simulated type I error rates summarized across 24-month confirmed progression probabilities 0.20, 0.21, ..., 0.30 in the control group. The following designs were considered: namely the fixed design with a total sample size of 1530 patients, and two BSSR designs with total sample sizes of 1836 and 2142 patients. A study is finished once a total of 374 events were observed. Per scenario 100 000 trials were simulated. For all three designs, the type I error rates are close to the nominal significance level of 0.05 (two-sided) and observed deviations from the nominal level are largely

TABLE 2 Simulated type I error rates summarized across 11 scenarios with proportions 0.20, 0.21, ..., 0.30 of patients with confirmed progression within 24 months for three designs, namely the fixed design with a total sample size of 1530 patients, and two blinded sample size reestimation designs with total sample sizes of 1836 and 2142 patients with a fixed total number of 374 events (100 000 simulation replications per scenario). The nominal significance level is 0.05 (two-sided). Details of the simulation scenarios are given in the text

Designs	Mean	SD	Minimum	Maximum
Fixed design with 1530 patients	0.0503	0.0004	0.0498	0.0510
BSSR with max. 1836 patients	0.0504	0.0006	0.0493	0.0512
BSSR with max. 2142 patients	0.0503	0.0008	0.0492	0.0517

within the simulation error. As additional patients are recruited into the designs with BSSR, those studies finished earlier than the fixed design studies. Most pronounced are these differences for the smallest proportion considered, namely 20% of patients with a confirmed disease progression within 24 months in the control group. Here, the fixed design studies need on average 49.6 months to completion, whereas the average completion times of the BSSR designs are with 43.3 and 39.9 months for maximum sample sizes of 1836 and 2142, respectively, much closer to the target of 39 months.

Simulated mean study length in months and mean number of additional patients under the planning alternative (ie, a hazard ratio of 0.7) are given in Figure 2 depending on the proportion of patients with confirmed progression within 24 months. Per scenario, 10 000 trials were simulated. As these are event-driven trials and the true treatment effect equals the alternative assumed in the planning, the target power of 90% is always achieved. In fact, as we use the parametric LRT for computational convenience, the power is on average across all scenarios and designs with 91.8% slightly higher than 90%; this effect is consistent across the designs considered. The average durations are close to the target of 39 months for all three designs when the event probability is 0.3 as assumed, with the BSSR designs being on average slightly shorter than with the fixed design at the price of higher mean sample sizes. However, the durations are markedly different between the fixed and BSSR designs with lower event probabilities. For instance, with an event probability of 0.25 in the control group, the BSSR design with the higher maximum sample size achieves a mean study length of 39.3 months, close to the target of 39 months whereas the fixed design takes on average 48.5 months to complete, ie, almost 10 months longer than the target. The shorter trial duration in the BSSR design comes at the cost of an increased sample size. The average total sample size is about 600 patients higher than in the fixed design (Figure 2, right panel). With an event probability of 0.2 in the control group, the fixed design needs on average 61.5 months to complete whereas the BSSR designs need means of 51.9 and 46.6 months to completion with maximum sample sizes of 1836 and 2142 patients, respectively. In order to achieve mean durations closer to the target in the BSSR designs, the maximum sample size would need to be increased.

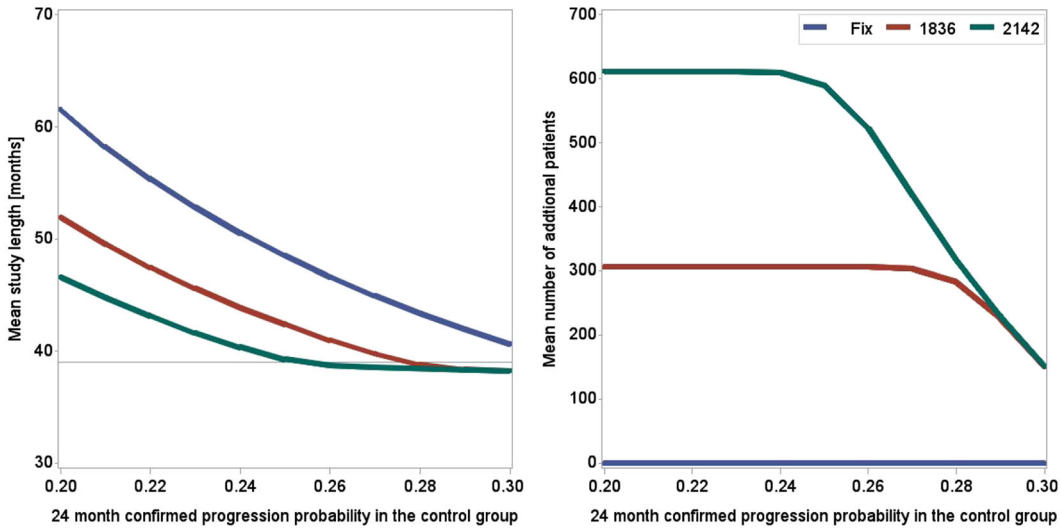

FIGURE 2 Simulated mean study length in months and mean number of additional patients depending on the proportion of patients with confirmed progression within 24 months in the control group for a study with a fixed total number of 374 events under the planning alternative of a hazard ratio of 0.7 (10 000 simulation replications per scenario). Three designs are considered: the fixed design (Fix) with a total sample size of 1530 patients and two blinded sample size reestimation designs with maximum sample sizes of 1836 and 2142 patients. Details of the simulation scenarios are given in the text

So far, we have considered the operating characteristics under the null hypothesis and under the planning alternative. Now, we will consider a range of alternatives where the effect size will be different from the planning assumption. The power varies greatly across the considered control rates and hazard ratios, but is always very similar for the three designs as the total number of events is always 374. For all scenarios considered, the mean durations are favorable for the BSSR designs in comparison with the fixed design (see Figure 3, panels in top row). Although the BSSR procedure uses the effect used in the planning to split the overall rates into group specific rates, the operating characteristics are still favorable in comparison with the fixed design even with this assumption being not correct. Again, the shorter trial durations in the BSSR designs come at a price of increased sample sizes (see Figure 3, panels in the bottom row).

6 | EXTENSIONS

Inspired by practical application of these procedures, we outline a number of extensions including application of reestimation procedures to interval censored data, and situations in which a confirmation of event is required leading to a certain lag time.

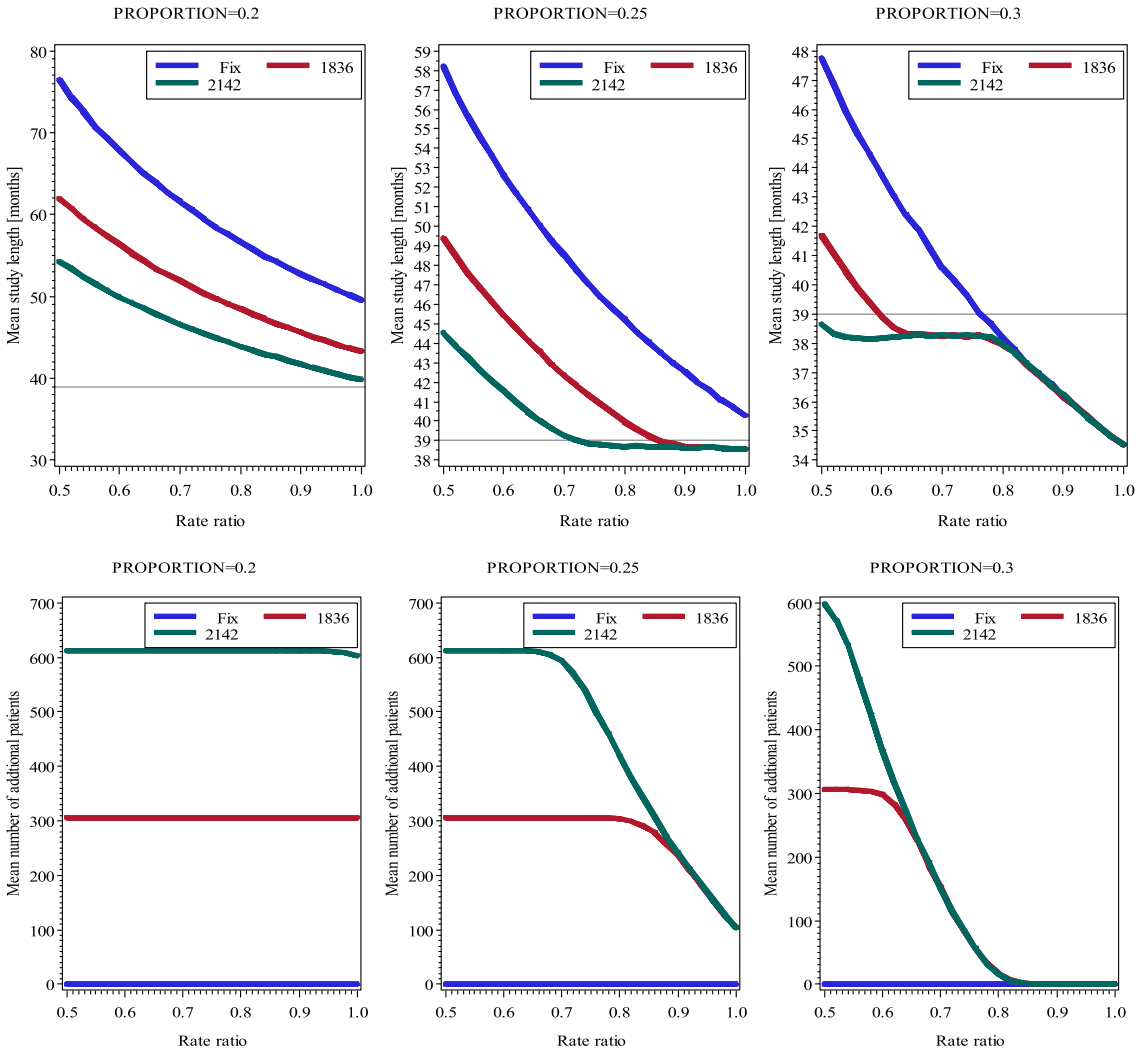

FIGURE 3 Simulated mean study length in months (top row) and mean number of additional patients (bottom row) depending on the hazard ratio and the proportion of patients with confirmed progression within 24 months in the control group for a study with a fixed total number of 374 events (2000 simulation replications per scenario). Three designs are considered: the fixed design (Fix) with a total sample size of 1530 patients and two blinded sample size reestimation designs with maximum sample sizes of 1836 and 2142 patients. Details of the simulation scenarios are given in the text

6.1 | Methods for extrapolation

In our example study, the expected follow-up was 42 months (39 months plus 3 months for confirmation), and therefore the event and study discontinuation processes need to be specified over the whole time period for power calculations. However, when reestimating the required sample size at the blinded review, the follow-up times were much shorter. Therefore, the estimated event and study discontinuation processes have to be extrapolated beyond the observed time period.

Different approaches to the extrapolation of survival functions beyond the observed time period in an interim analysis have been proposed. Here, we refer to Whitehead et al[40] who suggested to estimate at interim the deviation of the observed survival function across the treatment arms from the one assumed in the planning phase. The deviation was quantified as the average difference between the observed and assumed survival curves on the complementary log-log scale over the observed time period. The initially assumed survival curve is then shifted by the estimated deviation on the complementary log-log scale to inform the redesign of the study. Todd et al[43] suggested a modification of this approach by considering only the deviation (again, on the complementary log-log scale) at the latest time point rather than the average over the observed time period. In a comparison of both approaches, Todd et al concluded that both protect the power of the trial, but neither dominates the other for all scenarios considered.

Parametric models in principle lend themselves naturally to extrapolate beyond the observed time period, which has previously been noted by Hade et al (see Method B in the paper by Hade et al[41]). Relying on a parametric model requires some external data to verify the choice of parametric model. As is apparent from the review of the literature presented in Section 2, some data on the endpoint are available from SPMS and RMS. More flexible models will be less prone to bias at the cost of a larger variation in the parameter estimates and therefore in the resulting sample size.

Figures 4 shows Kaplan-Meier curves with 95% confidence intervals of the times to 3-month confirmed disease progression as observed in the placebo group in Kappos et al[62] which was a 2 year study assessing the efficacy and safety of fingolimod in relapsing MS. To these data, a variety of parametric models were fitted including exponential distributions, Weibull distributions, and piecewise exponential distributions. For the latter, the intervals of constant hazards were defined as 0-90, 91-360, and greater than 360 days. As can be seen from Figure 4, the relatively simple model of piecewise constant hazards over three intervals provides satisfactory fit whereas the exponential and Weibull distributions fit the data only well for the first 12 months or so. The Weibull model with the additional parameter improves the fit over the exponential model, although marginally.

6.2 | Confirmation of disability progression

As described in Section 2, disability progression has to be confirmed 3 months after a relevant increase in the EDSS score from baseline has been observed. At the time of the blinded review, an increase in EDSS may have been seen

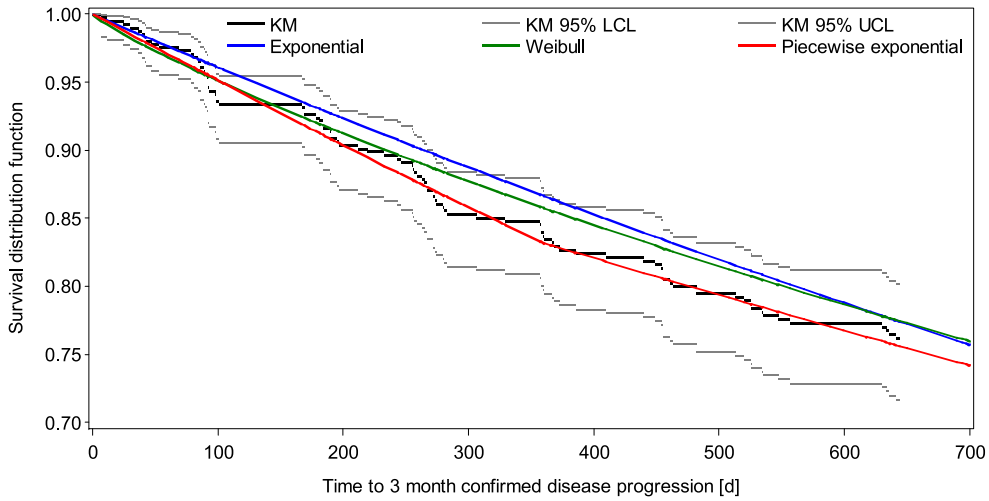

FIGURE 4 Kaplan-Meier curve (with 95% confidence intervals), exponential, Weibull and piecewise exponential fit to the time to 3 month confirmed disease progression of the placebo data in Kappos et al.[62] For the piecewise exponential fit the intervals of constant hazards were defined as 0-90, 91-360, and >360 days

for some of the patients, but without a confirmation. Censoring such patients with progression that is not yet confirmed (or equally counting them as events) would be informative and lead to bias when using a standard survival analysis.[69] The setting can be considered a missing data problem, where the confirmation status is missing for some of the patients. Wang et al[70] proposed a model-based analysis to handle such situations. An alternative would be to use multiple imputation,[71] which is in spirit similar to the approach by Cook and Kosorok,[69] where a progression with pending confirmation is weighted according to the probability that it will be confirmed. With a multiple imputation approach, the confirmation status for patients with unconfirmed progression is multiply imputed, resulting in several complete datasets. For each of these datasets, a standard survival analysis is then done, and the results are combined. A simple imputation model may assume that the confirmation status is distributed as *Bernoulli(π),* where π is the probability that a patient with observed progression will be confirmed. At the time of the interim analysis, suppose that m patients are available for which both an EDSS increase has been observed and the confirmation status is known (r of the m patients have a confirmed progression). Then, using a uniform prior for π, the Bayesian posterior distribution for π is *Beta($1 + r$, $m\text{-}r + 1$)*. The confirmation status is then imputed from the predictive distribution by repeatedly drawing first a value π^* from *Beta($1 + r$,$m\text{-}r + 1$)*, and then the confirmation status Y* from *Bernoulli(π^*)*. More sophisticated models could also be considered, for example, by including covariates in the statistical model for the confirmation status, similar to Brownstein et al.[72] The issue of delayed adjudication considered here is related but different from delayed ascertainment of events considered by Hu and Tsiatis[73] and Fine and Tsiatis.[74] If events are only ascertained with potential delay, an appropriate estimator of the survival function (such as the one proposed by Hu and Tsiatis[73]) needs to be employed in the sample size review to minimize bias in the assessment of the event probability.

7 | CONCLUSIONS AND DISCUSSION

The characteristics of study populations change over the years; This is particularly true for diseases, where a number of treatments were successfully trialed and licensed. MS is a good example, as now a range of treatments are available for its relapsing form and changes in study populations are well documented.[63,64,75] In progressive MS, a number of attempts to develop therapies have been undertaken which only very recently resulted in licensing of a drug. Also, for progressive MS, changes in study populations have been observed. However, the problem is not restricted to MS, but is typical for other chronic conditions, too. We cite here a recent systematic review in chronic obstructive pulmonary disease (COPD) as an example.[76] These changes often lead to a decrease in event probability, which means that trials are at risk to be underpowered when their design was based on the outcomes of past trials. BSSR procedures can mitigate these risks by adapting the sample size if event rates observed in the ongoing trial are lower than expected. Thereby, they raise the chances that a trial is robust and conclusive with usually no or practically irrelevant impact on the type I error rate. Here, we proposed a BSSR procedure for event-driven trials. Although event-driven designs are quite common in some disease areas such as oncology or cardiovascular disease, event-driven trials have not got the necessary attention in the context of BSSR. With this manuscript we close this gap.

Motivated by a recently completed phase 3 trial in SPMS, we conducted a simulation study to assess the operating characteristics of the proposed BSSR procedure and to compare them with those of a fixed sample size design. No inflation of the type I error rate due to the monitoring of the events was expected (see eg, Section 3.1 in the paper by Cook[47]), but the question was whether the reestimation of the sample sample size might result in an increased type I error rate. We did not find any indication that the BSSR inflates the type I error rate. Therefore, it fulfills regulatory requirements as set out by international guidelines.[26-29] Furthermore, the proposed blinded strategy has favorable properties as compared with the fixed design in terms of trial duration. When the event rate is lower than expected, the sample size will be increased accordingly in the BSSR design to compensate for the lower event rate. Then, the trial can be expected to be completed on time. If such strategies are employed across trials of a clinical development program, the timely completion of the entire program and with it, access to market can be expected.

A number of authors have considered design modifications in the context of time-to-event trials highlighting both the need and complexity of this kind of methodology. For instance, Cook[47] considers blinded and unblinded design adaptations informed by Markov models. These models and adaptations are more general than those considered here. We focus on completing event-driven trials within a given time frame, which is of high interest from a practical point of view as we argue above. Whereas the work by Cook[47] is more general, we provide specific insights of the operating characteristics of blinded adaptations through a range of simulation studies. Others including Hade et al[41] McClure et al[42] and Todd et al[43] have also focussed on BSSR procedures, but did not consider specifically event-driven designs.

In SPMS trials, the clinic visits are usually scheduled every 3 or 6 months, at which the EDSS as a measure of disability is assessed. This means that only the time interval is known in which the disease progression occurred and therefore the event times are, strictly speaking, interval censored. However, standard survival analyses not accounting for interval censoring are commonly applied. Given the number of visits within the follow-up time of typically 2 to 3 years, this approach appears to be reasonable as the visits are sufficiently frequent.[77] Nevertheless methods appropriate for interval censored data were recently applied in a primary progressive MS cohort.[78] Furthermore, the relationship between number of visits and length of follow-up changes if only shorter follow-up times are observed in interim analyses. This might indeed require the use of appropriate techniques for handling interval censoring (see eg, the article by Lindsey and Ryan[79]). The application of such methods is facilitated through the wider availability of software to carry out such analyses (eg, PROC ICPHREG and PROC ICLIFETEST in SAS).

If trial populations change over time, the utility of data from previous trials in informing the design of a new trial might be limited. In these settings, adaptive designs such as the proposed sample size reestimation procedure are useful tools to deal with the uncertainty. However, this requires that the trial population does not change much during the course of the study. This point has been considered in the context of adaptive designs, in particular with regard to homogeneity of the treatment effect over time.[80]

In developing the BSSR method and in the simulation studies, we assumed that the censoring mechanism is independent of the event process and the same across the treatment groups. In practice, however, one or both of the assumptions might not be valid. Blinded reviews would not reveal any differences in treatment arms. Whether and, if yes, under which circumstances, unblinded interim analyses and adaptation strategies could be a remedy is subject to future research.

Some extensions of the proposed procedure were sketched out. These included methods for extrapolations and the delayed confirmation of the events. The latter is similar to problems occurring with the adjudication of events, which is common practice in cardiovascular trials. Here, we used standard parametric models including exponential, piecewise exponential, and Weibull distributions to extrapolate beyond the observation period. A more flexible parametric framework for parametric analyses of time-to-event data and an extension to multistate models have been suggested recently,[81,82] which could potentially be applied for extrapolation purposes in sample size reviews and, of course, also in the final analysis. This will be considered in future research.

CONFLICTS OF INTEREST

T.F. provided consultancies to Novartis Pharma AG regarding sample size reestimation strategies for the MS study that served as an example in this paper. H.P. and H.S. are employees of Novartis Pharma AG.

ORCID

Tim Friede https://orcid.org/0000-0001-5347-7441
Heinz Schmidli https://orcid.org/0000-0003-0934-1066

REFERENCES

1. Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann Math Stat*. 1945;16(3):243-258.
2. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficacy of clinical trials. *Stat Med*. 1990;9(1-2):65-72.
3. Proschan MA, Wittes J. An improved double sampling procedure based on the variance. *Biometrics*. 2000;56(4):1183-1187.
4. Coffey CS, Muller KE. Exact test size and power of a Gaussian error linear model for an internal pilot study. *Stat Med*. 1999;18(10):1199-1214.
5. Denne JS, Jennison C. Estimating the sample size for a t-test using an internal pilot. *Stat Med*. 1999;18(13):1575-1585.
6. Kieser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Stat Med*. 2000;19(7):901-911.
7. Coffey CS, Muller KE. Controlling test size while gaining the benefits of an internal pilot design. *Biometrics*. 2001;57(2):625-631.
8. Miller F. Variance estimation in clinical studies with interim sample size reestimation. *Biometrics*. 2005;61(2):355-361.
9. Friede T, Kieser M. A comparison of methods for adaptive sample size adjustment. *Stat Med*. 2001;20(24):3861-3873.
10. Kieser M, Friede T. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Stat Med*. 2003;22(23):3571-3581.

11. Friede T, Kieser M. Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharm Stat*. 2011;10(1):8-13.
12. Proschan M. Two-stage sample size reestimation based on nuisance parameter: a review. *J Biol Res*. 2005;15(4):559-574.
13. Proschan MA. Sample size re-estimation in clinical trials. *Biom J*. 2009;51(2):348-357.
14. Friede T, Kieser M. Sample size recalculation in internal pilot study designs: a review. *Biom J*. 2006;48(4):537-555.
15. Phillips AJ, Keene ON. Adaptive designs for pivotal trials: discussion points from the PSI adaptive design expert group. *Pharm Stat*. 2006;5(1):61-66.
16. Chuang-Stein C, Anderson K, Gallo P, Collins S. Sample size reestimation: a review and recommendations. *Drug Inf J*. 2006;40(4):475-484.
17. Friede T, Kieser M. Blinded sample size re-estimation in superiority and non-inferiority trials: bias versus variance in variance estimation. *Pharm Stat*. 2013;12(3):141-146.
18. Friede T, Miller F. Blinded continuous monitoring of nuisance parameters in clinical trials. *J R Stat Soc Ser C*. 2012;61(4):601-618.
19. Zucker DM, Denne J. Sample-size redetermination for repeated measures studies. *Biometrics*. 2002;58(3):548-559.
20. Lake S, Kammann E, Klar N, Betensky R. Sample size re-estimation in cluster randomization trials. *Stat Med*. 2002;21(10):1337-1350.
21. Mütze T, Friede T. Blinded sample size re-estimation in three-arm trials with 'gold standard' design. *Stat Med*. 2017;36(23):3636-3653.
22. Schmidli H, Neuenschwander B, Friede T. Meta-analytic-predictive use of historical variance data for the design and analysis of clinical trials. *Computational Statistics & Data Analysis*. 2017;113:100-110.
23. Hartley AM. Adaptive blinded sample size adjustment for comparing two normal means—a mostly Bayesian approach. *Pharm Stat*. 2012;11(3):230-240.
24. Hartley AM. A Bayesian adaptive blinded sample size adjustment method for risk differences. *Pharm Stat*. 2015;14(6):488-514.
25. Mütze T, Schmidli H, Friede T. Sample size re-estimation incorporating prior information on a nuisance parameter. *Pharm Stat*. 2018;17(2):126-143.
26. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceutical for Human Use (ICH). ICH harmonised tripartite guideline: statistical principles for clinical trials. *Stat Med*. 1999;18:1905-1942.
27. Committee for Medicinal Products for Human Use (CHMP). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. Available from http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf
28. Food and Drug Administration (FDA). Guidance for industry: adaptive design clinical trials for drugs and biologics, draft, 2010. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201790.pdf
29. Food and Drug Administration (FDA). Guidance for industry and Food and Drug Administration staff: adaptive designs for medical device clinical studies, 2016. Available from: https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf
30. Cook RJ, Bergeron PJ, Boher JM, Lie Y. Two-stage design of clinical trials involving recurrent events. *Stat Med*. 2009;28(21):2617-2638.
31. Friede T, Schmidli H. Blinded sample size reestimation with count data: methods and applications in multiple sclerosis. *Stat Med*. 2010a;29(10):1145-1156.
32. Friede T, Schmidli H. Blinded sample size reestimation with negative binomial counts in superiority and non-inferiority trials. *Methods Inf Med*. 2010b;49:618-624.
33. Schneider S, Schmidli H, Friede T. Robustness of methods for blinded sample size reestimation with overdispersed count data. *Stat Med*. 2013;32(21):3623-3635.
34. Schneider S, Schmidli H, Friede T. Blinded and unblinded internal pilot study designs for clinical trials with count data. *Biom J*. 2013;55(4):617-633.
35. Schneider S, Schmidli H, Friede T. Blinded sample size reestimation for recurrent event data with time trends. *Stat Med*. 2013;32(30):5448-5457.
36. Asendorf T, Henderson R, Schmidli H, Friede T. Modelling and sample size reestimation for longitudinal count data with incomplete follow up. *Stat Methods Med Res*. 2019;28(1):117-133.
37. Asendorf T, Henderson R, Schmidli H, Friede T. Sample size re-estimation for clinical trials with longitudinal negative binomial counts including time trends. *Stat Med*. 2018. in press. https://doi.org/10.1002/sim.8061
38. Friede T, Häring DA, Schmidli H. Blinded continuous monitoring in clinical trials with recurrent event endpoints. *Pharm Stat*. 2018. in press. https://doi.org/10.1002/pst.1907
39. Whitehead J. Predicting the duration of sequential survival studies. *Drug Inf J*. 2001;35(4):1387-1400.
40. Whitehead J, Whitehead A, Todd S, Bolland K, Sooriyarachchi MR. Mid-trial design reviews for sequential clinical trials. *Stat Med*. 2001;20(2):165-176.
41. Hade EM, Jarjoura D, Wei L. Sample size re-estimation in a breast cancer trial. *Clin Trials*. 2010;7(3):219-226.

42. McClure LA, Szychowski JM, Benavente O, Coffey CS. Sample size reestimation in an on-going NIH-sponsored clinical trial: the secondary prevention of small subcortical strokes experience. *Contemp Clin Trials*. 2012;33(5):1088-1093.
43. Todd S, Valdes-Marquez E, West J. A practical comparison of blinded methods for sample size reviews in survival data clinical trials. *Pharm Stat*. 2012;11(2):141-148.
44. Lakatos E. Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Control Clin Trials*. 1986;7(3):189-199.
45. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*. 1988;44(1):229-241.
46. Shih JH. Sample size calculation for complex clinical trials with survival endpoints. *Control Clin Trials*. 1995;16(6):395-407.
47. Cook TD. Methods for mid-course corrections in clinical trials with survival outcomes. *Stat Med*. 2003;22(22):3431-3447.
48. Reich DS, Lucchinetti CF, Calabresi PA. Multiple sclerosis. *N Engl J Med*. 2018;378(2):169-180.
49. WHO (2006) Neurological disorders: public health challenges. World Health Organization.
50. Nicholas R, Giannetti P, Alanousi A, Friede T, Muraro PA. Development of oral immunomodulatory agents in the management of multiple sclerosis. *Drug Des Devel Ther*. 2011;5:255-274.
51. Nandoskar A, Raffel JR, Scalfari AS, Friede T, Nicholas RS. Pharmacological approaches to the management of secondary progressive multiple sclerosis. *Drugs*. 2017;77(8):885-910.
52. Kappos L, Bar-Or A, Cree BAC, et al. Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. *Lancet*. 2018;391(10127):1263-1273.
53. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*. 1983;33(11):1444-1452.
54. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceutical for Human Use (ICH) (2017). E9(R1) Estimands and sensitivity analysis in clinical trials. Available from: https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9-R1EWG_Step2_Guideline_2017_0616.pdf
55. European Study Group on Interferon beta-1b in Secondary Progressive MS. Placebo-controlled multicentre randomised trial of interferon b-1b in treatment of secondary progressive multiple sclerosis. *Lancet*. 1998;352:1491-1497.
56. Secondary Progressive Efficacy Clinical Trial of Recombinant Interferon-beta-1a in MS (SPECTRIMS) Study Group. Randomized controlled trial of interferon beta-1a in secondary progressive MS: clinical results. *Neurology*. 2001;56:1496-1504.
57. The North American Study Group on Interferon beta-1b in Secondary Progressive MS. Interferon beta-1b in secondary progressive MS: results from a 3-year controlled study. *Neurology*. 2004;63(10):1788-1795.
58. Pöhlau D, Przuntek H, Sailer M, et al. Intravenous immunoglobulin in primary and secondary chronic progressive multiple sclerosis: a randomized placebo controlled multicentre study. *Mult Scler*. 2007;13(9):1107-1117.
59. Freedman MS, Bar-Or A, Oger J, et al. A phase III study evaluating the efficacy and safety of MBP8298 in secondary progressive MS. *Neurology*. 2011;77:1551-1560.
60. Zajicek J, Ball S, Wright D, et al. Effect of dronabinol on progression in progressive multiple sclerosis (CUPID): a randomised, placebo-controlled trial. *Lancet Neurol*. 2013;12:857-865.
61. Nicholas RS, Han E, Raffel J, Chataway J, Friede T. Over three decades study populations in progressive multiple sclerosis have become older and more disabled, but have lower on trial progression rates: a systematic review and meta-analysis of 43 randomized placebo-controlled trials. *Mult Scler J*. (in press) https://doi.org/10.1177/1352458518794063. 2018.
62. Kappos L, Radue EW, O'Connor P, et al. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *N Engl J Med*. 2010;362(5):387-401.
63. Steinvorth SM, Röver C, Schneider S, Nicholas R, Straube S, Friede T. Explaining temporal trends in annualized relapse rates in placebo groups of randomized controlled trials in relapsing multiple sclerosis: systematic review and meta-regression. *Mult Scler J*. 2013;19(12):1580-1586.
64. Röver C, Nicholas R, Straube S, Friede T. Changing EDSS progression in placebo cohorts in relapsing MS: a systematic review and meta-regression. *PLoS One*. 2015;10(9):e0137052.
65. Nardi A, Schemper M. Comparing Cox and parametric models in clinical studies. *Stat Med*. 2003;22(23):3597-3610.
66. Schoenfeld DA. Sample size formula for the proportional-hazards regression model. *Biometrics*. 1983;39(2):499-503.
67. Benda N, Branson M, Maurer W, Friede T. Aspects of modernizing drug development using scenario planning and evaluation. *Drug Inf J*. 2010;44(3):299-315.
68. Friede T, Nicholas R, Stallard N, et al. Refinement of the clinical scenario evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. *Drug Inf J*. 2010;44(6):713-718.
69. Cook TD, Kosorok MR. Analysis of time-to-event data with incomplete event adjudication. *J Am Stat Assoc*. 2004;99(468):1140-1152.

70. Wang J, Ke C, Jiang Q, Zhang C, Snapinn S. Predicting analysis time in event-driven clinical trials with event-reporting lag. *Stat Med*. 2012;31(9):801-811.
71. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York: Wiley; 2002.
72. Brownstein NC, Cai J, Sladed GD, Bair E. Parameter estimation in Cox models with missing failure indicators and the OPPERA study. *Stat Med*. 2015;34(30):3984-3996.
73. Hu P, Tsiatis AA. Estimating the survival distribution when ascertainment of vital status is subject to delay. *Biometrika*. 1996;83(2):371-380.
74. Fine JP, Tsiatis AA. Testing for differences in survival with delayed ascertainment. *Biometrics*. 2000;56(1):145-153.
75. Nicholas R, Straube S, Schmidli H, Schneider S, Friede T. Trends in annualized relapse rates in relapsing remitting multiple sclerosis and consequences for clinical trial design. *Mult Scler J*. 2011;17(10):1211-1217.
76. Andreas S, Röver C, Straube S, Watz H, Friede T (2018) Reduction of COPD exacerbations in placebo groups of clinical trials over the past 15 years: systematic review, meta-analysis and meta-regression. (in preparation)
77. Carroll KJ. Analysis of progression-free survival in oncology trials: some common statistical issues. *Pharm Stat*. 2007;6(2):99-113.
78. Raghavan K, Healy BC, Carruthers RL, Chitnis T. Progression rates and sample size estimates for PPMS based on the CLIMB study population. *Mult Scler J*. 2015;21(2):180-188.
79. Lindsey JC, Ryan LM. Tutorial in biostatistics: methods for interval-censored data. *Stat Med*. 1998;17(2):219-238.
80. Friede T, Henderson R. Exploring changes in treatment effects across design stages in adaptive trials. *Pharm Stat*. 2009;8(1):62-72.
81. Crowther MJ, Lambert PC. A general framework for parametric survival analysis. *Stat Med*. 2014;33(30):5280-5297.
82. Crowther MJ, Lambert PC. Parametric multistate survival models: flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. *Stat Med*. 2017;36(29):4719-4742.

How to cite this article: Friede T, Pohlmann H, Schmidli H. Blinded sample size reestimation in event-driven clinical trials: Methods and an application in multiple sclerosis. *Pharmaceutical Statistics*. 2019;18:351–365. https://doi.org/10.1002/pst.1927